Faculty of Arts and Sciences
Dissertations, No. 743
Linköping Studies in Statistics No. 14

Scalable and Efficient Probabilistic Topic Model Inference for Textual Data

Måns Magnusson

Linköping University
Department of Computer and Information Science
Statistics and Machine Learning
SE-581 83 Linköping, Sweden

Linköping 2018

At the Faculty of Arts and Sciences at Linköping University, research and doctoral studies are carried out within broad problem areas. Research is organized in interdisciplinary research environments and doctoral studies mainly in graduate schools. Jointly, they publish the series Linköping Studies in Arts and Sciences. This thesis comes from Statistics and Machine Learning at the Department of Computer and Information Science.

Edition 1:1

ISBN 978-91-7685-288-0
ISSN 0282-9800
URL http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-146964

Typeset using LaTeX

Printed by LiU-Tryck, Linköping 2018

to Lisa and Jorun.

ABSTRACT

Probabilistic topic models have proven to be an extremely versatile class of mixed-membership models for discovering the thematic structure of text collections. There are many possible applications, covering a broad range of areas of study: technology, natural science, social science and the humanities.

In this thesis, a new efficient parallel Markov Chain Monte Carlo inference algorithm is proposed for Bayesian inference in large topic models. The proposed methods scale well with the corpus size and can be used for other probabilistic topic models and other natural language processing applications. The proposed methods are fast, efficient, scalable, and will converge to the true posterior distribution.

In addition, in this thesis a supervised topic model for high-dimensional text classification is also proposed, with emphasis on interpretable document prediction using the horseshoe shrinkage prior in supervised topic models.

Finally, we develop a model and inference algorithm that can model agenda and framing of political speeches over time with a priori defined topics. We apply the approach to analyze the evolution of immigration discourse in the Swedish parliament by combining theory from political science and communication science with a probabilistic topic model.

Acknowledgments

There are many people that I need to thank for their direct and indirect contributions to this thesis. People who have given their support and personal contributions, and also some that just put up with me through these five, very intensive, years.

First and foremost I want to thank my main supervisor Mattias Villani. It has been a privilege to be his student and I really want to thank him for all the ideas, time, and effort he put into me throughout the years. He has always pushed me to go further, accepting nothing less than high quality research from me. But he also helped me focus on the right things when so many exciting research projects were possible.

My co-supervisor Marco Kuhlmann has also been important during these years, helping me through the difficulties of Natural language processing and computational linguistics. Marco's advice and counseling has been invaluable to me.

I am also very grateful to David Mimno, who welcomed me to Cornell University and acted as my supervisor during the fall 2016. Doing research at Cornell for one semester really helped me to get different perspectives on the latent semantic analysis research field. The way I try to present the different parts of latent semantic analysis in this thesis is heavily influenced by discussions with David and David's course on advanced topic models.

The most important part of my graduates studies has been learning to be a researcher. I entered graduate school, knowing very little about how to do statistical research, especially in the field of probabilistic text modeling and natural language processing. But thanks to my many collaborators I now feel like I can actually do real research. My research collaborators on different projects have been extremely important. One of the longest collaborations has been with Leif Jonsson, who taught me all of the hidden knowledge of computer science and programming. Alexander Terenin contributed a lot to my deepened interest in the more theoretical parts of Bayesian inference. David Broman pushed my knowledge on computational concurrence and complexity. With Alexandra Schofield I enjoyed a lot of discussions on topic modeling in general and corpus curation specifically. Alexandra also made me feel very welcome to the Cornell group during my visit to Cornell. Finally, I want to thank my collaborators Richard Öhrvall and Katarina Barrling for the long and interesting discussions on how to combine probabilistic modeling with social science theory.

No man is an island. My fellow doctoral students at Linköping University, both at IDA and at other departments have been a great support. I want to give a special thanks to Josef Wilzén, with whom I have shared office and teaching duties for more than five years. I have also had the privilege of spending time with Per Sidén and Sarah Alsaadi, with whom I have had many exciting discussions. The discussions I've had with Johan Falkenjack and Johan Dahlin have also been really important for this thesis. I want to give a special thanks to my academic big brothers, Matias Quiroz and Feng Li, who have played an important role in getting my research started. In addition to the PhD students at Linköping University, I also want to thank the PhD

students in David Minmo's group at Cornell for the warm welcome and great discussions on the different flavors of topic models.

My colleagues in the Division of Statistics and Machine Learning have contributed to my professional time in graduate school with support, discussions, and feedback. The administrative staff at IDA has also been a tremendous support, helping me out in the administrative jungle.

During the final work with this thesis I got a lot of support. The proofreading made this thesis much better and I got tremendous help from both my mother, Inger Johansson, and Brittany Shahmehri in making the language better. Anders Nordgaard came with a lot of last minute improvements. Philosophical discussions with Karim Jebari on Bayesian epistemology were a great help, and Anne Moe helped out with all the practical arrangements in the final months.

Thanks also goes to my family. My mother and father have supported me enormously during this thesis. Pelle Gemmel and Åsa Ahrland also deserves to be mentioned for providing a lot of encouragements.

My last and most important thanks goes to Lisa and Jorun. My wife and daughter had to put up with endless research discussions, destroyed vacations, and late-night work. Had it not been for you, I'm not sure this thesis would have existed. Thank you.

Södermalm, April 2018

Contents

List of Figures

List of Tables

1 Introduction

This thesis develops new methods for the analysis of large collections of textual data and studies the use of these methods on problems in applied research fields. The main focus is on probabilistic topic models which are used to uncover the thematic structure in a collection of text documents. The contributions of the thesis include efficient algorithms for scaling topic models to considerably larger data sets, new techniques for interpreting the inferences from topic models in the context of applied problems, and an analysis of how different choices regarding data curation affect the inference in topic models.

1.1 Background

Written language is a relatively old invention, dating back as early as 3200 BC [28, p. 762]. Invented by the Sumerian civilization in Mesopotamia, written language then rapidly spread to neighboring civilizations. Since its invention, written language has increased in importance. When language was invented only a small part of society was literate in comparison with today, when more than 4 out of 5 people are literate [92]. This has made written texts one of the very foundations of modern society, and today it is one of the primary means of communication of human thoughts and ideas, both in private and public.

The computer is a much more recent invention. The first computer was constructed in the early 19th century and since then the computer has developed in an astonishing way [44]. The increase in *computational processing power* is enormous. From 1971 to 2016, computational processing power has increased by a factor of 400,000 [1]. During the same period, the world economy (seen as the gross domestic product, GDP) grew by a factor of 23 [5]. This impressive development in computational power has led to computers being an important part of modern society and has revolutionized society in a similar way to the steam engine and electricity.

The computational revolution in recent decades has in turn resulted in a situation where more and more communication and written language is stored digitally in different ways. This has led to an explosion of digital text *data*. Google Search alone indexes hundreds of billions of web pages, and just the index of these web pages is 100 petabytes in size [50]. Given that one piece of paper is roughly 2 kilobytes and 0.1 mm thick, printing out only this index would result in a pile 5 million kilometers high, or roughly 13 piles that each would stretch to the moon. And this is only Google's index of web pages for the current web.

The importance and increased availability of written text make it an important part of understanding humanity and the different aspects of human life and society. Many scientific subjects in the social sciences and the digital humanities use written text as the very foundation for scientific inquiry. Political science, history and literary science all rely heavily on written text as empirical material. The computational revolution and the increased accessibility of textual data give rise to new possibilities for these scientific fields. Many new *corpora*, or collections of documents or books, are available to researchers and practitioners. Social media platforms, web content, and online forums produce enormous amounts of textual data on a daily basis, data that can be used for scientific inquiry. In addition to the continuous stream of newly produced information, projects to turn previously written documents into digital text are also available to researchers. Millions of newspaper pages and books have been digitized and made available for research around the world [111, 51].

But what can researchers do with these new and huge corpora of text? "What do you do with a million books?", as the philologist Gregory Crane asked in the wake of the Google Book project [22]. Reading them would take 40 lifetimes. The sheer scale of the new data sources makes it difficult for researchers to approach the material with the standard approaches such as close reading, manual discourse analysis, or classical archive studies. Instead, new approaches have been proposed, such as literary scholar Franco Moretti's proposal of "Distant reading", which uses statistical approaches. By treating these large corpora as empirical data and employing statistical approaches to analyzing the data, new areas of research are possible in the humanities and the social sciences [79, 111].

Statistical inference concerns the problem of drawing conclusions for a *probabilistic model*, given observed *data*. The data is commonly assumed to have been generated in accordance with a probabilistic model, and by estimating the *parameters* of the model we can gain new knowledge. Given the ever-increasing amount of data available, there is a need for probabilistic models and methods of statistical inference that can produce scientific knowledge based on large-scale textual empirical data.

Textual data has been considered "unstructured" by many statisticians [72]. Given the history and the very purpose of written language, this perspective makes little sense. Writing is all about transferring information from the writer to the reader. Written text is hence highly structured, even though it is not numerically structured in a way common to the field of classical statistics. In the field of natural language processing (NLP), however, there is a long tradition of using probabilistic and statistical techniques to model textual data. The very foundation for statistical NLP consists of regarding a text, such as a sentence, paragraph or book, as a probabilistic model, $p(\mathbf{w})$, where $\mathbf{w}$ is the corpus to analyze and $p(\cdot)$ is a given probability model, with examples such as hidden Markov Models (HMM) [58], the Brown word cluster model [14], or models for Machine translation [15]. The statistical problem is hence to find good probabilistic models and methods to infer the parameters in these models efficiently.

Given the recent development of an ever-increasing amount of digital data and increased computational power, there is a need to enable researchers and scholars to use statistical methods on large corpora for the purpose of answering scientific questions. This thesis is one step on this path.

1.2 Motivation

The large increase in available digital textual data mentioned in the previous section has the effect that many contemporary corpora that are of scientific interest are quite large. Examples of corpora that have been studied in this thesis are the New York Times corpus and the Swedish parliament corpus. The New York Times corpus contains articles from the period 1987–2007, with approximately 1.8 million articles and roughly 500 million unique words. The New York Times was founded in 1851, making this corpus just a small (12%) slice of the newspaper's total historical output. In Sweden, multiple national newspapers have been made available from the 19th century and the Swedish parliamentary corpus contains speeches that were made in the Swedish parliament from 1994 to 2017, roughly 300,000 speeches that make up roughly 100 million words. As with the New York Times corpus, this is just a small slice of the whole corpora—protocols from the Swedish parliament exist from as early as 1627. The Google book project may be the most ambitious digitization project around today. The goal of the project is to scan all books in the world. According to Google, this would be roughly 120 million books, a total of four billion pages and two trillion words [51].

The scale of these large corpora is a challenge when using conventional statistical methods. A corpus of just 300 books can be as large as 35 million individual words. If we were to use a statistical method that took as little as 0.1 second per word, analyzing just 300 books would still take roughly 40 days. In such a situation, analyzing 30,000 books would be too time-consuming to be of any real value. The New York Times corpus is a corpus that will be used as an example in this thesis. With roughly 500 million words would take about 18 months to analyze. To be able to analyze these large corpora we need methods that are computationally efficient and can make use of modern computers.

However, the challenges of large corpora also bring rewards. The larger corpora that are available today enable researchers to do research that was not previously possible. Historians can analyze the public discourse in newspaper articles over decennia [112], literary scholars can uncover the macrostructure of literary history [52], and political scientists can experiment with social and political processes that drive discourse [90]. In addition to enabling researchers to analyze the data with standard methods, large digital corpora also enable researchers to study more complex and expressive models and capture more complex processes. Empirical material in the form of text can be complex in itself, and many different aspects may be of interest to individual researchers to study.

Probabilistic models are rigorous tools for statistical inference that are based on sound statistical and scientific principles. They can be powerful tools for research, but the problem with many probabilistic models is that they can be computationally costly to use. This is especially true as the models become more complex. So, given a situation where we have large corpora, the need for and interest in expressive models will lead to computationally costly models.

In the article "The free lunch is over: A fundamental turn toward concurrency in software" the author, Herb Sutter, concludes that concurrency of programs will become much more important [101]. This refers to the fact that the speed of individual central processing units (CPUs) has stopped increasing. Previously, CPU speed had been growing exponentially, in accordance with Moore's law [78], but as of 2003–2004, this development has stopped. Instead, multiple CPUs have been added to make up for the stagnation in computational performance. This will, according to Sutter, mean two things for software development and computation [101]:

- Applications will increasingly need to be *concurrent* if they want to fully exploit continuing exponential gains in CPU throughput.
- *Efficiency* and performance optimization of software will get more, not less, important.

The conclusion to draw from a statistical perspective is that in order to handle the large increase in data, we need to make statistical inference *concurrent*, i.e. it should be possible to perform efficiently in parallel, and *algorithmically efficient*, i.e. the computational cost of doing inference should be held to a minimum.

This leads to a trade-off between the complexity of probabilistic models (the computational cost of estimating the model) and the corpus size that can be analyzed in a reasonable amount of time. To enable more complex models, or to analyze larger corpora, it is important to derive probabilistic inferential methods that are concurrent and computationally effective.

Scaling statistical inferential methods connects modern statistics with techniques and theory from computer science, such as parallel algorithms and careful analysis of the computational complexity of inference algorithms. These issues have previously been considered to be solely a part of computer science, but statisticians can no longer ignore them when tackling the sizes of contemporary corpora. Statisticians need not only to study new inferential methods or new models, but at the same time they must explore models and techniques that can *scale*. The following quote from the National Research Council of the United States sums up the general problem of contemporary statistics, and the conclusions are even more important in the statistical analysis of textual data [21].

> (T)he challenges for massive data go beyond the storage, indexing, and querying that have been the province of classical database systems (and classical search engines) and, instead, hinge on the ambitious goal of *inference*. Inference is the problem of turning data into knowledge, where knowledge is expressed in terms of entities that are not present in the data per se but are present in models that one uses to interpret the data. Statistical rigor is necessary to justify the inferential leap from data to knowledge, and many difficulties arise in attempting to bring statistical principles to bear on massive data. Overlooking this foundation may yield results that are not useful at best, or harmful at worst. [21]

Given ever-increasing corpora, both contemporary and historical, we need inferential methods that can scale, but without giving up the statistical rigor.

Figure 1.1 summarizes the connections between statistics and other fields of importance for a thesis on probabilistic analysis of text. Fields such as mathematics, computer science, statistics and engineering, as well as domain fields such as social science, the humanities, and natural language processing, are all important when trying to model textual data. From a purely technical standpoint, we need to use theory from statistics for inference, numerical methods and applied mathematics for improving inferential speed, and computer science for complexity analysis and programming using (memory and computationally) efficient data structures. But we also require knowledge of more domain-specific fields, such as computational linguistics, and domain knowledge from the social sciences and the digital humanities, if this is the area of application.

The use of probabilistic models is gaining interest in applied research as a result of the increased availability of textual data sources and new methods. This has the effect that it is important to develop methods, best practices, and models that fit into the scientific research process. This, in turn, makes it important to examine, understand and improve properties that influence statistical conclusions, such as the effects of common corpus curating, model interpretability, and the connections between probabilistic models and the scientific models and theories of interest for applied researchers.

