Emergence of representations in natural data

Jaakko Väyrynen
Emergence of representations in natural data

Jaakko Väyrynen

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall F239a of the school on 22 September 2017 at 12.

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics
Supervising professor
Professor Mikko Kurimo, Aalto University, Finland

Thesis advisor
Doctor Mathias Creutz, University of Helsinki, Finland

Preliminary examiners
Doctor Magnus Sahlgren, Gavakai, Sweden
Professor Jörg Tiedemann, University of Helsinki, Finland

Opponent
Assistant Professor Filip Ginter, University of Turku, Finland

Aalto University publication series
DOCTORAL DISSERTATIONS 160/2017

© Jaakko Väyrynen

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

Unigrafia Oy
Helsinki 2017

Finland
Author
Jaakko Väyrynen

Name of the doctoral dissertation
Emergence of representations in natural data

Publisher School of Electrical Engineering
Unit Department of Signal Processing and Acoustics
Series Aalto University publication series DOCTORAL DISSENTATIONS 160/2017
Field of research Speech and Language Technology

Manuscript submitted 17 November 2016 Date of the defence 22 September 2017
Permission to publish granted (date) 20 June 2017 Language English

Monograph ☐ Article dissertation ☒ Essay dissertation

Abstract

This dissertation models natural image and language data with data-driven methods with focus in the interpretation of the emergent representation. Cognitive development and processing learns to handle input from the surrounding environment. Similarly, data-driven methods offer a flexible way to find exploratory views of the data.

Independent Component Analysis (ICA) is a proven unsupervised method especially in the field of neural signal processing. It can extract cognitively relevant source signals from seemingly garbled signal mixtures with the assumption of statistical independence. The concept is closely related to sparse coding, which is neurobiologically efficient and is a view of how sensory information is processed in the brain.

In the analysis of small video segments, another statistical concept, temporal coherence, is applied and the results are compared to those of ICA. The representations learned share major characteristics with those measured from the early processing in the visual cortex. A unified model which combines sparseness, temporal coherence and topological organization is introduced.

With similar methodological tools, the focus is shifted to natural language data with only minimal preprocessing in order to create language-independent methods. The meaning of words can be modeled with contextual co-occurrence information collected from a large corpus and vector space models. In contrast to classical methods utilizing second-order statistics, the ICA method can reveal the underlying sparse structure and make the representation more interpretable. In addition to validating the applied unsupervised methodology, the experimental results indicate that the parametrization of the data has a very large effect on the representation learned. With the developed analysis tools, the structure learned is matched to syntactic and semantic features at different levels. For translated sentence pairs, the result is a multilingual representation for words. The increased sparsity of the representations learned is validated by further nonlinear thresholding. The findings can be utilized to build distributional models for words which match better with semantic theories of word classes and relationships among word meanings in natural language processing tasks where more interpretability is desired.

Keywords lexical semantics, vision, language, meaning, computational modeling, vector space models, unsupervised learning, language independence, machine learning

ISSN-L 1799-4934 ISSN (printed) 1799-4934 ISSN (pdf) 1799-4942
Location of publisher Helsinki Location of printing Helsinki Year 2017

Ruippumattomien komponenttien analyysi (ICA) on ansioitunut ohjaamaton tilastollinen menetelmä erityisesti hermostollisten vasteiden käsittelyssä. Se pystyy erottamaan kognitiivisesti olennaiset lähteet näennäisesti sekoituneista signaaleista tilastollisen ruippumattomuusolotuksen avulla. Tämä konsepti liittyy läheisesti harvakoodaukseen, joka on neurobiologisesti tehokas ja edustaa näkemystä aistitiedon käsittelystä aivoissa.

Työssä sovelletaan myös toista tilastollista käsittettä, temporaalista koherenssia, videoanalyysissä ja verrataan sitä ICA-menetelmään. Löydetyt esitykset mallintavat samoja ominaisuuksia kuin primäärin näköaiakuvorengas yksittäissolut. Työssä esitetään yhteismalli, joka yhdistää harvuiden, temporaalisen koherenssin ja topologisen järjestymisen.

Samanlaisia menetelmällisiä työkaluja sovelletaan myös luonnollisen kielen mallintamiseen. Työssä käytetään pelkistettyä esikäsittelyä kieliirripumattomien menetelmien aikaansaamiseksi. Sanojen merkityksiä mallinnetaan kontekstualisten yhteiseisintymien avulla isosta korpuksesta vektoriavaruusmallellä. ICA-menetelmällä paljastetaan allaoleva harva rakenne ja tehdään esityksistä helpommin tulkittavia verrattuna klassisiin menetelmiin, jotka perustuvat toisen asteen tilastollisiin ominaisuuksiin.

Kokeelliset tulokset vahvistavat käytettyjen ohjaamattomien menetelmien toimivuuden ja osoittavat, että aineiston parametrisoinnilla on merkittävä vaikutus löydettyihin esityksiihin. Työssä kehitetänsä analyysimenetelmillä havaitaan löydettyjen piirteiden vastaavan eri tasoilla olevia perinteisiä syntaktisia ja semanttisia kategorioita sekä sanojen välisiä suhteita. Tulokset laajenevat myös monikieliseen aineistoon ja löydettyjen sanasijatyisten lisääntynyt harvus vahvistetaan epälineaarisella kynnystyksellä. Työn tuloksia voidaan käyttää rakentamaan distributionaalisia malleja luonnollisen kielen käsittelyn tehtävissä, joissa tulkittavuus on toivottu ominaisuus.
This work was conducted in Aalto University. It was started at the Department of Information and Computer Science and finalized at the Department of Signal Processing and Acoustics.

I was financially supported by the Helsinki Graduate School of Computer Science and Engineering (HeCSE), the Department of Information and Computer Science, and the Adaptive Informatics Research Centre (AIRC), located at the department. During my studies, I received personal scholarships from the Nokia Foundation and the Emil Aaltonen Foundation. The flexibility of Utopia Analytics Oy enabled me to finish this work.

I am very grateful for my supervisor Prof. Mikko Kurimo who provided excellent guidance in the finalization of this dissertation. I am indebted to my first supervisor Prof. Timo Honkela, who started me on this path, and to the Distinguished Prof. Emeritus Erkki Oja who not only guided and supported me but also provided the excellent working environment. I also want to thank my instructor Dr. Mathias Creutz for investing his time in supporting and encouraging me in the writing process.

Moreover, I am thankful for my co-authors, Prof. Timo Honkela, Prof. Aapo Hyvärinen, Dr. Mathias Creutz, Dr. Jarmo Hurri, Dr. Sami Virpioja, Dr. Tiina Lindh-Knuutila, and Lasse Lindqvist, for their contributions. This dissertation would not exist without your input. In addition to the ones already mentioned, I would also like to thank my friends and colleagues in the former Computational Cognitive Systems research group for the good times and the interesting discussions. To everyone at the Aalto University with whom I have had the pleasure to work with, and to those who facilitated the pleasant and inspiring workplace, thank you.

The pre-examiners of this dissertation, Prof. Jörg Tiedemann and Dr. Magnus Sahlgren, provided thoughtful comments and valuable suggestions.
Finally, I would like thank my wife for her support in this long process as we had to sacrifice many evenings and weekends. Noora, you are the constant that has endured this long process. This work is dedicated to you and to our shared project, Viljo.

Helsinki, August 31, 2017,

Jaakko Väyrynen
## Contents

Preface .......................................................... 1

Contents ......................................................... 3

List of Publications ........................................... 9

Author’s Contribution .......................................... 11

List of Abbreviations and Symbols .......................... 13

1. Introduction .................................................. 17
   1.1 Computational neuroscience applications .............. 20
   1.2 Linguistic and cognitive applications ................... 21
   1.3 Scope of the dissertation .................................. 23
   1.4 Contributions of the dissertation ......................... 24
   1.5 Structure of the dissertation ............................. 26

2. Structure in the visual system and language .......... 27
   2.1 Natural and computational emergence .................... 27
      2.1.1 Emergence in the visual system .................... 29
      2.1.2 Emergence in language ............................ 30
   2.2 Linguistic and cognitive representations ............... 31
3. Computational methodology

3.1 Statistical concepts

3.1.1 Random variables and probabilities

3.1.2 Quantitative measures of data

3.1.3 Central Limit Theorem

3.1.4 Entropy and negentropy

3.1.5 Sparseness

3.1.6 Correlation and independence

3.1.7 Temporal coherence

3.2 Linear models

3.2.1 Linearity

3.2.2 Linear system

3.2.3 Input features

3.2.4 Objective function

3.3 Machine learning

3.3.1 Learning algorithms

3.3.2 Eigendecomposition

3.3.3 Principal component analysis

3.3.4 Singular value decomposition

3.3.5 Preprocessing

3.3.6 Postprocessing

3.3.7 Independent component analysis
# 4. Models of natural images

4.1 Image representation and visual modeling

4.1.1 Early visual processing in the brain

4.1.2 Computational modeling of images

4.1.3 Linear generative model

4.2 Sparse models with independent component analysis

4.3 Topographic organization of basis vectors

4.4 Temporal coherence

4.5 Unified model

4.5.1 The bubble model

4.5.2 Interpretation of the model

4.6 Discussion

4.7 Conclusions

# 5. Linear generative models for words in natural language

5.1 Distributional linguistic vector spaces

5.1.1 Underlying assumptions

5.1.2 Text preprocessing

5.1.3 Data and feature selection

5.1.4 Co-occurrences of linguistic units

5.1.5 Weighting and normalization

5.1.6 Structural transformations

5.1.7 Post-processing

5.1.8 Similarity in the vector space

5.2 WordICA

5.2.1 Independent component analysis and linguistic vector spaces
<table>
<thead>
<tr>
<th>References</th>
<th>129</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publications</td>
<td>145</td>
</tr>
</tbody>
</table>
This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.


VI Jaakko J. Väyrynen, Lasse Lindqvist and Timo Honkela. Sparse distributed representations for words with thresholded independent component analysis.
In *Proceedings of the 2007 International Joint Conference on Neural Networks (IJCNN)*, Orlando, USA, pages 1031–1036, August 2007.


Author’s Contribution

Publication I: “Spatiotemporal receptive fields maximizing temporal coherence in natural image sequences”

The current author was jointly responsible for implementing the experimental part of the work: running the experiments and producing the quantitative measurements. The current author took part in editing the paper.

Publication II: “Bubbles: A unifying framework for low-level statistical properties of natural image sequences”

The current author was jointly responsible for implementing the experiments with natural data. The current author also took part in the editing.

Publication III: “Emergence of linguistic features: Independent component analysis of contexts”

The current author was responsible for designing and implementing some of the experiments, and contributed to the writing.

Publication IV: “Independent component analysis of word contexts and comparison with traditional categories”

The current author came up with the original idea, designed and implemented the experiments, and was the lead author in writing.
Publication V: “WordICA — Emergence of linguistic representations for words by independent component analysis”

The current author designed and implemented most of the experiments, and designed and implemented the automatic evaluation methods. The current author was a major contributor in writing.

Publication VI: “Sparse distributed representations for words with thresholded independent component analysis”

The current author developed the initial idea, designed and conducted most of the experiments and was the lead author in the writing.

Publication VII: “Semantic analysis in word vector spaces with ICA and feature selection”

The research idea was developed jointly. The current author was responsible for modeling and the feature selection experiments. The writing of the paper was a collaborative effort.

Publication VIII: “Emergence of multilingual representations by independent component analysis using parallel corpora”

The research idea was developed jointly, and the current author was responsible for implementing and running the experiments. The current author also participated in the analysis of the results. The writing of the paper was a collaborative effort.


The current author was responsible for the experiments with textual context and contributed to the writing.
List of Abbreviations and Symbols

Abbreviations

AI  Artificial Intelligence
BSS  Blind Source Separation
CLT  Central Limit Theorem
GA  Genetic Algorithm
HAL  Hyperspace Analogue to Language
IR  Information Retrieval
ICA  Independent Component Analysis
LDA  Latent Dirichlet Analysis
LSA  Latent Semantic Analysis
MDA  Maximum Distance Algorithm
ML  Maximum Likelihood
MI  Mutual Information
MSE  Mean Squared Error
NLP  Natural Language Processing
NMF  Non-negative Matrix Factorization
PCA  Principal Component Analysis
POS  Part-of-Speech
PPMI  Positive Pointwise Mutual Information
SOM  Self-Organanizing Map
SVD  Singular Value Decomposition
VSM  Vector Space Model

Symbols

\( a \)  value of random variable
\( A \)  random variable
List of Abbreviations and Symbols

a basis vector
A vector of random variables
A mixing matrix, synthesis
A(·) function
b value of random variable
B random variable
c category vector
C covariance matrix
f(·), F(·) functions
f feature vector
g(·), G(·) functions
h(·) function
h(i, j) spatial neighborhood function
H(·) entropy
i index
I(·) self-information
I(·; ·) mutual information
I(x, y, t) pixel value in image sequence
I identity matrix
j index
J(·) negentropy
JF Jacobian matrix
k index
K number of categories
l index
MA(·) moment-generating function of random variable A
n index
n noise vector
N scalar
N(µ, σ²) Gaussian distribution
p scalar
P permutation matrix
q weight vector
r scalar
s vector of independent sources
t scalar
ti word sequence
T(·) function
$u(t)$ random process
$U$ matrix of eigenvectors
$v(t)$ variance signal
$v$ eigenvector
$V$ matrix of eigenvectors
$w$ word
$w$ weight vector, receptive field, parameters
$W$ demixing matrix, analysis, filter weights
$x$ scalar
$y$ scalar
$x$ input, data vector
$X$ data matrix, observations
$y$ output, estimated sources
$Y$ uncorrelated data matrix
$z(t)$ underlying latent signal
$Z$ Gaussian variable
$z$ uncorrelated variables
$Z$ whitened data
$\alpha$ scalar
$\beta$ scalar
$\gamma$ aspect ratio of Gabor filter
$\epsilon$ small scalar
$\theta$ orientation of Gabor filter
$\Theta$ parameter vector
$\lambda$ wavelength of Gabor filter
$\Lambda$ diagonal matrix of eigenvalues
$\mu$ mean
$\rho$ Pearson’s correlation coefficient
$\sigma$ standard deviation
$\sigma^2$ variance
$\Sigma$ diagonal matrix of singular values
$\tau$ time variable
$\phi$ phase offset of Gabor filter
$\phi(t)$ temporal smoothing kernel
1. Introduction

The ever increasing transfer and storage of digital information has made it possible to study several interesting phenomena by leveraging unannotated data. This open-knowledge and data-driven research has produced recent advancements in bioinformatics, computational neuroscience and natural language processing (NLP). The vast amounts of data require automatic processing in order to efficiently take advantage of the possibilities, for instance, in facilitating communication and information retrieval (IR) within a multi-lingual and multi-modal environment.

The task of science is to explain complex natural phenomena, for which we only have partial observations. Computational linguistics tries to explain and process language. It deals with speech and text in computerized form. Computational neuroscience tries to explain cognitive processing, for instance, by trying to understand perception and by creating models based on physiological measurements. Both fields rely on mathematical modeling techniques and machine learning to build models which learn to represent and process the observed data. Large amounts of related unannotated data are available in both fields, whereas annotated data is much more limited.

This dissertation studies how general unsupervised methods (see, e.g., Hinton and Sejnowski [1999], Oja [2002]) learn to represent natural visual and textual data. The focus is on how statistical concepts beyond correlation produce emergent structures that are cognitively more meaningful and can be interpreted [Hansen et al., 2005]. Due to its simple calculation, linear correlation is a widely applied measure of dependence across many fields of research, but it does not represent higher-order statistics that could provide further insight into how cognition could organize and process information. Nevertheless, a transformation from observed variables into uncorrelated or latent variables is often the first step in finding structure in the data and also enables compression. This dissertation considers transformations based on a higher-order statistical concept, independence, which facilitates further processing and is at the core of fuzzy logic [Gaines, 1978] and causal reasoning [Barlow, 1989, Shimizu
Digital images and image sequences are already explicitly encoded as vectors of numerical pixel values. They can be processed with machine learning computationally, and several similarities between the outcomes of specific computational principles and known neural processes in the brain have been discovered. For instance, learning based on the concepts of sparsity, temporal coherence and topographic organization results in receptive fields similar to the neurons in the visual cortex (see, e.g., Hyvärinen et al. [2009]) and the auditory cortex [Körding et al., 2002]. These learned data-driven representations are usually different from the transforms used by lossy image and sound compression algorithms, as the goal of early sensory coding is suggested to be a transformation into a sparse distributed representation, which makes further processing easier [Field, 1994].

Similar numerical encoding processes are possible for textual units, such as words and documents, with the goal of capturing information or knowledge. For instance, the topical content of a document can be described by the words that occur in it, and the meaning of a word is related to how other words occur near it [Harris, 1954]. Both of these representations can be reduced to co-occurrences of textual units and contextual features defined from a text corpus. These vector space models (VSMs) [Salton et al., 1975] represent textual units as points in a space and the distance between two points measure the semantic relatedness between them. This dissertation shows how the independent component analysis (ICA) method [Comon, 1994] transforms a linguistic vector space into a sparse component representation, in which the emergent components encode interpretable linguistic information and knowledge. Specifically, the components of the word representation learned match with different known categories of words and encode cognitive information. The captured information is greatly influenced by how the vector space is constructed and how its structure is analyzed.

Machine learning can be divided roughly into supervised and unsupervised learning. Supervised computational methods find a mapping between input data and given associated outputs or labels, for instance, to categorize, rank or regress. This means that the method can ignore strong signals from the input variables if they do not help in the task. In contrast, an unsupervised method only tries to reproduce all of the observed data using a model. The brain has to learn to process the constant influx of sensory information mostly without links to reinforcement signals to guide the learning, and therefore neuronal processing can be argued to be more close to unsupervised learning than supervised learning [Barlow, 1989]. It is thus interesting to study how unsupervised machine learning concepts yield representations similar to
linguistic or cognitive knowledge collected from experts, cognitive experiments and human-annotated data. The emergent structure learned can be useful in other computational models or applications; it can give insights into cognitive processing, or it can be further processed to build more complex models or specific applications. For instance, deep neural networks combine features or embeddings learned at the lower levels of the network to solve various problems [Schmidhuber, 2015, LeCun et al., 2015].

The data selection and preprocessing have a large influence on the outcome with unsupervised learning methods, as the most prominent phenomena are usually modeled first. For instance, the word frequencies in text vary significantly, and special care has to be taken to highlight information which is expected or known to be relevant [Spärck-Jones, 1972]. Similarly, approximations of the data are convenient for computational reasons. For instance, the removal of high-frequency details from images makes them blurred, but still clearly recognizable [Koenderink, 1984]. If the removed signal portion has low energy, it should have little effect on the most prominent results found with unsupervised methods. Lossy compression of data is often based on removing information with minimal influence on perception.

Linear generative models explain observed variables as a weighted sum of activations for some underlying variables, and matrix algebra provides a compact mathematical language for the description. For instance, in the famous cocktail-party problem (see, e.g., Bronkhorst [2000]) the observations are recordings with multiple microphones, which pick up sounds generated by multiple sources. In the simplistic model, the weights are determined only by the distances between the microphones and the sound sources. In the mathematical representation each vector of observations is generated by multiplying a mixing matrix of weights with a vector of sound sources. The transformation learned from the estimated sources to the observed variables needs to accurately explain the data. The problem is called blind source separation (BSS), because very little is assumed to be known of the mixing process or the source signals. Similarly, models of natural images or words show one way of generating the data. Information derived from other disciplines, such as computational neuroscience and linguistics, can help both to design more plausible models and provide possible ground truths for the evaluation of the models.

In general, there are an infinite number of possible ways to generate the observations even in a simple linear model, and some restrictions to the model have to be assumed for it to be more unique. There are several general-purpose machine learning methods with different constraints. For instance, principal component analysis (PCA) learns uncorrelated sources, but it is based on second-order correlation and does not
Introduction

necessarily find the original components which generated the data. The additional assumptions, constraints or regularization may be suitable if the nature of the underlying data process is known and this can lead to better and more transparent models. This dissertation concentrates on assumptions on the latent variables and considers how the general concepts and principles of sparseness, independence, temporal coherence, and structural organization learn interpretable representations from natural images and language data. The studied concepts are tied to the current view of how sensory information is processed in the brain [Olshausen and Field, 2004].

1.1 Computational neuroscience applications

Unsupervised computational methods such as independent component analysis have successfully been applied to different biomedical signals [James and Hesse, 2004]. On one hand, digital images approximate visual sensory input through the eyes, and recorded sound imitates auditory sensory data. Different computational models can be taught with the data and the results can be compared to the knowledge of how the brain processes sensory information. On the other hand, direct signal measurements from the brain and the heart with electromagnetic sensors are available. These observations are mixed similarly to the cocktail-party problem and can be analyzed to infer perceived events. Next a brief introduction of the applications with modeling concepts related to this dissertation is given, and the specific developments and results are presented in more detail in Chapter 4.

Independent component analysis was first applied to monochromatic natural images and was found to learn representations which respond primarily to oriented edges and gratings and thus resemble simple cell neurons in the V1 area of the visual cortex [Bell and Sejnowski, 1997, van Hateren and Ruderman, 1998a]. Natural imagery contains photographs or video of natural scenes without man-made objects. Similarly, methods based on the concepts of sparsity [Olshausen and Field, 1996] and temporal coherence [Hurri and Hyvärinen, 2003] learn representations with similar characteristics. Extensions to image sequences [van Hateren and Ruderman, 1998b] as well as color and stereo images [Hoyer and Hyvärinen, 2000], which require encoding of the input signal appropriately, have provided physiologically interesting results. Different methodological developments of independent component analysis allow partial dependencies between components and can find a topographic organization which resemble those measured in the brain [Hyvärinen and Hoyer, 2001]. Specific applications in the image domain include image denoising [Hyvärinen et al., 2001] and recovering originals from mixed photocopies [Tonazzini et al., 2004].
It has also been demonstrated that an analysis of natural sounds with sparse coding learns a higher-order structure [Bell and Sejnowski, 1996]. Similarly, ICA finds components which are characterized in terms of their position and spread in the time-frequency plane [Abdallah and Plumbley, 2001], and reduce redundancy in speech representations [Lee et al., 2002].

Biomedical signal analysis with ICA has considered data from diagnostic tools such as electroencephalograms [Makeig et al., 1996], functional magnetic resonance imaging [McKeown et al., 1998], and electrocardiograms [Cardoso, 1998]. For instance, an electrocardiogram records information with multiple electric sensors from the skin to study the heart. Physical events, such as artifacts outside the heart and more interesting events such as valves opening, are recorded by different sensors at different signal strengths, analogous to the cocktail-party problem. Independent component analysis has been shown to be able to extract the individual events from the mixtures of signals, which can further extend our knowledge of brain structure [James and Hesse, 2004, Groppe, 2007]. The methodology has moved from strictly controlled experiments with only one type of stimuli to the use of rich, natural, multiple sensory stimuli in more real-life-like experiments (see, e.g., Ylipaavalniemi et al. [2009]) and to the modeling of the entire network [Kim et al., 2011].

In the context of neural circuit modeling, information processing is more scrupulously described with spiking neurons and also metabolic constraints have to be considered. This has led to the development of a biologically more plausible mechanism for ICA-like learning with spiking neurons [Savin et al., 2010].

Independent component analysis is clearly an established analysis method in computational neuroscience. The publications in this dissertation expand the research field by applying temporal coherence to natural image sequences and introducing a joint model for low-level statistical properties of natural image sequences. The compendium part of this dissertation provides a brief overview of the early research of finding the link between unsupervised modeling of natural imagery and the primary visual cortex. This gives a starting point for starting to apply independent component analysis to language data.

1.2 Linguistic and cognitive applications

The idea of applying ICA to text document analysis is first presented in Isbell and Viola [1999], where topic-centered representations were introduced to find projections for grouping related documents and highlighting important words. This section gives
a brief overview of recent developments in this research area. More detailed results and discussion of the central topics for this dissertation are presented in Chapter 5.

The use of ICA in text analysis idea has been extended from the original topics and documents [Kolenda et al., 2000, 2001, Bingham et al., 2002] to word types (Publication III), word tokens [Rapp, 2004, Šimon and Hong, 2007], morphemes [Lagus et al., 2005], and phonemes [Calderone, 2009], covering a wide range of levels in representations of written language. Independent component analysis has been shown to find meaningful or understandable representations at all levels. The explored applications for ICA in the field of written language analysis include language modeling [Kumaran et al., 2005]; text classification for different tasks [Yokoi et al., 2005, Sam et al., 2007, Pu and Yang, 2006, Yu and Ho, 2014]; finding translations in parallel (Publication VIII) and comparable texts [Hazem and Morin, 2012]; and induction of personality traits [Chagnaa et al., 2007] and word concepts [Chung and Pennebaker, 2008].

More generally, ICA has been suggested as a cognitive model that can learn structures that are well-aligned with those resulting from human cognitive activity [Kolenda, 2002, Hansen et al., 2005, Feng, 2008]. Various studies show that ICA can learn cognitive representations for text, discussions, images, combination of text and images, social networks and music. For example, Bartlett et al. [2002] showed that ICA learns to represent images of faces with localized features resembling facial features. ICA has also been applied to NLP tasks that examine other representations than written natural language, for instance, automatic speech recognition [Lee et al., 2000], fMRI analysis during word generation tasks [Dodel et al., 2000], sign language analysis [Bowden et al., 2004], and hand-written text analysis [Chen and Leedham, 2005]. In addition to the more unstructured and free-form data sources mentioned above, there are also structural information databases which can be analyzed with independent component analysis. Studies have considered, for example, XML documents [Wang et al., 2006], source code [Grant et al., 2008a], and semantic search [Ruotsalo and Frosterus, 2013]. These topics extend outside the written language are not covered in this dissertation.

The main topic in this dissertation in Chapter 5 is the goal to create and analyze unsupervised linear models of word meaning with independent component analysis. The data-driven and largely language-independent models learn from language usage data, without expert knowledge bases of the language. Especially, very little preprocessing is involved. For instance, no disambiguation systems or morphosyntactic analyzers have been applied. Similarly, contextual information is simply taken as the adjacent or neighboring words without applying syntactic analyzers to find
grammatically connected words in sentences. This simplified approach naturally affects the outcome, but with statistical methods and appropriately selected and preprocessed data and features, interesting events emerge from the data. Furthermore, this approach ensures that the methodology is robust and applicable to a variety of languages without extensive linguistic analysis tools.