Textual data from a probabilistic perspective

Written text is complex when we consider it as data for statistical analysis. Unlike "classical" data in statistics, such as the kind that results from randomized controlled trials or multivariate

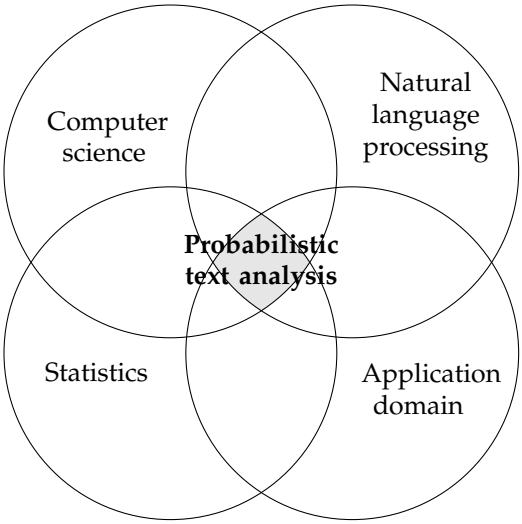


Figure 1.1: Overlap between the fields of importance in probabilistic analysis of text.

normal data, there are lots of different structures within the textual data itself that complicate inference. First of all, there is the meaning of the text, commonly described by the term *semantics*. This is what the writer tries to communicate to the reader using the text as a medium. Statistical inference is further complicated by the fact that the meaning of language is highly ambiguous. The same word can have multiple meanings. A simple example is that of "with" in the following two sentences. "I have a pizza with my friends" and "I have a pizza with olives" [37, p. 1], where the meaning of with can either be "together with (my friends)" or "with (olives) as topping.". This is a simple example, but this problem occurs in most situations when analyzing text. In Paper VI in this thesis we studied immigration, and the word "immigrant" had two slightly different meanings in the parliamentary speeches. In the context of immigration policy it simply meant "a person that immigrates to Sweden", while in the context of labor market policy, the meaning had a different connotation, a collective identity of immigrated people who have a difficult time in the Swedish labor market. So the meaning of individual words or the nuances of meaning may be different, depending solely on the context of usage.

In addition to the semantic content, written text also contains the *syntax* of the language, the linguistic rules, and principles of how to compose a readable textual utterance by ordering individual words [58]. The syntax makes it easier for a human reader to understand the content and grasp the meaning of the text, but it may be less relevant to a researcher who wants to analyze the corpus using computational methods.

These two aspects of text complicate how to model textual data probabilistically. It is difficult to come up with just one probabilistic model $p(\mathbf{w})$ that works in all situations. Instead, we need to use different models to capture different aspects or do different predictions. The purpose of our modelling also affects the level at which we will want to analyze our textual data. If we want to analyze the meaning of a collection of texts on the sentence level (such as what a fictional character is doing), we would probably choose models that can capture the syntactic structure of individual sentences, such as probabilistic context-free grammars [18]. If we are interested in meaning on a higher level, however, such as the meaning in individual documents, we may in-

stead be interested in capturing meaning as thematic concepts using a topic model. If we would like to analyze the individual meanings of words, we might choose a word embedding model [37]. These different facets of natural language force us to create different, often simplified, models for different purposes. To paraphrase the statistician George Box, all statistical models of language are wrong, but some are useful [12].

Even though we want to use simplified models to learn from large corpora, the structure of textual data has its own peculiarities. From a statistical perspective, textual data has the property of being *sequential*, *hierarchical*, *discrete*, *sparse* and *high-dimensional*.

The first, and most obvious, property of text is that it is sequential. A piece of text is meant to be read sequentially, one word, or sign, at a time, and the writer assumes that the reader is reading the text in this order – starting from the beginning and reading to the end. In most situations a word is dependent on the words that occurred previously in the sentence, or in the text as a whole. A sentence generally depends on previous sentences, and a paragraph is most often related to previous paragraphs [45].

The second property of textual data is that it is hierarchical, or compositional [37], in structure. A collection of text, such as a book or a document, is commonly made up of many smaller components, such as sentences and paragraphs. These components are hierarchically related. Sentences belonging to the same paragraph are probably more "similar" or "coherent" than sentences belonging to different paragraphs or different books. Paragraphs belonging to the same chapter or document are, in turn, generally more similar than paragraphs belonging to different books. Also, books by the same author are in turn more similar. To be able to interpret or analyze a selection of text, we need to take this structure into account and incorporate it into our models.

The third property is the discreteness of textual data. In most languages written language is made up of a lot of small components, such as individual alphabetical symbols or characters, signs, or more generally, different words or *tokens*. In this thesis, individual word tokens, composed of characters or signs, will be referred to as w_i and a sequence of N tokens will be denoted $\mathbf{w} = (w_1, \cdots, w_N)$. A *document* is a set of tokens $\mathbf{w}_d = (w_{d,1}, \cdots, w_{d,N_d})$ and a collection of documents will be referred to as a corpus, denoted $\mathbf{w}$. For example, consider the following sentence:

> The quick brown fox jumps over the lazy dog.

In the sentence above we have nine unique word tokens. These tokens are discrete, or categorical, in that they cannot be compared among themselves numerically in any obvious way. Without resorting to semantic models, it is hard to talk about a mathematical semantic distance between "jump" and "fox", even though *we know* that individual tokens can be more "similar" to each other in a semantic sense. We know that words like "jump" and "jumping" are semantically more similar than "jump" and "fox", but this is not captured in any way by the characters that are used.

Zipf's law, a quantitative law of linguistics, states that the number of occurrences of unique words in a corpus, called *word types*, is inversely proportional to the rank of the word type in a corpus of natural language [116, Ch. 2]. Zipf's law leads to very skew distributions of word frequencies. An example of Zipf's law is presented in Figure 1.2, where the distribution of the words makes up a distribution with a few, very common words, such as "the" and "and" and a large tail of very rare words such as "xylophone". Zipf's law implies the fourth property of text as data, which is that textual data is *sparse*. Only a very small subset of the *vocabulary*, the set of all word types of any given corpus, is actually used in a given document or text segment and the largest part of the vocabulary is made up of rare or uncommon words.

The last property of textual data is that it is high-dimensional, as a consequence of the discreteness and sparsity of text. The number of word types is very large in common applications,

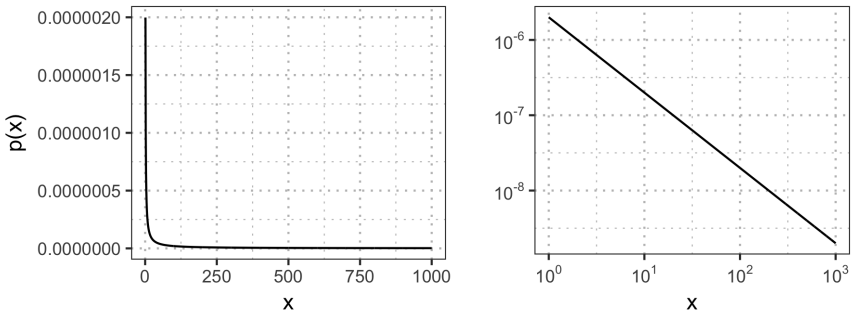


Figure 1.2: Zipf's distribution (pmf) with $s = 1$ (left) and the same distribution on a log-log scale (right).

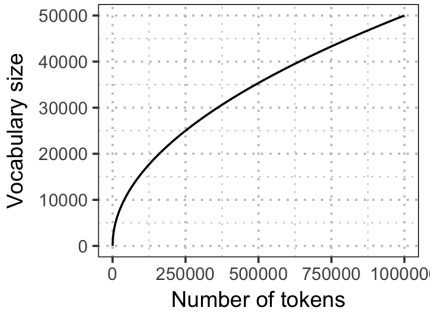


Figure 1.3: Heaps' law with $k = 50$ and $\beta = 0.5$.

ranging from tens of thousands to even millions of unique words. From a statistical perspective, where tokens are regarded as discrete entities, this results in models with a huge number of statistical parameters. The vocabulary of the corpus is, in practice, also continuously growing with the corpus size, something proposed by Harold Stanley Heaps in 1978 [47], known as Heaps' law. This law of quantitative linguistics states that the vocabulary size, V, is related to the total number of tokens, N, as $V \approx kN^{\beta}$, where α and β are corpus-specific parameters, but where $\beta < 1$ (see Figure 1.3). Hence, as the corpus grows, so does the vocabulary, with the result that original data is both high-dimensional and sparse. Most documents only make use of a very small part of the full vocabulary and the number of possible sentences or combinations of words is almost infinite. This also means that we can expect to get more high-dimensional as the size of the corpora increases.

1.3 Research questions

The purpose of this thesis is to contribute to fast and rigorous inference for different probabilistic models of text. This is done by developing new methods, approaches and knowledge that can be used in the probabilistic analysis of textual data. The main focus is Bayesian probabilistic topic models and the use, development and adaptation of these models to problems in the social sciences and the humanities. The research in this thesis can be summarized in the following research questions, which will be explained in more detail in Chapter 4.

1. How can we expand topic models to the larger data sizes that are now commonplace, without sacrificing soundness of the inference method?
2. How can we improve topic models, inference and interpretability to enable increased use in applied research settings?

Applications

Text as data is a promising field with many different applications. The following are examples of statistical inference based on large textual corpora, with an emphasis on topic models as a way of studying meaning in corpora, from which I have drawn inspiration.

Technology A common aspect of large-scale software maintenance is to handle bugs that have been identified. In large software systems, there is a system of reporting bugs to developers by using bug reports. These reports may consist of both technical information and natural language descriptions of the bug by the reporter. In this situation, we would like to use the text supplied by users as data in a bug localization system [67].

Probabilistic models for natural language can also be used to analyze computer languages or computer code [97]. As an example, in the paper "Mining concepts from code with probabilistic topic models" the authors analyze large-scale code repositories with the aim of increased understanding of software code structure and functionality, while also enabling increased code reuse and code refactoring [66].

Social science In political science, textual data has a long tradition as an empirical material for political analysis. The political scientist Justin Grimmer uses U.S. Senate press releases to study the expressed agenda of the different legislators in the U.S. Senate [41]. Recently a similar analysis of the European parliament proceedings was undertaken [38]. Grimmer models the corpus to identify the agenda of legislators in the U.S. senate. In this way he can study the policy areas that the different senators prioritize, something that has been used to study how party competition affects policy priorities [41]. He is able to show that when senators are from the same state, but from different parties, they tend to compete with each other by addressing the same topics in their press releases. However, if the senators belong to the same political party, they try to capture many different issues by dividing them between themselves, knowingly or unknowingly.

In the paper "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding" [25] the authors study 8,000 news articles on the theme of U.S. government arts funding over the period 1986–1997. The purpose is to increase understanding of the shift in public perception of arts funding that occurred during this period, as this formerly uncontroversial topic became increasingly divisive. The authors find that a shift occurred in the newspapers in 1989. They conclude that there were different ways, or frames, in which the controversy was presented in the material. It might focus on bureaucratic errors, congressional efforts to punish the National Endowment for the Arts (NEA) or seeing the debates over the NEA as a part of a larger culture war.

The (digital) humanities Literary text is the very foundation for the literary scholar, but treating the text as data, as in the case of "distance reading" proposed by Franco Moretti [79], is not as common as close, qualitative reading of individual pieces of art. However, for example, in the pamphlet "On Paragraphs: Scale, Themes and Narrative Form" [2] the authors approach the literary text by studying the paragraph as a literary component using topic models. They conclude that the paragraph is a forgotten aspect of literary corpora and can capture topical behavior better than other structures in literary text due to the inherent topical cohesion within paragraphs.

As has been pointed out by Yang et. al. [112], historical newspapers are a formidable source of information for historians, but the huge volume of newspaper articles makes it very difficult for historians to use this material efficiently for research. By analyzing newspapers in Texas during the period 1829 to 2008, both expected themes and new results were identified. In the newspapers, the heavy emphasis on the economics of cotton was expected, but using a computational approach the authors were also able to see that the battle of San Jacinto, the final battle in the Texas Revolution, was covered much more thoroughly in the newspaper than historians had previously argued. Using a topic modeling approach it was possible to both confirm previous historical knowledge and to cast new light on historical developments.

1.4 Thesis outline

The rest of the thesis is organized as follows. In Chapter 2 the general Bayesian probabilistic framework will be introduced, both from an epistemological point of view and as a method of statistical inference and analysis. In Chapter 3 an introduction to semantic modeling of textual data will be presented together with the role of Bayesian probabilistic methods in these settings. More practical aspects of inference from textual data will also be considered in this chapter. Finally, in Chapter 4, the introduction will be concluded with the guiding research questions, the overall contribution of this thesis and a summary of the papers.

2 Bayesian inference

This chapter gives a short introduction to Bayesian philosophical thinking and an introduction to Bayesian statistics, with emphasis on the purpose and role of the probabilistic model and its connection to scientific models and theory. The purpose is both to give the philosophical foundation for this thesis and to provide the necessary background for the coming chapters as well as the articles in this thesis.

2.1 Bayesian epistemology and confirmation theory

The philosophical foundation for scientific inquiry is the epistemology of science, a field of philosophy devoted to studying questions like: How can we know what we know? and How can we justify our belief? In epistemology, knowledge has traditionally been defined as *justified true belief*. This means that to be able to *know* x, we must believe that x is true, x must be true, and we need to have some kind of justification for believing that x is true [99].

Bayesian epistemology addresses the issue of how belief, or credence, should be defined. [46]. From a Bayesian epidemiological standpoint, we should formalize our credence in a given proposition x using the laws of probability. If we say that we believe x, we need to state the strength of our credence as $P(x)$, i.e. the probability of x being true. When we express our credence in terms of probabilities, we call it *subjective probabilities* [100].

A classical argument for using the laws of probability to quantify our credence is the *Dutch book argument*, initially proposed by Ramsey in 1926 [88]. The argument makes the case that if we do *not* conform to the laws of probability to express our credence, we would be willing to accept a wager that will guarantee a net loss, a so-called *Dutch Book*. So, to be rational, we need to let our beliefs or credences conform to the laws of probability. There is a lot of debate regarding the strength of the Dutch book argument and the argument has been criticized, but the Dutch book argument is one formal argument for the rational use of probability as a way of formalizing

one's credence. Alternative axiomatic approaches have also been presented by José Bernardo and Adrian Smith [7].

The benefit of stating our belief in the form of probabilities is that it gives a straightforward technical way of updating our credence in x as we observe new *evidence* E. We do this by using the conditional probability of an event A, given an event B, $P(A|B) = P(A \cap B)/P(B)$. In probability theory, the conditional probability is a definition [42]. From the perspective of Bayesian epistemology, the definition has an additional meaning. From an epistemological perspective, the definition leads to the conclusion that to be rational, we need to update our subjective probability based on new evidence, called *the simple principle of conditionalization*:

> If one begins with initial or prior probabilities P_i, and one acquires new evidence which can be represented as becoming certain of an evidentiary statement E (assumed to state the totality of one's new evidence and to have initial probability greater than zero), then rationality requires that one systematically transform one's initial probabilities to generate final or posterior probabilities P_f by conditionalizing on E [...]. [99]

The conclusion from the principle of conditionalization gives us a principle on how to update our subjective probability as we get new evidence. This leads us to Bayes' theorem. Bayes' theorem is a simple derivation from the axioms of probability and the definition of conditional probability, and can be seen as "inverting" a probability from $P(A|B)$ to $P(B|A)$, assuming $P(B) > 0$:[1]

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}.$$

Bayes theorem presents us with a formal way to update our credence in a proposition x, in the light of new evidence E. This is the foundation for Bayesian confirmation theory. Given our prior $P(x)$, our probability of the evidence $P(E)$ and our *likelihood* $P(x|E)$ we can update our subjective probability with Bayes theorem as

$$P(x|E) = \frac{P(E|x)P(x)}{P(E)}.$$

The philosophical implications of using Bayes theorem are that we can use evidence to improve our credence in a given hypothesis if the evidence supports the hypothesis. Such a position is different from the classical position of Popperianistic falsifiability. According to Popper [107], we can never verify a hypothesis, such as "all swans are white". The only thing we can do is to falsify the hypothesis by observing something that falsifies it, such as a black swan. But the Bayesian Confirmation theorist would instead, by observing an additional white swan, state that the subjective probability of the proposition "All swans are white" has increased with the new observations.

The ideas of Bayesian confirmation theory and its epistemological foundation carry over to Bayesian statistical inference [91]. To be able to use the Bayesian framework for statistical inference, we need a prior probability distribution, $p(\theta)$, for our parameters θ, together with the likelihood, $p(\mathbf{y}|\theta)$, that specifies how we assume the data, $\mathbf{y}$, has been generated, given the parameter θ. The questions of importance for Bayesian statistics concern the following two issues.

[1] Inverse probability is actually an older term. It is only recently that the term "Bayesian" has come to be common expression for inference using the inverse probability using Bayes Theorem [29].

1. How do we set up our observational model or *likelihood* $P(\mathbf{y}|\theta)$ and choose our prior distribution $p(\theta)$.
2. How do we calculate our posterior distribution $p(\theta|\mathbf{y})$, given observations?

These two problems will be addressed in the next section.

2.2 Bayesian statistical inference

Connecting epistemological reasoning with that of empirical data or observations is a statistical inferential problem. In the popular book "Bayesian Data Analysis" [34], the authors summarize Bayesian statistical analysis in the following three steps:

1. Set up a full *probability* model describing how data $\mathbf{y}$ has been generated, conditioned on (unobserved) model parameters θ through the *likelihood* $p(\mathbf{y}|\theta)$. The likelihood is combined with a *prior* over the set of parameters θ to set up the full probabilistic model.
2. Condition on the *observed data* using Bayes theorem to compute the posterior distribution for the parameters θ, $p(\theta|\mathbf{y})$.
3. Evaluate how well the model $p(\mathbf{y}|\theta)p(\theta)$ (or $p(y) = \int p(y|\theta)p(\theta)d\theta$) fits the data. Is the resulting model reasonable, or is there a need of changing the model to capture properties of importance in the data?

In addition to the three steps above, another step can be added, that of a decision based on the analysis made. These questions are addressed by the field of *Bayesian decision theory*, which is closely related to Bayesian statistical inference and Bayesian prediction theory. By combining prior knowledge and observations, the purpose is to make an optimal decision with regard to a *loss function*, $L(\theta, a)$, depending on both parameter θ and an action a [6]. The loss function measures the consequence of taking a specific action with a specific value of the parameter. In many situations, we are uncertain about the value of θ. In those cases we can represent our uncertainty of θ as a probability distribution $\pi^\star(\theta)$. Based on our uncertainty regarding θ either in the form of a prior or a posterior, we can compute the *Bayesian expected loss* for an action a as

$$E(L(\theta, a)) = \int L(\theta, a)\pi^\star(\theta)d\theta\,.$$

The expected loss for different actions guide us as a decision criterion, which allows us to choose among the different possible actions a. A common approach is to choose the action with the minimal expected loss, a decision principle called *the conditional Bayes principle*. Hence, Bayesian decision theory gives us the additional (optional) step in our Bayesian analysis [6].

- Use the final posterior $p(\theta|\mathbf{y})$ together with a decision-specific loss function $L(\theta, a)$ to compute the expected loss $E(L(\theta, a))$ for different actions a. Choose the action that minimizes the expected loss.

Approximating theory with a probability model

To do inference using Bayes theorem we need to have a full probability model that explicitly states how the data has been generated. To do this we need to specify both the likelihood $p(\boldsymbol{y}|\theta)$ and the prior $p(\theta)$ distribution. But how should we regard these model components, since they are such a crucial part of the inference we want to perform?

A "model" is a highly ambiguous concept in the scientific discourse. Models play a central role in most scientific disciplines, but at the same time it is not obvious what a scientific model is.

It is quite difficult to distinguish a scientific model from scientific theory and the concepts are often used interchangeably to describe the same thing [33]. In the social sciences it is even more complicated. In this context, "theory" can have an even broader meaning, such as taxonomies or general critical perspectives. Here I will focus on theory as a concept of (mechanical) explanatory theory (for an extended discussion, see [48]).

Based on the close connection between scientific models and scientific theory, we can use ideas about good scientific theory to define good scientific models. To judge the quality of a scientific theory, Colyvan proposes a couple of important properties [19, p. 78–79]. First, a scientific model should be able to explain our observations, i.e. both explain current observations and predict future observations. Secondly, a good scientific model should be consistent, both internally and with previous knowledge and theory. This property can be seen either as stating that the scientific model should have a mechanical explanation of observations [48] or that is should provide causal explanations [20]. Thirdly, a good scientific model needs to be simple, or parsimonious, and should attempt to explain events using a minimal number of theoretical components. Fourthly and lastly, a good scientific model should be fruitful. It should lead to new ideas and to possible new knowledge. These four aspects are all central to the quality of a scientific model [19, p. 78–79].

We can use a map as an example of a very simple (scientific) model of the world. A map should be able to do predictions on the locality of a certain point on the map - we should be able to predict what we will see in the geographical surroundings. It should be coherent with other maps of the same geographical region. Finally, it should be simple and should only present aspects that are of importance to explain the local surroundings.

A statistical, or probabilistic, model is not the same as a scientific model. A statistical model is a set of probability distributions $\mathcal{P}$ (commonly parametrized by parameters Θ) over a sample space $\mathcal{S}$ [74]. A probabilistic model specifies how observable data has been generated, given a set of parameters Θ.

Thus, a probabilistic model is, by definition, different from a scientific model or scientific theory. Unlike the scientific model, the definition of a statistical model is very clear. But similarly to a good scientific model, we can also judge the quality of different probabilistic models. First of all, a good probabilistic model should simplify a (probably) complex reality [7, p. 237]. In addition, some argue that the purpose of the probabilistic model is to approximate the *true* model that has generated the observed data. Others stress that it is the predictive performance of a statistical model that is the important quality of a probabilistic model. A model that can do a better job at predicting future (or current) observations should be preferred. A third perspective on the purpose of statistical models is that they play a central role in helping scientists to gain new insights and knowledge. Hence, from this perspective, a probabilistic model should not be evaluated by its predictive performance or whether it is *the true model*. The model is good if it is fruitful for understanding, communication between researchers, or if it can help in generating new knowledge, by [60]. Some define these as different goals as a distinction between exploratory probabilistic models and empirical probabilistic models [7, p. 238]. The main purpose of exploratory models is to explain the data using a model close to the true underlying (possible causal) process, focusing on $p(\mathbf{y}|\theta)$. The main purpose of empirical models is instead to achieve high predictive performance, focusing on $p(\mathbf{y})$.

The different purposes of the probabilistic models are connected to the ideas of good scientific theory. Both scientific models and probabilistic models should be parsimonious. Both should be able to explain both current and future observations. Both should have the role of creating new knowledge and insights.

The thought that the probabilistic models should be "true" can be regarded as closely connected to the idea of coherent scientific theory as mechanical or causal explanations of observations. We want both a scientific model and a probabilistic model to be coherent with other beliefs, theory

and knowledge. The problem is that it is practically impossible to define a *true* probabilistic model. The famous quote by George Box, "All models are wrong, but some are useful" [13], points to the fact that a true model cannot exist. Instead we need to consider a model to be a more or less of a good approximation [102]. Similar ideas are discussed by Kass and Raftery [59]:

> "Though one rarely believes a scientific law in an absolute sense, it is a great convenience to speak and to act as if laws are valid. When one says that a certain theory is correct, one means that deviations from it are sufficiently minor to be irrelevant for all practical purposes at hand." [59]

Therefore, there is a close connection between how we should define scientific theory and how we should set up our probabilistic models. Some authors even go so far as to represent the probabilistic model as a direct, formal representation of the scientific theory [11, p. 71].

In Paper VI in this thesis we try to connect political science theory on radical right parties with a probabilistic model to measure semantic meaning in the Swedish parliamentary speeches corpus.

Exchangeability and the representation of the Bayesian probabilistic model

From a Bayesian perspective, the probabilistic model has two components, the prior and the likelihood [74]. The distinction between the two can be formalized using the representation theorem, which in turn relies on the idea of exchangeability.

If we want to set up a probabilistic model for our data, we have to specify a probability distribution over our observed data as $p(y_1, y_2, ..., y_n)$. In addition, we can assume our observations to be *exchangeable*, i.e. that $p(y_1, y_2, ..., y_n) = p(y_{\omega(1)}, y_{\omega(2)}, ..., y_{\omega(n)})$ for any permutation of the index ω [7]. Based on the assumption of exchangeability, we can use the de Finetti *representation theorem* to show that we can represent our distribution over our observations $p(y_1, y_2, ..., y_n)$ in the form of a likelihood and a prior. As an example, the representation theorem for $p(y_1, y_2, ..., y_n)$ being a joint probability distribution of 0–1 random variables, states that

Theorem 1 *If $y_1, y_2, ...$ is an infinite exchangeable sequence of 0–1 random quantities with probability measure P, there exists a distribution function Q such that the joint mass function $p(y_1, y_2, ..., y_n)$ for $y_1, y_2, ...$ has the form*

$$p(y_1, y_2, ..., y_n) = \int_0^1 \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} dQ(\theta)\,,$$

where,

$$Q(\theta) = \lim_{n\to\infty} P[x_n/n \leqslant \theta]\,,$$

with $x_n = y_1 + \cdots + y_n$, and $\theta = \lim_{n\to\infty} x_n/n$. [7, p. 172]

So, given exchangeability of our observations, in this case binary numbers, we can represent our data in the form of independent and identically distributed Bernoulli random variables conditional on a parameter θ, together with a prior distribution $Q(\theta)$. This gives us a formal

argument of representing our probability distribution over observations as a likelihood and as a prior - based only on the assumption of exchangeability of our observations. The representation theorem shows that we can factorize our probability model into a model of the data generation process, the likelihood, and to a model of our belief, the prior. [7, p. 237]

2.3 Examples of probabilistic models

The examples below give a more concrete example of Bayesian probabilistic models and Bayesian inference for these models. The purpose is both to provide an introduction to the Bayesian inferential paradigm and to give examples that will extend to the situation of Bayesian modeling of textual data.

A Multinomial-Dirichlet model example

As a simple example of how to conduct a Bayesian analysis, we can study the analysis of political affiliations or vote intentions. A common approach to studying the current voting intentions of a population is to ask a random sample of the population the question "What party would you vote for if today was election day?". The answer to this question is discrete in the form of different political parties. Based on the answers, the voting intentions of the whole population are then inferred. This example focuses on political parties, but in later chapters we will see the same modeling approaches used for textual data, but with word types instead of political parties and word frequencies instead of the number of party votes.

To answer the question of current voting intentions in the population, we need to follow the three steps defined above. Our first step would be to define a full probability model for our observed data. We would, in accordance with Box' process, start with a very simple model for our observations. The simplest model would be to assume that the data from our survey, $\mathbf{y}$, can be seen as counts of voting intentions for different political parties. A good probability model for this simple model would be a multinomial probability distribution with an underlying vector of (unobserved) party preference proportions θ, the proportions that are of interest to us. This probability distribution has the probability mass function (pmf):

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{n!}{y_1! \cdots y_k!} \theta_1^{y_1} \cdots \theta_k^{y_k} ,$$

where

$$n = \sum_{k=1}^{K} y_k .$$

Together with our model for the data, we also need to specify a prior distribution for θ. A common choice would be a Dirichlet distribution for θ. The Dirichlet distribution is a multivariate probability distribution over the simplex with the following probability density function [83]:

$$p(\theta|\boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} ,$$

where

$$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}$$

is the generalized Beta function, and

$$\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)\,.$$

Now, when we have specified our likelihood and prior, the second step in our analysis would be to use our model, prior and condition on observed data. In our case, this would be a sample of voting intentions for different political parties, $\mathbf{y}$. Given Bayes' theorem we can now compute the posterior distribution of the voting intentions $\mathbf{p}$ for the different political parties as:

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta) \cdot p(\theta)}{p(\mathbf{y})}$$

In this case we can compute the posterior distribution analytically:

$$\begin{aligned} p(\theta|\mathbf{y}) &= \frac{p(\mathbf{y}|\theta) \cdot p(\theta)}{p(\mathbf{y})} \\ &= \frac{\mathrm{B}(\boldsymbol{y}+1)^{-1}\theta_1^{y_1}\cdots\theta_k^{y_k} \cdot \mathrm{B}(\boldsymbol{\alpha})^{-1}\theta_1^{\alpha-1}\cdots\theta_k^{\alpha-1}}{\mathrm{B}(\boldsymbol{\alpha})^{-1}\mathrm{B}(\boldsymbol{y}+1)^{-1}\mathrm{B}(\boldsymbol{\alpha}+\boldsymbol{y})} \\ &= \frac{1}{\mathrm{B}(\boldsymbol{\alpha}+\mathbf{y})}\theta_1^{y_1+\alpha-1}\cdots\theta_k^{y_k+\alpha-1} \end{aligned} \tag{2.1}$$

since

$$\begin{aligned} p(\mathbf{y}) &= \int_\theta p(\mathbf{y}|\theta)p(\theta)d\theta \\ &= \frac{\Gamma(\sum y_k+1)\Gamma\left(\sum\alpha\right)}{\Gamma\left(\sum\alpha+y_k\right)}\prod_k^K\frac{\Gamma(y_k+\alpha)}{\Gamma\left(y_k+1\right)\Gamma(\alpha)} \\ &= \mathrm{B}(\boldsymbol{\alpha})^{-1}\mathrm{B}(\boldsymbol{y}+1)^{-1}\mathrm{B}(\boldsymbol{\alpha}+\boldsymbol{y})\,. \end{aligned} \tag{2.2}$$

We see from Equation 2.1 that the posterior distribution for the party support proportions is a Dirichlet distribution with parameters $\alpha + \mathbf{y}$, a standard result in Bayesian statistics. This property, that the posterior distribution is of the same family as the prior, or is closed under sampling, is called *conjugacy* or conjugate priors [24]. The fact that the Dirichlet is a conjugate prior of the multinomial distribution simplifies computations. This will be an important building block in the probabilistic modeling of textual data, where we can make use of parameters that are *conditionally conjugate* given other model parameters.

The third step concerns how well this model fits our data. Do we need to revise the model? Do we need to extend it? In this case, we could expect survey sampling effects and errors, such as effects of non-response and other survey imperfections. This could then be added to the model to handle such imperfections in our model [35].