1.3 Scope of the dissertation

This dissertation is a multidisciplinary work in computer science. It considers topics in the fields of computational neuroscience and computational linguistics, with a strong focus on the latter. Unsupervised machine learning methods are applied to natural data in order to learn emergent representations which describe the observations. The main focus is to study what kind of interesting emergent representations can be learned with simple and non-structured assumptions of the data. To support that goal, data or task-specific preprocessing and tools are applied only if they are generally adopted by the field and deemed necessary. The utilized vector space models [Salton et al., 1975] provide an algebraic representation of the natural data suitable for efficient computational processing. The structure of the vector space is analyzed with the unsupervised independent component analysis [Comon, 1994] method. The work is applied in two different but central research fields, computational neuroscience and language technology, which are actively researched and there exists theories and models which provide interpretable representations at the lower levels. This enables the comparison of the emergent representations learned from the natural data to known neurobiological, linguistic and cognitive knowledge.

The data-driven paradigm is built on learning from data instead of relying on handcrafted models and structure. Furthermore, in this work the data is assumed to be natural, i.e., similar to something which everybody encounters. Especially, task specific external knowledge is not provided when the model are learned. For instance, publicly accessible digital images and videos of natural scenes are assumed to provide statistically relevant and typical visual sensory information. Similarly, collections of natural text should be representative of normal observed written language. This dissertation works with digitized images and text instead of more realistic models of the vision or speech recordings, because already existing publicly available data can be utilized without making too many assumptions, and the results from the data-driven methods can be compared to known knowledge of human cognitive processing.

The learning models studied in this work are unsupervised, which means that labeled
data is not utilized when the models are trained. Structure of the model learned, the learning method and the data will specify what kind of representation will emerge. Specifically, the training does not take advantage of manually built knowledge resources, such as dictionaries or labeled image banks.

In this dissertation lexical semantics [Pustejovsky, 1995] is approached from a statistical point of view with vector space models [Salton et al., 1975] which provide a distributed and numerical representation with statistics computed from a text corpus for the analyzed symbolic textual elements. The vector spaces are created in a pragmatic and simple manner in line with common procedures in the research field. This work especially investigates the structures that can be found with the unsupervised independent component analysis method from vector spaces. The models learned are linear to facilitate fast learning and to remove the additional assumptions present in more complex models.

Similarly, small image patches sampled from digital images and image sequences are analyzed with unsupervised models to learn linear models which can represent the image data, which is already in a format more suitable for numerical processing. The work shows how the known properties of visual sensory information processing in the visual cortex emerge from the statistics of natural images.

### 1.4 Contributions of the dissertation

Publication I explores how the computational principle of temporal coherence learns simple-cell-like receptive fields from natural data. This is the first time temporal coherence has been used to model spatiotemporal receptive fields of simple-cells. The experiments with natural image sequences show that while the spatial characteristics of the emergent receptive fields are shared with the simple-cell receptive fields, there are differences in the temporal characteristics. The properties are measured quantitatively and compared against similar results with independent component analysis and physiological measurements from adults cats from literature.

Publication II proposes a unified framework for the low-level statistical properties of natural image sequences. It joins the computational principles of sparseness and temporal coherence with correlation of energies. The temporal integration model with the sparseness and temporal coherence properties is shown to separate signals better than either principle alone. This supports the statement that the proposed model is a better model of the structure of natural stimuli than previously proposed models. The unified model presents an explanation to why both sparseness and temporal
coherence learn similar receptive fields from natural data.

The WordICA approach, which I have developed in its current form, is described in Publication III with manual and visual analysis of the emergent features. It is the first method to specifically structure word representations, and present the emergent structure learned as cognitively meaningful and interpretable. The emergent structure learned is visualized and shown to match with syntactic part-of-speech (POS) categories. The correlation between the components learned and syntactic part-of-speech categories is examined in Publication IV. It introduces a correlation-based measure for the similarity between linguistic word categories encoded as binary vectors and the emergent independent components.

The justifications for the WordICA model are discussed in more detail in Publication V. It also introduces an evaluation measure for the separability of the emergent components and known linguistic categories. It is shown that the WordICA components separate syntactic part-of-speech categories significantly better than components learned with the singular value decomposition method based on second-order correlations.

An extension of the WordICA model to a bilingual word representation based on a parallel corpus is introduced for the first time in Publication VIII. The created bilingual word space is explored by considering the distances between emergent word vectors between known translations of words, and by a visualization of the emergent bilingual components. Publication IX presents a method for learning and visualizing interlingual lexical mappings using the self-organizing map (SOM) algorithm taught on a bilingual word space.

In Publication VI, representations found with WordICA are studied with multiple-choice test of synonyms and related words. The vector representations for each word are thresholded, and it shown that the new representation retains the information of the relations between words much better than a representation learned with singular value decomposition. The automatic generation of multiple-choice test from lexical resources is presented.

Publication VII explores how semantic information is present in linguistic vector spaces. Several vector space models are explored with different evaluation tests. For the first time, the semantic nature of features extracted with WordICA is thoroughly examined both manually and quantitatively.

The introductory part of this dissertation represents the attached papers in relevant scientific contexts. Some background information is given on the knowledge of oper-
Introduction

ation of the visual system and relevant linguistic theories, which provide prior information for model selection and represent structures towards which the unsupervised methods are guided and evaluated against. The relevant methodological foundations for machine learning are discussed briefly. These make it possible for people with different background to read this dissertation. The research of image representations is placed in a larger research context and visualize the similarity of the representations learned with the different computational methods. A summary of recent developments in written text processing with independent component analysis is given, with a focus on word representations. Several practical implications of different preprocessing methods are considered.

1.5 Structure of the dissertation

Chapter 1 gives an introduction to the dissertation topic and scope, lists the scientific contributions and the current author’s contribution to each publication. The physiological and cognitive background information related to natural data processing is briefly discussed in Chapter 2, and the relevant background in computational methodology for data-driven processing is introduced in Chapter 3. The specific models for representation of visual data are presented in Chapter 4, and the textual models in Chapter 5 with a literature survey to methodologically related scientific studies. A summary of the dissertation is given in Chapter 6.
2. Structure in the visual system and language

Terms such as emergence and self-organization are often linked with descriptions of both natural and artificial complex systems [Gershenson and Fernandez, 2012]. Emergence will be discussed later in Section 2.1. On one hand, it is a widely held belief that cognitive representations and categorization are highly affected by experience which is reflected in the statistics of sensory experiences and observed language usage [Kelly and Martin, 1994]. On the other hand, computational models can both be emergent and a way to model emergent properties of natural systems [Symons, 2008]. In this dissertation, computational methods learn to analyze and structure observed information.

In order to understand what could be interesting results from the computational methods, it is essential to know what other existing representations there are for benchmarking and evaluation. This will contribute to selecting suitable input data as well as the analysis of the representations found. Two different themes are considered: emergent visual representations measured from the mammalian brain, and linguistic word categories. This chapter aims to give a short overview of these topics from the perspective of computational modeling.

2.1 Natural and computational emergence

Emergence is an important concept in this dissertation. The term ‘emergent’ was coined by Lewes [1875, p. 412] and can be summarized as the popular notion of the whole being more than the sum of its parts. A modern definition of emergence is given by Goldstein [1999] as ‘[T]he arising of novel and coherent structures, patterns and properties during the process of self-organization in complex systems’. There are several natural processes that involve emergence, such as evolution. These have influenced the development of computational models, which are discussed later. Self-organization itself is a related concept which is often described as self-assembly or
self-organizing behavior in deterministic dynamic systems [Ashby, 1947].

The concept of emergence is further divided into strong and weak forms. Weak emergence explains the new properties as unexpected qualities, whereas strong emergence proposes that the new qualities are not even in principle deducible from the system’s components. For instance, thermodynamic temperature is a weak emergent macroscopic quality which arises from the underlying vibrational motions of matter’s particle constituents. The strong form of emergence is much more demanding and it has been argued that consciousness is possibly the only instance [Chalmers, 2006].

The ultimate goal of artificial intelligence research is to create a machine with capabilities matching or surpassing those of humans. Machine learning, computer vision and natural language processing are sub-tasks of artificial intelligence research related to this dissertation. Similarly to the consciousness being emerged by the simple building blocks of neurons and repeating simple structures in the brain, it is hoped (and feared) that an artificial intelligence will finally emerge when the computational capabilities of software and hardware have matured enough. Recent developments in computer-based reasoning, computer vision, models and learning algorithms have produced significant advances in several tasks [Montemerlo et al., 2008, Ferrucci et al., 2010, Silver et al., 2016].

There are several examples of weak emergence in nature which have inspired methodological development in computational sciences. In complex systems theory, high-level patterns emerge in cellular automata [Wolfram, 1983]. These discrete systems contain simple identical components on a grid with local interactions. A related notion for continuous systems is swarm intelligence, which can model, for instance, the behavior of ant colonies or bird flocking [Garnier et al., 2007]. Artificial neural networks, artificial immune systems and evolutionary computation have been inspired by nature as powerful methods for knowledge representation and learning. Artificial neural networks are built from simple processing elements and connections inspired by biological neural networks. Algorithms are used to modify the connection weights between elements to learn complex patterns from data. Artificial immune systems solve problems with models inspired by observations and abstractions of natural immune systems [de Castro and Timmis, 2002, p. 58]. Evolutionary computation includes iterative algorithms for meta-heuristic optimization based on principles of Darwinian evolutionary systems [de Jong, 2006].

The notion of weak emergence is relevant in this dissertation because it is assumed that human cognitive processing learns from observations of the world to process sensory information. The structure learned provides an abstraction level that is utilized
in further higher-level processing. This can be contrasted with the view of nativism, where beliefs, ideas and skills are considered to be hard-wired in the brain, or the brain contains specific cognitive modules instead of only general purpose cognitive capabilities [Fodor, 1975]. In this dissertation, it is assumed that statistical inference with unsupervised methods is sufficient to learn interesting phenomena and knowledge representations related to visual information processing and language modeling. Interaction with the world, which is not discussed in this dissertation, would help unsupervised methods to focus on data more relevant to, for instance, communication and survival. Computationally, this can be accomplished with reinforcement [Sutton and Barto, 1998] or supervised learning.

2.1.1 Emergence in the visual system

The visual system in the brain is able to quickly assimilate highly discriminative and wide range information optically through the light receptors in the eyes. The evolution of vision has been the subject of significant study and discussion because of its complex structure and pervasiveness across species [Lamb et al., 2007]. The focus in this dissertation is on how the first layers of the visual cortex in mammals learn to process visual information by adapting to the sensory information it receives. There are several specific areas and pathways in the brain which are connected to the visual input from the eyes. The retina contains cones and rods, which detect light and color. Visual processing such as contrast adaptation starts already at the retina [Heinrich and Bach, 2001]. The simple cells in the V1 area of the visual cortex have been identified to respond to edges and lines, which are required to perceive objects by their boundaries. Furthermore, the simple cells are also organized on a two-dimensional lattice in which receptive fields that respond to similar stimuli are close to each other [Hubel, 1995]. The visual processing system in the brain is influenced by the input it receives. This is demonstrated by, for instance, deprivation of natural input during the critical development stage, which leads to different simple cell receptors compared to unrestricted natural input [Hubel and Wiesel, 1959].

On the conscious level of perception, the Gestalt effect [Palmer, 1990] in vision describes how the perceived and recognized forms and objects are different from the individual patterns that are recognized and processed by the lower levels, i.e., we perceive whole forms instead of just a collection of local element activations. In the spatial domain we tend to perceive complete objects even when they are partially occluded or incomplete, for instance, we tend to see squares and circles instead of discontinuous lines and curves. In the temporal domain, a series of slowly chang-
ing still pictures with high enough frame rate creates the illusion of moving objects. The representations at the first levels of visual processing, which will be described in Section 4.1, could be related to the observed properties of the conscious level of perception, such as the Gestalt effect.

2.1.2 Emergence in language

Language is a complex system well suited especially for communication, but is also used for other purposes, including expression of emotions and social grooming [Östman et al., 2007]. Similarly to visual processing, there are specific areas in the brain specialized to language and speech processing. For instance, the Broca’s and Wernicke’s areas were identified before modern brain imaging capabilities by speech comprehension and generation aphasias, respectively [Christensen, 2001]. The existence of language itself is an emergent phenomenon [Logan, 2007]. Furthermore, emergence can be seen in the continuing language change and the capacity of children to learn a language by being exposed to it without external and explicit control of the grammar. A specific theme here is that emergent categories enable efficient knowledge representation as well as productive creation of new utterances in a language.

In this dissertation, knowledge of language is considered to emerge from language use [Croft and Cruse, 2004]. Especially children learn to speak and understand a language by being exposed to it at the critical learning age without explicit knowledge of grammar. Specifically, the linguistic norms and conventions of the society are neither passed innately by genetic material nor by explicit teaching, but can be seen as a system emerging from long-time participation in communicative problem-solving in various social circumstances. Language acquisition is therefore linked to general processing mechanisms in the brain. These ideas have been formalized in the connectionist and emergentist theories of language acquisition by Elisabeth Bates and Michael Tomasello (see, e.g., Bates [1999] and Tomasello [2003]).

Language change can also be argued to be an emergent phenomenon. Each language speaker is able to modify language or use it in a new way, and some of the variations can be adopted by other users. When enough speakers have adopted the modification, the language can be seen to have changed organically [Keller, 1994]. In contrast to modeling based on observation and experimentation by experts, computational modeling is able to model change as their parameters can be adjusted more easily when new data is available.
2.2 Linguistic and cognitive representations

Language represents a view of how cognition operates, but both are observed directly only on a very high level, as the final actions or produced language. The underlying structure and representations are not directly observed but understanding what they are and how they are learned would clearly help with both creating and analyzing computational models of language.

Each language has different conventions for combining smaller units into more naturally communicated utterances. Language speakers learn to understand and produce language, and especially children learn to generalize without explicit teaching. Linguists have construed grammars as models of languages based on observations, experimentation and analysis of language data over long periods of time. The grammars can be prescriptive (normative) grammars defining the proper use of language in a top-down fashion, or bottom-up descriptive grammars for actual language usage, including dialects and colloquial language. Grammars rely heavily on different abstractions, such as categories, rules and generalizations in order to describe languages compactly.

Abstractions such as categories help both in understanding previously unheard utterances as well in the creative production of new utterances. Language structure is often researched and analyzed separately at different levels of grammar: phonology studies how individual words are formed from sounds, morphology inspects the internal structure of words, syntax examines how the words are joined to form sentences, semantics is the study of meaning, and pragmatics considers how the context influences the meaning. Another viewpoint known as the cognitive grammar theory states that the study of meaning cannot be strictly separated from the study of the other levels of the grammar [Saeed, 2003].

Knowledge of the world should be compatible between people for successful communication. Any separation of encyclopedic and linguistic knowledge can be meaningless when semantics is discussed. The hypothesis of abstract concepts provides the notion of a mental representation which stores the accumulated knowledge of real world objects with their properties and relations. This links reality, thoughts and meaning together. When people refer to objects with words, the recipient usually has a very similar idea of the typical properties of the object, i.e., their mental representations are similar from comparable experiences. Brown [1973] suggested that all children seem to express categories for agents, objects, actions, and locations. For instance, the categories of agents and objects make a distinction between animate and
inanimate objects.

In addition to similarities, there are also differences especially across languages. Varying conditions and traditions have pushed languages to adopt different ways of naming parts of the world. Examples of such variation can be seen in how body-parts are identified and named across languages [Andersen, 1978]. As a whole, similar sensory-motor experiences, shared history and cross-cultural interaction probably have contributed to the at least coarsely compatible conceptualization across languages. This relates to the philosophical embodied mind theory, which states that the human body largely shapes all aspects of cognition (see, e.g., Lakoff [1990] and Varela et al. [1992]).

The Gestalt effect, discussed in Section 2.1.1 for visual perception, can also be recognized in linguistics, for instance, in language learning, language understanding, and cognitive grammar. For example, children first associate objects to heard utterances (e.g. ‘it’s a box|clock’) as single units, and only later the individual words are separated [Palmer, 1990]. Moreover, small typos in written language or lapses in speech are often not perceived at all, and meaning of the words and sentences are understood on the conscious level as a whole instead of the underlying elements such as characters and words. This view is considered especially in the cognitive grammar theory, which proposes that language consists of elementary, meaningful chunks of language called constructions [Lakoff, 1990].

Linguistics is the study of human languages and aims to describe language form, meaning and its use in context, giving a diverse field of research from different viewpoints. One of the major goals of language is to communicate meaning. This dissertation concentrates on the study of word meaning or lexical semantics based on data-driven distributional vector space models [Salton et al., 1975] originally developed for information retrieval. There are also models of textual representation inspired by natural systems, such as artificial neural networks [Honkela et al., 1995, Bengio et al., 2003] and artificial immune systems [Pöllä and Honkela, 2010].

This dissertation studies words, the smallest unit of speech [Bloomfield, 1984, p. 178] which represent concrete objects, abstract ideas and actions, providing a link between linguistic and cognitive representations. Semiotics is the more general study of symbols and signs both within and without language [Chandler, 2007]. Linguistic grammars utilize categories or features in various forms in order to compactly describe languages. The segmentation of utterances into words is also a form of categorization, which is often explicitly marked in the textual representation of language.

The statistical and computational approach in this dissertation relies on models of
word meaning based on text corpora of actual language usage. This is in contrast to traditional symbolic models, which rely to a large extent on the intuition of linguists and manual definition of linguistic rules and categories. Moreover, the structural categories of grammar have been argued to be language-specific [Haspelmath, 2007] even if universal annotation schemes are being developed [Nivre et al., 2016]. However, different existing manually constructed word categorizations provide a point of comparison and a way to analyze the statistically emergent categories directly. Direct manual evaluation still has to resort to analyzers’ opinion or intuition. In this dissertation, different word categories defined by linguists are compared to emergent categories or features learned with computational methods from texts representing examples of word usage. This can be viewed as a necessary first step of verifying the methodology and towards more comprehensive evaluation. The emerging structures could further be evaluated in real-world applications, psychological experiments or manually by linguistics. Similarly, the models in different tasks could be analyzed from an information theoretic perspective.

The term ‘word’ in this dissertation usually refers to orthographic words, but the work is not limited to them. The orthographic words are readily available in many languages and corpora with very little processing. Additional information or analysis could extend the work into lexemes (sets of forms taken by single words) or lexical items (possibly multi-word semes, the smallest unit of meaning recognized in semantics). Similarly, the list of possible words can be called a vocabulary, lexicon or lexis, depending on the elements. There is also ambiguity in the word meaning as words can have different senses. The term ‘token’ refers to single instance of word.

2.2.1 Categorization theories

Linguistic and cognitive theories of categories are based on observation, linguistic intuition, hypothesis-testing and analysis. These create competing but also partially incompatible theories. Similarly, when computational models are later applied in Chapter 5 several implicit and explicit assumption are made which influence the result. Different known possible categories for words are applied as an evaluation resource towards which the computational methods could be directed.

Categorization of similar words into classes or categories allows abstraction in the description of language but also provides a possible way of representation and extraction of knowledge. Semantic categories (birds, animals) tell more about the structure of the world, whereas grammatical categories (noun, preposition, etc.) define the words usage in sentences and are applied in grammar rules which model the gram-
matically acceptable sequences of words. It should be noted that there is a semantic link between members of the same category (e.g. nouns vs. adjectives) and the separation of semantics and grammar is not clear but rather a continuum. For instance, not all verbs are linked to all nouns, but there is an interplay of semantics. Compared to traditional linguistic grammars, the widely used data-driven $n$-gram language model in its basic form completely discards the use of categories and operates solely on observed language usage.

There is clear evidence for the existence of categories in the mental representation [Vigliocco et al., 2011]. Some of the categories have been suggested to be innate (e.g., face detection, see Bednar and Miikkulainen [2003]), but mostly they are learned from normal exposure to natural stimulus instead of special training. In general, humans have a tendency to perceive smaller distances between instances of sensory experiences from the same object than between different objects, i.e., there is sharp change of perception with smooth identity change. For example, visual views of the same face from different angles are perceived as more similar than views of different faces. This phenomenon of categorical perception was originally identified in speech perception, especially in the classification of a continuum of vowel sounds into abstract vowel categories [Harnad, 1990]. The Sapir-Whorf hypothesis, which suggests that language shapes the way one thinks, has been proposed as one of the influences for categorical perception, but is has been shown to be induced by learning alone [Goldstone, 1994].

Traditional semantic categorization theories, such as logic and set-theoretic approaches, propose mental representations with categories using necessary and sufficient features. Words (e.g. sparrow, penguin) can be assigned different features (+can-fly, +has-wings) and they can be collected into categories (flying-things, nouns, birds) based on these features. Set operations can mathematically modify these categories (flying-things $\cap$ birds = flying-birds). Semantic categories (birds) describe more the properties of the things they refer to.

Alternatively, prototype theory [Rosch et al., 1976] in cognitive science suggests that some members of a category are more central than others, without demanding shared features for all members. For instance, four-legged chairs are more often encountered than three-legged chairs, and the word ‘chair’ is more frequently cited as an instance of furniture than ‘stool’. The theory helps to build graded categories and shared meanings for words.

The Conceptual Spaces theory [Gärdenfors, 2004] has been proposed as an explanation for relations between prototypes and mapping from the perceptual space to
categories. The theory represents natural categories as convex regions in a multidimensional feature space, where the features are quality dimensions such as weight and color. The possibility to measure similarity between words and concepts is also important. Prototype theory enables this by graded categorization and the Conceptual Spaces theory makes it computable, but does not directly explain where the quality dimensions come from.

Statistical data-driven linguistic representations make assumptions of what is relevant and consider selected computable statistics [Salton et al., 1975, Lund et al., 1995] or prediction tasks [Collobert et al., 2011] to create a distributed representation. In contrast to the abstract and clearly defined quality dimensions in the Conceptual Spaces theory, the features are nebulous and shaped by the data. The multidimensional representation is continuous without clear boundaries and categorizations have to be produced by further computational processing, such as clustering [Pereira et al., 1993]. The main power of the continuous multidimensional representation is the ability to measure similarity or relatedness of the elements by distance in the space learned.

Different linguistic theories provide intuition and starting points for deciding how the features are computed in data-driven modeling. Furthermore, an unsupervised representation can be analyzed by comparing its properties directly with those of a manually defined categorization. Alternatively, a supervised method can be taught to utilize the unsupervised representation directly in modeling the linguistic representation.

### 2.2.2 Category elements

Written language is essentially a sequence of discrete symbols. The symbols can be identified as characters from different alphabets, syllables and words, or even ideograms in Chinese or pictograms in ancient Egyptian. These small abstract units can have some specific meaning or function and can be combined to form phrases, sentences and longer text segments with compounded or collocative meanings. The smaller units are selected from a very limited set (such as the alphabet) and repeat often, whereas there is basically no limit for the longer units (such as sentences) and they typically appear only once. Even though this dissertation focuses on the analysis of words, the methods and ideas presented are not limited to those units.

Words consist of one or more morphemes, which are the smallest meaning bearing units. For instance, the word ‘singing’ can be thought to consist of the morphemes ‘sing’ and ‘ing’. Unlike orthographic words, written languages do not mark morphemes and they have to be identified by other means. For agglutinative languages,
statistical morpheme segmentation can provide practical results, but for fusional languages other methods may be required in order to identify different distinct orthographic representations of morphemes [Creutz and Lagus, 2007b].

Words are related to each other in different ways, and combinations of words into phrases can limit or alter the meaning of the individual words. Syntactic models of language try to explain how words are connected to each other in a sentence, as the connection between words is not determined by the number of other words between them. Moreover, words and morphemes can have multiple meanings which sometimes need to be disambiguated based on the textual and physical context as well as world knowledge. Word meaning is therefore closely associated with the context it appears in.

2.2.3 Word resources

The chosen data-driven approach leverages existing general data collections of word usage. The output of the unsupervised learning is contrasted against existing linguistics and cognitive resources for words, such as semantic and syntactic word categorizations and different relations between words.

The examined statistical methods require textual training data for the computational models. There are abundant resources for unsupervised methods which rely solely on examples of language usage without specific annotations. However, even then the training data is focused on some specific domain and cannot contain all possible styles, phrases or even word forms, however huge the data set is. Some resources of language usage have been collected for research purposes, for instance the British National Corpus [Burnard, 1995]. More often, publicly available resources such as the Wikipedia Encyclopedia or institutional publications have been collected into text corpora for research and development purposes. Parallel resources, such as the Europarl corpus [Koehn, 2005] and the OPUS resource [Tiedemann, 2012] are widely applied in state-of-the-art research and applications. In this work, for instance, in Publication VII the Wikipedia Encyclopedia is used as a text corpus.

Sometimes texts contain some additional explicit structural information, such as topic, author or style information, that is interesting for research purposes. However, in contrast to the more general unannotated resources, annotated resources are focused on some specific question and encode additional structural information about it, for instance, labels for the elements or relations between the elements. Linguistic annotation tries to make linguistic information explicit in the data according to the chosen theory that provides certain abstractions and regularities. For instance, the British
National Corpus contains syntactic part-of-speech tagging with the CLAWS4 tagset [Leech et al., 1994]. However, it should be noted that there will always be alternative ways of annotating data and each of them will have their own shortcomings with respect to the description of linguistic phenomena and specific language use.

In contrast to unannotated resources, annotated resources for supervised methods are much harder to come by as the methods need also the correct response for each example. Linguistic experts are usually required for the creation of grammatically annotated resources, but members for some of the semantic categories can be elicited from layman subjects, for example, by asking them to list words belonging to given concepts [Battig and Montague, 1969]. This results in dissertation rely on annotated resources in evaluation of the unsupervised methods, not for training the models.

The information present in the representations learned can be directly or indirectly evaluated with the annotated data. Indirect evaluation could entail a task in which the representation learned is used as data in a supervised task with known labels. This can provide clear information on how inclusion of the representations changes the results in the high-level task, for instance, machine learning, information retrieval or part-of-speech tagging. Direct evaluation usually deals with comparing the match between the emergent and the linguistic representation. This is not as straightforward, especially if the representations are distributed and the labels are not. As discussed above, the linguistic representation has to be selected and, furthermore, the data has to presented in way that the unsupervised method learns relevant representations. The publications in this dissertation are focused on direct evaluation and the analysis of the emergent representations. Relevant research with indirect evaluation are discussed.
3. Computational methodology

A vast array of computational methods have been applied to unsupervised learning from natural data and it is impractical to try to list all of the methods comprehensively here. This chapter gives a short overview of relevant concepts in statistics, probability theory and machine learning. Important learning methods based on these concepts are introduced, including independent component analysis which is the main methodology and the focus in this dissertation.