This simple example, summarized below, shows the basic steps of a Bayesian analysis in the case of discrete data or counts. Here the example is just a simple toy example, but it is an important building block for more elaborate models in probabilistic modeling of textual data, presented in Chapter 3.

$$\begin{aligned} \text{Prior} &: \theta \sim \text{Dirichlet}(\alpha) \\ \text{Likelihood} &: \mathbf{y}|\theta \sim \text{Multinomial}(\theta) \\ \text{Posterior} &: \theta|\mathbf{y} \sim \text{Dirichlet}(\alpha + \mathbf{y}) \end{aligned}$$

Some special cases of the multinomial and Dirichlet distribution are commonly referred to with other names. In the case $K = 2$, the multinomial distribution is reduced to the Binomial distribution and the Dirichlet distribution is reduced to the Beta distribution. In the case with a multinomial distribution with only one draw (i.e. $n = 1$), the multinomial distribution is reduced to the Categorical distribution.

Simulating from Dirichlet and Multinomial distributions

Simulating from well-known probability distributions is important in many situations in Bayesian inference, especially in simulation-based inference (see Section 2.4). One aspect of this thesis is to show that by using clever tricks to generate samples for Dirichlet and multinomial distributions, it is possible to improve performance in simulation-based inference for large-scale textual analysis problems.

There are multiple methods for generating a draw from the Dirichlet distribution, such as using draws from the Polya-Urn distribution, Stick-breaking and, as is probably the most common approach, a normalized vector of independent $\Gamma(\alpha_i, 1)$ draws where $\Gamma(\alpha, \beta)$ is the Gamma distribution. We show in Paper II in this thesis that it is also possible to approximate a draw from the Dirichlet distribution using a normalized vector of Poisson random variables as $\text{Po}(\alpha_i)$. This approximation will converge in distribution to the Dirichlet distribution as $\sum \alpha_i \to \infty$. The approximation has the benefit of being a *discrete* approximation of the continuous Dirichlet distribution and can be used to reduce computational complexity in models using Dirichlet priors such as large-scale topic models.

Drawing samples from a multinomial distribution with parameters $\theta_1, ..., \theta_K$ can, as with the Dirichlet distribution, be done in many different ways. The simplest approach would be to draw a uniform random variable u_i and then iterate over the cumulative sum of the θ:s to determine the category, or slot, into which the random value u_i falls. The problem with this naive approach is that the cost of drawing a categorical variable will increase with the number of categories. In a paper from 1977, Alastair Walker propose an alternative approach [109]. By precomputing probability tables (so called Alias tables) for a given value of θ, we can improve the computational efficiency in sampling categorical variables from the multinomial distribution. If we have a computed Alias table we can draw a uniform random variate u_i and then simply look up the category corresponding to this value, at a computational cost that is independent of the number of categories. This can be of huge importance if the number of categories is large, such as in large topic models.

Finite mixture models

A slightly more complex model example is *mixture models*, which represent unknown subpopulations, or clusters, in a population. A finite mixture model assumes that there is a finite number of sub-populations and that the observed data is a mixture of these sub-populations. This mixture distribution can be expressed probabilistically as follows

$$p(y|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_k^K \pi_k p(y|\theta_k)\,,$$

1. $\boldsymbol{\pi} \sim \text{Dirichlet}(\beta)$
2. For each component k to K:
   a) $\boldsymbol{\theta}_k \sim \text{Dirichlet}(\alpha)$
3. For each observation i:
   a) $z_i|\boldsymbol{\pi} \sim \text{Categorical}(\boldsymbol{\pi})$
   b) $\mathbf{y}_i|z_i \sim \text{Multinomial}(\theta_{z_i}, n_i)$

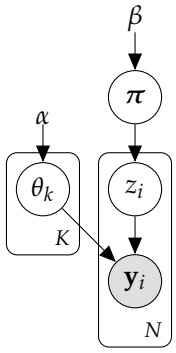


Figure 2.1: A multinomial finite mixture model generative process (left) and graphical model (right).

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, π_k is the proportion of mixture component k and $p(y|\theta_k)$ is an arbitrary probability distribution for the observed data y in the kth component with parameter θ_k.

We can use a mixture model to extend the simple Multinomial-Dirichlet model in the previous section. By assuming that each $p(y_k|\theta_k)$ is a multinomial distribution we can build a more complex model for the political example. Figure 2.1 shows a *generative* mixture model of multinomial distributions where $\boldsymbol{\theta}_k$ is the proportions of the multinomial distribution and n_i is the total number of observations from that draw. For an example of data generated from this model, see Figure 2.2.

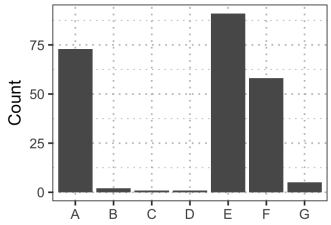
This mixture model is a simple example of how the multinomial and Dirichlet distributions can be combined into a more expressive model. Analogously to the previous example, this model could model voter intentions in different voting districts where there are different voting patterns, which has been done in the paper "A Bayesian cluster analysis of election results" [86]. This model, though simple, has worked well in the area of text clustering, when used to cluster tweets. We will return to this in Chapter 3 [114].

Finally, this model assumes that we know the number of clusters K a priori, which is seldom the case in practical applications. Instead, we would like to infer K as well by generalizing the Dirichlet mixture model to the Dirichlet process (DP) mixture model.

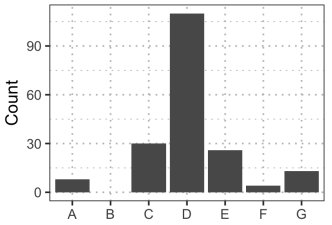
Infinite mixture models

The Dirichlet distribution is a distribution over a finite number of categories and can be used in finite mixture models. The finite mixture model can be generalized to an *infinite* number of components using the Dirichlet process. As has been shown by Radford Neal [81], we can arrive at the Dirichlet process mixture model by parameterizing the Dirichlet prior over the components as $\text{Dir}(\gamma/K)$ and let $K \to \infty$ (see Figure 2.1 and set $\beta = \gamma/K$)).

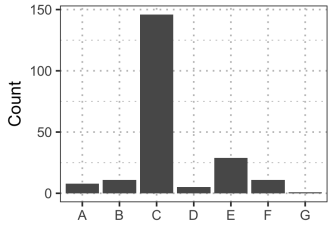
Another way of characterizing the DP in infinite mixture models is to use a stick-breaking construction of the Dirichlet process [104]. Using the stick-breaking representation of the model we can construct a $\text{DP}(\gamma, G_0)$ as:



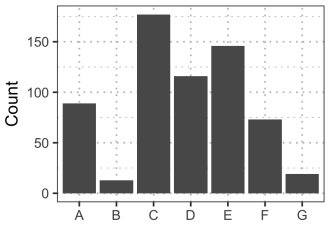
(a) Cluster A



(b) Cluster B



(c) Cluster C



(d) The full data, from all clusters.

Figure 2.2: Example of data generated from a multinomial finite cluster model with $K = 3$ clusters and $\alpha = 0.5$.

$$
\begin{aligned}
\beta_k &\sim \text{Beta}(\gamma, 1) \\
\pi_k &= \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \\
\theta_k^\star &\sim G_0 \\
G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^\star}
\end{aligned}
$$

Dirichlet process mixture models [3, 27] are popular *infinite* mixture models due to their simplicity, especially when G_0 is a conjugate to the observations . As a simple example, we can extend the finite mixture model in Figure 2.1 to a situation where we do not know the number of components, K, a priori. Instead, we obtain the posterior distribution for K conditional on our observations. The Dirichlet process infinite multinomial mixture model (DPMMM) is presented in Figure 2.3 where a Dirichlet prior G_0 is used. This model can then be inferred using different algorithms [81, Ch. 3]. In the case of twitter message clustering, such a model would infer the number of different clusters of tweets that exist, or in the case of vote district structure, the number of different clusters of voting structures that exist. None of these problems have, to my knowledge, been studied using DP multinomial mixture models.

One important property of the Dirichlet process is that we can approximate the expected number of clusters or components as a function of data size. The expected number of clusters in a Dirichlet process mixture model can be approximated as

1. $G_0 \sim \text{Dirichlet}(\beta)$
2. $G \sim \text{DirichletProcess}(G_0, \alpha)$
3. For each observation i:
   a) $\theta_i | G \sim \text{G}$
   b) $\mathbf{y}_i | \theta_i \sim \text{Multinomial}(\theta_i, n_i)$

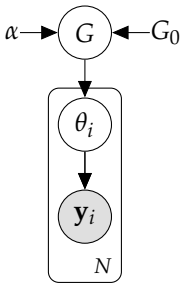


Figure 2.3: A multinomial infinite Dirichlet process mixture model generative process (left) and the graphical model (right).

$$E(K) \approx \gamma \cdot \log\left(1 + \frac{N}{\gamma}\right),$$

where γ is the concentration parameter of the Dirichlet process and N is the number of observations [104]. This property gives us some theoretical justification on how we can expect clusters to grow in the case of increasing data, a result used in Paper I and Paper II in this thesis.

2.4 Simulation-based statistical inference

When we are doing Bayesian statistical analysis we are interested in the posterior distribution $p(\theta|\mathbf{y})$. How can we compute this posterior distribution in more complex models? Let us study the example with a finite mixture model presented in Figure 2.1. We are, in this case, interested in the posterior distribution

$$p(\boldsymbol{\pi}, \mathbf{z}, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K | \mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\pi}, \mathbf{z}, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K) \cdot p(\boldsymbol{\pi}, \mathbf{z}, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K)}{p(\mathbf{y})}.$$

If we want to analyze this posterior distribution we will have to compute

$$p(\mathbf{y}) = \sum_{z_1} \cdots \sum_{z_N} \int_{\theta_1} \cdots \int_{\theta_K} \int_{\pi} p(\mathbf{y}|\boldsymbol{\pi}, \mathbf{z}, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K) d\boldsymbol{\pi} d\boldsymbol{\theta}_K \cdots d\boldsymbol{\theta}_\mathbf{1}.$$

In this case, we see that this expression needs to sum over K^N, a sum that is complicated in most practical situations. This is an example of an integral that is practically impossible to solve for real problems, where N is sufficiently large. The problem is further complicated by the fact that we are not only interested in the joint posterior distribution as such, but also in different properties of the posterior distribution, such as the expected value or the variance of the posterior parameters. To compute these properties we need to compute the integral

$$\int f(\theta) p(\theta|\mathbf{y}) d\theta,$$

where $p(\theta|\mathbf{y})$ is the posterior distribution for the parameter θ and $f(\theta)$ is an arbitrary function of the posterior parameter θ. To be able to do Bayesian inference, we need to *approximate* the posterior distribution $p(\boldsymbol{\pi}, \mathbf{z}, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K | \mathbf{y})$ in a way that will allow us to both compute the joint

posterior distribution and compute integrals, such as the expected value, for the posterior parameters.

Monte Carlo inference

In the situation where we want to compute an integral of the type $\int f(x)p(x)dx$ where $p(x)$ is a probability distribution from which we can draw samples, we can approximate the integral as:

$$g = \int f(x)p(x)dx \approx \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)}) = \hat{g}\,,$$

where $x^{(s)}$ is a sample drawn from the probability distribution $p(x)$. This approach to integration has the benefit that the approximation $\hat{g}$ of g is unbiased (i.e. $E(\hat{g}) = g$) and the variance of $\hat{g}$ decreases with the number of samples ($Var(\hat{g}) = \frac{1}{S}Var(g)$), independent of the dimension of x. This property is very appealing since it means that we can get an increasingly better and better approximation of g by simply drawing more and more samples from the posterior distribution, the dimensionality of x being of no consequence.

But the question of how we can draw samples from the posterior distribution $p(\theta|\mathbf{y})$ still remains. In our example in Figure 2.1 we have a complex joint posterior distribution with an integral that is an exponential sum with the number of data points, $p(\mathbf{y})$.

Markov Chain Monte Carlo

If we want to sample from a complex posterior distribution $p(\theta|\mathbf{y})$, we can do this using a *Markov chain*. A Markov chain is a sequence of random variables, $X_0, X_1, \cdots$ that satisfies

$$P(X_n, A) = P(X_n \in A|X_{n-1}, \cdots, X_0) = P(X_n \in A|X_{n-1})\,,$$

where $P(X_n \in A|X_{n-1})$ is the *transition kernel* from X_{n-1} to X_n and

$$P(X_n \in A|X_0) = P^n(X_0, A)$$

is the n-step transition distribution from X_0 using the transition kernel P [108]. The probability distribution π is a *stationary*, or *invariant*, distribution of a Markov chain if

$$\pi(A) = \int P(x, A)\pi(x)dx\,,$$

or $\pi P = \pi$.

A Markov chain where $P^n(x, A) > 0$ for a finite n is called *irreducible*, meaning that the chain $P^n(x, A)$ can always reach a state x in a finite number of steps.

The periodicity of a Markov chain can be informally explained as: if the periodicity of a Markov chain is i, the Markov chain can only return to the starting state x_0 in i steps, or multiples thereof. If a Markov chain has periodicity 1 the chain is acyclic (or aperiodic). Another property of Markov Chains is recurrence, such as positive recurrence and Harris recurrence. These are, informally, Markov chains that can return to any given region infinitely often with positive probability according to π (for a formal definition of recurrence, Harris recurrence and acyclic Markov chains, see [108] or [63]). If a Markov chain is both aperiodic and positively recurrent it is called an *ergodic* Markov chain.

The total variation distance between the two distributions $\tilde{\pi}$ and π is defined as

$$\|\tilde{\pi} - \pi\| = \sup_{A \in \mathcal{F}} |\tilde{\pi}(A) - \pi(A)|,$$

where $\mathcal{F}$ is a sigma-algebra on some sample-space Ω.

Based on the described properties it is possible to present the following theorem [108]:

Theorem 2 *Suppose P is π-irreducible and $\pi P = \pi$. Then P is positive recurrent and π is the unique invariant distribution of P. If P is also aperiodic, then, for π-almost all x*

$$\|P^n(x, \cdot) - \pi\| \to 0\,,$$

where $\|\cdot\|$ denote the total variation distance. If P is Harris recurrent, then the convergence occurs for all x.

Theorem 2 gives us the guarantee that we can generate samples from our posterior distribution if we set up an irreducible and recurrent Markov Chain P with our posterior distribution as the invariant distribution π and run the Markov chain long enough. Based on the draws from the posterior distribution (or rather the Markov chain samples) we can then do Monte Carlo inference for our parameters of interest. The problem, then, has to do with constructing a Markov Chain P with the posterior as the invariant distribution.

Theorem 2 is very important for this thesis. In Paper I and Paper II we develop large-scale Markov chains that are computationally efficient, concurrent, and will converge to the true posterior distribution. Many other approaches to parallelizing algorithms that are currently in use do not come with this guarantee, and in Paper I we show some of the effects this can have. The importance of this theoretical guarantee is even more profound in large unsupervised models, such as probabilistic topic models, where it can be difficult to evaluate the models.

The Metropolis-Hastings algorithm

There are many different approaches to setting up a Markov Chain that will converge to the invariant distribution that is the posterior distribution of interest. One common approach is to use the Metropolis-Hastings algorithm to construct an ergodic Markov Chain. We do this by sampling values x_n from a *proposal* distribution $q(x_n|x_{n-1})$ and then, as the final step, we decide if we should accept the new draw by computing the acceptance probability a of the transition from x_{n-1} to x_n

$$a = \min\left(1, \frac{\pi(x_n)}{\pi(x_{n-1})} \frac{q(x_{n-1}|x_n)}{q(x_n|x_{n-1})}\right),$$

where $\pi(x_n)$ is the unnormalized posterior density evaluated at x_n. This will result in an ergodic Markov chain with the stationary distribution π, and in this way we can generate samples from the posterior distribution.

Gibbs sampling

Another approach to constructing a Markov chain for a posterior distribution π is to sample each group of parameters iteratively from its respective full conditional posterior distribution

$p(x_{\mathcal{I}}|x_{-\mathcal{I}})$, where $\mathcal{I}$ is the set of parameter coordinates that is updated and $x_{-\mathcal{I}}$ are the remaining parameters. This approach, called *Gibbs sampling*, will also result in an ergodic Markov Chain that targets the joint posterior distribution π [36].

One of the benefits of the Gibbs sampler is that it can exploit conditional conjugacy in inference. For example, in a mixture model we can make use of the fact that, given latent cluster indicators, other parameters may be conditionally conjugate, like in a normal finite or infinite mixture model. This makes it simple to build more complex models but still have a straightforward inference algorithm. The idea of using the Gibbs sampler in mixture models is closely connected to the idea of the EM algorithm where the expectation over latent cluster distributions can simplify otherwise complex maximum likelihood estimates [23].

Gibbs sampling for the multinomial mixture model

As an example, we construct a Gibbs sampler for the posterior of the multinomial mixture model presented in Figure 2.1.

The full joint posterior distribution is

$$p(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K|\mathbf{y}) .$$

To sample from this distribution we can sample a subset of parameters at a time, conditioning on the other parameters and using conditional conjugacy. We can do the sampling from the posterior in the following three steps, by first initializing the parameters $\boldsymbol{\pi}, \mathbf{z}, \boldsymbol{\theta}_1 \cdots \boldsymbol{\theta}_K$ to a random starting state X_0.

1. Sample $\boldsymbol{\pi}|\mathbf{z}$
   We start by sampling $\boldsymbol{\pi}$ using the conditional conjugacy property.
   $$p(\boldsymbol{\pi}|\mathbf{z}) \sim \text{Dir}\left(\alpha + \mathbf{n}^{(c)}\right) ,$$
   where $\mathbf{n}^{(c)}$ is a vector of length K containing the number of observations belonging to the different components.
2. Sample $\boldsymbol{\theta}_1 \cdots \boldsymbol{\theta}_K|\mathbf{z}, \mathbf{y}$
   For each component $1, ..., K$ we then sample each vector $\boldsymbol{\theta}_k$ of length L, again using the conditional conjugacy as follows:
   $$p(\boldsymbol{\theta}_k|\mathbf{z}, \mathbf{y}) \sim \text{Dir}\left(\beta + \mathbf{n}_k^{(y)}\right) ,$$
   where $\mathbf{n}_k^{(y)}$ are the frequencies of counts of $\mathbf{y}$ that, conditioned on cluster assignment $\mathbf{z}$, belongs to component k.
3. Sample $\mathbf{z}|\boldsymbol{\pi}, \boldsymbol{\theta}_1 \cdots \boldsymbol{\theta}_K, \mathbf{y}$
   Finally we need to sample the component assignments for each observation i, $\mathbf{z}$ conditioned on $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K$. Since this is a discrete parameter, the sampling is reduced to computing the probabilities of each component 1 to K and then draw a component k from a Categorical distribution with probabilities
   $$p(z_i = k|\boldsymbol{\pi}, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K, \mathbf{y}_i) \propto \pi_k \cdot \prod_l^L \theta_{k,l}^{y_{l,i}} ,$$
   where $y_{l,i}$ is the multinomial frequency of the lth category of observation $\mathbf{y}_i$.

These three steps make up an ergodic Markov chain that will converge to the stationary distribution $p(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\theta}_1 \cdots \boldsymbol{\theta}_K | \mathbf{y})$ that is our joint posterior distribution. Sample averages converge to posterior expectation, just as in Monte Carlo sampling. This is true even if the generated samples from the Markov chains are auto-correlated. But auto-correlated samples makes the estimation of the expectations less efficient (higher variance).

3 Probabilistic latent semantic modeling of text

This chapter introduces probabilistic models for textual data, with the focus on latent semantic modeling and, more specifically, probabilistic topic models.

3.1 Modeling semantics

One famous quote from the area of quantitative linguistics comes from Firth in his paper *Studies in linguistic analysis*:

> "You shall know a word by the company it keeps." [31]

This quote summarizes the idea that the meaning of words, their *lexical semantics*, are defined by the textual context in which they appear. The word "cold", for instance, has very different meanings in different contexts. The sentence "I have a cold" employs a very different meaning of "cold" than the sentence "The soup is cold."

Semantic analysis is the analysis of the meaning of words. Many different approaches and methods exist, from purely linguistic or logical analysis to computational and statistical methods [89]. In [62] the linguist Geoffrey Leech proposes three main groups of semantics.

1. The *conceptual* meaning is the core functionality of language, to express the explicit meaning of individual words or symbols to convey a message from a writer to a reader. The conceptual meaning of a word is the definition of the word we look up in a dictionary.
2. The *associative* meaning is additional meaning that is connected to a word, but which is not directly related to its conceptual meaning. In [62], multiple types of associative meanings are proposed. One such meaning is the *connotative* meaning, which is the way a reader connotes words to other words. The word "cow" has the conceptual meaning of

a large domesticated mammal with the scientific classification Bos taurus. The connotative meaning may, on the other hand, be different for different people. For some people, the word "cow" may relate to words such as "milk" or "beef", while others might relate "cow" to "green fields" or "summer". Another aspect of the associative meaning is the *social* meaning of words. By using a word from a specific dialect, the word gives information about the origin of the writer or character and about the context of the word. Another example of social meaning is the provenance of a word, for example if it is in the context of a news item or a scientific article. Using the word Bos taurus for cattle conveys information about the provenance of the word, in this case that it is probably being used in some kind of scientific context.

3. The *thematic* meaning of words depends on the ordering, emphasis, and focus. As an example, words can often have different meanings in active sentences than in their passive counterparts.

As we can see from the categorization of [62], the way meaning is expressed can be quite complex. Different aspects of a word may present different types of semantics. We can also see that the associative meaning is a concept that is very vague, opaque, and dependent in large part on the *context*, be it the context of other words or the context of the writer or reader. As is proposed by [62, p. 18f.] these aspects lend themselves better to be understood approximately using statistical methods.

In this thesis emphasis will be placed on *topical* semantics or meaning, which is very closely related to the associative meaning that is defined by Leech [62]. As has been discussed by [40], topical meaning can be seen as a *contextual associative relationship*. As an example, taken from [40], the word "bird", might give us associations to words such as "sing", "fly", and "nest". However, if the context, such as the document where we find the word "bird", contains words such as "turkey" and "dinner", we would probably get another connotative meaning of the word "bird", to use Leech's definition [62], to other, more food-oriented words. This thesis concerns issues within the area of statistical, or probabilistic, methods for topical semantic analysis.

3.2 Probabilistic modeling of textual data

Treating text as data in a probabilistic framework means that we need to define a probabilistic model for our corpus $\mathbf{w}$. Based on this model, we can then use the inferential engine of Bayes' theorem in Chapter 2 to compute our posterior distribution, given our corpus. Then, using our posterior distribution we can answer research questions that are of interest to us. To be able to analyze our corpus using statistical methods, we need to cast our data in the form of a probabilistic model. One common approach is to treat each word, or token, as an individual *discrete* observation, w_i, with the collection of the word tokens, $\mathbf{w}$, making up the data or observations. In this way, we can specify a probabilistic model over the individual words or tokens as $p(w_1, \cdots, w_N)$ where N is the total number of tokens in the corpus.

The idea of using probabilistic models for text is far from new and a multitude of different models exist. During the 21st century, Bayesian methods for probabilistic model inference have become more popular and have shown progress in a multitude of different text modeling areas, such as unsupervised text segmentation [87], named entity recognition [30], unsupervised coreference resolution [43], probabilistic context-free grammars [53] and unsupervised language learning methods [98]. As the examples show, Bayesian inferential methods have shown very good performance in unsupervised settings, settings where we try to learn underlying structures rather than focusing on predicting observations with high accuracy. Unsupervised methods are appealing in the field of modeling text, due to the very large amount of unlabeled "unstructured" textual data, and they are even more appealing in the case of semantic analysis where supervised "true" labels are very rare [62].

1. $\boldsymbol{\pi} \sim \text{Dirichlet}(\beta)$
2. For each category k to $|\mathcal{C}|$:
   a) $\theta_k \sim \text{Dirichlet}(\alpha)$
3. For each document d:
   a) $c_d \sim \text{Categorical}(\boldsymbol{\pi})$
   b) $\mathbf{w}_d \sim \text{Multinomial}(\boldsymbol{\phi}_{c_d}, n_d)$

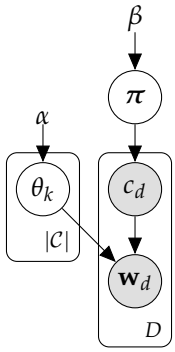


Figure 3.1: The multinomial Naive Bayes generative model.

In the following sections, we will focus more on probabilistic models for documents or larger text segments. This can be contrasted against the NLP applications that were presented previously, models that put more emphasis on the sentence as the unit of interest. This moves us in the direction of document classification, clustering, and eventually latent semantic modeling of documents.

A classical model for modeling documents or text segments is the Naive Bayes classifier. The purpose of the model is to classify a document $\mathbf{w}_d$ to a class $c_d \in \mathcal{C}$, a common example being the classification of spam e-mails vs. ham (or good) e-mails. This is an example of supervised classification. We know the classes, c_d for a number of documents or text segments. The idea is to use Bayes' theorem to do the classification of a given document d as [58, Ch. 6]:

$$p(c_d|\mathbf{w}_d) = \frac{p(\mathbf{w}_d|c_d)p(c_d)}{p(\mathbf{w}_d)}\,.$$

A common approach is to regard each document as a *bag-of-words*, i.e. the order of the words does not matter. One way of modeling this is to let the likelihood be a generative model where each category c_i is a multinomial distribution over the vocabulary, with parameter $\boldsymbol{\phi}_{c_d}$. The assumption that words are independent is the naive assumption that gives the classifier its name. Even though non-generative text classification algorithms such as `fastText` [56] are popular today, Naive Bayes approaches are still popular and effective for large-scale text classification [57].

The multinomial Naive Bayes model has a close connection with the multinomial mixture model presented in Figure 2.1. In this model, we assume that our observations are made up of a mixture of multinomial distributions. If we treat our multinomial distribution as a distribution over our vocabulary and treat the mixture components $\mathbf{z}$ as known or observed, we will actually end up with the generative model of the Naive Bayes classifier [84]. The generative model of the Naive Bayes classifier is shown in Figure 3.1, with the simplifying assumption that the number of words per document, n_d, is known.

In the multinomial Naive Bayes model we treat the classes as being observed. If we relax this assumption, and regard the underlying classes, or clusters, as unknown, we will end up with a simple model for document clustering, the multinomial mixture model. In this clustering model we regard the corpus $\mathbf{w}$ as a mixture of document clusters, but unlike the Naive Bayes model, we do not know which cluster each document belongs to. This is exactly the model presented in Section 2.1 and it is a model that has worked well to cluster tweets [114]. Using the

Gibbs sampler in Section 2.4 we can do inference to learn the underlying clusters to which the individual documents belong.

The model has, as previously stated, been shown to perform well in the case of clustering tweets [114] and very short documents, but it has been applied to many other different tasks. As an example we can use the same model, due to its close connection with the Naive Bayes classifier, to augment the Naive Bayes classifiers with unlabeled documents [84]. But as has been shown by Nigam et. al. [84] there may be a need for multiple clusters to capture the structure within each category used by the Naive Bayes classifier, indicating that it is necessary to use multiple multinomial distributions to represent documents, even within known categories. This document clustering model is one example of an approach to modeling latent topical semantic structure, by capturing a structure of words that belongs to the same cluster.

3.3 Latent semantic modeling

Latent semantic modeling methods allow for the modeling or analysis of the latent, or unobserved, meaning in a given corpus. The idea of latent semantic modeling has a long history and originated with the distributional hypothesis proposed by Firth [31]. This relatively simple idea, that the meaning of the words is defined by their contexts, has generated many different approaches to estimating the semantics from a given corpus. The main idea of most approaches to distributional semantic modeling is to reduce a large, sparse, word-context matrix into a lower, more dense, representation. This lower representation, hopefully, captures the latent semantic structures in the corpus. As an example, consider the following two sentences:

1. "A friend in need is a friend indeed."
2. "She is my friend indeed."

If we do a bag-of-words assumption, similarly to the Naive Bayes example, we can represent these sentences in the two matrices in Table 3.1 and Table 3.2, by our definition of context.

	a	friend	in	indeed	is	my	need	she
Sentence 1	2	2	1	1	1	0	1	0
Sentence 2	0	1	0	1	1	1	0	1

Table 3.1: A word–context matrix using sentence as context (document–term matrix).

The basic idea of many latent semantic representations is to reduce these matrices of co-occurring words in different contexts to a lower, denser representation. From a linear algebra perspective we can regard this as doing a matrix factorization of the co-occurrence matrices in Table 3.1 and Table 3.2 into matrices of lower rank that approximate to the original co-occurance matrix.

Modeling latent semantic meaning can be done in a variety of ways. We can choose different definitions of context, we can choose different approaches to estimating the semantical representations, and we can choose different approaches to represent the latent semantic dimensions. Table 3.3 summarizes different approaches to latent semantic modeling based on the aspects *context*, *representation*, and *estimation*.

The first aspect, contexts, concern how we define the semantic context when modeling. Here we can identify two common approaches, which have also been expressed in the two different co-occurrence matrices above. Word2vec [75] and Random indexing [94] are examples of methods

	a	friend	in	indeed	is	my	need	she
a	2	2	0	0	1	0	0	0
friend	2	3	1	2	0	1	0	0
in	0	1	1	0	0	0	1	0
indeed	0	2	0	2	0	0	0	0
is	1	0	0	0	2	1	1	1
my	0	1	0	0	1	1	0	0
need	0	0	1	0	1	0	1	0
she	0	0	0	0	1	0	0	1

Table 3.2: A word–context matrix using words as context, with a word window size of 1 (term–term matrix).

that define the context of a word as a *word window*, where the context of a word w_i consists of the words $w_{i-L}, \cdots, w_{i-1}, w_{i+1}, \cdots, w_{i+L}$ where L is the window, or context, size. Topic models and Latent Semantic Analysis (LSA) [61], however, define the context of a word as the *document* in which the word occurs, which can be seen roughly as the largest possible word window for documents. As mentioned in Section 3.1, the semantic meaning of a word is defined by the context. Different contexts will affect the aspect that is captured. Topic models and LSA, with larger contexts, will generally capture more meaning, while models using smaller contexts, such as Brown clusters, will capture more syntactic structure [58, Ch. 16].