The information in this chapter is collected from various sources, for instance, Hinton and Sejnowski [1999] for unsupervised learning, Hyvärinen et al. [2001] for independent component analysis, and Hyvärinen et al. [2009] for tools for natural image analysis.

3.1 Statistical concepts

In statistics and probability theory, a random variable $A$ is a variable whose value is subject to chance. The values can be scalars ($A \in \mathbb{R}$) or vectors ($A \in \mathbb{R}^d$). The concepts here are defined for discrete random variables, but they exist also for continuous-valued random variables. Instead of tabulating all possible values and their probabilities, statistical measures that condense the behavior and provide a concise approximative description of the data are typically computed. They can also be incorporated into objective functions when models are fitted to data. Naturally, random variables are not observed directly, but a finite sample of observations is assumed to represent a larger population for which statistical estimates can be computed.

3.1.1 Random variables and probabilities

The chance of observing value $a$ for the random variable $A$ is the probability $P(A = a)$ or the shorthand $P(a)$. Probabilities are non-negative ($P(a) \geq 0$) and sum to one.
Computational methodology

\( \sum_{a \in A} P(a) = 1 \). A probability distribution assigns a probability to each of the possible outcomes of \( A \) and is usually written in the shorthand form \( p(a) \). The joint probability \( P(a,b) = P(A = a, B = b) \) is the chance of observing \( a \) and \( b \) together. The conditional probability \( P(a|b) = P(A = a|B = b) \) represents the chance of observing \( a \) when \( b \) has already observed. The Bayes’ theorem

\[
P(A|B) = \frac{P(B|A)P(A)}{P(B)}
\]

connects the joint and conditional probabilities.

A random variable \( A \) can be completely defined by a probability density function (pdf) \( p_A(a) = P(A = a) \), a moment generating function \( M_A(t) \) or a cumulant generating function, if they exist. They are discussed more in Section 3.1.2. Parametric probability distributions define the shape of the distribution with only a few parameters instead of listing all possible outcomes and probabilities, making it possible to make mathematical operations and derivations on a higher abstraction level. An important class of probability distributions is the exponential family:

\[
p(x) = h(x) \exp(\theta \cdot T(x) - A(\theta))
\]

which includes, for instance, the Gaussian, Bernoulli, Exponential, Poisson and Dirichlet distributions. The vector \( \theta \) contains the parameters, \( h(\cdot) \) is the ‘underlying measure’, \( T(\cdot) \) is a vector of ‘sufficient statistics’, and the log normalizer \( A(\cdot) \) ensures that the density integrates to one. The properties of the exponential family facilitates and simplifies several machine learning methods, such as maximum likelihood (ML) estimation explained in Section 3.3.1. For instance, the important Gaussian (or Normal) probability distribution:

\[
N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|x - \mu\|^2}{2\sigma^2}\right)
\]

is completely defined by the mean \( \mu \) and variance \( \sigma^2 \) parameters.

A Gaussian distribution assumes the least information and has the maximum entropy of all distributions with a specified mean and variance. Measures of nongaussianity are important in the development of algorithms for independent component analysis. Nongaussian distributions can further be divided into subgaussian and supergaussian distributions, which are illustrated in Figure 3.1. A subgaussian distribution, such as the uniform distribution, typically has a lower peak and weak tails in the pdf compared to a supergaussian distribution, such as the Laplace distribution, which has a higher peak and fat tails. Many natural signals, such as image and sound have been found to be supergaussian [Torralba and Oliva, 2003], which motivates the use of algorithms which assume nongaussianity.
### 3.1.2 Quantitative measures of data

The properties of probability distributions, or observed data, can be measured with moments and cumulants. The expectation operator $\mathbb{E}[A] = \sum_{a \in A} a \cdot p(a)$ can be interpreted as the mean $\mu$ of a random variable, and the moment-generating function $M_A(t) = \mathbb{E}[e^{tA}] = \sum_{k=0}^{\infty} \frac{t^k \mathbb{E}[A^k]}{k!}$, if it exists, defines the pdf of $A$ with the moments $\mathbb{E}[A^k]$. The cumulants $\kappa_k$ are defined as the coefficients of the power series expansion of the cumulant-generating function, which is the logarithm of the moment-generating function. The $k$th moment about the mean (or $k$th central moment) is $\mu_k = \mathbb{E}[(A - \mathbb{E}[A])^k]$ and is independent of translation, i.e., the value does not change when the distribution is shifted by changing the mean. For instance, the second central moment for a random variable $\mu_2 = \sigma^2 = \mathbb{E}[(A - \mathbb{E}[A])^2]$ is called variance, and covariance $\sigma_{A,B} = \text{cov}(A, B) = \mathbb{E}[(A - \mathbb{E}[A])(B - \mathbb{E}[B])]$ for two variables. A natural generalization of variance for multidimensional random variables, $A$, is the covariance matrix $C = \mathbb{E}[(A - \mathbb{E}[A])(A - \mathbb{E}[A])^T]$ which

![Figure 3.1](image_url)  
**Figure 3.1.** Pdfs of a Gaussian distribution (solid line) and two different types of nongaussian distributions. A supergaussian distribution (dashed line) has a higher peak and fatter tails, whereas a subgaussian distribution (dotted line) has a lower peak and weaker tails. All distributions have the same mean and variance.
contains the pairwise covariances of all variables. For uncorrelated variables the matrix is diagonal, and the variables are said to be white, when for each variable the mean is zero and the standard deviation is one. The data will then not have any first or second order correlations and the covariance matrix is a unit matrix. Specific matrix factorizations of the covariance matrix can be used for variable decorrelation, which is a typical preprocessing step.

The \( k \)th standardized moment is \( \frac{\mu_k}{\sigma_k} \), which is a dimensionless quantity and is independent of any linear change of scale. The fourth standardized moment \( \frac{\mu_4}{\sigma_4} \) is called kurtosis, and measures the heavy tails of the probability distribution. Compared to the exponential distribution, a heavy-tailed distribution goes slower to zero far around the mean. An adjusted value \( \kurt (A) = \frac{\mu_4}{\kurt^2} = \frac{\mu_4}{\sigma_4} - 3 \) is the (excess) kurtosis, and it is positive for supergaussian distributions and negative for subgaussian distributions. The normalization factor \(-3\) comes from a simplification based on unit-variance [Hyvärinen and Oja, 2000]. Thus the absolute value or the square of the (excess) kurtosis is a possible measure of nongaussianity, but the rapidly growing function is sensitive to outliers [Mardia, 1970]. More robust approximations of nongaussianity based on higher-order statistics are discussed in Section 3.1.4.

### 3.1.3 Central Limit Theorem

Computational modeling is based on observations from which statistical properties are estimated. The law of large numbers states that statistical properties such as averages can be estimated accurately from a finite but large enough sample set. An important property of random variables is that the mean of sufficiently large number of random variables will be approximately normally distributed under certain conditions [Rice, 2010]. The theorem is not valid, however, if the variables have distributions with infinite variances, such as the power-law distribution. In this case the mean is not necessarily normally distributed and is related, for instance, to economical modeling. The Central Limit Theorem (CLT) provides the intuition that motivate the ICA algorithms [Hyvärinen et al., 2009]. The CLT intuitively states that a linear mixture of informative random variables with non-Gaussian probability distributions becomes less informative and more Gaussian.

### 3.1.4 Entropy and negentropy

In information theory, there are several entropy-based concepts, such as negentropy and self-information, which quantify information content. The quantification of information has provided important theoretical results in several fields and applications,
such as communication, compression and neurobiology. The basic measure, entropy, measures the randomness of a random variable

\[ H(A) = \mathbb{E}[I(A)] = -\sum_{a \in A} p(a) \log_b p(a) \] (3.4)

where the self-information \( I(A) = -\log_b P(A) \) is measured in bits (\( b = 2 \)) or nats (\( b = e \)). For this dissertation, negentropy

\[ J(A) = H(Z_A) - H(A) \] (3.5)

is more interesting. \( Z_A \) is a random Gaussian variable with the same mean and covariance matrix as \( A \). The measure is always non-negative and zero only for a Gaussian variable with the same mean and covariance matrix as \( A \). Therefore it is a viable measure of non-Gaussianity, i.e., distance to normality. Within the context of compression, it represents the amount of information that can be saved when \( A \) is represented in an efficient way, for instance, in terms of a parametric distribution. However, computation of negentropy would require the estimation of \( p(a) \) and simpler approximations are more practical.

**Approximations of negentropy**

Truncated polynomial extensions of higher-order moments are the classical method for approximating the negentropy of a random variable \( A \) with zero mean and unit variance [Jones and Sibson, 1987]:

\[ J(A) \approx \frac{1}{12} \mathbb{E}[A^3]^2 + \frac{1}{48} \text{kurt}(A)^2. \] (3.6)

However, it is not considered a very robust measure against outliers and approximations of the form

\[ J_G(A) \approx \sum_{j=1}^{m} k_j [\mathbb{E}[G_j(A)] - \mathbb{E}[G_j(Z)]]^2 \] (3.7)

can be used instead. The variable \( A \) is assumed to be zero mean and have unit variance, and \( Z \) is a Gaussian variable also with zero mean and unit variance. With the simplest case of only one function (\( m = 1 \)), the measure is maximised when \( \mathbb{E}[G(A)] \) is maximized. The functions \( G_j \) have to conform to some constraints, for instance, they should grow no faster than quadratically [Hyvärinen, 1998]. This clearly makes the approximation computationally less sensitive to outliers than the kurtosis-based one in Equation 3.6. The functions can be chosen to capture some known non-Gaussian properties, such as asymmetry and sparseness, by basing them on the log-densities of some known distributions with those properties. For instance,
the functions
\[
G_1(A) = \frac{1}{\alpha} \log \cosh \alpha A, \quad 1 \leq \alpha \leq 2 \\
G_2(A) = -\exp(-A^2/2)
\] (3.8)

have proved very useful [Hyvärinen et al., 2001] and have been used with the FastICA algorithm, as discussed in Section 3.3.7.

### 3.1.5 Sparseness

Sparseness is a property of a random variable. In contrast to a Gaussian random variable, a sparse random variable takes more often both very small and very large values. These represent the fat tails and the high peak of a supergaussian distribution, shown in Figure 3.1. To compensate, the variable has to take middle values less often than a normally distributed variable. A sparse signal can be thought to be more interpretable as it activates rarely instead of providing constant activations at various levels. Similarly, each instance of a high-dimensional but sparse multidimensional random variable would only have a few active components, making it easier to interpret.

The Laplace distribution is a typical example of a sparse distribution and Bayesian models can induce sparsity with the Laplace prior. The most interesting property is that the random variable is activated, i.e., is significantly non-zero, only rarely, as illustrated with random signals in Figure 3.2. As the probability distribution of a sparse variable is supergaussian, higher-order moments such as kurtosis are related to sparsity.

A computationally cheap method to obtain a sparse representation is to threshold an already existing representation [El Ghaoui et al., 2013]. Also, variable selection methods find a form of sparser representation by completely removing some variables. These typically have the downside of increasing the reconstruction error, and several machine learning methods include sparsity in the model with regularization, such as LASSO [Tibshirani, 1996] for regression and sparse non-negative coding [Hoyer, 2004] for non-negative matrix factorization (NMF). Similarly, sparseness is the key statistical concept in sparse coding [Olshausen and Field, 1996] and compressed sensing [Donoho, 2006b], which are frameworks for finding a sparse representation of data.

Sparsity can be measured, for instance, with sparsity inducing norms. The pseudo-norm, \( l_0(x) = |\{i : x_i \neq 0\}| \), counts the non-zero entries and has frequently been
Computational methodology

Figure 3.2. An example of a normally distributed random signal (top) and a sparse signal (bottom) with zero means and the same variance. The sparse signal gets values near zero more often and correspondingly has few absolutely high values, which show up as spikes.

used in classification. The more flexible $l_1(x) = \sum_i ||x_i||$-norm induces sparseness under certain conditions [Donoho, 2006a] by punishing values that deviate from zero.

3.1.6 Correlation and independence

Correlation and dependence measure the relationship between two or more random variables. Statistical observations of co-occurring events are one of the methods for making associations and drawing conclusions, but significantly correlation does not imply causality or provide the direction of the causal effect.

Correlation, or linear dependence, between two random variables suggests that one variable can be predicted with a linear model from the other. Cross-correlation introduces an additional time-delay for variables with time signature. The amount of
correlation can be measured, for instance, with Pearson’s correlation coefficient

\[ \rho(A, B) = \frac{\text{cov}(A, B)}{\sigma_A \sigma_B}. \] (3.9)

Variables with zero covariance are said to be uncorrelated. The standard analysis methods, principal component analysis and singular value decomposition, represent the correlated variables in terms of new uncorrelated variables. This is illustrated in Figure 3.4 on page 54.

Independence is a stronger assumption than uncorrelatedness, as it indicates that an observation of one variable does not affect the probability of the other variable. Two random variables are independent if and only if the joint probability distribution can be factored into marginal distributions:

\[ p(A, B) = p(A)p(B). \] (3.10)

Independence of variables can be measured with mutual information (MI)

\[ I(A; B) = H(A) - H(A|B) = H(B) - H(B|A) = H(A) + H(B) - H(A, B) \] (3.11)

which states how much information is shared between the variables. Equation 3.10 and Equation 3.11 can be extended into any number of variables.

Uncorrelated variables are a special case of independent variables as independent variables are always uncorrelated, but the other way around is not true. Because correlation coefficients only detect linear dependencies. Figure 3.7 on page 62 shows an example of two correlated variables which are projected onto uncorrelated variables first and then onto independent variables. Additionally, independent random variables satisfy the property

\[ \mathbb{E}[h_1(A)h_2(B)] = \mathbb{E}[h_1(A)] \mathbb{E}[h_2(B)], \] (3.12)

for all functions \( h_i \). It is utilized in the derivation of the FastICA algorithm in Section 3.3.7.

### 3.1.7 Temporal coherence

Temporal data proposes an order to the observations, which is present in several natural observations, for instance, in a speech signal and frames in a video. The concept of coherence is related to stability and predictability. A simple idea is to assume that outputs of target values in subsequent time points should be ‘coherent’. For instance, autocovariance for a stationary signal, \( C_{AA}(\tau) = \mathbb{E}[(X(t) - \mu)(X(t + \tau) - \mu)] \),
measures covariance of the signal between a $\tau$-delayed version of itself. Figure 3.3 illustrates a random signal with temporal coherence. For temporal data, the concept of autocovariance is related to correlation but using a time delay and a single variable instead of two variables. Models based on temporal coherence add nonlinear functions to catch higher-order statistics.

![Figure 3.3](image)

**Figure 3.3.** An example of a random normally distributed signal without temporal coherence (top) and a signal with temporal coherence (bottom) in which the variance at each point is controlled by a smooth, continuous signal. Both signals have equal variance.

Physically, the existence of continuity can be intuitively motivated by inertia, conservation of mass and momentum. In language, some properties such as a word sense, topic or sentiment could be assumed not to change (or to change slowly) during an utterance. For instance, in word sense disambiguation, the *one sense per discourse* assumption is often applied to finding different representations for the different senses of the word [Gale et al., 1992]. A related idea is signal decay in cache-based $n$-gram language models which gives greater influence to recently seen words [Clarkson and Robinson, 1997].

Hinton [1989] proposes temporal coherence as a learning principle in backpropaga-
tion neural networks, which is implemented by Mitchison [1991] and Földiák [1991]. The idea behind temporal coherence has also been introduced as slow feature analysis [Wiskott and Sejnowski, 2002] and temporal stability [Kayser et al., 2001, Körding et al., 2004]. In this dissertation, Section 4.4 describes how temporal coherence learns cognitively meaningful representations from video.

3.2 Linear models

The purpose of a mathematical model is to provide an explanation and formalize the theory behind the studied phenomenon. A model is based on a set of assumptions, for instance, a description of the underlying probability distributions and how the observations are generated. One expected aspect of a model is predictive capability, i.e., it should generalize to previously unseen observations instead of simple memorization of the history and description of the properties of the observations. In general, the parameters of the model are estimated from a sample of observations, which are assumed to originate from the underlying population.

3.2.1 Linearity

The linear aspect of a model may refer to different things, such as the structure of a linear system or the mathematical form of the final objective function. Here a linear model is described as a system based on the use of a linear operator which satisfies the superposition principle, which states that the output is directly proportional to the input.

In reality, many studied natural systems in engineering, physics and other sciences are inherently nonlinear because of the sheer complexity. Nonlinear models are more expressive, but there are often computational challenges. However, linear modeling is often the first step when a system is examined in a data-driven manner. At the least, linearization provides a local approximation which is both practical and can be more easily interpreted. The application of simpler linear models to complex nonlinear problems would cause errors in term of bias, whereas too complex models tend to lead to higher variance and overfitting. This bias–variance tradeoff is discussed more in Section 3.3.1.
3.2.2 Linear system

Let the linear system be defined by a stochastic vector $x$ for which there are observations to compute statistical estimates. In a supervised setting the general model $y = f(x; w)$ would try to explain the output $y$ using a model family $f$ with parameters $w$. The explanatory power of the model is measured by the error $e(f(x, w), y)$ between the prediction of the model and the known output value.

A linear system satisfies the superposition principle and can be modeled with a set of linear equations, described compactly with matrix algebra: $y = Wx$, where the transformation matrix $W$ contains the coefficients of the linear equations. For supervised models, the error between the predicted output and the known output for given input provides a function to be optimized for the optimal coefficients $W$. The error is often measured in the least squares sense by minimizing the sum of the squares of the errors: $\sum e^2$.

For unsupervised learning algorithms, the wanted output $y$ of the system is not given but may be inferred by the structure that can be found by modeling the data directly. For linear models, the output would be the reconstruction $\hat{x}$ of the data with an analysis step $s = Wx$ followed by a synthesis step $y = As$:

$$y = \hat{x} = As = AWx.$$ \hfill (3.13)

The analysis step maps the observations onto features $s$, whereas the synthesis step tries to reconstruct the observation from the internal representation as a superposition of basis functions $a_i$ scaled by latent activations $s_i$ in the internal representation:

$$\hat{x} = \sum_i a_i s_i.$$ \hfill (3.14)

Now the reconstructed input can be compared to the original input. However, the optimization now includes both the parameters $A$ and the unobserved latent variables $s$. The two phases can also be considered as encoding and decoding processes, such as compression and decompression. The latent representation is a very typical approach with unsupervised methods, such as the k-means clustering [Lloyd, 1982] and the Expectation-Maximization algorithm [Dempster et al., 1977].

3.2.3 Input features

The input features play an integral part in any data-driven model. If the input consists solely of noise or irrelevant information, a supervised model cannot be expected...
predict the output. Typically only a small set of all available features can be included for computational reasons. Feature selection [Guyon and Elisseeff, 2003] is even more crucial for unsupervised models, as statistically strong but irrelevant input features can cause the model to focus on those instead of more interesting features. Manual data selection can often lead to implicit feature selection which is not always taken into account. Statistical inference of the model parameters can also include automated feature selection.

The input can be raw signal measurements or processed higher-level features and several applications rely on specific hand-crafted features based on extensive research and domain knowledge. Examples of such features include Mel-frequency cepstral coefficients in automatic speech recognition [Mermelstein, 1976], as well as texture and histogram features in computer vision [Li and Allinson, 2008]. In text processing, the output of linguistic parsers and analyzer provide very high-level features. Especially for lower-level features, standard preprocessing steps, such as feature combination, normalization or weighting of the input is often integral in modeling. For instance, several natural phenomenon are not linear but exponential in nature, but logarithmic scaling makes a simple linear model suitable [Newman, 2005]. Additional weighting can be based on the theory of how the data is generated and what is interesting. Similarly, nonlinear combinations of features can contain relevant information that a simple linear model cannot capture.

### 3.2.4 Objective function

The previously defined, typically multidimensional model, does not yet explain how to estimate optimal model parameters. Learning is often based on a scalar-valued objective function which can be minimized or maximized with machine learning. In supervised learning the objective is often to minimize the average or cumulative error between the predictions of the model and the desired output values. For unsupervised methods the learning could rely on the optimization of a function based, for instance, on the statistical measures such as information or sparseness in order to make the search problem more unique.

The model and the objective function derived from the error can contain additional constraints, for instance, on the desired properties of the model parameters. Given an objective function with suitable properties, machine learning techniques can be applied to learn optimal values for the parameters of the model (more details below).
3.3 Machine learning

Machine learning attempts to explain empirical data without explicitly given models. Within the probabilistic framework this can be understood as inferring the underlying probability distribution from a number of samples. Similarly, the fit between the data and the model can also be considered to be based on the discrepancy between the data and the predictions made by the model. In both cases the goal of machine learning is to find parameters which optimize an objective function that describes the fit between the model and the observations.

Parametric models and distributions, such as the Gaussian distribution, are important in machine learning as the shape is defined by a few parameters instead of describing the whole probability distribution. Independent component analysis is an unsupervised generative machine learning model. An unsupervised model estimates the probability distribution of the data $p(x)$ without known label or output information attached to the samples $x$ [Hinton and Sejnowski, 1999]. A generative model allows the generation of new samples with the model, whereas a discriminative model $p(y|x)$ focuses on generating the output $y$ for the observations.

In this section, the relevant machine learning methods for this dissertation are briefly described. The important unsupervised principal component analysis and singular value decomposition models are based on second-order moments and can be solved with eigenvalue decomposition. They can be used for dimensionality reduction and for finding uncorrelated components, which are typical preprocessing steps in independent component analysis and other data-driven analysis methods. There is no general analytical solution for independent component analysis and numerical optimization techniques (with iterative algorithms) need to be applied. This section describes the popular FastICA algorithm and its derivation based on the previously described statistical concepts.

3.3.1 Learning algorithms

The supervised modeling problem can be represented as an optimization of an objective function based on the loss $L(\theta, \hat{\theta})$, which represents the scalar cost of the approximation $\hat{\theta}$ given the underlying truth $\theta$. For unsupervised modeling, there is generally no underlying truth and the objective function has to be based on some other measure, for instance, the maximization or minimization of some wanted property.

Many traditional analysis methods, for instance, the least squares method, principal
component analysis, and several other analysis tools can be understood to minimize the reconstruction error or expectation of the norm $l_2(x - \hat{x}) = E \left[ (x_i - \hat{x}_i)^2 \right]$. This concept of mean squared error (MSE) can be split into bias and variance terms, which are the two reasons why supervised learning algorithms cannot generalize outside the training set. Different learning algorithms have different ways of balancing the two sources of error and lead to different models.

Maximum likelihood estimation

With probabilistic models, learning can be viewed as maximization of the posterior or the likelihood of the data. Maximum likelihood estimation finds the parameters $\theta$ which maximizes the likelihood $\mathcal{L}(\theta; x) = p(x|\theta)$ over the observations $x = x_1, \ldots, x_n$. If the observations are assumed to be independent and identically distributed (i.i.d), the joint probability decomposes into a product of individual probabilities: $\prod_{i=1}^{n} p(x_i|\theta)$. Furthermore, the optimization can be done for the log-likelihood $\mathcal{L}(\theta; x) = \log \mathcal{L}(\theta; x) = \sum_{i=1}^{n} \log p(x_i|\theta)$. This is especially useful for the exponential family of distributions, because the exponential and the logarithm cancel each other out.

Usually the models do not fit the data exactly, and there is always some discrepancy between the data and the model. A perfect fit of the model to noisy data usually indicates overfitting with a model with too many degrees of freedom or parameters. Overfitting can be reduced with regularization and separate observations for improving and testing the parameters.

Optimization

Many optimization problems cannot be solved in closed form and iterative and approximative algorithms are needed. The curse of dimensionality makes it impractical in higher dimensions to try out all possible parameter values even approximately, because the number of points to consider grows exponentially with the number of dimensions. Iterative solutions can be based on the shape of the objective function. Especially for continuous and differentiable objective functions, parameter values can be improved by following the gradient. However, it is often the case that only an equal or an improved solution can be found, but it cannot be guaranteed that the global optimum has been reached, unless the function to be optimized has a specific known property or shape, such as convexity.

With a real-valued function $F(x)$, the shape of the function at a given single point can be approximated using derivatives, given that the function is continuous and the derivatives can be approximated. A local minimum or maximum of the function can
be found from an initial guess and following the gradient. Another approach is to find the stationary points where the gradient is zero. A gradient descent algorithm moves towards the negative of the gradient $\partial F/\partial x$, which gives the linear approximation for the direction of the steepest local growth. For vector-valued functions, the best local approximation is defined by the Jacobian matrix $J_F$ which contains all first-order partial derivatives $\partial F_i/\partial x_j$ of the component functions against the variables. A related algorithm is the Newton’s method, which finds the roots of a function $F(x)$ by improving the initial guess $x$ by following the gradient to the point where it intercepts the zero plane

$$J_F(x)(x^* - x) = -F(x).$$  \hspace{1cm} (3.15)

Large data sets and complex models often require further approximations for the computation of the gradient and higher-order derivatives.

### 3.3.2 Eigendecomposition

Eigenvalue decomposition represents data with linearly independent eigenvectors and corresponding eigenvalues, and provides a solution to principal component analysis and singular value decomposition. A non-zero vector $v$ is an eigenvector of the matrix $X$ if and only if $Xv = \lambda v$ where the scalar $\lambda$ is the eigenvalue corresponding to the eigenvector $v$.

Every $N \times N$ real-valued non-singular symmetric matrix has $N$ linearly independent real eigenvectors and can be factorized as

$$X = V \Lambda V^{-1} = V \Lambda V^T$$  \hspace{1cm} (3.16)

where the eigenvectors $V = v_1, \ldots, v_N$ are real and mutually orthogonal, and the diagonal matrix $\Lambda$ has the corresponding eigenvalues $\lambda_1, \ldots, \lambda_N$ in the diagonal. The eigenvalue decomposition can be numerically solved with several methods, such as power iteration or the QR algorithm [see, e.g., Saad, 2011].