The second aspect that differentiates between approaches is how the latent vectors are being represented. There are two primary ways that latent models represent dense vectors, either as real vectors on $\mathbb{R}$ or as positive vectors on $\mathbb{R}_+$, the simplex being a special case. This is an example of the difference between LSA and the common topic model Latent Dirichlet Allocation (LDA). LSA represents meaning, using real-valued vectors, while LDA uses simplex (or Dirichlet-distributed) vectors. This difference between LSA and LDA can also be viewed from a linear algebra perspective, where LSA can be seen as a general matrix factorization (in LSA, singular value decomposition is used to do a Matrix factorization), while many topic models, such as LDA, have a close connection to non-negative matrix factorization [26, 77]. The two representations show a trade-off in latent semantic modeling. Using non-negative representations of meaning makes the representations more interpretable, but can also make estimation potentially more complex or drag down performance [76, 68].

The last aspect of choosing an approach to latent semantic modeling is the estimation of the latent semantic representations. The main distinction here is between probabilistic methods and more general computational approaches. Probabilistic approaches using Bayesian estimation or maximum likelihood methods rely on a generative probabilistic model for which inference is then conducted. This can, in turn, be done using different estimation strategies, such as MCMC or Variational inference [8]. Probabilistic methods can also use other estimation strategies, such as greedy heuristics, to find maximum likelihood estimates of word class partitions [14]. A common non-probabilistic approach to estimation is to use methods from linear algebra, such as SVD in the case of LSA [61]. Levy and Goldberg show that predictive approaches, such as word2vec [75], can be seen as an implicit matrix factorization of a pointwise mutual information matrix, making a close connection between matrix factorizations and predictive methods [64]. In many situations, the same representation and context can be computed with both approaches. As an example, it is possible to infer the parameters of topic models using either a linear algebra approach (with some conditions, see [4]) or by using a probabilistic generative approach and Bayesian inference [10, 39].

1. For each component k to K:
   a) $\phi_k \sim \text{Dirichlet}(\beta)$
2. For each document d:
   a) $\theta_d \sim \text{Dirichlet}(\alpha)$
   b) For each token i:
      i. $z_{id} \sim \text{Categorical}(\theta_d)$
      ii. $w_{id} \sim \text{Categorical}(\phi_{z_{id}})$

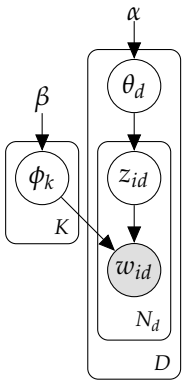


Figure 3.2: The generative model for the Latent Dirichlet Allocation (LDA) topic model (left) and the graphical model (right).

The different approaches and examples can be found in Table 3.3. This table is by no means an attempt to grasp the whole area of latent semantic approaches, but should rather be seen as a rough conceptual model of different approaches to latent semantic modeling.

Context	Repr.	Estimation	Example
Word	Reals	Linear algebra	Random indexing [94] word2vec [75]
Word	Reals	Probabilistic	Exponential family embeddings [93]
Word	Simplex	Linear algebra	Interpretable word embeddings [80, 68]
Word	Simplex	Probabilistic	Brown clusters [14]
Document	Reals	Linear algebra	Latent Semantic Analysis [61]
Document	Reals	Probabilistic	-
Document	Simplex	Linear algebra	Anchor-word topic models [4]
Document	Simplex	Probabilistic	LDA [10], pLSA [49]

Table 3.3: Approaches to latent semantic modeling

The main focus in this thesis is on topic models, and more specifically on probabilistic topic models.

3.4 Probabilistic topic models

Probabilistic topic models are a class of models that generally uses the context of documents to infer underlying themes or topics. One of the most popular approaches in this class of models is the Latent Dirichlet Allocation model, or LDA [10]. In LDA, each topic is a probability distribution over the vocabulary, ϕ_k, and each document d is modeled as a probability distribution over the topics, θ_d. The full generative model of the LDA model is shown in Figure 3.2.

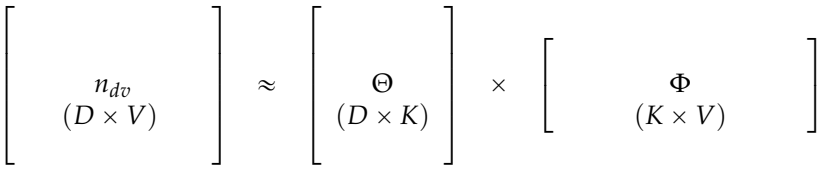


Figure 3.3: Conceptual depiction of LDA as a matrix decomposition.

As shown by the generative model in Figure 3.2, we have two parameter blocks, $\Phi = (\phi_1, \cdots, \phi_K)^T$ of size $K \times V$ and $\Theta = (\theta_1, \cdots, \theta_D)^T$ of size $D \times K$. Together Θ and Φ make up the two matrices that can be seen as a decomposition of a word-document matrix into two matrices Θ and Φ of lower rank. Figure 3.3 shows the connections to the ideas presented in the previous section.

As an example, a probabilistic topic model has been run with 100 topics on a corpus of New York Times articles during the period 1987 to 2007. Below is an article from the 25th of Febuary 1999.

> Closing arguments were heard yesterday in the Federal bankruptcy fraud trial of Stephen J. Sabbeth, whose legal problems have raised doubts about his ability to continue as leader of the Nassau County Democratic Party.
>
> Mr. Sabbeth is charged with trying to conceal $750,000 from his bank creditors by hiding the money in a secret account in his wife's maiden name, rather than use it to pay creditors when his lumber business went into bankruptcy 10 years ago.
>
> "This is a case about greed," the prosecutor, an assistant United States attorney, Seth Marvin, told the jury. He described Mr. Sabbeth as "a constant control freak" who "knowingly and fraudulently conspired to conceal and transfer money into a hidden account, then lied about it."
>
> Mr. Sabbeth's lawyer, Gustave Newman, said his client had not concealed any of his business dealings. "Mr. Sabbeth's financial records were all available to bank officials and auditors," he said. "He had nothing to hide. He did nothing illegal."
>
> If found guilty, Mr. Sabbeth could be sentenced to a maximum of 20 years in prison. A guilty verdict would also cost Mr. Sabbeth his position as Nassau Democratic Party chairman and his $104,000-a-year post as a co-commissioner on the Nassau County Board of Elections.

By estimating Θ and Φ we can study the topics that are used in this article. The posterior Dirichlet distribution for this article, i.e. θ_d, is presented in Figure 3.4.

From Figure 3.4 we can see that this article contains a few larger topics, namely topics 2, 10, 11 and 53. If we study the top words for these topics, i.e. the words with the 10 highest values of ϕ_2, ϕ_10, ϕ_11, and ϕ_53, we can get an idea of what the topic represents. The top words of the different topics are presented in Table 3.4.

As we can see from Table 3.4 and Figure 3.4 the article above can be represented by a small number of topics, in this example financial, political and criminal topics. These topics can thus represent the underlying thematic or semantic content of the article. The model itself, presented

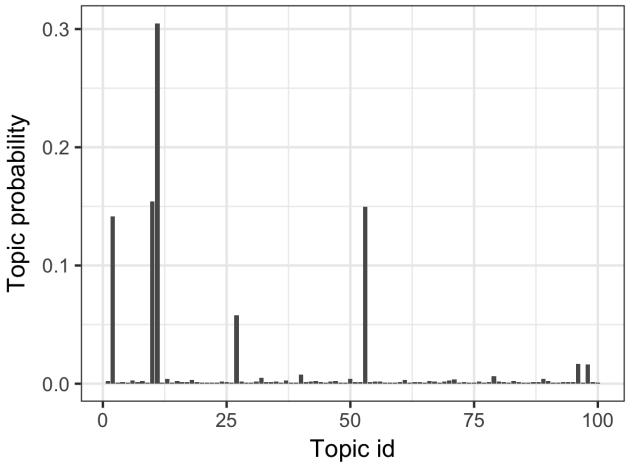


Figure 3.4: The computed $E(\theta_d)$ for the cited article above.

Topic	Top words (by ϕ_{kv})
2	party election voters campaign democratic vote candidates
10	bank banks loans loan insurance savings banking credit
11	trial prison jury prosecutors convicted guilty charges case
53	investigation inquiry documents investigators officials report

Table 3.4: The words with highest probability ($p(w|k)$) for topic 2, 10, 11 and 53.

in Figure 3.2, does not make much sense as a model for language, but the example indicates that the model can actually capture latent topical semantics efficiently.

Inference for the LDA topic model

There are many different inference approaches for the LDA model , but in general, the two most common approaches are MCMC samplers and variational approximations of the posterior distribution [10, 39]. In this thesis, the main focus will be on MCMC approaches.

The most straightforward MCMC approach for the LDA model is to use the conditional conjugacy of the model and construct a sampler that samples the different parameter blocks $\mathbf{z}$, Θ, and Φ in a fashion that is similar to the multinomial mixture model in 2.1.

1. For all k in 1 to K, sample
$$\phi_k|\mathbf{z} \sim \text{Dir}(\mathbf{n}_k^{(v)} + \beta)$$
where $\mathbf{n}_k^{(v)}$ is the counts of topic indicators z by topic k and word type v.
2. For all d in 1 to D, sample
$$\theta_d|\mathbf{z} \sim \text{Dir}(\mathbf{n}_d^{(d)} + \alpha)\,,$$

where $\mathbf{n}_d^{(d)}$ is the counts of topic indicators z by document d and topic k.

3. For all topic indicators $\mathbf{z}$

$$\text{Sample } z_{w_i}|\Theta, \Phi \propto \theta_{d(w_i),k} \cdot \phi_{k,v(w_i)} \,,$$

where $d(w_i)$ is the document index of word w_i and $v(w_i)$ is the word type of w_i.

Due to the conjugacy between the Dirichlet distribution and the multinomial distribution, it is straightforward to integrate over both Θ and Φ in the model (in a similar way as in Equation 2.2 in Chapter 2). Doing this, we only need to infer the topic indicator parameters $\mathbf{z}$. Using this type of MCMC samplers for $\mathbf{z}$ is commonly called *collapsed* Gibbs sampling. Collapsed inference for the LDA model can be conducted either by variational inference [103] or by Gibbs sampling [39].

The collapsed Gibbs sampling algorithm has been one of the more popular inference methods for LDA due to its simplicity and perceived better mixing (something that we discuss in Paper II). The sampling for LDA using the collapsed Gibbs sampler is conducted by sampling every single topic indicator z_i conditional on all other topic indicators $\mathbf{z}_{-i}$ as follows:

$$p(z_i = k|w_i, \mathbf{z}_{\neg i}) \propto \frac{n_{k,v(w_i)}^{(v)} + \beta}{\sum_v^V n_{k,v}^{(v)} + \beta} \cdot \frac{n_{d(w_i),k}^{(d)} + \alpha}{\sum_k^K (n_{d(w_i),k}^{(d)} + \alpha} \,, \tag{3.1}$$

where $n_{k,v(w_i)}^{(v)}$ is the number of topic indicators for topic k with the word type of token w_i and $n_{d(w_i),k}^{(d)}$ is the number of topic indicators for topic k in the document that token w_i belongs to.

This is a straightforward and elegant Gibbs sampler, and it has been a popular approach for both the standard LDA model and for many different extensions. There are two things to note with this popular sampler. First, we can see that the collapsed sampler, unlike the naive sampler, is serial. If we want to sample a topic indicator z_i we need to do that conditionally on *all* other topic indicators $\mathbf{z}_{-i}$. It is not possible to sample multiple topic indicators at the same time. The second aspect is that the sampling complexity of sampling one topic indicator is $O(K)$, meaning that the sampling time for one topic indicator z_i is proportional to the number of topics in the model. So if we assume that larger corpora need more topics, the computational complexity of the sampler will increase with the size of the corpus. Combining these two aspects makes the standard collapsed Gibbs sampler for LDA a poor sampler from a scaling perspective.

In the simple example of the NYT corpus above, the total number of topic indicators is roughly 500 million tokens in total. Since the collapsed Gibbs sampler in Equation 3.1 needs to iterate over every individual topic and if we use 1000 Gibbs iterations, the small computation in the collapsed sampler in Equation 3.1 needs to be performed roughly 50,000 billion times. Even for a modern CPU, this is time-consuming. Given the fact that individual CPUs are no longer increasing in performance [101], the conclusion from a statistical perspective is that to handle large corpora we need to make use of computational parallelism and smart algorithms for large-scale inference, which is explored in papers I and II in this thesis.

The basic LDA topic model has been extended in various different directions, such as the time dimension [9], supervised classes [73], word order [110] and segmentation [87]. And in many cases the extended model builds upon the idea of adding and combining different multinomial distributions and Dirichlet distributions to model additional aspects of interest [65, 16]. The idea of extending the topic model to capture other aspects of interest is used in Paper VI to connect empirical textual data and political science theory.

3.5 Practical curation of corpora and the implications for inference

Text is complex, in that it contains many different structures that are used for human understanding. But to model textual data, we need to define and use simpler models for computational reasons. In previous sections, we saw that the use of the multinomial distribution in the LDA model means that we assume that word order does not matter, a heavily simplifying assumption, but one that is used for computational reasons.

In practical analysis, we make a lot of assumptions before we actually do any inference in our models. This is often referred to as pre-processing or corpus curation, and the effect of these decisions rests (implicitly) on assumptions regarding how these choices affect the inference that is made. Below are common curation techniques and some rough implicit assumptions they rely on in the case of the standard LDA model. The purpose of these approaches is motivated by practical arguments, such as reducing the vocabulary of the text or removing noise. Below are common curation choices when doing topic modeling.

- *Punctuation*, such as ?!.,", is commonly removed. The implicit assumption is that punctuation is assumed not to contain any semantic information.
- *Lowercasing* can be done by lowercasing all uppercase characters. This choice can be regarded as an assumption that there is no semantic difference between words with upper and lowercase characters.
- *Numbers* can either be removed or changed to a general "NUMBER" token. This is an implicit assumption that numbers, as such, may contain semantic information, but the exact number does not.
- *Tokenization* and *collocation combination* concerns the way in which a sequence of characters is converted into individual tokens. It is possible to do this in a number of ways. How, for example, should "New York City" be treated, as one token or three? If we combine this into one token we assume that there is a semantic difference between "New York City" and the three tokens "city", "New" and "York", and we also implicitly assume that we have one, rather than three, observations.
- *Stemming* and *lemmatizing* are techniques for converting a word into its lexical stem or lemma. The word "running" will be converted into "run" and "cars" will be converted to "car". When using these techniques we assume that the inflections of individual words do not contain any information of importance in identifying underlying semantic themes.
- *Stop word removal* is the technique of removing very common words, such as "a" and "the". This is commonly done both to reduce the size of the corpus and to remove words with little or no semantic information. According to Zipf's law [116] a few word types make up a large part of all tokens in a corpus, so in many situations removing stop words can reduce the size of the corpus by 50%, speeding up inference.
- *Removal of rare or infrequent terms* can be done to reduce the vocabulary. Again, motivated by Zipf's law, we know that a large number of word types only occur a few times. By removing these rare words we assume that these words are so rare that they would not affect the overall inferential performance.
- *Document segmentation* concerns the issue of how we define a document for a given topic model. As mentioned previously, topic models use the document as context. But how we define context can vary and is essentially up to the analyst to decide. We can define documents as being individual paragraphs, sections, chapters or books. If we assume a paragraph to be a document in the standard LDA model, we implicitly assume that each paragraph is (conditionally) independent.

All of these pre-processing or corpus curation choices imply some assumptions, and are in essence additional modeling assumptions that are made due to the complexity of the textual documents. However, they are a part of the curating process rather than the modeling process itself. A small contribution to this field of research is made in Paper V.

4 Research Questions and Summary of Contributions

In previous chapters, Bayesian statistical inference was introduced together with applications and models, specifically for text and latent semantic modeling. This sets the stage for a more elaborate description of the thesis research questions and the contributions of this thesis.

4.1 Research questions

The purpose of this thesis is to contribute to fast and rigorous inference for different probabilistic models of text. This has been done by developing new methods and approaches, and by increased knowledge that can be used in the probabilistic analysis of textual data. The main focus is Bayesian probabilistic topic models and applications thereof, such as in the social sciences, humanities and technical fields.

1. How can we expand topic models to the larger data sizes that are now commonplace, without sacrificing soundness of the inference method?
2. How can we improve topic models, inference and interpretability to enable increased use in applied research settings?

The first research question addresses the issue of large-scale inference. As has been explained in previous sections, there is an increased need for large-scale analysis of textual corpora, and probabilistic topic models have been one important and popular approach to modeling these corpora. There are multiple approaches to doing inference in large-scale corpora. Variational Bayes (VB) has been a popular approach due to the fact that it is generally considered to be faster than MCMC approaches. But methods relying on VB come with few theoretical guarantees that the method will actually approximate the posterior efficiently [8]. Blei et. al. [8] use large-scale corpora as examples where variational inference may be preferred. By relaxing theoretical guarantees we can compute posteriors for large corpora in a reasonable time. If, instead,

we want the theoretical guarantees of MCMC, we risk having to handle the problem of the standard way of doing Gibbs sampling for topic models. The popular collapsed Gibbs sampler is not parallelizable and it does not scale well with the number of topics. There have been improvements to inference performance (such as [113, 115]), but these approaches still rely upon the serial collapsed Gibbs sampler, making it impossible to use multiple computers or CPUs and still provide theoretical guarantees. Newman et. al. [82] propose an approximation to the collapsed Gibbs sampler to make it work in parallel, but this approximation means that the sampler will no longer have any theoretical guarantees [106]. This first research question addresses the question of doing large-scale probabilistic topic model inference with theoretic guarantees.

The second research question addresses the issue of how we can extend knowledge about modeling textual data in a way that can be used in applied research situations, and more specifically in improving interpretability. This research question has a multitude of different aspects. First of all, it concerns interpretability in modeling textual data. Text classification algorithms, such as Naive Bayes, can be one way of classifying documents, but modeling individual word tokens can make it difficult to interpret how the classification has been conducted. This might not be of interest in a pure machine learning setting where accuracy may be the main quality dimension, but as is discussed in Section 2.2, internal cohesion and interpretability are more relevant in research when we want to understand and explain. This makes interpretability in modeling one aspect that is of importance. Another important issue is that of corpus curating choices. Corpus curation is very commonly done, but the effects of specific choices about what to curate and how are still poorly understood. When the effects of these choices have been studied, the resulting conclusions are sometimes very different from what is commonly thought (see [96] as an example). The final issue is that of connecting scientific research questions and probabilistic models. To be able to produce new scientific knowledge by modeling textual data, it is important to connect the scientific questions and the probabilistic models in ways that make it possible to draw conclusions about social science and humanities theory from empirical textual data.

4.2 Summary of Contributions

The rest of the thesis consists of the following journal papers and conference proceedings, exploring and contributing to the research questions presented above.

Paper I Måns Magnusson, Leif Jonsson, Mattias Villani, and David Broman. "Sparse Partially Collapsed MCMC for Parallel Inference in Topic Models". In: *Journal of Computational and Graphical Statistics* (2017)

Paper II Alexander Terenin, Måns Magnusson, Leif Jonsson, and David Draper. "Pólya Urn Latent Dirichlet Allocation: a sparse massively parallel sampler". Minor revision, Transactions on Pattern Analysis and Machine Intelligence

Paper III Måns Magnusson, Leif Jonsson, and Mattias Villani. "DOLDA—a regularized supervised topic model for high-dimensional multi-class regression". Revision resubmitted to journal

Paper IV Leif Jonsson, David Broman, Måns Magnusson, Kristian Sandahl, Mattias Villani, and Sigrid Eldh. "Automatic Localization of Bugs to Faulty Components in Large Scale Software Systems using Bayesian Classification". In: *Software Quality, Reliability and Security (QRS), 2016 IEEE International Conference on*. IEEE. 2016, pp. 423–430

Paper V Alexandra Schofield, Måns Magnusson, and David Mimno. "Pulling Out the Stops: Rethinking Stopword Removal for Topic Models". In: *EACL 2017* (2017), pp. 432–438

Paper VI Måns Magnusson, Richard Öhrvall, Katarina Barrling, and David Mimno. "Voices from the far right: a text analysis of Swedish parliamentary debates". Submitted to journal

Extending topic models to larger corpora

As has been noted in Section 3, the standard collapsed Gibbs sampler for LDA, proposed by Griffiths and Steyvers [39], works very well for smaller corpora. However, the computational complexity and the serial nature of the sampler make it difficult to scale this sampler to larger corpora, which is becoming more and more important. In Paper I we propose a sparsity-aware partially collapsed parallel sampler for the basic LDA model. The idea can also be extended to other similar models. We also prove theoretically that this sampler, even though it also samples additional parameters Φ, will still be dominated by sampling the topic indicators $\mathbf{z}$. The sampling of $\mathbf{z}$ is also done faster by using the sparsity of the collapsed Gibbs sampler together with Walker–Alias tables for multinomial sampling, reducing the overall complexity of the sampler from $O(K)$ to $O(K_d)$ amortized, where K_d is the number of topics occurring in document d. If we assume that document size does not grow with the size of the corpus, this effectively reduces the complexity of the sampler compared to the basic collapsed Gibbs sampler for LDA.

These ideas are used in Paper II to further improve the computational speed and performance for LDA topic models [105]. In Paper II we use the idea of approximating the continuous Dirichlet distribution with a discrete distribution based on the Pólya urn distribution. This simple approximation, which we prove to converge to a Dirichlet distribution by using the central limit theorem, has three implications for inference using MCMC and sampling Dirichlet distributions, like in partially collapsed LDA.

The first implication is the reduced memory footprint. If the Dirichlet distribution is sparse, we only need to store elements greater than 0. This was our initial purpose with the Polya-Urn approximation, since we wanted to do massively parallel inference on Graphical Processing Units (GPU), which often have limited memory.

The second implication is computational efficiency. Sampling a Dirichlet distribution using the ordinary approach of normalized Gamma variates is costly. Using the Pólya urn instead, we can sample from the Poisson distribution. Using Alias table categorical sampling together with normal approximations of the Poisson distribution, we show that we can reduce the sampling cost considerably compared to the standard Gamma variate approach.

The third and last result of sparsity in Φ is that it makes it possible to use the additional sparsity in sampling $\mathbf{z}$. This additional sparsity further reduces the sampling complexity of the partially collapsed Gibbs sampler of Paper I from $O(K_d)$ to $O(\min(K_d, \phi_v))$ where ϕ_v is the length of non-zero elements in Φ for word type v [70]. For larger corpora with a larger number of topics, we can show that this further improves the sampling speed of $\mathbf{z}$ for the LDA model—but still with theoretical guarantees. The Gibbs sampler presented in Paper II is, to the best of our knowledge, the Gibbs sampler with the lowest computational complexity, which improves the possibility of using MCMC inference for large-scale topic models as well as for similar models. It also enables us to create a sampler for LDA to use with GPUs that guarantees convergence to the true posterior distribution.

Interpretable topic models for applied research

Interpretability of models and results is important, both to facilitate the scientific process and to understand the underlying models. In Paper III we introduce a linear supervised topic model that combines the Diagonal Orthant (DO) probit model, the LDA topic model with a horseshoe prior, to the regression coefficients [69]. The purpose of the model is to create a supervised topic model that can do multinomial classification with a model that is easy to interpret for applied researchers. The topics are themselves a type of abstraction that is easier for humans to grasp than a large number of individual word type parameters, such as in Naive Bayes. The horseshoe prior to the model reduces the number of "signal" topics that affect a given class by shrinking the regression coefficients efficiently towards zero, making the classification even easier to interpret. Finally, the DO probit model does not do classification with regards to a reference class, further

simplifying the interpretation of the model parameters. We also show that this is a model that scales well and provide a parallel implementation that is able to handle large corpora.

With the same idea of interpretable topic models, the model in Paper VI focuses on extending current topic models for ideological modeling to the area of parliamentary speeches in the Swedish parliament [71]. The model includes both political framing of topics and a new idea of word seeding to define a topic a priori that is of interest to study. Using this model we study how the discourse on immigration has changed over time in the Swedish parliament. We are especially interested in how the immigration discourse is affected by the Sweden Democrats entering the parliament in 2010 by using a topic model to capture what is of interest from a political science research perspective.

As mentioned in Section 3.5 a common approach when modeling text is to curate the corpus before actual inference is conducted. In Paper V we study the effect of stop word removal on topic models and conclude that the previously held belief that removing stop words will improve inference is probably not true [95]. Removing stop words does not seem to affect model inference much at all, neither by improving the model performance nor by harming inference.

4.3 Extensions and future research

This thesis has studied different applications and developments for analysis of probabilistic methods of textual data. These results have, in turn, opened up new research questions in this field.

Most topic models proposed in the literature are more or less unsupervised or exploratory in nature, one example being the original LDA model. But in many applied research situations, there is interest in using topic modeling for confirmatory analysis of text. As an example, applied researchers may be interested in a very specific part of a posterior, such as individual topics that are of particular interest to a given research question. An example of this is presented in Paper IV, where immigration, as a topic, was studied over almost 25 years in the Swedish parliament.

This idea of confirmatory modeling of latent structures is not new. In the seminal paper [55] from 1969, a general approach for constraining latent concepts in factor analysis is proposed, which would later come to be known as confirmatory factor analysis. These ideas and this knowledge should be used to enable similar approaches in the field of topic models and text analysis. As in exploratory factor analysis, all free parameters are estimated while for confirmatory modeling, a number of restrictions are made to the covariance matrices based on given hypotheses [32].

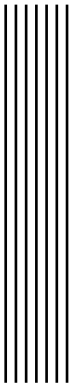
Similar approaches are needed to enable researchers to not just explore the topical contents of their corpora, but also to use the model for scientific hypotheses testing. Together with posterior diagnostics and model evaluation techniques, enabling these kinds of ideas should make analyzing latent constructs in large corpora a great methodological tool in the contemporary social sciences and the humanities.

The second relevant offshoot of this thesis is the importance of estimating large-scale corpora efficiently using Bayesian methods. In this thesis, new methods for large-scale topic models have been studied and new methods for parallel and computationally efficient samplers have been derived. Even though it may not be possible to increase the efficiency of sampling much more, other approaches to improving inferential speed may be of relevance and should be studied further.

Combining ideas from sub-sampling and simulated annealing may enable us to find better initial states for our MCMC or to do correct inference directly [17, 85]. By combining efficient samplers with similar ideas on sub-sampling and simulated annealing we would be able to fur-

ther increase the speed to infer large-scale text models under constraints—but still be able to do statistical inference.

Finally, the need for better and faster inferential techniques is seldom of much use if practitioners, such as researchers in the social sciences and the humanities, cannot access the algorithms in a user-friendly fashion. This calls for further work on implementing these methods, diagnostics and model extensions in more researcher-friendly software, such as R and Python libraries, making it practical and easy to analyze large corpora on ordinary laptops and workstations. Although this thesis has provided statistical methods that would enable such implementations, there is currently a need for good, practical, user-friendly implementations.



Bibliography

[1] *The future of computing*. The Economist. Mar. 12, 2016.

[2] Mark Algee-Hewitt, Ryan Heuser, and Franco Moretti. *On Paragraphs: Scale, Themes and Narrative Form*. Literary Lab, 2015.

[3] Charles E Antoniak. "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems". In: *The annals of statistics* (1974), pp. 1152–1174.

[4] Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. "A practical algorithm for topic modeling with provable guarantees". In: *International Conference on Machine Learning*. 2013, pp. 280–288.

[5] The World Bank. *GDP (current US$), 1973-2016*. data retrieved from World Bank national accounts data, and OECD National Accounts data files, `https://data.worldbank.org/indicator/NY.GDP.MKTP.CD`. 2017.

[6] James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[7] José M Bernardo and Adrian FM Smith. *Bayesian theory*. IOP Publishing, 1994.

[8] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: A review for statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877.

[9] David M Blei and John D Lafferty. "Dynamic topic models". In: *Proceedings of the 23rd international conference on Machine learning*. ACM. 2006, pp. 113–120.

[10] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[11] Kenneth A. Bollen. *Structural Equations with Latent Variables*. Wiley series in probability and mathematical statistics. Applied probability and statistics section. Wiley, 1989. ISBN: 9780471011712. URL: https://books.google.se/books?id=Vr3rAAAAMAAJ.

[12] George EP Box. "Robustness in the strategy of scientific model building". In: *Robustness in statistics* 1 (1979), pp. 201–236.

[13] George EP Box. "Science and statistics". In: *Journal of the American Statistical Association* 71.356 (1976), pp. 791–799.

[14] Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. "Class-based n-gram models of natural language". In: *Computational linguistics* 18.4 (1992), pp. 467–479.

[15] Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. "The mathematics of statistical machine translation: Parameter estimation". In: *Computational linguistics* 19.2 (1993), pp. 263–311.

[16] Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. "Modeling general and specific aspects of documents with a probabilistic topic model". In: *Advances in neural information processing systems*. 2007, pp. 241–248.

[17] Changyou Chen, David Carlson, Zhe Gan, Chunyuan Li, and Lawrence Carin. "Bridging the gap between stochastic gradient MCMC and stochastic optimization". In: *Artificial Intelligence and Statistics*. 2016, pp. 1051–1060.

[18] Shay Cohen. "Bayesian analysis in natural language processing". In: *Synthesis Lectures on Human Language Technologies* 9.2 (2016), pp. 1–274.

[19] Mark Colyvan. *The indispensability of mathematics*. Oxford university press, 2001.

[20] Thomas D Cook, Donald Thomas Campbell, and William Shadish. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, 2002.

[21] National Research Council et al. *Frontiers in massive data analysis*. National Academies Press, 2013.

[22] Gregory Crane. “What do you do with a million books?” In: *D-Lib magazine* 12.3 (2006), p. 1.

[23] Arthur P Dempster, Nan M Laird, and Donald B Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the royal statistical society. Series B (methodological)* (1977), pp. 1–38.

[24] Persi Diaconis, Donald Ylvisaker, et al. “Conjugate priors for exponential families”. In: *The Annals of statistics* 7.2 (1979), pp. 269–281.

[25] Paul DiMaggio, Manish Nag, and David Blei. “Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding”. In: *Poetics* 41.6 (2013), pp. 570–606.

[26] Chris Ding, Tao Li, and Wei Peng. “On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing”. In: *Computational Statistics & Data Analysis* 52.8 (2008), pp. 3913–3927.