### 3.3.3 Principal component analysis

Principal component analysis, also known as the Karhunen-Loève transform or the Hotelling transform, is a method of projecting a vector $x$ of variables into uncorrelated variables $z$. PCA is a classical method in exploratory data analysis and data preprocessing. The principal components are ordered so that the first principal component accounts for the direction of largest variance in the data. The subsequent principal components are orthogonal to the previous ones and account for the largest variance left. This is illustrated in Figure 3.4.
PCA can be solved with different methods. For instance, eigenvalue decomposition of the real and symmetric covariance matrix $C = \mathbb{E} \left[ (\mathbf{x} - \mathbb{E} [\mathbf{x}]) (\mathbf{x} - \mathbb{E} [\mathbf{x}])^T \right] = \mathbf{V} \Lambda \mathbf{V}^T$ is a solution if the covariance matrix has rank equal to the number of variables. A related solution is given by singular value decomposition of the centered variables $\mathbf{x} - \mathbb{E} [\mathbf{x}]$. Principal components can also be learned from the least squares perspective by finding a vector $\mathbf{v}$ which minimizes the reconstruction error in the $l_2$-norm sense: $\arg\min_{\|\mathbf{v}\|^2 = 1} \mathbb{E} \left[ \| \mathbf{x} - \mathbf{v}^T \mathbf{z} \|^2 \right]$. The biologically inspired Oja’s learning rule utilizes Hebbian learning [Oja, 1982]. More recent probabilistic models have been developed for handling partially missing observations and large data sets (see, e.g., Ilin and Raiko [2010]).

In addition to removing correlation between variables, PCA is useful for dimensionality reduction. If only the eigenvectors with the largest eigenvalues are retained, the variables left out correspond to the least variance and can be thought to be noise or other weak uninteresting signals.

![Figure 3.4. An example of removing correlation with PCA.](image)

3.3.4 Singular value decomposition

Singular value decomposition (SVD) is a matrix factorization and dimensionality reduction method which is closely related to principal component analysis (see, e.g., Manning and Schütze [1999]). Let the data matrix $\mathbf{X} = (\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(T))$ have zero sample mean. Then the singular value decomposition $\mathbf{X} = \mathbf{U} \Sigma \mathbf{V}^T$ represents the data with the eigenvectors $\mathbf{U}$ of the estimated covariance matrix $\frac{1}{T} \mathbf{X} \mathbf{X}^T$, the eigenvectors $\mathbf{V}$ of the matrix $\frac{1}{T} \mathbf{X}^T \mathbf{X}$, and the diagonal matrix $\Sigma$ of singular values. The eigenvalues $\Lambda = \frac{1}{T} \Sigma^2$ correspond to both eigenvectors. Principal component analysis projects the data onto the eigenvectors so that the new covariance
matrix $C_Y$ is diagonal for uncorrelated data $Y$:

$$X = U\Sigma V^T, \text{where } U^TU = V^TV = I \text{ and } \Sigma \text{ is diagonal},$$

$$Y = U^TX = \Sigma V^T,$$

$$C_Y \approx \frac{1}{T}YY^T = \frac{1}{T}(\Sigma V^T)(\Sigma V^T)^T = \Sigma(V^TV)\Sigma^T = \Sigma^2.$$

Reduced or truncated SVD gives a $p$-rank approximation of the data matrix $\hat{X}_p = \hat{U}\hat{\Sigma}\hat{V}^T$ by selecting only the first $p$ singular vectors and singular values, illustrated in Figure 3.5 on page 58. This approximation is optimal in the Frobenius norm sense.

### 3.3.5 Preprocessing

**Centering**

In ICA, the sample mean $\bar{x}$ can be removed without any loss of generalization, since the mean can be returned after the ICA algorithm:

$$x(t) \leftarrow x(t) - \bar{x}, \quad \bar{x} = \frac{1}{T}\sum_{t=1}^{T}x(t). \quad (3.18)$$

Typical statistical concepts, such as variance, kurtosis, and covariance matrices will have much simpler forms for zero-mean (centered) data, and these are typically assumed when algorithms are being developed.

**Decorrelation**

Observed variables are often correlated, i.e., they depict the same information. Decorrelation of variables is often an analysis method of its own, and has been shown to be useful in visualization, compression, and feature extraction. The covariance matrix for uncorrelated variables is diagonal, which makes computation of the matrix inverse extremely efficient. There are several classical methods for decorrelation, but most relevant to this work are PCA and SVD. Essentially standard decorrelation procedures solves half of the estimation with ICA [Hyvärinen et al., 2001].

**Whitening**

An ICA algorithm can also assume whitened data $Z$ for computational reasons. Whitening requires variance normalization in addition to decorrelated variables. These reduce the data covariance matrix $C_Z$ to an identity matrix. PCA can be used to compute both the variance normalization and variance decorrelation, and thus ICA can
Computational methodology

be seen as an extension of PCA:

\[
Z = \left( U \Lambda^{-1/2} \right)^T X = \Lambda^{-1/2} Y = \sqrt{T} V^T \\
C_Z \approx \frac{1}{T} ZZ^T = \frac{1}{T} (\sqrt{T} V^T)(\sqrt{T} V^T)^T = V^T V = I.
\] (3.19)

In the whitened space, uncorrelated components are independent, which makes it possible to force a set of components to be independent of each other by forcing them to be uncorrelated. The variances of the components are also set to one as the scaling properties of the linear ICA model cannot estimate the variances of the components. Thus, several independent components are found by forcing the estimated components to be orthogonal, and therefore the demixing matrix \( W \) is a rotation matrix, illustrated in Figure 3.7 on page 62. The symmetric orthogonalization

\[
W \leftarrow (WW^T)^{-1/2} W
\] (3.20)

can be obtained with eigenvalue decomposition or by an iterative algorithm.

Dimensionality reduction

The rank or intrinsic dimensionality of the data tells how many components are actually required to represent the data in exact terms. Typically it is possible even to go lower than that, because not all components contribute equally to the data or wanted information. Some component representations, such as PCA and SVD, provide inherent dimensionality reduction because the components are ordered. This is not the case with ICA, even though components can be estimated one at a time. A low-dimensional representation can be found before, during or after finding the structure.

Noise is something that one typically wants to remove, whereas very weak signals are something that explain very little of the data.

The decorrelated signals often contain weak noise and other uninteresting signals, because the co-occurrences of kept rare words are still unreliable. Retaining only the strong signals with the highest variance is a simple way to ensure that ICA concentrates on reliable data [Pu and Yang, 2006]. The decorrelated components produced by PCA and SVD are ordered in a way that the first component are responsible for most of the variance. Dimensionality reduction proceeds by selecting only the components with largest eigenvalues of singular values. The best 1-dimensional representations of the decorrelated Gaussian signal in Figure 3.6 would be the x-axis, which is the direction of largest variation.

In general, dimensionality reduction makes many algorithms faster to compute with very little loss in representation accuracy.
For instance, vector space representations discussed in Section 5.1 can have hundreds of thousands of features, but the dimension can be significantly reduced as long as the amount of signal energy is not increased significantly. Furthermore, dimensionality reduction to the extracted number of components simplifies the equations in ICA algorithms. Dimensionality reduction is also important if the algorithm assumes whitened data. Variance normalization in whitening makes all input features equally important in terms of variance. For instance, ICA would learn also very rare events without the removal or weak uncorrelated signals.

### 3.3.6 Postprocessing

Postprocessing of the representation that has been learned has aims similar to preprocessing: The representation should be as useful as possible. Linear postprocessing after linear ICA does not make sense and any postprocessing operation should therefore be nonlinear.

General postprocessing techniques for blind source separation include deflation, denoising, filtering, ranking and detection [Choi et al., 2005]. Deflation is related to learning multiple components one at a time and is usually incorporated in the algorithms, for instance, as a Gram-Schmidt orthogonalization method and normalization.

The main advantage of the ICA method over many classical analysis methods is that the ICA-induced representations are sparser for supergaussian sources. This makes it easier to both manually and automatically process the representation [Barlow, 1989]. The sparse representation are easier to interpret as only a few of all the features are active at the same time, for instance, this resembles more natural classification where typically only a small set of features are needed for defining most things. Additional computational processing can be easier, as statistically independent features are easier to modify and combine, for instance, single features can be adjusted as statistical independence says that the value of one feature does not affect the value of the other. Publication VI shows how a non-linear thresholding operation is more suitable with ICA-based representation than SVD-based representation.

### 3.3.7 Independent component analysis

The main algorithmic method in this dissertation is independent component analysis. Here the method is explained and its relation to principal component analysis and singular value decomposition is shown. For more information, see Hyvärinen et al. [2001] for a textbook description, Choi et al. [2005] for a general review of ICA, and
Hyvärinen [2013] for a review of recent algorithmic developments.

Independent component analysis is often presented as a blind source separation method, but it has also been described as a feature extraction and clustering method. As a blind source separation method, the goal in ICA is to separate a multivariate signal \( x(t) \) into additive subcomponents assuming the mutual statistical independence of the non-Gaussian latent source signals. The original idea of independent component analysis is suggested already in Ans et al. [1985] and later clarified in Comon [1994].

The mathematical description is as follows. The observed random data vectors \( x(t) \) are assumed to be generated as an instantaneous linear mixture. In vector format the equation

\[
x(t) = \sum_{k=1}^{n} a_k s_k(t) = A s(t)
\]

shows how the observations are created as a linear mixture of the independent sources \( s = (s_1, \ldots, s_k, \ldots, s_n)^T \), where the sample index \( t \) has been dropped for simplicity. The columns \( a_k \) of the mixing matrix \( A \) are the basis in which the observations are represented. The observations can be assumed to be centered (\( E[x] = 0 \)) without any loss in generalization, which also implies that the sources are centered (\( E[s_k] = 0 \)).

The task in ICA is to estimate both the components and the mixture weights when only the vectors \( x \) are observed. In general, there is no unique solution, because any matrix \( A \) of full rank could be used to represent the observations. The assumption of statistical independence of the components \( s_k \) fixes the solution within a few ambiguities derived from the linear model, which are discussed shortly.

Independent component analysis can also be interpreted as a matrix factorization method similar to PCA. Figure 3.5 shows how principal component analysis can be used as a preprocessing step before independent component analysis, especially if the ICA algorithms require decorrelated or whitened data.

**Figure 3.5.** Matrix factorization with SVD (top) and ICA (bottom). Zero-mean data can be whitened with SVD (or with PCA) before it is analyzed with ICA. The dashed lines show the possibility of dimensionality reduction with SVD.
Identifiability

The ICA model in Equation 3.21 is linear, which makes the independent components identifiable only to a degree and the sign, scale and order of the components cannot be estimated. Moreover, the independent components can be estimated if at most one of the components is Gaussian and the number of observed mixtures must be at least as large as the number of estimated components. The latter requirement is equivalent to saying that the mixing matrix is invertible if it has full rank.

The linear mixing indicates that the order of the components is arbitrary. Any reordering of the components with a permutation matrix \( s' = Ps \) (\( P \) has a single one on each row and each column, i.e., \( P^T P = I \)) can be countered by the same reordering of the columns of the mixing matrix to generate the same observations:

\[
x = A's' = (AP^T)(Ps) = A(P^TP)s = AIS = As. \tag{3.22}
\]

Principal component analysis and singular value decomposition are also linear, but the eigenvectors are ordered by the magnitudes of the eigenvalues. In contrast to PCA, the order of the independent components varies for each run of the algorithm. This should be taken into account when the ICA representation is evaluated, for example, the consistency of the results can be verified with multiple runs of the algorithm [Himberg and Hyvärinen, 2003]. For this dissertation, the reported qualitative scores are average scores over multiple runs of the algorithm.

Similarly, any scaling \( \alpha_k s_k \) (\( \alpha_k \neq 0 \)) of the components can be matched with its inverse in the columns of the mixing matrix:

\[
x = A's' = \sum_{k=1}^{n} (\alpha_k^{-1} a_k)(\alpha_k s_k) = \sum_{k=1}^{n} a_k(\alpha_k^{-1} \alpha_k)s_k = \sum_{k=1}^{n} a_k 1s_k = As. \tag{3.23}
\]

A partial solution is to set the variances of the components to a specified value, such as \( \text{Var}[s_k] = E[s_k^2] - E[s_k] = E[s_k^2] = 1 \), which simplifies the algorithms. This still leaves the signs of the components to be ambiguous. The ambiguities in the order and signs of the components have to be taken into account in the analysis of the estimated components.

ICA algorithms

Several approaches and algorithms have been developed to solve the ICA problem (see, e.g., Acharya and Panda [2008]). Most of the algorithms implement numerical solutions, but the algebraic ICA [Yamaguchi and Itoh, 2000] gives a closed-form solution, but it is feasible only for separating two sources. The algorithms measure independence or higher-order information in various ways. An InfoMax-based
Figure 3.6. Examples of the preprocessing stages in ICA for signals with different distributions. Each plot shows the estimated probability density of a two-dimensional variable. The first row is for a uniform, sub-Gaussian density, the second is a Gaussian density, and the third is a sparse signal with a super-Gaussian Laplacian distribution. In the first column are the original (centered) signals, in the second are the decorrelated signals, in the third are the whitened signals, and on the fourth are the estimated independent components. Decorrelation orients the axes to the directions with most variance in the second column. Whitening makes all dimensions have equal variance in the third column. Independent component analysis rotates the whitened signals in the fourth column and thus serves as the final step in revealing the underlying basic structure of the data.
ICA algorithm [Bell and Sejnowski, 1995] is based on entropy, whereas tensor-based approaches, such as JADE [Cardoso, 1998], are based on higher-order cumulants. Kernel-based approaches move the search from input space into non-linear feature space, where lower-order methods such as PCA can be applied to finding independent components [Bach and Jordan, 2002]. One very popular algorithm that maximizes nongaussianity and takes advantage of the Central Limit Theorem is FastICA [Hyvärinen and Oja, 1997]. It is applied in this dissertation and described in detail in Section 3.3.7.

The task of estimating both the unknown mixing matrix $A$ and the independent components is often represented as the inverse problem of finding the demixing matrix $W = (w_1, \ldots, w_n)^T \approx A^{-1} = A^{-1}$ which maximizes the independence of $y$:

$$y = Wx.$$ (3.24)

Maximization of the independence of $y$ has been approached with higher-order cumulants and gradient-descent of non-linear activation functions. Algorithms have been derived, for example, with maximization of nongaussianity, minimization of mutual information, maximum likelihood estimation, maximization of information flow and tensorial methods.

The link between independence and nongaussianity can be derived intuitively with the Central-Limit Theorem. The estimated independent component $y$ is a linear combination of the actual independent sources $s$ with weights $q$:

$$y = w^T x = w^T (As) = (A^T w)^T s = q^T s.$$ (3.25)

While the CLT is defined only at the limit, in practice a sum of even a few variables with any distribution looks decidedly more like a Gaussian distribution [Raol, 2015]. In fact, the estimate $y$ in Equation 3.25 is least Gaussian when it is one of the independent components $s_i$ and only one of the elements in $q$ is nonzero [Hyvärinen and Oja, 2000]. The search for independent components is then reduced to the search of maximally non-Gaussian components, which can be done with classical measures of nongaussianity such as kurtosis and approximations of negentropy. Under some additional assumptions, simple and robust approximations of nongaussianity are based on the maximization of

$$J(y) \propto E[G(y)]$$ (3.26)

where $G$ is practically any non-quadratic function, such as

$$G_1(y) = \log \cosh y \quad G_2(y) = -\exp\left(-y^2/2\right)$$ (3.27)
which have more convenient statistical properties than the kurtosis-related $G_3(y) = y^4$ [Hyvärinen et al., 2001]. The fourth power in the kurtosis means that outliers have a big effect on the total value. For a more rigorous justification of ICA, see, e.g., Hyvärinen et al. [2001].

Figure 3.7. An example of a finding independent components. The data is on the left. The middle figure shows that for whitened data, the PCA projections (solid and dotted lines) are aligned by the axis, and the variance along each dimension is equalized. The rightmost figure shows that ICA finds a rotation after PCA and variance normalization. It aligns the data along the axis which maximizes the independence of the components.

Connection to sparseness

Independent component analysis is closely connected to sparseness and sparse coding [Hornillo-Mellado et al., 2005, Hyvärinen, 2010]. On one hand, Comon [1994] has shown that the estimation of the ICA model (Eq. 3.24) can be reduced to the search for uncorrelated directions in which the components are as non-Gaussian as possible. On the other hand, if the independent components have sparse probability distributions, which is true if the distributions are super-Gaussian [Hyvärinen and Oja, 1997], the search is for uncorrelated and sparse projections. Therefore, on the condition that the components are constrained to be uncorrelated, the ICA model estimation for sparse data is roughly equivalent to sparse coding [Hyvärinen et al., 1998]. If the distributions have some other very non-gaussian distribution, the search is for independent components which are also interesting. Section 5.2.2 gives more details of the distribution of learned components for text.

FastICA algorithm

A brief derivation of the FastICA algorithm [Hyvärinen and Oja, 1997] is given here to tie together the given background in machine learning. The FastICA algorithm for one component can be derived as a fixed-point iteration for the maximization of an approximation of negentropy. The goal is to find a unit length weight vector $\mathbf{w}$ which maximizes the nongaussianity of $\mathbf{w}^T \mathbf{x}$, where the data $\mathbf{x}$ has been whitened, i.e., it has zero mean, decorrelated variables and unit variance. The removal of the data mean is trivial and can be restored after the linear model has been found. The requirement of decorrelation can be accomplished with several standard methods, such
as principal component analysis, and the original structure can be restored with the pseudoinverse matrix even when the dimensionality has been reduced. The equalization of the variances for all dimensions has an effect on the learned components, but adverse effects can be minimized by removing the dimensions with the smallest variances.

The maxima of the approximation of negentropy in Equation 3.7 under the constraint $E[(w^T x)^2] = ||w||^2 = 1$ are found at the optima of

$$E[G(w^T x)]$$

(3.28)

With Lagrange multipliers, the solutions are found at points where

$$F(w) = E[xx^T g'(w^T x)] - \beta w = 0$$

(3.29)

where $g$ is the derivative of $G$ and $\beta$ is a constant [Hyvärinen and Oja, 1997]. A diagonal approximation of the Jacobian matrix of the objective function in Equation 3.29

$$J_F(w) = E[xx^T g'(w^T x)] - \beta I \approx E[xx^T] E[g'(w^T x)] - \beta I = E[g'(w^T x)] I - \beta I$$

(3.30)

can be easily inverted. Newton’s method in Equation 3.15 with the Jacobian approximation gives the fixed-point step for computing the new direction for $w$

$$w^* = w - \frac{E[xx^T g'(w^T x)] - \beta w}{E[g'(w^T x)] - \beta} = \frac{E[xx^T g(w^T x)] - E[g'(w^T x)] w}{E[g'(w^T x)]}$$

(3.31)

This gives the one-unit FastICA algorithm for whitened data $x$:

**Require:** Data $x$ is whitened.

Choose an initial normalized weight vector $w$

while not converged do

Let $w^+ \leftarrow E[w g(w^T x)] - E[g'(w^T x)] w$

Normalize $w^+ \leftarrow w^+ / ||w^+||$

end while

where converge can be measured, for instance, by the change in the angle of the weight vector. It should be noted that both $w$ and $-w$ maximize the objective function if $G$ is symmetric. The function $g$ is the derivative of a non-quadratic nonlinearity $G$. Some examples are given in Equation 3.8.

All independent components can be found iteratively or at the same time. A deflationary algorithm finds the components one by one and restrict each new one to be orthogonal to all previously estimated weight vectors, for instance with Gram-Schmidt orthogonalization. This finds independent components because the data is whitened in which case decorrelation implies independence. The components can be
found also with a symmetric algorithm which stores all estimated weight vectors to a demixing matrix $W$ and uses the matrix form of the FastICA algorithm [Hyvärinen et al., 2001]. The normalization step will force the set of weight vector to be orthogonal as well as of unit length.

The FastICA algorithm has been proven to be a useful tool for signal processing in many fields of research. It is reasonably simple and has fast convergence. The algorithm has been derived from different perspectives: maximization of nongaussianity, minimization of mutual independence, and maximum likelihood estimation [Hyvärinen et al., 2001] It has also been shown to be a special case of the denoising source separation framework [Särelä and Valpola, 2005].
4. Models of natural images

This chapter considers unsupervised models of natural images and video based on specific statistical concepts which learn from image statistics, similar to how the visual system is receptive to the statistics of input it receives [Hubel, 1995]. It is shown how simple linear models with unsupervised learning can find representations with major characteristics similar to specific neuronal representations by considering only the statistics of natural images. The results of this specific and well motivated application of the independent component analysis method are reasonably easy to evaluate because there is a direct correspondence to the measured function of the visual cortex. These studies provide a starting point when independent component analysis is applied in the language domain in Chapter 5, where the best practices learned from the image domain give an excellent starting point for further experiments.

Naturally, not all computational learning methods produce similar representations, as the data and the selected method influence the output, especially for unsupervised methods. Here the visual sensory information is substituted by leveraging existing digital images and movies in the computational experiments. The statistical concepts of independence, sparseness and temporal coherence, introduced in Section 3.1, have been successfully applied to finding representations from image data in both the spatial and the spatio-temporal domain [Olshausen and Field, 1996, Bell and Sejnowski, 1997, Hurri and Hyvärinen, 2003]. Unsupervised computational models based on these concepts learn properties similar to known neuronal representations, whereas methods based only on simpler second-order correlations result in different characteristics [Hancock et al., 1992]. Additionally, structures similar to the cortical organization in the brain can be induced by allowing some dependencies [Hyvärinen and Hoyer, 2001].

In this chapter, the relationship between the digital visual data representation and sensory information is discussed. The linear generative model for the data is introduced and the representations based on the studied statistical concepts are visualized.
The contributions of this dissertation to computational neuroscience are extensions of previous research results. Publication I shows how a representation with characteristics similar to independent component analysis and measured spatiotemporal receptive fields in the visual cortex are found with temporal coherence. This is the first time temporal coherence is used to model spatiotemporal receptive fields. Publication II proposes a unified framework which combines the principles of sparseness, temporal coherence and topographic organization into a unified model. It provides a hypothesis why similar representations can be learned with different computational principles.

4.1 Image representation and visual modeling

The visual system in the brain processes incoming information to infer knowledge about the environment. The sensory information represents three-dimensional physical objects projected onto two dimensions, which the brain has to analyze to infer the state of the real world. The lower levels of the visual cortex are quite well mapped and the functionalities of different levels are known. Computational models which intend to imitate the visual system require reasonable approximations of what kind of sensory input is received and how it is processed. Fortunately, an eye and a camera operate on the same principles of light focused with lenses and captured with photosensitive elements on a surface. Thus, computational experiments are greatly simplified by substituting actual physical sensory information with available digitized images or video. This work considers only the lowest level of the visual processing which are well mapped, and therefore the computational results are more easily compared to the measured properties in the literature.

4.1.1 Early visual processing in the brain

The visual processing in the brain begins with the eyes from which the signal is carried to the visual cortex. The eyes capture light and color information with the cone and rod receptors on the surface of the retina, which is then passed through neural connections into the brain. Nearly all visual sensory information enters the visual cortex through the area V1, located in the occipital lobe at the back of the brain. This region has been extensively studied (see, e.g., Hubel [1995]) and the responses of individual cells to visual stimuli have been mapped in cats as early as 1950’s [Hubel and Wiesel, 1959]. The neurons in the V1 area, classified into simple cells and complex cells, receive input from the retina which covers only a
Models of natural images

small part of the whole field-of-view. The neurons respond best to elongated stimuli, usually described as bars and edges. The simple cells are sensitive to the orientation, position and size of the input. The complex cells are partly invariant to some of the characteristics, such as location or orientation, thus making the neural wiring more complicated than with simple-cells [Hyvärinen and Hoyer, 2001]. From a research perspective, it is interesting to see computational principles which result in similar characteristics exhibited by the neurons in the visual cortex as it can provide evidence or further insight of how the brain operates.

The simple cells constitute one of the first levels of pattern recognition and act as receptive fields, whose activations are passed on as input for further processing. In essence, the task is to represent the data in a way which is useful for additional processing at the higher levels of cognition. The neurons in the visual systems learn to pick up important details and to reduce irrelevant ones. The simple-cell receptive fields have been shown to act similarly to Gabor transforms, which are local oriented edge and bar detectors [Hubel, 1995].

Figure 4.1 illustrates measured properties of spatio-temporal receptive fields in adult cats [DeAngelis et al., 1993]. The properties measure, for instance, directionality, spatial size and temporal duration. They provide a quantitative basis for evaluating the results of the computational models.

The statistical properties of the visual system are strongly influenced from what is seen during the critical learning stage [Hensch, 2004], but it is also adaptive to further changes in the statistics [Dragoi et al., 2000]. It is natural to think that the system will not tune to process information which is not present or is not considered important at higher levels. The former was demonstrated in an experiment where kittens which were only able to see horizontal bars learned simple cell organization that only activated with horizontal input and not with vertical input [Hubel and Wiesel, 1959].

In general, the current view of the functionality of sensory neural networks emphasizes the relationship between the structure of the cells and the statistical properties of the input they process [Field, 1994]. Moreover, it has been shown that simple-cell like receptive fields emerge when natural image data is processed with specific computational methods and principles, such as sparse coding [Olshausen and Field, 1996], independent component analysis [Bell and Sejnowski, 1997] and temporal coherence [Hurri and Hyvärinen, 2003], which are considered in this work.
4.1.2 Computational modeling of images

Existing natural photographs are typical data that has been used to emulate visual sensory input [Olshausen and Field, 1996, Bell and Sejnowski, 1997]. Very rarely the data has been gathered specifically for the research, and the existing multitude of digitized photographs and video are often leveraged. Some notable exceptions are the recordings from an adult cat’s point of view [Betsch et al., 2004] and studies on human infants [Aslin, 2009] in which the data collection was done specifically for research purposes.

Abstract pictures, manufactured objects and computer-generated graphics are typically excluded because they contain different statistics. For instance, computational analysis of power spectrums shows that images of man-made environments emphasize horizontal and vertical directions compared to natural scenery [Torralba and Oliva, 2003]. These statistical differences can be utilized in the recognition of man-made objects with independent component analysis [Boutell and Luo, 2005], and
stress the importance of using natural images for the learning in computational neuroscience.

Digital images and video have been designed to fool the human visual system into ignoring the representational limitations of spatial and temporal aspects of the discrete representation with pixels and video frames. This allows computational methods to approximate photosensitive receptors in the retina with pixels. This simple model does ignore some aspects of the visual system, for instance, the irregular distributions of cones and rods on the retina. However, the object of this work is not to build a model of the complete visual system but to explore more general principles of learning. A more careful consideration might be needed when more specific properties or higher-order visual processing is studied as the choices can have more impact on the characteristics learned. For instance, higher-level processing includes as input the output activations of a lower-level process to recognize invariant or more complex shapes, such as contours [Hoyer and Hyvärinen, 2002].

The approximate visual field received by a neuron in the V1 area of the visual cortex is depicted by a small image patch $I(x, y)$ which corresponds to a small area of adjacent pixels $(x, y)$ in the natural images. The small field of view ensures that the image patches contain mostly only small portions of actual objects and not complete objects with much detail such as trees or faces, which makes it possible to approximate the images with only a small number of very general basis vectors. For simplicity, the image patches are typically rectangular and most effects such as cortical magnification and other retinal processing are ignored. Spatio-temporal patches $I(x, y, t)$ are formed by considering an additional time variable $t$ which corresponds to frames in a video.

For vector and matrix computations, each monochrome image patch $I$ is vectorized into a single column vector $x = v(I)$. This encoding does not include the location of the patch in the images nor pixel positions or the temporal time-stamp in the video. Specifically, the knowledge of neighboring pixels is not encoded and not used in the models. This ensures that the structure learned stems for the statistics of the data. The dimension of the data vector $x$ is equal to the number of pixels in the patch.

A large number of image patches are sampled from random positions in the images, which produces i.i.d. samples suitable for statistical processing. The samples will be used to learn the parameters of the unsupervised computational models which try to represent the observations. The representations learned will depend on the data and the assumptions in the model and the learning algorithm.

The field of image processing has produced several methods for general image rep-
resentation and processing. For instance, discrete Fourier transforms and discrete cosine transforms are practical and efficient for general image compression because they have known mathematical representations and a large body of theoretical work to build upon. It is known, however, that the visual cortex learns to recognize local patterns similar to another well known form, namely the Gabor transform [Jones and Palmer, 1987]. Instead of guiding the representation learned in a top-down supervised manner, this work studies what structures emerge when simple unsupervised generative models are taught with natural image data.

In contrast to the global vertical and horizontal wave patterns with discrete Fourier Transforms, Gabor transforms are local oriented edge and bar detectors [Hubel, 1995]. The orientation selectivity has been utilized in texture processing [Jain and Farrokhnia, 1990]. A spatial Gabor filter is defined as a sinusoidal wave multiplied by a Gaussian envelope signal. The real part is given by

$$gabor(x, y) = \exp\left(\frac{x'^2 + \gamma y'^2}{2\sigma^2}\right) \cos\frac{2\pi x'}{\lambda} + \phi$$

(4.1)

$$x' = x \cos \theta + y \sin \theta$$

$$y' = -x \sin \theta + y \cos \theta$$

with parameters wavelength $\lambda$, orientation $\theta$ and phase offset $\phi$ for the sinusoidal factor. The Gaussian envelope is governed by the aspect ratio $\gamma$ and the bandwidth computed from the standard deviations of the Gaussian envelope $\theta$ and the Gabor function $\sigma$. The properties of the functions are similar to those measured from simple-cells in Figure 4.1. The properties of the Gabor filters enable quantitative comparison with the representations learned with computational models and the characteristics of the simple-cells.

The raw image patches $I(x, y, t)$ are often preprocessed to make the data more sensory-like, to reduce computational costs and to remove low-levels statistics which are thought to be uninteresting. Contrast adaptation can be approximated by taking the logarithm of pixel luminosity values and by normalization. Removal of the mean and second-order statistics with whitening are typical preprocessing steps which remove the redundant static part of the images. It has has been suggested that ganglion cells in the retina perform a process similar to whitening with spherical local center-surround receptive fields [Field, 1987, Dan et al., 1996]. This not only equalizes the responses of the neurons and removes redundant information, but also produces more sparse responses [Graham et al., 2004]. The above-mentioned preprocessing steps can be applied both in the spatial and in the temporal domain. For computational purposes, the image data is often projected onto a low-dimensional space which removes high-frequency information such as noise without significant loss of signal energy. This
filtering which blurs the image input is similar to the inability to focus in objects at long distances in human infants.

### 4.1.3 Linear generative model

The simplest image generation model is a linear model, which is known to roughly approximate the characteristics of most simple cells [DeAngelis et al., 1993]. In the model, each image patch \( x \) is explained as a superposition of at most \( n \) basis vectors \( a_i \) with weights \( s_i \):

\[
x = \sum_{i=1}^{n} a_i s_i. \tag{4.2}
\]

The model is the same for both spatial image patches \( I(x, y) \) and spatio-temporal image sequences \( I(x, y, t) \), where \( x \) and \( y \) are the pixel positions and \( t \) is a time index. Different assumptions of the weights \( s_i \) and the basis \( a_i \) give rise to different linear models and learning algorithms. For instance, the independent component analysis model in Equation 3.21 assumes statistical independence of the activations \( s_i \), whereas temporal coherence in Equation 4.4 assumes the activations are statistically coherent with small time delays.

The linear generative model explains how images might be generated as a superposition of the basis images. The algorithms for learning the receptive fields usually learn filters \( w_i \) which recognize specific patterns. With the assumption that the linear system can be inverted, the stochastic coefficients \( s_i \) representing the simple-cell responses can be approximated as

\[
y_i = s_i = W x_i \tag{4.3}
\]

with filter weights in the rows of \( W \) act as the receptive fields emulating simple cells. A receptive field measures the presence of the pattern indicated by the corresponding basis image. The basis vectors are basically low-pass filtered versions of the receptive fields [Hyvärinen and Hoyer, 2001].

The basis vectors \( a_i \) can be visualized as a small number of image patches which help to encode all images by only giving the weights \( s_i \) as the new representation. This is illustrated in Figure 4.2. Similar visualization is possible for the receptive fields learned \( w_i \). An implicit set of basis vectors would contain only one active pixel per basis vector, which would contain no relations between pixels. Without any assumptions, the data could be represented with any set of basis vectors which span the image space. Therefore, some additional assumptions have to be made to make the basis vectors more unique.
Assumptions based on second-order variance implicitly assume Gaussian distributions and find basis vectors similar to the Fourier transform. Second-order basis images learned with principal component analysis contain sinusoidal frequencies similar to the Fourier transform, which are useful for lossy image compression with approximate image reconstruction in which high-frequency details and noise are removed. However, second-order information is a dense representation in the sense that it normally requires many active basis vectors to represent the original image patches and does not create an intuitive image basis which could be understood to generate natural images.

The assumption of independence can be motivated by sensory coding [Barlow, 1989], because the important information is not spread around the representation. Significantly, it enables component-wise processing, as shown in Equation 3.12. The assumption of independence typically produces sparse representations for supergaussian natural data. Sparse representations typically represent the data with only a few active basis vectors, similar to sensory encoding [Field, 1994].

4.2 Sparse models with independent component analysis

This section briefly discusses the background of data-driven sparse models for images and image sequences. The representations found serve as a comparison for the subsequent models which utilize the principle of temporal coherence.

Experiments with cats and monkeys have revealed that neurons in the visual cortex act as localized edge detectors [Hubel and Wiesel, 1959, 1968], which has led to the discussion of neurons creating a sparse representation based on information theory [Atick, 1992]. Studies with principal component analysis, which utilizes only second-order statistics, have not been able to predict similar features from natural image data [Hancock et al., 1992], whereas Olshausen and Field [1996] showed that sparseness maximization can. Similarly, independent component analysis has been shown to find also receptive fields resembling simple cells in the visual cortex [Bell and Sejnowski, 1997]. Independence and sparseness are closely related especially with the supergaussian image data [Simoncelli, 2005]. The relationship is discussed in Section 3.3.7.
Independent component analysis of small image patches finds a representation of the visual data in which each image patch can be reconstructed as a linear composition of basis images. The columns of the mixing matrix $A$ of the ICA Equation 3.21 are the image basis, and the vectors $s_i$ gives the activity of each sample image patch.

A comparison between the computed spatial ICA filters learned from natural image data and measurements of simple-cell receptive fields in the visual cortex was conducted by van Hateren and Ruderman [1998a]. They concluded that there is a good match in the properties of spatial frequency bandwidth, orientation tuning bandwidth, aspect ratio and length of the receptive field. However, they did not find a match in the spatial scale distributions as the ICA filters were limited close to the finest allowed scale by the sampling lattice.

An expansion from spatial image patches $I(x, y)$ to spatio-temporal image sequences $I(x, y, t)$ is straightforward. Several image patches at subsequent frames in the same image position are vectorized, and the computations with ICA are similar to the spatial case.

Van Hateren and Ruderman [1998b] qualitatively compared computed spatio-temporal ICA filters which were learned from natural video data to measurements of spatio-temporal simple-cells receptive fields in the primary visual cortex. They found that both are localized in space and time, bandpass in spatial and temporal frequency, tuned in orientation, and commonly selective of the direction of the movement.

### 4.3 Topographic organization of basis vectors

This is a background section. The previous section showed how the statistical independence assumption results in a representation with characteristics similar to receptive fields of simple cells with the basic linear generative model. Further representations of higher-order structure can be considered to be dependent of receptive fields with partially shared properties, for instance, to increase rotational invariance by pooling together receptive fields with varying orientations. The receptive fields in the V1 area are known to be organized on a two-dimensional lattice with slowly varying properties [Hubel, 1995]. A possible reason for the organization can be further processing, in which receptive fields invariant to location or orientation could be computed from the neighboring simple-cell activations. Similar characteristics emerge with computational models when the receptive field outputs are allowed to be partially dependent.
Independent component analysis has been extended to allow some relations between neighboring components, for instance, topographic ICA [Hyvärinen et al., 2001] assumes a fixed topology of the components and lets the neighboring component activities correlate. A neighborhood function \( h(i, j) \) is set to identify the proximity of the basis vectors \( a_i \) and \( a_j \). It introduces partial dependencies between the components. Each filter \( w_j \) can be thought to contribute to the activations of neighboring filters as \( y_i = \sum_j h(i, j)w_j^T x_i \). In the generative model this can be specified through independent random signals \( u_j \) which are joined to influence the variances \( v_i = \sum_j h(i, j)u_j \) of the components. The topology and also all dependencies between the components disappear when the neighborhood is defined with the Dirac delta function \( h(i, j) = \delta(i - j) \), showing that the topographic ICA models is a generalization of the standard ICA model.

The relaxation of the independence using a set topology results in an image basis where the major properties of the found basis vectors change gradually between neighboring basis images. This is similar to measured simple-cell organization in the visual cortex [Hubel, 1995]. More complicated models can take advantage of the topological ordering. For instance, elongated patterns have been shown to emerge from pooled outputs of neighboring receptive fields [Hoyer and Hyvärinen, 2002].

### 4.4 Temporal coherence

This section contains contributions from Publication I. In addition to sparseness, temporal coherence is another principle for learning visual representations similar to simple-cell receptive fields [Hurri and Hyvärinen, 2003]. The use of temporal coherence can be intuitively motivated by continuity in the physical world. For instance, due to physical constraints, the visual information transforms smoothly instead of objects instantly materializing, changing and disappearing. This is true both spatially and temporally. Pixels in the same position in an image sequence should therefore correlate which each other, similar to spatially neighboring pixels. Normal linear correlation in time (autocorrelation) finds filters similar to PCA, but correlation of energies, for instance, learns filters with characteristics similar to those found with sparseness or independence [Hurri and Hyvärinen, 2003].

The objective function in the FastICA algorithm in Equation 3.28 is based on the expected value of squared filter activations \( y = w^T x \) modified by a non-linear function: \( \mathbb{E}[G(y^2)] \). Temporal coherence considers temporal data \( x(t) \) and it can be assumed that there is little change in the activation levels with a small time delay \( \Delta t \) and the
objective function has the form

$$E[ G(y(t)) G(y(t - \Delta t)) ] .$$

Clearly, selecting the nonlinearity and the time delay appropriately, the objective function for temporal coherence has the form used for sparseness maximization through kurtosis or independence maximization through negentropy. The objective function, however, can be shown to work also when the static part of the image is removed and thus operate beyond sparsity [Hurri and Hyvärinen, 2003].

The nonlinearity $G$ is strictly convex, even (rectifying), and differentiable, and $\Delta t$ denotes the delay in time. The nonlinearity emphasizes large responses over small ones. A rectifying nonlinearity $G_1(x) = x^2$ could be replaced by $G_2(x) = \ln \cosh x$, which does not grow as fast and might be more robust against outliers. Similar to ICA, orthogonality constraints can be added to bound each component and to keep the filters from converging to the same the solution. The objective function in Equation 4.4 can be optimized with gradient based method.

Some spatio-temporal receptive fields found with temporal coherence are shown in Figure 4.3. Similar to the spatial case, the computed receptive fields are localized, oriented and multi-scale. They also share some properties of physiologically relevant temporal properties. Some of the receptive fields seem to be space-time separable, where the receptive field can be expressed as a product of a one-dimensional temporal profile and a two-dimensional spatial profile. The separable receptive fields have both constant and changing time profiles. The inseparable receptive fields respond to different velocities.
Quantitative measurements of receptive fields found with ICA and temporal coherence are shown in Figure 4.4. The operational definitions of the quantities can be found in DeAngelis et al. [1993]. The histograms are very similar, showing that an unsupervised method based on the statistical concept of temporal coherence finds receptive fields which share properties with simple-cell measurements from the visual cortex. Compared to physiological measurements made in adult cats, there are similarities in spatial measurements but differences in the temporal characteristics. The differences might imply that the temporal aspects of the data or the applied temporal preprocessing differ from the physiological processing.

![Figure 4.4](image)

**Figure 4.4.** Quantitative measurements of the characteristics of spatiotemporal classical receptive fields obtained with temporal coherence (black bars) and ICA (gray bars) show quite similar distributions (Publication I). A comparison with the measurements from the visual cortex in Figure 4.1 share major characteristics.

### 4.5 Unified model

The statistical structure of natural images has been modeled with different statistical properties. Publication II introduced a generative model which combines sparse coding, temporal coherence and topographic components into a single model. It is based
on sparse activation ‘bubbles’ which represent the contiguous activation of the linear filters spanning both space and time.

4.5.1 The bubble model

The unified model introduced in Publication II assumes a higher-order process which creates events that spread through time and space to create bubble-like structure similar to temporal coherence in the time domain. For simplicity, only the bubble model for spatial filters \( w_j \) is described here. Publication II derives also models for spatiotemporal filters.

The linear generative model for image sequences assumes that the filter activations \( s_i(t) \) are influenced by the underlying bubble processes \( u_j(t) \). When the process is non-zero it generates a bubble both in time and spatially along the topology of the filters. This is an extension of the temporal coherence model. The bubble influences the variance of the filter activations.

\[
v_i(t) = f \left( \sum_j h(i,j) [\phi(t) * u_j(t)] \right) \quad (4.5)
\]

\[
s_i(t) = v_i(t) z_i(t) \quad (4.6)
\]

\[
I(x,y,t) = \sum_{i=1}^{n} a_i(x,y)s_i(t) \quad (4.7)
\]

The variance signal \( v_i(t) \) is assumed to be sparse and have temporal coherence. The sparsity is generated by sampling a very sparse non-negative random process \( u_j(t) \). Temporal coherence is modeled by a convolution with a simple temporal smoothing filter \( \phi(\tau) \), for instance the low-pass Gaussian kernel \( \exp \left[ -\tau^2/(2\sigma^2) \right] \). The spatial organization of the filters is defined by the neighborhood function \( h(i,j) \) which provides spatial coherence. The scalar function \( f \) is a technical addition which allows simple approximations of the probability densities involved. For simplicity, the underlying latent signal \( z_i(t) \) can be assumed to be Gaussian noise with unit variance.

An approximation for the probability density can be derived similarly to topographic ICA [Hyvärinen et al., 2001]. Publication II shows how the log-likelihood \( \log L \) can be approximated with bubble activations \( b_i(t) \):

\[
b_i(t) = \sum_\tau \sum_{j=1}^{n} h(i,j) \phi(\tau) \langle w_j, I_{t-\tau} \rangle^2 \quad (4.8)
\]

\[
\log L(w_1, \ldots, w_n; I(x,y,t)) \approx \sum_{t=0}^{T} \sum_{i=1}^{n} G(b_i(t)) \quad (4.9)
\]

\[
G(b) = -\alpha \sqrt{b + \epsilon} + \beta \quad (4.10)
\]
in which the filters \( w_i \) are constrained to be orthogonal. The nonlinear function \( G \) depends on \( u \) and \( f \), and similar convex approximations as in ICA can be used. The stabilization parameter \( \epsilon \) avoids the singularity of the derivative, and the parameters \( \alpha \) and \( \beta \) do not effect the maximal points of the function. A gradient descent algorithm shown in Publication II can be used to optimize the log-likelihood and to learn filters based on natural image data.

### 4.5.2 Interpretation of the model

The bubble model explains the output of filter activations in space and time, integrating sparseness, temporal coherence and spatial organization of the filters. The bubble detector \( b_i(t) \) pools the energies of the filters \( w_i \) locally over space and time.

The spatial integration is governed by the neighborhood function \( h(i, j) \) for the filter topology, and the low-pass filter \( \phi(\tau) \) integrates the activities over time. The detector can be interpreted as a very simple estimator of the underlying variance process \( u_i(t) \). Different simplifications of the model in Figure 4.5 illustrate the bubbles in space and time.

When the neighborhood function for the spatial organization of the filters is ignored by using the Dirac delta function, \( h(i, j) = \delta(i - j) \), the model combines only sparseness and temporal coherence. The generative model creates temporal bubbles where filter activations are temporally coherent but are independent of the other filters. This model suggests why both maximization of the sparseness and temporal coherence of linear filter outputs results in receptive fields that have the properties of simple cells. The receptive fields are clearly localized, and sensitive to position, size and phase.

Similarly, the temporal integration function \( \phi(\tau) \) can be set to the Dirac delta function and the model learns sparse filters with topographic organization only. When both the neighborhood function and the temporal integration are ignored, the model finds only a sparse structure, similar to independent component analysis.

The log-likelihood of the bubble model can be minimized with a gradient based method described in Publication II. Examples of the spatial filters are shown in Figure 4.6 and they display characteristics similar to the models based on sparseness, temporal coherence and topographic organization.

The unified model can be expanded to spatio-temporal filters in a manner similar with independent component analysis and temporal coherence. Figure 4.7 shows the spatio-temporal filters learned and their activations for natural video patches. The positions of the filters in the topology are restricted to one dimension only for the pur-
4.6 Discussion

As shown in this section, unsupervised learning methods guided by the concepts of sparse coding, temporal coherence and independence give rise to Gabor-like representations similar to simple cells in the visual cortex. Extensions to those methods learn topographic organization [Hyvärinen et al., 2001] and subspaces similar to complex cells [Hyvärinen and Hoyer, 2001, Berkes and Wiskott, 2005]. Further expansions with, for instance, convolutional architectures and deep neural networks, can further deal with translation invariance and model higher-order features [Lee et al., 2008]. The lower levels of the deep networks often then act as feature extractors similar to the simpler models, and can benefit from an unsupervised pre-training phase.
Basis vectors

Figure 4.6. Example of spatial basic vectors estimated by the unified model from natural image sequences Publication II. The results are very similar to what can be found with previously presented models.

[Erhan et al., 2010, Le et al., 2010] or by stacking the simpler models [Le et al., 2011]. In general, unsupervised feature extraction and data representation feature extraction have a significant impact on the performance of machine learning methods [Bengio et al., 2013, Le, 2013].

4.7 Conclusions

The chapter showed how computational modeling can find representations similar to known neuronal processing units in the visual cortex by considering only natural image statistics. The shared characteristics include locality and sensitivity to position and size. It has been reviewed how the statistical properties of sparseness and temporal coherence find representation with similar characteristics for natural image patches. The temporal coherence model was expanded to spatio-temporal image patch sequences. Specifically, the representations learned with all the models using higher-order statistics resemble simple-cell measurements from the brain, whereas
Models of natural images

Figure 4.7. Example of bubbles that emerge from image sequences (Publication II). (a) A representation with one-dimensional topography was learned in order to be able to visualize the results. Each row is one filter. (b) Output of the filters for two different image sequences, coded as gray-scale values (gray=0). The vertical axis is the filter index, and the horizontal axis is the time index. One can clearly see the bubble-like quality of the activations. (c) Image sequences uses as the input in (b).

popular image analysis tools based on second-order statistics are quite different.

The presented unified model for natural images combines the statistical concepts of sparseness, temporal coherence and topographic ordering in the same model. The bubble model can be taken as a useful prior model for further visual processing, image denoising by bubble thresholding, better rate coding, improved minimum wiring length and more invariant features. Separately, sparseness and temporal coherence result in very similar receptive fields from natural image data. The unified model suggests that a possible reason is that the spatial proximity makes independent subspaces and temporally stable subspaces coincide.

Statistical independence is closely related to sparseness, especially for natural data.
The statistical notion of independence is well known to be well suited for information processing, which learns to represent the sensory information in a way that higher-level processing is easier. This is a clear sign that second-order statistical information is not sufficient for more complex models and information processing. Furthermore, the local receptive fields learned with higher-order statistics are much more intuitive to understand than the global receptive fields learned with second-order statistics. The next chapter considers a similar problem of learning a representation in an unsupervised manner in the language domain.
5. Linear generative models for words in natural language

Natural language, especially in the written domain, might seem to have quite different representation and structure from the natural images considered in the previous section, especially if classical symbolic approaches to language are compared to sensory information. There is evidence that cognitive information processing in the language and in the vision domain might share learning mechanisms that allow statistically observed elements in the environment to shape cognitive information processing (see, e.g. Kirkham et al. [2002]). This chapter shows how independent component analysis can be applied to natural language processing. In order to utilize the same statistical tools as in the vision domain, written language has to be encoded appropriately. Existing linguistic practices and theories are important information in the capture of relevant distributional information.

Text is only one of the modes of natural language, in addition to speech in spoken languages and gestures in sign languages. The two main modes, text and speech are very different from a computational point of view. Speech can be represented as a continuous signal recorded by a microphone or equivalently as a multidimensional spectral representation of the same signal. Similar to most analog signals, speech contains noise, and the underlying physical processes that produce the sound create constraints and variation which are evident in the acoustic signal even though they are not distinguished in from the perceived communication. For instance, the vowels in speech are partly overlapping and they must typically be interpreted in the context in which they appear. Similarly, the signal is significantly affected by the speaker’s dialect, age, gender and emotional status. The conversion of speech to text is its own field of research.

Written text already encodes language using discrete symbols that are generally non-overlapping, e.g. orthographic characters are combined into words and sentences which are separated by spacing and punctuation. The typical noise sources in natural speech, such as hesitation, are mostly not present in text in electronic format. Most
of the possible noise in the signal is usually ignored, but sometimes tools such as spell-checking are useful (see, e.g., [Meystre et al., 2008]). Existing texts provide a starting point for data-driven linguistic and cognitive computational analyses, because the texts have been created using shared meaningful elements. Unlike with speech signals, individual elements such as characters, words and sentences can be identified with great accuracy with simple automatic methods or rules, and the focus of the analysis can be on the content or other deeper analysis.

Several successful applications have been developed in language technology through statistics and unsupervised methods, for instance automatic speech recognition of highly inflecting languages [Hirsimäki et al., 2006] and statistical machine translation [Koehn, 2009]. However, computational processing of the textual units is not enough for understanding. It requires information which is not explicitly stated and require cognitive processing. Human-level natural language understanding can be thought to require first solving the hard AI problem, which would produce generally intelligent computer programs [Shapiro, 1992]. There are also several smaller challenges, for instance, linking inflected word forms to each other and separating ambiguous words. Moreover, the meaning of a sentence depends on the meaning of the words in the sentence and how they are combined and related to each other. On the other hand, the statistics of text can be used to infer knowledge and the relationship between textual elements.

Words are symbolic in the sense that the meaning of a word is not encoded in the representation itself and has to be stored in a separate knowledge or meaning representation. Only a very limited amount of information can be inferred from meaningful or symbolic sub-part units, such as morphemes, syllables or characters. Rather, context of the word provides information from which a possible meaning can be inferred. Distributional semantics [Cohen and Widdows, 2009] is the study of linguistic items based on their distributional properties in a large sample of language data. In essence, it infers a meaning representation from language use.

Independent component analysis is one of the general unsupervised machine learning methods which has been applied to language data. It can provide structural analysis of linguistic vector spaces which represent symbolic language tokens with contextual features from actual language use. The ICA method learns a linear model which represents the observed features as a mixture of statistically independent components. Originally, it was applied to the analysis of documents [Isbell and Viola, 1999] similar to previously developed latent semantic analysis (LSA) [Deerwester et al., 1990]. Both methods find an underlying representation for the topics, but independent component analysis has the additional benefit of finding an emergent
representation which is more interpretable, i.e., not only finding an alternative representation from a computational point of view, but also finding the underlying topics or other sources. Lately, the structural analysis of SVD, the underlying method behind LSA, has also been considered [Bullinaria and Levy, 2012]. This dissertation expands on the previous research from document analysis with ICA to word analysis, with extensions and applications. Concurrent research has also showed how ICA can be applied to the analysis of word tokens as well as sub-word elements [Rapp, 2004, Calderone, 2008b, Watts, 2012]. On a more general level, ICA has been shown to be able to find cognitively interesting structures from different domains, such as music [Hansen et al., 2005, Feng, 2008]. Other methods are briefly discussed 5.1.6.

Similar to all unsupervised methods, data selection is important with independent component analysis, if the goal is to find specific structure at wanted level of detail. Specifically, the selection of analyzed units, the used contextual features, and all additional data processing influences the outcome of any unsupervised modeling method. Thus, depending on the input, ICA can analyze general or specific phenomena in language.

This dissertation includes contributions in the development and study of the word vector spaces with independent component analysis. It also reviews and analyzes relevant research on independent component analysis on text in order to provide a larger context for the work.

This chapter is structured as follows. Section 5.1 will first discuss linguistic vector spaces with a special focus on how the construction will influence the prominent information. This is important as any unsupervised model will only represent the given data. The result is then very dependent on the data selection and processing. The WordICA method for modeling words with independent component analysis is described in detail in Section 5.2 with consideration on how to tackle different problems with unsupervised methods by data selection. The rest of the chapter includes a survey of recent applications of independent component analysis in linguistic analysis with a focus on word-level analysis. This dissertation has contributed to the development of the WordICA method and several applications.

5.1 Distributional linguistic vector spaces

This section provides of overview of how to apply distributional vector spaces from start to finish, covering and explaining the most common techniques. Texts are explicitly formed as a hierarchy of conventionalized structural elements including dis-
crete documents, paragraphs, utterances, sentences, words, characters and sequences of those. Manual or automatic linguistic analysis can provide additional structure that is implicitly present in the texts. Classical natural language processing relies on symbolic models, but distributional representations based on statistics such as frequencies provide a powerful yet simple representation that has been applied in several linguistic and cognitive modeling tasks [Turney and Pantel, 2010].

Next the underlying assumptions in linguistic vector space models are considered. The assumptions related to independent component analysis are discussed in Section 5.2.2. There, I will explain how textual processing can be used to both simplify and enrich the data before the distributional representations are formed. The resulting numerical data can be fed to computational analysis methods, such as independent component analysis. Some caveats regarding the frequencies are discussed.

5.1.1 Underlying assumptions

Statistical semantics [Weaver, 1955] is the study of word usage patterns to discern the meaning of words. It is based on the distributional hypothesis which states that words that occur in the same kinds of contexts tend to have similar meanings. The bag-of-words model [Harris, 1954] was introduced as a distributional representation for documents. The bag-of-words model ignores word order and represents each document as a multiset of words, which can be encoded as a vector with constant length. Vector space models [Salton et al., 1975] apply the distributional hypothesis and represent documents as the occurrences of words in it with the bag-of-words model. The transition from a sequence of symbols in a text corpus to a numerical matrix enables the use of linear algebra and numerical machine learning algorithms. Vector spaces can naturally be built for any units for which co-occurrences can be computed, and the models have been utilized in several fields of research, such as information retrieval [Salton et al., 1975], natural language processing [Schütze, 1993] and cognitive science [Landauer and Dumais, 1997, Petersen et al., 2010].

Distributional hypothesis

As mentioned above, the distributional hypothesis assumes that words which occur in the same contexts tend to have similar meanings. The context does not have to be restricted to the whole text, such as a document, but can include more specific contextual features such as another word in some window around the word, or a specific position in relation to the analyzed word. The validity of the distributional hypothesis has been thoroughly discussed (see, e.g., McDonald and Ramscar [2001],
The distributional hypothesis has been expanded by moving from modeling attributional similarity with word-feature matrices to relational similarity with pair-pattern matrices. For instance, the word pairs ‘mason/stone’ and ‘carpenter/wood’ fit the patterns ‘X cuts Y’ or ‘X works with Y’ [Turney and Pantel, 2010]. Instead of counting word-feature co-occurrences in contexts as an implicit is-a relation, explicit lexical patterns which relate two words together are included as pair-pattern co-occurrences. This enables the model to capture higher abstractions such as analogies, but requires tackling the emerging combinatorial problems and increasing sparsity of the data.

Bag-of-words model

The original bag-of-words model was developed in the field of information retrieval to represent a document as the frequencies of words in it [Harris, 1954]. The model makes it possible to represent and compare the similarity of two documents by the shared words, regardless of their position and the lengths of the documents. Mathematically, this can be implemented with sets or vectors. In set theory, the bag-of-words representation for a document is a multiset, which ignores the order of the words but stores the number of times each word occurs. A possible vectorial representation is based on each word having a vector, with the constraint that all vector pairs are uncorrelated. A document vector is then constructed as a sum of all the word vectors in the document.

The vectorial representation is simple and enables the use of linear algebra. The geometric interpretation is that each document is a point in a space and points next to each other are similar. The word vectors are constructed in a way that all points are equally apart from each other. Two documents can be compared by measuring the distance between their respective points.

The more general bag-of-features model describes the analyzed unit (e.g. word, document or word-pair) as frequencies of different features which co-occur with the units in some contexts. The features can identify the document, a specific word, a phrase or other information in the text [Virpioja et al., 2012]. The context defines which co-occurrences are counted for each instance of the analyzed unit. Structure, such as partial word order information can be encoded by concatenation of different contexts, for instance, separate left and right contexts [Lund and Burgess, 1996]. This representation allows the comparison of words as well as other analyzed units.

The assumption of uncorrelated word vectors can be flexed for computational reasons or in order to incorporate information of the relationship between the words. For in-
Linear generative models for words in natural language

In high-dimensional random vector spaces are nearly orthogonal [Ritter and Kohonen, 1989, Kaski, 1998, Kanerva et al., 2000]. They allow, for instance, to encode order information within the same dimensionality with shifted random vectors [Jones and Mewhort, 2007].

The bag-of-words model does not directly model word compositionality. This has been directly considered by representing word-pairs [Turney and Pantel, 2010]. Extensions based on functionality between words instead of only distances (e.g. matrices as linear functions) on word vectors has been explored with the specific cases of verb-noun [Mitchell et al., 2008] and adjective-noun modifications [Baroni and Zamparelli, 2010]. In addition to WordICA, interpretability has also been target of compositional semantic spaces [Fyshe et al., 2015].

**Vector space models**

Vector spaces represent a collection of analyzed units using the bag-of-words model to represent each unit as a vector. The geometric interpretation is that the units are points in the created vector space, and semantic similarity or relatedness between the units is measured with the distance between the respective points, i.e., a short distance between points corresponds to some form of similarity. The information in the vector space models is thus directly dependent on how the vector space model is created.

With linguistic vector spaces, the form of the representation is typically a word-feature, a word-word or a pair-pattern matrix, but other combinations are also possible. In addition to the mode of the representation, weighting and algebraic processing plays a crucial role in the formation of a linguistic vector space. Finally, the similarity measure has to be selected [Turney and Pantel, 2010].

The raw matrix of counts is often not directly usable because the underlying distributions of many linguistic elements. Zipf’s law states that the frequency of a word is inversely proportional to its rank [Powers, 1998]. The consequence is that there is a very large difference between the occurrence counts of frequent and less frequent words. If the counts are interpreted as a signal, the frequent words contain a majority of the the energy. This usually does not reflect the semantic importance of frequent function words, for instance, in information retrieval. A language-specific approach relies on excluding specific stop-words, which are problematic for the task at hand. Several language-independent statistical approaches have been proposed to reduce the impact of statistically uninteresting values, including tf-idf weighting and mutual-information based methods (see, e.g., Salton and Buckley [1988] and Bulli-
Additionally, the linguistic vector space is often processed with machine learning, in order to make the representation more usable, either computationally or in order to find structure. Finally, the similarity metric has to be selected. This has been shown to be at least partially task-specific [Paukkeri et al., 2011], but the cosine distance measure is commonly selected. Compared to the Euclidean distance, it measures the angle between two vectors, which is length invariant to the number of words in the document.

The above-mentioned selections have led to the development of several different geometrical vector space models, such as LSA [Deerwester et al., 1990], Word Space [Schütze, 1993] and HAL [Lund et al., 1995]. The WordICA method described in this Chapter finds structure from a word vector space with the independent component analysis [Comon, 1994] method. In addition to the different structural transformation, WordICA is able to find meaningful components instead of focusing only on similarity of word vectors.

The information captured by linguistic vector spaces can be influenced in all steps of the procedure: training data selection, textual pre-processing, selection of analyzed units and contextual features, computation of co-occurrences, weighting and normalization, structural transformations, dimensionality reduction, post-processing, and selection of measurement tools. An optimal choice is to make the selections depend on both the data and the application in question, while taking care to tune hyper-parameters and processing decision on a separate development set. Below some possible steps are discussed.

5.1.2 Text preprocessing

Texts are explicitly formed as a hierarchy of conventionalized structural elements including discrete documents, paragraphs, utterances, sentences, words, characters and sequences of those. The vector space models typically utilize additional processing of the text corpus and the mathematical vector space, which is especially important for the unsupervised models. The text is processed to both enrich the data and to remove information which is thought irrelevant. In this dissertation the focus is on unsupervised methods with minimal processing and texts are not enriched with additional processing.

Manual or automated linguistic analysis provides additional structure which is not explicitly marked in text, such as sub-word units like syllables and morphemes, or...
multi-word units such as collocations and chunks. Similarly categorical information (e.g. lemmas, part-of-speech tags, word senses and named entities), and relations (e.g. parse trees, thematic and grammatical relations, co-reference structures, and ontology mappings) can be added with structural analysis (see, e.g. Jauhar et al. [2015]). The structure beyond the sequence of tokens provide more details on the relations between the tokens and thus more informative contexts. However, expanding from the explicitly marked structure to augmented data is a complication as there are multiple, even conflicting, linguistic theories and tools for each level of analysis. Therefore, there is no single method for the analysis and no general gold standard for the evaluation of the results found with unsupervised methods.

Language processing methods can also be used to filter the text. The goal can be the removal of assumed unrelated information to the task, such as punctuation and typographic elements in standard information retrieval. Otherwise distinct word types can be unified, for instance, with lowercasing, stemming, lemmatization and spell checking, in order to reduce data sparseness. In this dissertation, the choice was to exclude punctuation and to lowercase the characters.

5.1.3 Data and feature selection

Statistical methods learn a model to represent the training data, and the models learned should be utilized in similar text domains as the training data. For instance, research with a targeted focus application, such as personality traits [Chung and Pennebaker, 2008] or well-being [Honkela et al., 2012], are best performed on texts related to those topics. The feature selection procedure limits the vector space and can be viewed as a form of dimensionality reduction. This dissertation focuses on general features and, therefore, more generic data sources are utilized.

The analyzed linguistic units can be, for instance, documents, sentences, phrases, words, morphemes or phonemes. The contextual features can be the same units or other information present in the data, such as part-of-speech tags, author, genre or sentiment information. A typical goal is to learn a relatively small model for the data, for instance, a latent model with a limited number of components. This implies that the analyzed units should not include units which are not interesting for the possible task, for instance, analysis of words for colors should include the words for colors but maybe no other words. Otherwise the model will first model the most frequent words which may not be of interest, and the details of the color words might not be modeled at the wanted granularity. Similarly, expert or prior knowledge of informative features can be leveraged in data selection. The possible units and features are
often restricted with stop-word-lists, frequency criteria or other pruning techniques relevant to the analysis purpose. Research related to general word analysis without specific unit selections is reviewed in Section 5.3, and research related to unit selection in Section 5.4.

5.1.4 Co-occurrences of linguistic units

The bag-of-features model was originally applied in information retrieval to represent documents with words as the features in a document space [Salton et al., 1975]. Then the model was transposed to represent words with documents as the features vector space [Deerwester et al., 1990]. Later it was discovered that paradigmatic word relations such as synonyms were better modeled with features that represent nearby words [Lund and Burgess, 1996], which led to the construction of second-order co-occurrences. A syntagmatic association (first-order co-occurrence) occurs if two words typically are in the same context, for instance, two words that are in the same phrase, sentence or document share a syntagmatic relation. A paradigmatic association (second-order co-occurrence) between two words implies that they share the same context, but do not co-occurr. For example, synonyms can easily be substituted for each other, but only one of the words is present at a time. Independent component analysis has been applied in this dissertation to both paradigmatic and syntagmatic linguistic spaces.

**Syntagmatic vector space**

A syntagmatic vector space with the bag-of-features representation can be computed for any selected units and features, as a unit-feature matrix. A typical example is a word-document matrix $X_P$, where the rows give a vector representation for words and the columns give a vector representation for the documents. In the representation, two document vectors are the more similar the more the same words occur in both the documents. This a standard representation for documents in information retrieval. The frequencies can be thought to list the most prominent words in the document, especially after the counts have been weighted appropriately to count of the average pervasiveness of the words. The raw counts can be computed by going through the contexts once and accumulating the word-feature co-occurrences in a matrix. A carefully chosen representation, such as a sparse matrix, greatly reduces the space requirements.
Paradigmatic vector space

The syntagmatic vector space does not fully capture the semantics of words, because not all related words occur often in the same documents. A paradigmatic vector space contains co-occurrences of units in some contexts, which are typically much shorter than the document. A typical example is a word–word matrix $X_S$, which computes how many times words co-occur in some context, for instance, next to each other. In this representation two word vectors are more similar if they often occur with some third word. Especially, the two words do not have to occur in the same contexts. The counts are accumulated across all texts. A typical context is a window of specific length centered on the unit word. These can be generated with a sliding window method by going through the text once and accumulating the word co-occurrences in a matrix.

From syntagmatic to paradigmatic

The above-mentioned syntagmatic and paradigmatic spaces capture different types of similarity or association between words (see, e.g., [Sahlgren, 2006]). A paradigmatic word-word matrix with document contexts can be generated from a syntagmatic word-document matrix by matrix multiplication: $X_P \approx X_S^T X_S$. The diagonal values might differ depending on how self word-occurrences are computed, but this property clearly demonstrates the main relationship between the two main word vector spaces types. Moreover, this algebraic relation explains why singular value decomposition has been shown to improve synonym modeling of a syntagmatic word-document space [Landauer and Dumais, 1997], because the singular vectors are based on the eigendecompositions of the matrices $X_S^T X_S$ and $X_S X_S^T$ with centered data.

5.1.5 Weighting and normalization

Frequent words or tokens can prevent other more interesting linguistic phenomena from being clearly present in the data. The frequency for a single word follows the power law [Powers, 1998], where a small number of words occur very often and a large number of words occur rarely. These are problematic in statistical and unsupervised approaches because a very small set of words occur much more frequently than the majority of words. Especially unsupervised modeling methods concentrate on the most frequent events, which is not necessarily where the targeted information resides [Caron, 2001, Bullinaria and Levy, 2012, Osterlund et al., 2015].

Several classically used weighting methods originate from information retrieval [Spärck-
Jones, 1972] and information-theoretic ideas, including tf-idf and point-wise mutual information. A simple logarithmic transformation of the frequencies dampens the most frequent words but does not change the ranking order. It can be also motivated with models of perception where quite many sensory events are perceived logarithmically rather than linearly [Dehaene, 2003].

In document modeling, the dominance of long texts can be removed by normalization with text length or by vector normalization. Features which are present nearly always are not informative and can be dampened. Normalization of very rare features can create modeling artifacts, because an event seen only once will get the maximum weight whereas the weight is spread with frequent events. The maximum weight will have a high value compared to other values and modeling methods may squander features for modeling the very rare events. In general, the optimal weighting method is always both data and task dependent, and therefore no single weighting method can be recommended. It is important to note that the standard cosine similarity measure includes vector normalization, but it may be beneficial to apply vector normalization independently in order to help the structural transformation model the data appropriately.

The vector space model transforms a text from a sequence of symbols to a numerical matrix and enables the use of linear algebra and numerical machine learning algorithms. When each analyzed textual unit is represented by a vector, the semantic similarity or relatedness of any two textual units in the vector space can be measured by standard distance or similarity measures, such as the Euclidean distance or the cosine similarity measure. All analyzed vectors for textual units can be collected into a matrix for further processing with standard matrix factorization methods or other machine learning tools.

In this dissertation, classical and simple weighting schemes are applied when deemed appropriate. To goal has been to use general methods in order to make the methods and tools generic. Used methods include, for instance, tf-idf weighting, PPMI-weighting and logarithmic frequency dampening. The effect of rare events has been limited by the application of mutual information-based weighting or skipping vector normalization altogether in favor of frequency dampening. Nonlinear weighting schemes based on frequency and Pearson’s correlation are also possible [Rohde et al., 627–633].
5.1.6 Structural transformations

Structural transformations take the weighted matrix and give another way of representing the same information. The goal can be smoothing the counts, dimensionality reduction, or finding latent features. Several general-purpose methods, such as principal component analysis, singular value decomposition, non-negative matrix factorization and independent component analysis have been applied to the tasks. Additionally, natural language specific topic models have been developed based on Bayesian inference and count data. Dimensionality reduction is often feasible because the intrinsic dimension of linguistic vector spaces has been reported to be very low [Karlgren et al., 2008]. Dimensionality reduction aims at an optimal approximation of the original weighted co-occurrence data which minimizes reconstruction error, whereas additional assumptions set the structure as will be discussed further down. The dimensionality reduction methods can also be seen as smoothing methods which first project the data onto the lower-dimensional latent features and then back to the original space. In addition to the standard method of removing low-energy latent features, other weighting schemes for the latent features have been shown to provide beneficial [Caron, 2001, Bullinaria and Levy, 2012, Osterlund et al., 2015].

The matrix representation of a linguistic vector space can be factorized to a linear model in which the vectors are explained as a sum of components and weighted by their activities, as illustrated in Figure 5.1. The general-purpose methods of singular value decomposition, principal component analysis, independent component analysis, non-negative matrix factorization and random projection all fit this framework. Probabilistic vector space models can also be described in this manner.

Figures 5.1. The feature-unit matrix $X$ is explained as a matrix product of the feature-component matrix $A$ and the component-unit matrix $S$. The factorization gives a lower-rank dimensionality reduction if the number of components is smaller than the original number of features (solid lines). If the dimensionality is not reduced (dashed lines), the factorization transforms each original vector (gray column in matrix $X$) as the components learned as the basis vectors. In both cases the activation of the components (gray column in matrix $S$) gives a new vector representation to the analyzed units.

The matrix factorization is not unique, because the data can be represented in any basis that spans the original data. Additional assumptions can be used to add structure
on the representation. With both PCA and SVD the component vectors are orthonormal, and they are equal for centered data. An orthonormal set of component vectors can be interpreted geometrically as a rotation. Linguistic vectors spaces are often composed as high-dimensional but very sparse matrices with count data, so they can be efficiently stored in memory with special data structures. Explicit centering of the data would remove the sparse structure and produce computational difficulties. Sometimes the mean of the data is not removed for computational reasons with the reasoning that the mean is very close to zero and can be captured by one of the principal components. PCA and SVD are important methods because of their ability to reduce dimension. However, as explained in Hansen et al. [2005], the directions of maximum variance are usually not well-aligned with human cognition. Factor analysis is a related methodology that tries to find a small number of unobserved latent variables which explain the observed correlated variables. These methods can be considered classical and general methods which have have been applied in various research fields.

Several NLP-specific methods have been developed with additional assumptions. The non-negative matrix factorization method requires that the factored matrices are both non-negative. Probabilistic factorization methods such as probabilistic LSA [Hofmann, 2000] and latent Dirichlet allocation (LDA) [Blei et al., 2002] require that the matrices are non-negative and scale to probabilities. These constraints can be very useful with positive count data or probabilistic input data, but the model estimation is often computationally more intensive than with PCA or SVD.

5.1.7 Post-processing

A vector space captures information but finding or visualizing relevant information in it might still be problematic. Additional post-processing or using a supervised learning method or an unsupervised visualization method can help to describe the captured structure. Post-processing can also exclude some of the features if they are considered irrelevant for the task [Bullinaria and Levy, 2012, Osterlund et al., 2015]. Publication VI considers thresholding small values to zero to emphasize the sparse structure learned with independent component analysis.

5.1.8 Similarity in the vector space

Finally, the selected distance or similarity measure has an effect on how the captured information is measured in the vector space. The cosine similarity and the Euclidean distance are classical but effective methods even today [Bullinaria and Levy, 2007].
There are many alternatives that can be more suitable for a specific task [Lee, 1999, Paukkeri et al., 2011] but they are not considered here because of simplicity.

5.2 WordICA

The WordICA method is the main focus of this dissertation. The initial concept is presented in Honkela et al. [2003] but the main formulation and major developments after that are introduced in the publications in this dissertation. This section covers important historical, theoretical and practical issues that I have come across in my research. WordICA applies independent component analysis as the structural transformation to a word vector space. One of the conclusions of this work is that the latent components found are much sparser and that the components match better with linguistic knowledge than with classical methods which will be discussed in detail in later in this section. In this work, it is shown that the sparse models created with ICA are, indeed, more transparent, interpretable and easier to visualize than dense models. The publications in this dissertation include development of methodology for dealing with sparse linguistic data in ICA, development of analysis methods for the emergent features and applications to different datasets. Additionally, relevant work is introduced to create a complete view of the applications and research of ICA in word vector spaces.

The WordICA method is closely linked to other word space methods [Schütze, 1993, Lund et al., 1995, Sahlgren, 2006]. In contrast to related methods, it uses independent component analysis for finding the latent structure. It can be seen as an expansion of the SVD-based methods, as the closely related principal component analysis is traditionally used as a dimensionality reduction method with ICA. Specifically, for simplicity and efficiency, ICA can be implemented as whitening followed by rotation. From a conceptual point of view, this is very important with comparison to SVD-based methods as the rotation does not change the Euclidean distances between points and any further processing relying solely on distances would show no real change [Vicente et al., 2007]. Independent component analysis by itself can also be used for dimensionality reduction, but the properties of linguistic vector spaces can make it problematic.

Next the underlying assumptions and some practical details of working with linguistic vector spaces and ICA are discussed. Then the overall WordICA research is reviewed, in which the topics are divided into general word analysis and targeted word analysis. The goal is to give an overview of all relevant research in an organized
manner and my word and related work are discussed together. General word anal-
ysis is viewed more as exploratory research which considers either a large set of
words, or limits the vocabulary mostly by frequency for computational reasons. In
targeted word analysis, a specific limited lexicon of analyzed units is selected in or-
der to extract features of expected type. Word sense induction is a special case of
targeted word analysis, where the analysis is limited only to one word. Finally, a
closely related area, sub-word analysis, is briefly discussed before considering more
application-oriented topics.

5.2.1 Independent component analysis and linguistic vector spaces

The success of latent semantic analysis in corpus linguistics has produced much re-
search on applying unsupervised mathematical models for modeling documents and
words. Independent component analysis is a blind signal separation method, that has
been successfully applied to several tasks with natural continuous signals, and it nat-
urally was one of the methods to be applied to the creation of latent document spaces
[Isbell and Viola, 1999, Kolenda et al., 2000].

Isbell and Viola [1999] were the first to suggest independent component analysis as
the method to find structure for textual data for information retrieval. The compo-
nents found were interpreted as the underlying topics. As the topic structure was
not given to the method, the components had to be analyzed manually or compared
to existing topic structures. Manual inspection of the documents could be done by
inspecting words that are most active in each component or inspecting in which doc-
uments the component is active. For instance, if the active words clearly depict some
area, such as sports, the component could be labeled accordingly. This is illustrated in
Table 5.1. The most active components for each document can be interpreted as the
discussion topics. The first columns contains professions, the second nationalities,
the third verbs in the past tense for third person singular, and the fourth superlatives.
The WordICA method follows the same methodology, but applies it to the analysis
of words instead of documents. Additionally, the selection of the computed usage
information, such as documents or neighboring words, plays a large part in what
information the underlying vector space model captures.

5.2.2 Theoretical and practical considerations

Here the theoretical and practical considerations of applying ICA to linguistic vector
space models are discussed. First of all, the nature of the input data has to be con-
sidered as it has the statistical characteristics of language data. The assumptions of
Table 5.1. Examples of components found with ICA from a word co-occurrence matrix computed from Finnish Wikipedia articles (with English glosses below in parentheses). Only the most active words for selected component are shown.

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition 1</th>
<th>Definition 2</th>
<th>Definition 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>näyttelijä</td>
<td>(actor)</td>
<td>yhdysvaltalainen</td>
<td>(American)</td>
</tr>
<tr>
<td>jalkapalloilija</td>
<td>(football player)</td>
<td>suomalainen</td>
<td>(Finnish)</td>
</tr>
<tr>
<td>kirjailija</td>
<td>(writer)</td>
<td>ranskalainen</td>
<td>(French)</td>
</tr>
<tr>
<td>poliitikko</td>
<td>(politician)</td>
<td>saksalainen</td>
<td>(German)</td>
</tr>
<tr>
<td>jääkiekkoilija</td>
<td>(hockey player)</td>
<td>brittiläinen</td>
<td>(British)</td>
</tr>
<tr>
<td>muusikko</td>
<td>(musician)</td>
<td>ruotsalainen</td>
<td>(Swedish)</td>
</tr>
<tr>
<td>laulaja</td>
<td>(singer)</td>
<td>italialainen</td>
<td>(Italian)</td>
</tr>
<tr>
<td>taidemaalari</td>
<td>(painter)</td>
<td>englantilainen</td>
<td>(English)</td>
</tr>
<tr>
<td>säveltäjä</td>
<td>(composer)</td>
<td>japanilainen</td>
<td>(Japanese)</td>
</tr>
<tr>
<td>elokuvaohtja</td>
<td>(movie director)</td>
<td>kanadalainen</td>
<td>(Canadian)</td>
</tr>
</tbody>
</table>

The linear and noiseless ICA model were already discussed in Section 3.3.7, but here they are considered from the perspective of language data.

The co-occurrence counts for linguistic elements have to be considered for multiple reasons. The count data are discrete and non-negative, which are not the expected input for classical signal processing methods such as ICA, PCA or SVD. However, it is known that several practical datasets have the same properties, and that the count values can be transformed to real values with weighting methods. Other approaches, such as the discrete component analysis method, can model the counts directly [Buntine and Jakulin, 2006]. The outcome will contain real-valued and negative values, whose interpretation has to be considered. For instance, the resulting components can have their signs changed, similar to negative images in the image analysis domain. Furthermore, the counts for word occurrences and co-occurrences vary significantly which will affect the results if not handled.
One of the challenges when working with language data is the dimensionality that needs to be handled. In contrast to SVD, PCA requires centering which turns the sparse co-occurrence matrix into a dense representation that is difficult to handle [Manning and Schütze, 1999]. Therefore, we had to develop alternative methods to avoid computational problems.

Independent component analysis assumes the components to be statistically independent which is often measured by nongaussianity of individual components. Variance and correlation utilized in PCA and SVD are second-order statistics, whereas measures of nongaussianity are based on higher-order statistics, such as kurtosis and nongaussianity. Text statistics are typically supergaussian and independent component analysis will find a sparse representation. This is discussed more below.

Independent component analysis is a linear model, which assumes that the data is generated by a linear process. Most of the interesting actual processes probably are not strictly linear, but is a common approximation that works well in practice. Outside the scope of this dissertation, the ICA model has been extended to a nonlinear model, but learning is much more computationally intensive and the risk of getting stuck in a sub-optimal local maximum is high. Another possibility is to project the data into a higher-dimensional space and compute the linear model there, similar to kernel methods (see, e.g. Harmeling et al. [2001] and Buch [2011]).

The basic ICA model assumes the mixing matrix to be invertible, i.e., the rank of the input data is at least the number of sources. This has to be taken into account with linguistic vector spaces because the intrinsic dimensionality of co-occurrences data is quite low [Karlgren et al., 2008], even though the data matrices are very high-dimensional. The overcomplete case with more sources than input features has been studied and applied with other natural signals, but the algorithms are naturally more complicated. Sparsity is one of the approaches to find a unique set of overcomplete basis [Donoho, 2006b]. In addition to just limiting the number of extracted components, dimensionality reduction is very important preprocessing step, especially for algorithms which assume equal variance of all components. With SVD and PCA, this is not a problem, because the components are estimated in order of decreasing variance. However, with ICA and language data this is important, because the original data is very high-dimensional and typically contains several very sparse, non-informative features. Whitening of the data will make all features equally important, so it is crucial that the non-informative features have been dropped before.

The noiseless ICA model assumes that there is no extra noise component. The noisy ICA model $x = As + n$ would assume additive noise for the observations, which is a
typical approach with signals from measurement instruments. A text corpus probably contains typos and other mistakes which transfer to the co-occurrences, but in general no extra noise component can be identified. Rather, the selection and weighting of features and contexts can amplify relevant statistical information. Dimensionality reduction with principal component analysis is a normal preprocessing step with independent component analysis to remove low-intensity noise components.

At most one of the independent components can be normally distributed, which is in strict contrast to several classical statistical models such as principal component analysis that at least implicitly assume Gaussian variables. Similarly, the implicit assumption of Gaussian sources in PCA can make it an inadequate method when the true sources are non-Gaussian. Figure 3.6 shows how all rotations of two whitened Gaussian variables are equivalent for ICA, i.e., the estimated Gaussian source is a linear mixture of all the Gaussian sources. Blanchard et al. [2006] propose a method that finds interesting non-Gaussian components as orthogonal to a Gaussian subspace learned. The nature of textual features found by ICA is first investigated in Isbell and Viola [1999], where it is inferred that the components found are not Gaussian. In general, many natural signals, such as sound sources, have been shown to have a supergaussian distribution [Bell and Sejnowski, 1995]. This can be relevant, as some ICA algorithms assume knowledge or estimates for the super- or subgaussianity of the signals. This is collaborated by Figure 5.2 which shows that FastICA method finds supergaussian signals, and that the signals found with singular value decomposition are much more normally distributed. This also signifies that the signals found with ICA are sparser than those found with with SVD. The representations were estimated from positive point-wise mutual information (PPMI, Niwa and Nitta [1994]) weighted word co-occurrences within short text fragments in Wikipedia text.

![Figure 5.2](image)

**Figure 5.2.** The activity distributions of ICA and SVD representations compared to a Gaussian and sparse Laplace distributions. All variables have been scaled to unit variance. The plot on the left shows how the ICA activity distribution has a higher peak. The plot on the right shows on the logarithmic scale how ICA also has heavier tails.

Many of the algorithms require sampling for the estimation of statistical properties.
The samples are assumed to be independent and identically distributed. The selection of texts for the training corpus might be very different from each other, with different distributions. Moreover, the sliding window approach in computing word-word co-occurrences might break the independence assumption. This could be remedied by sampling without overlapping context windows. However, these are not typically considered with linguistic vector space models but are generally applied in the image domain.

ICA assumes that the sources are statistically independent. In real life, this strict assumption is very rarely fulfilled exactly. However, it is not problematic even if the components are not exactly independent, since in that case ICA reduces to projection pursuit and tries to find interesting, non-Gaussian directions [Friedman and Tukey, 1974]. There are several generalizations of ICA which relax the independence assumption, typically in a local and structured manner. For instance, tree-dependent component analysis also learns a tree structure for dependencies between the components [Bach and Jordan, 2003]. For text data one can find semantic relationships between the neighboring components, similar to subtopics [Sasaki et al., 2012].

5.2.3 Evaluation

The outcome of the WordICA model includes the word-by-component mixing matrix and the component-by-context loading matrix. In the image domain, it is possible to easily visualize hundreds of basis vectors as small images. In the language domain, such visualization are not possible as each dimension takes more space than a pixel and the components learned have to be analyzed in a different manner. A classical method is to display and analyze the words with the highest (absolute) values for each component independently and manually label the topic or the content [Isbell and Viola, 1999]. In most cases, the independent components are clearly interpretable. This is quantified with multilingual texts in Section 5.3.3 by annotating the components learned and analyzing the labels.

Similarly, the signs of the components had a natural interpretation in the image domain, and changing the sign would at worst only produce the negative of the image. With WordICA, the resulting components tend to be skewed and the sign can be automatically changed by inspecting the skewness of the components. The skewness of the components can be increased with skewed nonlinearities. Clearly, changing the signs of the components does not affect the ICA model.

Publication IV describes an automatic evaluation measure in which a correlation-based score between each component and predefined lists of known topics. Even
though it is limited to the known lists, it has the possibility to speed up development. A joint measure over the maximum component-wise correlations gives a single score for the set of components. The measure is useful for analysis of both syntactic and semantic information, assuming that the lists are available.

The correlation score is very useful, but gives only a partial view of the content of the components. Publication V describes an evaluation measure, which quantifies the separability of known topics by considering pairs of components and pairs of topics. This is a generalization of the previous correlation-based score and it can, for instance, tell whether related categories such as adjectives and adverbs are mixed together in the representation or can be identified by considering two components.

In a vector space, the relatedness or similarity of analyzed elements is evaluated by considering the distance between the respective points in the low-dimensional latent space. With independent component analysis, specific care should be taken so that comparison to SVD-based methods is justified. This should be considered for all distance-based evaluation and processing, including clustering [Vicente et al., 2007].

A classical evaluation measure for word spaces is a multiple choice synonym test [Landauer and Dumais, 1997], which is very limited in the number of samples. Publication VI introduces automatically created multiple-choice tests from much more extensive and publicly available linguistic resources. They lack the finesse of the manually created tests, but cover a much larger field and can be created for several existing resources spanning different knowledge and languages. Publication VI demonstrates how post-processing of the components has an effect of the results, and how ICA-based and SVD-based methods converge when the distances between points are computed in the whole latent space. It shows the sparser structure found with ICA compared to SVD-based methods.

The latent spaces created with WordICA, other vector space method, or word embedding methods can also be used as input to supervised methods or natural language processing tasks. Several examples are given in Section 5.7.

5.3 General word analysis

This section describes predominantly my contributions to the field I call general word analysis. Citations to background research are provided as required. Word spaces can be generated for different vocabulary, texts and contexts. Here research in which the emergent features for words are analyzed without a specific task in an exploratory
manner is presented. The analyzed vocabulary is limited only for computational reasons or to simplify the manual analysis. Additional filtering may be performed by selecting only words in specific part-of-speech categories. First, an overview of the research topics is given and then each is discussed in more detail.

Honkela et al. [2003] and Publication III originally introduce the general WordICA method, which applies ICA for the extraction of word features from a word co-occurrence data matrix. The papers provide a qualitative analysis of the extracted feature vectors. Borschbach and Pyka [2007a,b] further manually investigate the grammatical and linguistic nature of the emergent features in a similar experimental setting. A method to visualize the match between individual ICA features and part-of-speech categories is provided in Publication IV. This match is quantified in Publication III. The methodology is extended in Väyrynen and Honkela [2005] in which a new separation measure for ICA features and set of known categories is introduced. Publication V gives more background and motivation to the WordICA method and draws together the previous research.

The WordICA method has been further developed and analyzed by me and others. Publication VI explores the sparsity of the components with thresholding small values to zero, and evaluates how it compares to singular value decomposition in different multiple-choice tests of synonyms and related words in English. The same methodology is tested with other languages in Väyrynen et al. [2007]. Publication VIII considers multiple languages at the same time using a parallel text corpus. Not only does the vector space include translations, but the emergent components are multilingual as well. Hazem and Morin [2012] explore a similar idea with comparable texts which are not parallel but discuss the same topic. Sasaki et al. [2012] introduces an ICA model which allows linear and higher-order correlations between the components and elicits on ordering on the emergent features, similar to the topographic extensions of ICA for image data in Section 4.3. Their method finds groups of topically related features, such as different units of time and types of media.

5.3.1 Learning POS information

Honkela et al. [2003] and Honkela and Hyvärinen [2004] originally introduce the WordICA method. They computed a word co-occurrence matrix from posts to the connectionist e-mail list with manually selected 100 analyzed words and 2 000 most frequent context words. The co-occurrences were collected only for one immediately preceding word from the analyzed words. A similar data matrix based on the newspaper Times text is built in Borschbach and Pyka [2007a,b]. The frequency
counts were dampened by taking the logarithm of frequencies increased by one. Importantly, no other weighting or normalization scheme was applied, such as tf-idf or entropy weighting in LSA. As only a very small number of features (10–30) were extracted, it was possible to visualize the word vectors found and provide manual analysis for the possible interpretation of the features.

The experiments are confirmed and extended in Publication III where a larger text collection of English e-books from project Gutenberg is used. The number of analyzed words was increased to 50,000, which were analyzed with 1,000 context words. Both words sets were then selected as the most frequent words. Significantly, no stop-word lists were applied to restrict the list of analyzed words in these experiments, which is not typical in all language technology approaches. This illustrates the robustness of the WordICA method. The context window covered the immediately preceding and following words. The experiments reveal that morphosyntactic features can be identified, which resemble POS tags in this setup rather than topical information. The extracted 100 components could not easily be analyzed visually, and instead correlation between features and known POS tags were computed. Visualization of the correlations and further manual analysis of the components with high correlations with POS tags revealed that some components clearly represented words belonging to specific POS categories. The results also suggest that several components were learned for the highly abstract open class POS categories, such as nouns and verbs, in which each component modeled some semantic aspect.

This general analysis discovers broad categories that are prominent in the data. For instance, including all frequent function words and not using stop-word lists most likely helps to find syntactic information related to POS categories, which could be undesirable with topic-level analysis. A more detailed analysis of a single POS category is possible by limiting the analyzed words to the analyzed category, such as adjectives [Honkela et al., 2010].

Components as categories

Vector spaces have traditionally been applied only to measure the relatedness of words as distances between the corresponding word vectors, where the distances have been computed in the whole vector space without considering the structure that can be found from the word space. The traditional structural transformations (e.g. SVD or ICA), without dimensionality reduction, only represent the vectors in a new base which does not effectively change the distances between vectors. Dimensionality reduction does change the distances, for instance, truncated SVD find a representation that tries to minimize the distortion between the representations in some sense [Man-
ning and Schütze, 1999]. It should be noted that this is often a global criterion which
does not guarantee that all local pairwise distances are preserved. The distances in
the vector space are thus mostly derived from what contextual information is included
and how the data is preprocessed and weighted.

The standard linear analysis methods (e.g. SVD, ICA, and NMF) learn a component-
based representation in which the components can be analyzed individually, i.e., dis-
tances are computed in a sub-space by considering one or more of the components.
The components found depend on the data but also the computational method. This
dissertation studies how independent component analysis is suitable for the analysis
of words in contexts. Publication III showed how the most active words in a category
can be visualized for manual analysis of the independent components. Publication
IV introduced correlation as a measure between single components and known word
categories. Finally, Publication V extended this to component pairs and considers
how well-known word categories can be separated by component pairs learned.

Separability of component categories

Publication IV showed how correlation can used both to visualize and quantitatively
measure how well an emergent component matches with a known category. This does
not take into account how an emergent representation can find components which
model a mixture of categories. To provide a better analysis, Publication V proposes a
separation measure which considers how well two features explain two known cate-
gories. Ideally, a single component would model one of the categories. Typically this
is not the case and considering a pair of categories gives an idea of how much each
category is spread across the components.

In general terms, measurements for cohesion inside a cluster does not tell whether the
clusters are well separated. The developed ad-hoc separation measure for categories
quantifies the match between features found and part-of-speech tags.

The evaluation measure considers two categories with indices $k$ an $l$ and two compo-
nents with indices $i$ and $j$. On one hand, the words $w_n$ in the $N$-word vocabulary can
be divided into four non-overlapping sets:

1. words belonging to category $k$ and not to $l$,
2. words belonging to category $l$ and not to $k$,
3. words belonging to both categories $k$ and $l$, and
4. words belonging to neither category $k$ nor to $l$,

where most category pairs have only a few words in the overlapping type (3). In
some cases, there can be overlap, especially when considering hierarchical or related categories, such as verbs in general and verbs in different tenses. On the other hand, the words in two features, \( f_i \) and \( f_j \), can be plotted two-dimensionally with different markers for each of the four types. If the features separate the categories well, type (1) words would be concentrated on one axis and type (2) words on the other axis. Furthermore, if a feature represents well a category, the words in the category would be away from the origin and words not in the category would be near origin.

The evaluation measure quantifies the separability of the two components given the two categories. A category is defined as an \( N \)-dimensional vector \( c_k \), where \( c_{kn} \) tells whether the \( n \)th word in the vocabulary is in the \( k \)th category. In a similar fashion, the values of the components, \( x_n \) and \( y_n \), have the strengths of components for the same word. The evaluation measure gives a positive or negative reward depending on the category of the word and the coordinates of the word related to the axis and the origin as

\[
sep(k, l, i, j) = \frac{1}{N}(r_1 + r_2 + r_3 + r_4)
\]

with the rewards split into four parts corresponding to the four disjoint types:

\[
r_1 = \sum_{n=1}^{N} c_{kn}(1 - c_{in})(|x_n| - |y_n|)
\]

\[
r_2 = \sum_{n=1}^{N} (1 - c_{kn})c_{in}(|y_n| - |x_n|)
\]

\[
r_3 = \sum_{n=1}^{N} c_{kn}c_{in}(|y_n|^p + |x_n|^p)^{\frac{1}{p}}
\]

\[
r_4 = -\sum_{n=1}^{N} (1 - c_{kn})(1 - c_{in})(|x_n|^p + |y_n|^p)^{\frac{1}{p}}
\]

where distance from the origin is calculated using the \( p \)-norm distance metric. A brute force search finds the pair of components \((i, j)\) which best separates each pair of categories \((k, l)\):

\[
sep(k, l) = \max_{i,j} sep(k, l, i, j)
\]

\[
(i_{kl}, j_{kl}) = \arg \max_{i,j} sep(k, l, i, j).
\]

Figure 5.3 shows the pair of independent components that best separate adjective and verb categories. Each category is clearly more active only on one of the components. A comparison to components from SVD shows how both components are more mixtures of the categories.

The mean over the best feature pairs for each category pair provides the overall ca-
pacity of separation over all $K$ categories:

$$sep = \frac{2}{K(K-1)} \sum_{k,l} sep(k,l).$$

(5.8)

Figure 5.3. The two components which best separate adjectives and verbs. The first row is for components estimated with ICA, the second for SVD components. The first column shows the activations of only adjectives, the second column only verbs, the third column is for words which are neither adjectives nor verbs, and fourth column is for words which can be both adjectives and verbs.

### 5.3.2 Thresholded WordICA

The separation measure introduced in Section 5.3.1 takes advantage of the found independence of the features. Publication VI introduces thresholded WordICA, which considers the sparseness of the features and applies a nonlinear postprocessing to the selected features. The nonlinearity of the postprocessing is important, as a linear processing could take place before the linear model.

The feature activations with WordICA are sparse, and only a few of the features are very active for each word. The thresholded WordICA takes advantage of this and sets feature activations with small absolute values to zero. This can be understood as noise reduction and is related to shrinkage methods (see, e.g., [Hyvärinen et al., 2001]. Specifically, the number of features set to zero was varied and the created word representations were used in multiple-choice test of related words. As expected, removing information from the representation did not improve the results. A comparison of the ICA and SVD features showed that on average the ICA representation suffered much less than the SVD representation in most cases, demonstrating the sparser structure found with ICA compared to SVD. Väyrynen et al. [2007] provide extended analysis with more languages and evaluation sets with similar conclusions.

The proposed thresholding shows how the sparseness of the components found is an asset and provides a way to quantify the sparsity of the representation. This is be-
cause sparse components are easier to interpret manually. Furthermore, a sparse representation can be useful in higher cognitive processing [Field, 1994]. Computational approaches include deep neural networks, which find an unsupervised representation in the lower layers. Yokoi et al. [2006] utilize the sparsity of independent components by pruning irrelevant components in a information recommendation application.

![Figure 5.4. Example of features activations before (all bars) and after (filled bars) non-linear thresholding.](image)

### 5.3.3 Multilingual word spaces

Publication VIII extends the WordICA methodology to multilingual analysis. Word co-occurrences are computed across languages with the parallel bi-lingual sentence-aligned English–Finnish part of the Europarl corpus [Koehn, 2005]. Here the context window for a word is defined to span the whole parallel sentence in the other language, and the frequencies were dampened with a logarithmic transformation. A manual analysis of neighboring words in the created space showed that translations of words can be found close to each other. Some examples are given in Table 5.2, showing that the underlying vector space can be a source for finding translation candidates as words which are close to each other in the vector space. Furthermore, the emergent features with ICA were also multilingual and most active words in each feature were related.

The choice of included contextual information and the analyzed units shapes the nature of the components found, as discussed in Section 5.1.4. If the analyzed units includes words from different languages and the sentential context is a translation, in-
dependent component analysis finds components that are multilingual. Specifically, the components contain similar information to monolingual analysis but components can have words from different languages. Three example components are given in Table 5.3.

Table 5.3. Example of most active words in three features learned with ICA from multilingual data. The English gloss is given in parenthesis.

<table>
<thead>
<tr>
<th>word</th>
<th>English gloss</th>
<th>values</th>
<th>eroja (differences)</th>
</tr>
</thead>
<tbody>
<tr>
<td>saksan (of Germany)</td>
<td></td>
<td>values</td>
<td>eroja (differences)</td>
</tr>
<tr>
<td>ranskan (of France)</td>
<td></td>
<td>rauhan (of peace)</td>
<td>different</td>
</tr>
<tr>
<td>germany</td>
<td></td>
<td>demokratian (of democracy)</td>
<td>different</td>
</tr>
<tr>
<td>france</td>
<td></td>
<td>vapauden (of freedom)</td>
<td></td>
</tr>
<tr>
<td>french</td>
<td></td>
<td>democracy</td>
<td></td>
</tr>
<tr>
<td>german</td>
<td></td>
<td>ihmisoikeuksien (of human rights)</td>
<td>differences</td>
</tr>
<tr>
<td>sweden</td>
<td></td>
<td>arvoja (values)</td>
<td>erot (differences)</td>
</tr>
<tr>
<td>netherlands</td>
<td></td>
<td>solidarity</td>
<td>toisaan (each other)</td>
</tr>
<tr>
<td>ranska (France)</td>
<td></td>
<td>peace</td>
<td>disparities</td>
</tr>
<tr>
<td>belgian (of Belgium)</td>
<td></td>
<td>arvojen (of values)</td>
<td>eri (different)</td>
</tr>
<tr>
<td>ruotsin (of Sweden)</td>
<td></td>
<td>kunnioittaminen (respect)</td>
<td>erilaiset (different)</td>
</tr>
<tr>
<td>saks (Germany)</td>
<td></td>
<td>oikeusvaltion (of country of justice)</td>
<td>differ</td>
</tr>
<tr>
<td>italian (of Italy)</td>
<td></td>
<td>principles</td>
<td>differing</td>
</tr>
<tr>
<td>kingdom</td>
<td></td>
<td>continent</td>
<td>eroavat (differ)</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

The multilinguality of the WordICA components was manually analyzed in a previously unpublished experiment, in which 100 Finnish–Greek components for words were evaluated. Native laymen in the respective two languages described the ICA components in English based on the most active words in each component. In a ne-
gotiation phase the descriptions where discussed and amended jointly. The parallel
descriptions of the components where analyzed on a three-level scale for agreement.
The results in Table 5.4 show that a clear majority of the components can be analyzed
and described bilingually. Further experiments outside this dissertation have shown
that this can be extended to more than two languages with multi-language corpora
such as the Europarl Corpus [Koehn, 2005]. Additionally, the experiment demon-
strates that a human description based on the most active words in each language in
the component is a viable way for describing the multilingual topic or subject on a
majority of the independent components.

<table>
<thead>
<tr>
<th>No. of components</th>
</tr>
</thead>
<tbody>
<tr>
<td>agreement</td>
</tr>
<tr>
<td>partial agreement</td>
</tr>
<tr>
<td>unclear meaning</td>
</tr>
<tr>
<td>87</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>100</td>
</tr>
</tbody>
</table>

Table 5.4. Analysis of the thematic agreement of the most active words with Finnish-Greek WordICA
components.

Publication IX experiments with a bilingual word vector space for creating a con-
ceptual map of words. The bilingual vector space was constructed with the help of
the parallel sentences in the German–English part of the Europarl corpus [Koehn,
2005]. The self-organizing map [Ritter and Kohonen, 1989] was applied to creat-
ing a non-linear mapping between the words. The context for a word was based on
the co-occurrences of context words in the sentence and the parallel sentence in the
other language. The analyzed words were manually selected as the 150 most frequent
nouns in both languages. This enables the visualization of the resulting map and en-
sures that there are at least some translations of words present in the vocabulary. The
context words were the most frequent 3889 words in both languages. The variances
of the context variables were normalized to one, and the L2 norm of the word vec-
tors was normalized to one. An analysis of the results shows that translations and
semantics are the most prominent information in the underlying vector space, not the
division between the two languages. This was possible because of careful selection
of the contexts and suitable weighting of the data matrix.

5.4 Targeted word analysis

This section provides mainly background research on a field I call targeted word
analysis. My contribution is in Section5.4.4 in the evaluation of semantic informa-
tion captured by the WordICA method. Linguistic vector spaces can be constructed with large, open vocabulary for tasks such as information retrieval, part-of-speech induction, or exploratory search. It is very practical to reduce the dimensionality of the vector space and this can often be done without significant reduction in the information stored [Karlgren et al., 2008]. Unlike supervised learning, unsupervised learning first models the main characteristics of the data, which is not always the focus of the study. One method to change the focus of the analysis is to weight the data or select different features. This section gives an overview of how the analyzed vocabulary influences the resulting structure of the model. Especially, the motivation is to learn components which are still emergent but more relevant to the given task than components computed for the whole lexicon.

Chagnaa et al. [2007] perform concept analysis for a short list of verbs co-occurring with statistically relevant nouns. Chung and Pennebaker [2008] approach psychological trait induction in a data-driver manner from relevant texts without predefined traits, but with different small, selected lexicons. Honkela et al. [2012] analyze well-being texts in topic modeling with specific vocabulary. Publication VII considers semantic analysis and demonstrates how limiting the analyzed words helps to focus the low-dimensional representation to the task. Lindh-Knuutila and Honkela [2015] deepens the analysis. These studies show that data selection has a clear impact on the emergent features, and an unsupervised method can learn different points of view to the same text data.

5.4.1 Concept induction for verbs

The preliminary work by Chagnaa et al. [2007] investigated ICA in a cognitive task concerning latent concepts and noun memberships for those concepts. The work assumes that concepts are formulated as a linear combination of verbs. This approach shows how selection of the index and context words enables different interpretations of the linear model by choosing the analyzed units and contextual information appropriately.

The experiments were conducted with 15 manually chosen common Korean verbs and 91 nouns. Both the verbs and the nouns were selected with the help of frequency and mutual information thresholds. Estimates of mutual information between the verbs and the nouns were stored in a verb-noun data matrix. The analyzed verbs were manually identified to have 12 distinct meanings, which was set as the number of extracted independent components with the FastICA algorithm.

The resulting mixing matrix was evaluated qualitatively and found to contain roughly
the expected 12 distinct meanings, with jointly active components for verbs with shared meaning. The concepts found with ICA were compared to linkage-based hierarchical clustering of the verbs from a latent semantic space built with SVD, and ICA was concluded to be more effective in extracting hidden concepts from the data.

The experiment setting is very similar to sense induction in Section 5.5 expect that there the expected number of components is lower than the number of words, because some of the verbs were expected to have very similar meaning and there is no need for disambiguation. Additionally, the contextual features with nouns were manually limited to a very small number without disclosing the exact methodology.

5.4.2 Personality induction

Chung and Pennebaker [2008] approached personality induction in a data-driven manner using self-descriptive narratives, which can be seen as the natural input for how people create coherent models of the people they meet. Formal personality assessments greatly rely on experts to create and analyze itemized questionnaires. These trait level questionnaires have been criticized of restricting defined traits responses to numerical values. The study uses ICA as a factor analysis tool for word occurrences to provide a straightforward way of data acquisition of self assessments, which is not limited to a predefined set of psychological traits. The research attempts to find the basis dimensions of self-concepts based in adjectives or most frequent content words. Limiting the analysis to personality-based adjectives helps the unsupervised method to find features relevant in the task and focuses the analysis.

Chung and Pennebaker [2008] collected a total of 1 165 open-ended self-descriptive narratives over a period of three years. Experiments were conducted with lexicons selected both manually and automatically. The dimensionality of the created binary term-document matrices was first reduced with principal component analysis, followed by a varimax rotation. With the manually selected lexicon, emergent components grouped together adjectives that are psychologically meaningful and coherent, with some of the components matching traditional trait-like dimensions.

Similarly to the above-mentioned concept analysis with term-term vector spaces, this research is similar to sense induction [Rapp, 2004, Šimon and Hong, 2007]. However, the personality-trait analysis is performed with a term-document matrix, whereas the concept-analysis was computed with term-term data.
5.4.3 Well-being informatics

Honkela et al. [2012] presented a methodology which utilize ICA as a topic modeling component in the area of well-being informatics. A word-document matrix of size $186 \times 2570$ was computed from texts collected from a social media site, and both the analyzed words and documents were related to the well-being theme. The dimensionality of the data matrix was reduced to 20 and the same number of independent components were extracted with FastICA.

A number of the most active words in each emergent feature could be associated with themes related to well-being. Furthermore, the absolute value of the correlation coefficient between the ICA features for positive and negative topics and automatic sentiment analysis scores were shown to be high in several cases.

5.4.4 Semantic analysis

Publication VII experiments with semantic categories and shows how the selection of the lexicon can affect the data. Especially, the importance of selecting the lexicon before dimensionality reduction is demonstrated.

The considered semantic category test [Patel et al., 1998] contains multiple categories and a total few hundred words. An initial word co-occurrence matrix from a subset of English Wikipedia articles for the 200,000 most frequent words covers all the words in the semantic category test. The data matrix is weighted with positive point-wise mutual information.

The selection of the subset of words in the analyzed lexicon can be done either before or after dimensionality reduction. It is obvious that if ICA or SVD is applied after the subset selection (subset+ICA or subset+SVD), the components will find a more detailed representation for the limited set of words. In the case of word analysis, the subset selection will be carried out after a dimension reduction for the complete matrix of 200,000 word vectors (ICA+subset and SVD+subset).

The results, shown in Table 5.5, were evaluated with strict and lax criteria [Sahlgren, 2006] which required that at least 9 or 6 words, respectively, in the most active 10 words must share the same category to be counted. With the focused model ICA was able to find 17 categories out of 53 with the strict criterion, and 37 categories with the lax criterion. In comparison, SVD found only two categories which passed the strict test, and 19 categories passed the relaxed condition. The results show that ICA is much better than SVD at finding semantically focused components when the lexicon
has been reduced.

As expected, in the dimensionality reduction+subset case, the number of matching categories and differences between the methods were much smaller. The unsupervised methods model the whole lexicon and the analyzed semantic categories are not prominent in the co-occurrences, which makes it unlikely that dedicated features would be learned for them. Even the computationally more expensive SENNA word vectors learned with a supervised method [Collobert et al., 2011] do not have components related to these considered semantic categories.

Table 5.5. Fraction of categories which filled the strict and lax condition for ICA, SVD and SENNA features.

<table>
<thead>
<tr>
<th></th>
<th>Strict</th>
<th>Lax</th>
</tr>
</thead>
<tbody>
<tr>
<td>subset+ICA</td>
<td>17/53</td>
<td>37/53</td>
</tr>
<tr>
<td>ICA+subset</td>
<td>1/53</td>
<td>12/53</td>
</tr>
<tr>
<td>subset+SVD</td>
<td>2/53</td>
<td>19/53</td>
</tr>
<tr>
<td>SVD+subset</td>
<td>3/53</td>
<td>8/53</td>
</tr>
<tr>
<td>SENNA</td>
<td>0/52</td>
<td>4/52</td>
</tr>
</tbody>
</table>

Lindh-Knuutila and Honkela [2015] continue the analysis with different evaluation sets, manual analysis and comparison to latent Dirichlet allocation. They show that with smaller models their evaluation favors ICA over LDA, and that the ICA features are easier to interpret as meaningful.

5.5 Word sense induction

This section provides mainly background research on word sense induction, with the exception of a single experiment to provide a counter-argument to one of the papers. Word sense induction tries to extract the possible senses for a word type, which could then be utilized in word sense disambiguation for word tokens. Rapp [2004] computed word co-occurrences for the analyzed word type in different corpora and report satisfactory results, whereas Šimon and Hong [2007] used a single corpus and clustered word tokens encoded with ICA features unsuccessfully. An experiment in Section 5.5.2 shows that ICA is able to extract features which contain sense information with word tokens.
5.5.1 Senses from multiple corpora

Rapp [2004] conducted word sense induction by applying ICA to contextual information of words in three different corpora with a word-corpus matrix. This enables the use of aggregated contextual information over word tokens, but still finds several different vector representations for a single word type. This study focused on the use of ICA to learn word senses for a single word from multiple corpora, whereas the WordICA research analyzed many words with statistics from one corpus.

Rapp [2004] analyzed twelve ambiguous English words from Yarowsky [1995] by extracting two independent components for each word and determining the nature of the components. The most similar words corresponding to the most similar word vectors were used to describe the semantic context of the components. In 67% of the cases the expected results were obtained, in almost 17% of the cases the result was unexpected but nevertheless plausible, and in another 17% of the cases the result was erroneous.

5.5.2 Senses from clustered contexts

In contrast to the above-mentioned approach, Šimon and Hong [2007] used only one Chinese corpus. ICA features were first computed for word types, which where used to encode contexts for word tokens. These were then clustered and compared to known sense information. In the first step, they built a word co-occurrence matrix for nouns, verbs, adjectives and adverbs. In the second step, they encoded 4+4 windows for word tokens with the ICA features and used a maximum-linkage hierarchical algorithm to find the same number of clusters as there are senses for each word. In overall, the results were not satisfying.

It should be noted that clustering is based on distances between word vectors, and independent component analysis can be presented as data decorrelation and variance normalization followed by the ICA rotation, which does not affect the distances in specific cases [Vicente et al., 2007]. The results should therefore be not that different from similar approaches which use SVD and clustering for sense induction.

Further investigation of sense features

The unsatisfactory results in Šimon and Hong [2007] for the extraction of word senses with ICA were not expected. I analyzed the used methodology and propose an explanation why the clusters found did not match with the word senses, even if as many as one thousand components were extracted. The extracted features with ICA represent
the statistical information present in the input data, and in contrast to supervised or discriminative learning, the desired output is not present to guide the model to focus the modeling to the word sense information. Therefore, even if ICA is able to find features relevant for the word sense information, they most likely will be lost when a coarse clustering is performed with all the features.

To show that ICA is able to find interesting features, I ran a simplified experiment. The experiment studied the single adjective HARD from sense-tagged Senseval data [Leacock et al., 1998]. The data consisted of PPMI-weighted instances of the lexeme in sentence contexts and 50 independent components extracted from it. Correlation between the three known senses of the lexeme and the individual components measures the match between the senses and components. A majority of the components found did not seem to correspond to the senses, but there were clearly some which did. The best matching individual components are shown in Figure 5.5. Therefore, ICA was able to find components that could help in the word sense induction and disambiguation tasks. However, an unsupervised clustering algorithm is not a suitable classifier because most of the structure found can be irrelevant to the task. This finding agrees with the proposal using supervised procedures after unsupervised feature extraction [Feng and Hansen, 2008].

Figure 5.5. Examples of ICA features correlating with known senses. The instances of the lexeme HARD are ordered by the three word senses separated by the vertical lines. The plots show ICA component values for each instance of the lexeme with the highest correlated individual components. There is a clear structure matching the three senses.
5.6 Sub-word analysis

This section briefly describes background research for the analysis of linguistic sub-word units, i.e., morphemes and phonemes. The reviewed studies apply the same WordICA principles to sub-word units.

Independent component analysis has been applied to the feature analysis of known morphemes in Finnish [Lagus et al., 2005] and to the more relaxed morpheme analysis without known morpheme segmentation of Italian verbs [Calderone, 2008a]. Experiments for the utilization of ICA features for morphemes in language modeling are described in Section 5.7.1. Analysis on the phoneme level is considered in Calderone [2009].

5.6.1 Morpheme analysis

Lagus et al. [2005] analyze morphemes analogously to word analysis and learns the ICA decomposition for a morpheme co-occurrence matrix computed from Finnish newspaper text. The segmentation into morphemes is assumed to be known, i.e., the inflected words are segmented into a sequence of morphemes which are used as tokens in the analysis.

The data matrix was computed for the 3759 most frequent morphemes and the 506 most common morphemes (including the word boundary) were used as context morphemes. A manual analysis of 50 components extracted with the FastICA algorithm shows representations which on a coarse level model main syntactic categories in Finnish and on a more detailed level depict potential thematic roles of the morphemes. The components represent verbs (inflected and derived), inflected nouns, derived nouns, locations and persons, adjective stems, with several components for each class. Most components appear to signify some interesting property, and each morpheme has only a few active components, similar to sparse coding.

5.6.2 Phoneme analysis

Calderone [2009] analyzes phonemes analogously to word and morpheme analysis. The significantly lower number of distinct phonemes compared to words and morphemes guarantees good coverage of the distributional properties even with moderate corpus sizes. There are also clear phonological theories and categories, for instance phonotactics and vowel/consonant categories, against which the found emergent features can be matched. The examined issues include the identification of phonologi-
Linear generative models for words in natural language
cally motivated constraints (e.g. vowel harmony) and the distinction of vowels and consonants.

Calderone [2009] reports experimental results with English, Italian and Finnish text data. English and Italian are phonologically transcribed and Finnish text is used without transcription because the pronunciation is very close to its orthographic representation. The texts include 39 phonemes analyzed for English, 26 phonemes for Italian, and 21 for Finnish. The phonotactic context is computed as a phoneme-phoneme co-occurrence matrix with additional contextual phonemes for the beginning and the end of the word. The results with all three languages show that the most prominent distinction found by ICA is the separation of phonemes into vowels and consonants.

5.6.3 Morpheme induction

Calderone [2008a] tries to overcome the use of linguistic knowledge incorporated by the use of predetermined primitive types such as stems, suffices and signatures employed, for instance, in Creutz and Lagus [2007a] and Goldsmith [2001]. Instead of finding morpheme elements for optimal segmentation of the lexicon, the goal is also to find features which encode morphosyntactic attributes. The data matrix was constructed to hold co-occurrences of morphological segments and morphological contexts, where segments were constructed as bag-of-character models for orthogonal vectors for each distinct 30 orthographic characters in Italian. The experimental results for 31 verbs in all 51 simple verbal forms are described as satisfactory, yielding both grammatical or morphosyntactic features and attributes for paradigmatic information. Especially, verbs of the same grammatical classes shared the same features, and verbs of the same paradigmatic classes had common features.

5.7 Applications

This section provides background research and describes a number of application oriented studies utilizing ICA features for words and other linguistic units. The applications include language modeling, language identification, and text classification. ICA features have been reported to be tested in an unsupervised manner in language modeling as additional topic models for increased accuracy [Kumaran et al., 2005], as class-based language models to achieve smaller model size [Virpioja, 2005], and as a classifier for short-text classification [Pu and Yang, 2006]. Moreover, ICA has been applied to feature extraction and dimensionality reduction before supervised classification methods in different tasks, such as language identification [Selamat and Lee,
5.7.1 Language modeling

A literature survey found two distinct approaches to taking advantage of ICA components in language modeling. Both approaches utilize either incorporate or utilize ICA features in existing frameworks. Kumaran et al. [2005] augments a standard n-gram model with an ICA-based topic model, and Virpioja [2005] creates equivalence classes for statistical morphemes from ICA features in a class-based n-gram model.

N-grams models with additional topic models

Statistical language models are generic in the sense that they learn statistics from the training data, which can contain a mixture of texts from different domains. A typical approach in statistical modeling is to consider known or otherwise coherent partitions of the training data to learn an optimized model for each part. Kumaran et al. [2005] experiment with incorporating a topic model into a probabilistic n-gram language model in order to increase the predictive capability. They apply a framework very similar to Bellegarda [1998], but they extend it by applying ICA after creating the topic model with LSA.

In their experiment, a word-document matrix is built with the 9371 largest documents from a subset of the Wall Street Journal corpus [Paul and Baker, 1992], in which each document consists of a few short paragraphs. The 10,000 most frequent words were selected in the vocabulary, without stop-word lists. The matrix was weighted with an entropy-based weighting, dimensionality was reduced with SVD, after which ICA was used to extract the same number of components. The cosine distance was used to measure similarities between pseudo-documents of each word in the vocabulary and the current history, i.e., all the words seen in the test document so far. The distances were scaled to probabilities and incorporated with the n-gram probabilities. The perplexity of the proposed method on a separate test is reported for a few values for the dimensionality reduction and different n-gram lengths. The maximum reduction in perplexity achieved was 45.45% compared to a baseline bi-gram model, and 38.09% for a tri-gram model.

The critique to the experiment can be directed to the use of cosine distance, which measures angles between vectors. Perplexities are compared only for the baseline without any topic model, and it not clear how the contribution for the reduction in perplexity is divided between SVD and ICA, or if ICA had any impact on the result.
Linear generative models for words in natural language

Class-based language models with quantized features

Virpioja [2005] reports some preliminary experiments for clustering ICA features into equivalence classes for class-based language modeling in order to produce models with a smaller lexicon. The experiments are run on a Finnish text corpus with language models based on statistical morphemes [Creutz and Lagus, 2007a]. Different morpheme–feature matrices were computed with 89 368 analyzed morphemes and different locations of nearby morphemes with a low-frequency cutoff to limit the total dimensionality to 1 000. The features included, for instance, the immediately following morpheme and the two morphemes nearest to both left and right, in which each location was encoded separately. The dimensionality of the data was reduced with PCA before ICA. The sparsity of the estimated independent components was utilized by creating a binary code with a manually selected threshold, and all units with the same code were set to the same cluster. No clear advantage was reported with the tested class-based language model neither by itself nor by an interpolation with a standard n-gram model.

5.7.2 Text classification

Independent component analysis has been proposed as a feature extraction method in several text classification tasks, such as language identification [Selamat and Lee, 2008], short-text classification [Pu and Yang, 2006], illicit web page classification [Sam et al., 2007], and information filtering [Yokoi et al., 2006]. These studies demonstrate how dimensionality reduction before ICA is necessary. Additionally, the unsupervised ICA features can improve supervised classification.

5.7.3 Document models

Since the publication of the review by Kolenda [2002], new results in the applications of ICA-based document analysis have been published. For instance, the document representations have been applied in new classification tasks, such as language identification and several text classification tasks.

Topic analysis with ICA represents a word-document matrix as a product of a word-component matrix and a component-document matrix. The former describes each component found, and the latter gives a distributed representation for each document with the activity of each component learned. The components are assumed to be the
Linear generative models for words in natural language underlying independent topics and can be visualized with the most active words in each component [Isbell and Viola, 1999]. If the documents are ordered according to known topic or by time in chat analysis, the activities of each component can be visualized [Kolenda et al., 2001, Bingham et al., 2002].

Some of the research topics for document-centered analysis include topic detection (topic classification), topic induction (topic summarization), and topic clustering. The most recent ICA-based papers in document analysis will be briefly discussed, because the main focus of this dissertation is in how ICA has been extended to other areas of textual analysis.

With latent topic models, the documents are assumed to be generated as a mixture of topics. One of the benefits of ICA-based methods is that the representations found should not only model the data, but individual components should be interpreted as meaningful. Moreover, Sevillano et al. [2007] has shown that ICA and NMF outperform LSA in the extraction of latent topics from term-document matrices.

Similar to Bingham et al. [2002], the studies in Grant et al. [2008b] show how independent component analysis can identify threads of related conversations in an artificial news repository. Specifically, the authors find a list of topics used in each document, and a list of the documents that best fit each of the topics. The independent components found are manually analyzed against the known topics in the document collection. Their experiment used the IEEE VAST 2007 Contest dataset with 1,500 news stories from a fictitious newspaper which were preprocessed. The resulting document-word matrix has 1,455 documents (rows) and the frequencies of 7,402 words (columns) in each document. The values in each row are normalized between zero and one.

Sevillano et al. [2004] propose two new ideas for improved ICA-based classification. The first is to use the cumulative distribution functions (cdf) of the independent components as a relative measure of relevance. Their reasoning is that previously applied selection methods have been based on raw values, magnitudes, and probabilities based on soft-max normalized magnitudes, which can be problematic if the components have different scales or probability distributions. The second is to introduce confidence regions for each known class. The classified documents are only classified to a class if the relevance values for each component are inside the corresponding confidence region computed from training data, and otherwise rejected. The proposed methods are evaluated with a tf-idf weighted $3400 \times 180$ term-document matrix computed for four thematic domains from a Catalan newspaper text. The results are reported to indicate modest classification accuracy increase with the cfd-
based relevance measure on separate test data. The confidence regions for each class are tested on additional documents from two new domains, and show that the method is able to reject classification for most of the out-of-domain documents. The rejected in-domain documents increase classification precision.

**Language identification**

Selamat and Lee [2008] experiment with language identification with ICA features as input to a neural network classifier. A word-document vector space is constructed from 300 web documents in Arabic, Persian and Urdu, which all are written with the Arabic script and share words for historical reasons. Dimensionality of the data is reduced with PCA to three and then whitened. The same number of components is extracted with FastICA. The ICA components are visually shown to roughly correspond to the known languages of the documents, and are used as inputs to a back-propagation neural network classifier with reasonably high precision (93%) but low recall (35%).

All details of the experiment setup are not given, which makes it hard to place the work the proper context. Especially, the proposed method is not compared to any existing baseline or method. Moreover, the details of the neural network topology are not described which makes it very difficult to consider the impact of ICA after dimensionality reduction and whitening. The low-dimensional representation is useful for visualization, but including more dimensions would most likely have improved the supervised classification accuracy.

**Short text classification**

Different online communication channels and e-commerce produce a lot of short messages, which have different characteristics than traditional written sources and new approaches have been explored [Song et al., 2014].

Pu and Yang [2006] propose ICA as a classifier in a short-text classification task in Chinese. Specifically, the computed ICA features are not fed into a supervised classification system, but used to cluster documents directly. The experiments with ICA are conducted with and without SVD-based dimensionality reduction. In the combined SVD+ICA model the dimensionality is reduced to 17 before four independent components are estimated with FastICA. Separate test documents are projected with the estimated model and assigned to the component with the highest activation. The assignments are evaluated as clusters, and the SVD+ICA combination is reported to produce clearly more salient clusters.
Pu and Yang [2006] demonstrates that SVD based LSA as a preprocessing step improves the results as it creates more overlap between the short document vectors. However, the proposed method is not compared to any standard classifier, and utilizing the ICA features as input to a supervised classifier could have improved the results.

Illicit web page classification

Sam et al. [2007] propose and evaluate a general web page classification into illicit and healthy pages, in which unsupervised feature extraction from a term-document matrix with tf-idf weighting is done with PCA and ICA before classification with a supervised neural network.

They report experiments for PCA and ICA as dimensionality reduction methods to a combined PCA+ICA approach where PCA is used for dimensionality reduction and ICA to only restructure the low-dimensional space, similar to WordICA. The number of extracted features is selected as the number of principal components corresponding to largest eigenvalues that explain at least 95% of the total variance. Experiment are reported on different combinations of the known labels. The ICA features learned are used as input to a neural network classifier.

Average accuracies over the different test sets are at the 83–95% range, but are not compared to any other feature extraction or classification method. The best average results are achieved with the combined PCA+ICA approach in which ICA is applied after dimensionality reduction with PCA. ICA without the dimensionality reduction step is reported to work clearly the worst.

Information filtering

Information technology can help users to find relevant information from the vast amount of available information. Information filtering systems require a topic model which contains information relevant to the recommendation task, which is a binary classification task of documents into interesting and uninteresting to each user. Unsupervised latent topic models describe all information present in the given data and thus can contain many latent features that are irrelevant to the task. Yokoi et al. [2005] prune topic-centered ICA features to those that are useful in information filtering with a clustering algorithm and evaluate the method in a recommendation system.

In the reported experiment in Yokoi et al. [2005], a word-document matrix is analyzed with ICA to estimate 623 independent components which are pruned with the maximum distance algorithm (MDA) to retain only relevant ones. A user profile for
information recommendation is learned with a genetic algorithm (GA) and evaluated. The authors find the ICA+GA combination to be slightly better than GA alone, but pruning of the irrelevant independent components in the ICA+MDA+GA variant with supervised MDA gave superior results.

Yokoi et al. [2006] extend the previous work by reducing dimensionality with unsupervised SVD before ICA instead of supervised MDA after ICA. Dimensionality is first reduced with SVD to 409 features, which corresponds to 80% of total variance in the data. They report improved precision for the SVD+ICA+GA over ICA+GA, and it is also clearly better than the ICA+MDA+GA reported in Yokoi et al. [2005], even though the data is slightly different.

**Emotion label identification**

Yu and Ho [2014] compare latent semantic analysis, independent component analysis, and their combination as feature extraction methods for supervised classification of Chinese text documents into emotional classes with support vector machines (SVMs). The data consists of 1711 psychiatric social texts manually assigned to 7 emotion classes representing different depressive problems. An evaluation based on IR-based measures on a separate test set shows that the combination of LSA+ICA outperforms the individual methods and the raw bag-of-words features.

### 5.8 Discussion

The continuing success of deep neural networks in many areas (see, e.g., Section 4.6) has also produced new approaches in language technology, such as recurrent neural language models for language modeling [Bengio et al., 2003] and Long Short-Term Memory models in machine translation [Sutskever et al., 2014]. This has also inspired new approaches to learning word representations. The continuous bag-of-words model [Mikolov et al., 2013] operates on instances of local word context windows and captures syntactic and semantic information and shows the capacity compute with the relationships between the word vectors.

Recently, there has been a shift back towards traditional and simpler distributional models. Pennington et al. [2014] propose a weighted least squares regression model which incorporates words co-occurrence statistics and shows improvements over neural models. Lebret and Collobert [2014] show that word embeddings based on Hellinger PCA of the word co-occurrence matrix compare well to embeddings gained with neural network language models. Furthermore, Levy et al. [2015] show that the
reasons behind the improvements with the neural models can be adapted and applied in count-based methods.

The development of the WordICA methodology is covered in the publications of this dissertation. Naturally, the initial analysis was exploratory and consisted of manual inspection of the features to understand the model, similar to the analysis of early document topic models or image bases in applying ICA for images. Manual analysis of both the model and the evaluation results is important and can reveal why the representation behaves as it does, for instance, finding the sources of good or bad results. Required resources for evaluation can also be inaccessible due to language or content. A very important reason for exploratory analysis is the fundamental task of finding new or surprising phenomena which usually requires looking outside the box or on how the model work inside. This also motivates the use of interpretable features in word representations.

The applied quantitative evaluation measures contain both methods specifically developed for assessing the individual features and standard methods applied in the research field. The representations learned with WordICA are always compared to those learned with SVD, which is the basis for the more popular LSA method. Publication VII provides comparison to word embeddings and Lindh-Knuutila and Honkela [2015] to LDA. Comparison to SVD is a natural choice, as ICA can be seen as a generalization of SVD, and the higher computational complexity should show in the quality of the results. Non-negative matrix factorization is another method capable of extracting interpretable features, but comparison to it is not made because the method has typically been applied only to topic modeling and not to word representation learning. Not all of the applicable quantitative evaluation methods are relevant, because the main goal is to learn interpretable single components whereas the full set of components basically contain the same information as the SVD representation.

The experiments were conducted mostly with English text in order to have access to training corpora, evaluation resources as well as to make the results accessible to most researchers. The conducted multilingual experiments include other languages and partly validate the methodology in a larger content. Additional validation is provided by the reviewed related work in Section 5.7, which covers multiple applications and more languages.
5.9 Conclusions

This chapter described how independent component analysis has been applied to the task of learning structure from linguistic vector spaces in an unsupervised fashion. Linguistic vector spaces are constructed from actual word usage statistics with the assumptions that word similarity is reflected in how the words co-occur.

On one hand, classical factor representation methods, especially singular value decomposition, are typically used for dimensionality reduction. On the other hand, independent component analysis structures the word space in a way that reflects linguistic and cognitive representations. In contrast to hard clustering methods, the structure found is continuous and does not rely solely on word distances but overall statistical properties.

The independent components found are emergent in the sense that they are found from the statistical properties of the language usage data instead of manual classification or supervised learning. Previously, independent component analysis has been shown to be a useful tool for finding an underlying interpretable structure from different signal measurements [Feng, 2008].

The sparse word features learned enable more interpretable representation than the dense low-dimensional representation learned with classical methods such as principal component analysis and singular value decomposition. This was demonstrated with several evaluation data sets. Methods maximizing statistical independence provide a more suitable computational model for word representations than methods based on second-order correlations, especially if interpretability is a required property of the model learned. There is increasing interest in finding sparse, interpretable semantic word representations (see, e.g., Fyshe et al. [2015], Faruqui et al. [2015], Vyas and Carpuat [2016]). Novel evaluation methods were developed in this dissertation for assessing the interpretability of the components learned. A related evaluation measure has been presented in Tsvetkov et al. [2015].

This chapter also underlines how an unsupervised method can model the given data, but it will not necessarily find a representation that is suitable to a specific task. Supervised methods and feature selection can help in selecting the most important features for each task. Properly applied normalization and dimensionality reduction can also focus the unsupervised methods to find the most salient features. For independent component analysis, the results show that with different selections of language data the unsupervised method focuses on different linguistic phenomena.
6. Summary

This dissertation explores the use of unsupervised methods based on statistical concepts as analysis tools suitable for higher-level analysis with interpretable features. Independent component analysis is a linear generative model which assumes features which are statistically independent. In practice, this is often measured as maximally non-Gaussian sources, whereas classical methods based on simpler statistical concepts of finding features maximizing variance implicitly assume Gaussian distributions. The latter are very useful as preprocessing tools, for instance, in dimensionality reduction, but are not intended to find dimensions which can be more suitable for higher-level processing such as logical operations on feature representations.

Computational neuroscience and linguistics are fields which try to explain how brains process and learn to process information with mathematical models. Computational neuroscience starts on the lowest levels and considers sensory information and how it is processed before it is passed on to higher levels of cognition, for instance in vision. Similarly in linguistics, the processing starts from the physical acoustic signal followed by the syntax and semantics which identify linguistic units such as words and start to map different relationships between them. Existing collections of digital images and texts can be leveraged as approximations of natural data which contain information similar to that available during cognitive processing in the brain.

In computational neuroscience, unsupervised methods based on different statistical concepts have been shown to learn patterns with characteristics similar to simple-cell properties. This dissertation has studied the concepts of sparseness, independence and temporal coherence. Moreover, organization of the pattern recognizers similar to our knowledge of the cortical structure of the brain can be found with fairly simple statistical models. In this dissertation the view is expanded and links between the statistical concepts are made. Temporal coherence is shown to find spatio-temporal features which are similar to spatio-temporal features found with independent component analysis, and also share properties with simple-cell measurements. It is already
known that there is a close link between sparse models and independent component analysis. The proposed unified generative model combines properties of sparseness, temporal coherence and organization. It is a step towards explaining how sparseness and temporal coherence are able to learn similar structure.

In language technology, there has been increasing interest in measuring semantic relatedness in artificial intelligence, psychology and cognitive science. One specific methodology has been the extraction of semantic representations from word co-occurrence statistics with linguistic vector spaces and further dimensionality reduction with machine learning. This is thematically closely related to learning from the statistics of sensory information in the visual domain. This dissertation focuses on restructuring the vector space using independent component analysis, which gives the underlying dimensions meaning. Similar methodology has been applied successfully in the visual domain.

Independent component analysis is introduced as a method to finding emergent structure in linguistic vector spaces for words. Linguistic vector spaces represent linguistic units in a distributional model of co-occurrences with other units, such as words or documents. Previous work on vector spaces has utilized, for instance, singular value decomposition for variance modeling and especially dimensionality reduction. From a modeling point of view, different projections and dimensionality reduction try to preserve the distances that are already in the original vector space. The proposed WordICA method utilizes dimensionality reduction as a preprocessing step, after which structure is found. Dimensionality reduction is crucial as algorithms such as FastICA operate on whitened data which will model uninteresting weak signals if there are signals with very varying variance. This is often the case in linguistic vector space in which words with very different frequencies are included. Specifically, it is shown how creation of the vector space and selection of the analyzed units has a huge impact on information which is modeled by an unsupervised method. In general, the applied methodology is very language and domain independent and can easily be adopted to new tasks.

Different evaluation measures between the emergent feature from WordICA and known word categories and word relations are considered. The presented qualitative and quantitative results support the hypothesis that the higher-order statistics used in independent component analysis finds features which are more interpretable and match better with linguistic knowledge than classical methods based on second-order statistics. This is in line with the research that purports ICA as an unsupervised method capable of finding cognitively relevant features from different data.


M. Borschbach and M. Pyka. Specific circumstances on the ability of linguistic feature extraction based on context preprocessing by ICA. In *Proc. of ICASS*, pages 689–696, 2007a.

M. Borschbach and M. Pyka. Maintaining the influence of context matrix construction principles on the ability of linguistic feature extraction by ICA. In *Proc. of ICSPC*, pages 1199–1202, 2007b.


D. L. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006a.


References