[27] Michael D Escobar and Mike West. “Bayesian density estimation and inference using mixtures”. In: *Journal of the american statistical association* 90.430 (1995), pp. 577–588.

[28] Brian M Fagan and Charlotte Beck. *The Oxford companion to archaeology*. Oxford Companions, 1996.

[29] Stephen E Fienberg et al. “When did Bayesian inference become" Bayesian"?” In: *Bayesian analysis* 1.1 (2006), pp. 1–40.

[30] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. “Incorporating non-local information into information extraction systems by Gibbs sampling”. In: *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics. 2005, pp. 363–370.

[31] John R Firth. “A synopsis of linguistic theory, 1930-1955”. In: *Studies in linguistic analysis* (1957).

[32] Richard J Fox. *Confirmatory factor analysis*. Wiley Online Library, 1983.

[33] Roman Frigg and Stephan Hartmann. “Models in Science”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2017. Metaphysics Research Lab, Stanford University, 2017.

[34] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Vol. 3. CRC press Boca Raton, FL, 2014.

[35] Andrew Gelman, Sharad Goel, Douglas Rivers, David Rothschild, et al. “The mythical swing voter”. In: *Quarterly Journal of Political Science* 11.1 (2016), pp. 103–130.

[36] Stuart Geman and Donald Geman. "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". In: *IEEE Transactions on pattern analysis and machine intelligence* 6 (1984), pp. 721–741.

[37] Yoav Goldberg. "Neural Network Methods for Natural Language Processing". In: *Synthesis Lectures on Human Language Technologies* 10.1 (2017), pp. 1–309.

[38] Derek Greene and James P Cross. "Exploring the political agenda of the European parliament using a dynamic topic modeling approach". In: *Political Analysis* 25.1 (2017), pp. 77–94.

[39] Thomas L Griffiths and Mark Steyvers. "Finding scientific topics". In: *Proceedings of the National academy of Sciences* 101.suppl 1 (2004), pp. 5228–5235.

[40] Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. "Topics in semantic representation." In: *Psychological review* 114.2 (2007), p. 211.

[41] Justin Grimmer. "A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases". In: *Political Analysis* 18.1 (2009), pp. 1–35.

[42] Allan Gut. *An Intermediate Course in Probability*. 2nd. Springer Publishing Company, Incorporated, 2009.

[43] Aria Haghighi and Dan Klein. "Unsupervised coreference resolution in a nonparametric bayesian model". In: *Annual meeting-Association for Computational Linguistics*. Vol. 45. 1. 2007, p. 848.

[44] Dan Halacy. *Charles Babbage: father of the computer*. Crowell-Collier, 1970.

[45] Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. *Cohesion in english*. Routledge, 1976.

[46] Stephan Hartmann and Jan Sprenger. "Bayesian epistemology". In: *Routledge companion to epistemology* (2010), pp. 609–620.

[47] Harold Stanley Heaps. *Information retrieval: Computational and theoretical aspects*. Academic Press, Inc., 1978.

[48] Peter Hedström. *Dissecting the social: On the principles of analytical sociology*. Vol. 10. Cambridge University Press Cambridge, 2005.

[49] Thomas Hofmann. "Probabilistic latent semantic analysis". In: *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. 1999, pp. 289–296.

[50] *How Search organizes information*. `https://www.google.com/search/howsearchworks/crawling-indexing/`. Accessed: 2018-01-11. 2018.

[51] Joab Jackson. "Google: 129 Million Different Books Have Been Published". In: *PCWorld* (2010). Accessed: 2018-01-11.

[52] Matthew L Jockers. *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.

[53] Mark Johnson, Thomas L Griffiths, and Sharon Goldwater. "Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models". In: *Advances in neural information processing systems*. 2007, pp. 641–648.

[54] Leif Jonsson, David Broman, Måns Magnusson, Kristian Sandahl, Mattias Villani, and Sigrid Eldh. "Automatic Localization of Bugs to Faulty Components in Large Scale Software Systems using Bayesian Classification". In: *Software Quality, Reliability and Security (QRS), 2016 IEEE International Conference on*. IEEE. 2016, pp. 423–430.

[55] Karl G Jöreskog. "A general approach to confirmatory maximum likelihood factor analysis". In: *Psychometrika* 34.2 (1969), pp. 183–202.

[56] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. "Bag of tricks for efficient text classification". In: *arXiv preprint arXiv:1607.01759* (2016).

[57] Young Gyo Jung, Kyung Tae Kim, Byungjun Lee, and Hee Yong Youn. "Enhanced Naive Bayes Classifier for real-time sentiment analysis with SparkR". In: *Information and Communication Technology Convergence (ICTC), 2016 International Conference on*. IEEE. 2016, pp. 141–146.

[58] Daniel Jurafsky and James H. Martin. *Speech and language processing: An introduction to natural language processing*. 3rd ed. draft. Prentice Hall, 2016.

[59] Robert E Kass and Adrian E Raftery. "Bayes factors". In: *Journal of the american statistical association* 90.430 (1995), pp. 773–795.

[60] Sadanori Konishi and Genshiro Kitagawa. *Information criteria and statistical modeling*. Springer Science & Business Media, 2008.

[61] Thomas K Landauer, Peter W Foltz, and Darrell Laham. "An introduction to latent semantic analysis". In: *Discourse processes* 25.2-3 (1998), pp. 259–284.

[62] Geoffrey N. Leech. *Semantics: The study of Meaning*. 2nd. 1981.

[63] David A Levin, Elizabeth L Wilmer, and Yuval Peres. "Markov chains and mixing times". In: (2009).

[64] Omer Levy and Yoav Goldberg. "Neural word embedding as implicit matrix factorization". In: *Advances in neural information processing systems*. 2014, pp. 2177–2185.

[65] Wei-Hao Lin, Eric Xing, and Alexander Hauptmann. "A joint topic and perspective model for ideological discourse". In: *Machine Learning and Knowledge Discovery in Databases* (2008), pp. 17–32.

[66] Erik Linstead, Paul Rigor, Sushil Bajracharya, Cristina Lopes, and Pierre Baldi. "Mining concepts from code with probabilistic topic models". In: *Proceedings of the twenty-second IEEE/ACM international conference on Automated software engineering*. ACM. 2007, pp. 461–464.

[67] Stacy K Lukins, Nicholas A Kraft, and Letha H Etzkorn. "Source code retrieval for bug localization using latent dirichlet allocation". In: *Reverse Engineering, 2008. WCRE'08. 15th Working Conference on*. IEEE. 2008, pp. 155–164.

[68] Hongyin Luo, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. "Online learning of interpretable word embeddings". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 1687–1692.

[69] Måns Magnusson, Leif Jonsson, and Mattias Villani. "DOLDA—a regularized supervised topic model for high-dimensional multi-class regression". Revision resubmitted to journal.

[70] Måns Magnusson, Leif Jonsson, Mattias Villani, and David Broman. "Sparse Partially Collapsed MCMC for Parallel Inference in Topic Models". In: *Journal of Computational and Graphical Statistics* (2017).

[71] Måns Magnusson, Richard Öhrvall, Katarina Barrling, and David Mimno. "Voices from the far right: a text analysis of Swedish parliamentary debates". Submitted to journal.

[72] Robert Malone. *Structuring Unstructured Data*. Forbes. Apr. 5, 2007.

[73] Jon D Mcauliffe and David M Blei. "Supervised topic models". In: *Advances in neural information processing systems*. 2008, pp. 121–128.

[74] Peter McCullagh. "What is a statistical model?" In: *Annals of statistics* (2002), pp. 1225–1267.

[75] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[76] David Mimno and Laure Thompson. "The strange geometry of skip-gram with negative sampling". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 2873–2878.

[77] Andriy Mnih and Ruslan R Salakhutdinov. "Probabilistic matrix factorization". In: *Advances in neural information processing systems*. 2008, pp. 1257–1264.

[78] Gordon E. Moore. "Cramming more components onto integrated circuits". In: *Readings in computer architecture*. Ed. by Mark Donald Hill, Norman Paul Jouppi, and Gurindar Sohi. Gulf Professional Publishing, 2000.

[79] Franco Moretti. *Distant reading*. Verso Books, 2013.

[80] Brian Murphy, Partha Talukdar, and Tom Mitchell. "Learning effective and interpretable semantic models using non-negative sparse embedding". In: *Proceedings of COLING 2012* (2012), pp. 1933–1950.

[81] Radford M Neal. "Markov chain sampling methods for Dirichlet process mixture models". In: *Journal of computational and graphical statistics* 9.2 (2000), pp. 249–265.

[82] David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. "Distributed algorithms for topic models". In: *Journal of Machine Learning Research* 10.Aug (2009), pp. 1801–1828.

[83] Kai Wang Ng, Guo-Liang Tian, and Man-Lai Tang. *Dirichlet and related distributions: Theory, methods and applications*. Vol. 888. John Wiley & Sons, 2011.

[84] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. "Text classification from labeled and unlabeled documents using EM". In: *Machine learning* 39.2 (2000), pp. 103–134.

[85] Fritz Obermeyer, Jonathan Glidden, and Eric Jonas. "Scaling nonparametric Bayesian inference via subsample-annealing". In: *Artificial Intelligence and Statistics*. 2014, pp. 696–705.

[86] Xavier Puig and Josep Ginebra. "A Bayesian cluster analysis of election results". In: *Journal of Applied Statistics* 41.1 (2014), pp. 73–94.

[87] Matthew Purver, Thomas L Griffiths, Konrad P Körding, and Joshua B Tenenbaum. "Unsupervised topic modelling for multi-party spoken discourse". In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. 2006, pp. 17–24.

[88] Frank P Ramsey. "Truth and probability (1926)". In: *The foundations of mathematics and other logical essays* (1931), pp. 156–198.

[89] Nick Riemer. *Introducing semantics*. Cambridge University Press, 2010.

[90] Margaret E Roberts, Brandon M Stewart, and Edoardo M Airoldi. "A model of text for experimentation in the social sciences". In: *Journal of the American Statistical Association* 111.515 (2016), pp. 988–1003.

[91] Jan-Willem Romeijn. "Philosophy of Statistics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2017. Metaphysics Research Lab, Stanford University, 2017.

[92] Max Roser and Esteban Ortiz-Ospina. *Literacy*. Published online at OurWorldInData.org. 2018. URL: https://ourworldindata.org/literacy.

[93] Maja Rudolph, Francisco Ruiz, Stephan Mandt, and David Blei. "Exponential family embeddings". In: *Advances in Neural Information Processing Systems*. 2016, pp. 478–486.

[94] Magnus Sahlgren. *An introduction to random indexing*. 2005.

[95] Alexandra Schofield, Måns Magnusson, and David Mimno. "Pulling Out the Stops: Rethinking Stopword Removal for Topic Models". In: *EACL 2017* (2017), pp. 432–438.

[96] Alexandra Schofield and David Mimno. "Comparing apples to apple: The effects of stemmers on topic models". In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 287–300.

[97] Bunyamin Sisman, Shayan A Akbar, and Avinash C Kak. "Exploiting spatial code proximity and order for improved source code retrieval for bug localization". In: *Journal of Software: Evolution and Process* 29.1 (2017).

[98] Benjamin Snyder and Regina Barzilay. "Unsupervised Multilingual Learning for Morphological Segmentation." In: *ACL*. 2008, pp. 737–745.

[99] Matthias Steup. "Epistemology". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2017. Metaphysics Research Lab, Stanford University, 2017.

[100] Michael Strevens. "Notes on Bayesian Confirmation Theory". 2017.

[101] Herb Sutter. "The free lunch is over: A fundamental turn toward concurrency in software". In: *Dr. Dobb's journal* 30.3 (2005), pp. 202–210.

[102] T. Tarpey. "All Models are Right...Most are Useless". JSM Proceedings: Papers Presented at the Joint Statistical Meeting. 2009. URL: http://corescholar.libraries.wright.edu/math/211.

[103] Yee W Teh, David Newman, and Max Welling. "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation". In: *Advances in neural information processing systems*. 2007, pp. 1353–1360.

[104] Yee Whye Teh. "Dirichlet process". In: *Encyclopedia of machine learning*. Springer, 2011, pp. 280–287.

[105] Alexander Terenin, Måns Magnusson, Leif Jonsson, and David Draper. "Pólya Urn Latent Dirichlet Allocation: a sparse massively parallel sampler". Minor revision, Transactions on Pattern Analysis and Machine Intelligence.

[106] Alexander Terenin, Daniel Simpson, and David Draper. *Asynchronous Gibbs Sampling*. 2015. eprint: arXiv:1509.08999.

[107] Stephen Thornton. "Karl Popper". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2017. Metaphysics Research Lab, Stanford University, 2017.

[108] Luke Tierney. "Markov chains for exploring posterior distributions". In: *the Annals of Statistics* (1994), pp. 1701–1728.

[109] Alastair J Walker. "An efficient method for generating discrete random variables with general distributions". In: *ACM Transactions on Mathematical Software (TOMS)* 3.3 (1977), pp. 253–256.

[110] Chong Wang, Bo Thiesson, Chris Meek, and David Blei. "Markov topic models". In: *Artificial Intelligence and Statistics*. 2009, pp. 583–590.

[111] Gregor Wiedemann. "Opening up to big data: Computer-assisted analysis of textual data in social sciences". In: *Historical Social Research/Historische Sozialforschung* (2013), pp. 332–357.

[112] Tze-I Yang, Andrew J Torget, and Rada Mihalcea. "Topic modeling on historical newspapers". In: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Association for Computational Linguistics. 2011, pp. 96–104.

[113] Limin Yao, David Mimno, and Andrew McCallum. "Efficient methods for topic model inference on streaming document collections". In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2009, pp. 937–946.

[114] Jianhua Yin and Jianyong Wang. "A Dirichlet Multinomial mixture model-based approach for short text clustering". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2014, pp. 233–242.

[115] Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric Po Xing, Tie-Yan Liu, and Wei-Ying Ma. "Lightlda: Big topic models on modest computer clusters". In: *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2015, pp. 1351–1361.

[116] George Kingsley Zipf. *The psycho-biology of language: An introduction to dynamic philology*. 2nd ed. M.I.T. Press, 1968.

Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-147613

Department of Computer and Information Science
Linköpings universitet

Dissertations

Linköping Studies in Science and Technology
Linköping Studies in Arts and Science
Linköping Studies in Statistics
Linköping Studies in Information Science

Linköping Studies in Science and Technology

No 14 **Anders Haraldsson:** A Program Manipulation System Based on Partial Evaluation, 1977, ISBN 91-7372-144-1.

No 17 **Bengt Magnhagen:** Probability Based Verification of Time Margins in Digital Designs, 1977, ISBN 91-7372-157-3.

No 18 **Mats Cedwall**: Semantisk analys av process-beskrivningar i naturligt språk, 1977, ISBN 91- 7372-168-9.

No 22 **Jaak Urmi:** A Machine Independent LISP Compiler and its Implications for Ideal Hardware, 1978, ISBN 91-7372-188-3.

No 33 **Tore Risch:** Compilation of Multiple File Queries in a Meta-Database System, 1978, ISBN 91- 7372-232-4.

No 51 **Erland Jungert:** Synthesizing Database Structures from a User Oriented Data Model, 1980, ISBN 91-7372-387-8.

No 54 **Sture Hägglund:** Contributions to the Development of Methods and Tools for Interactive Design of Applications Software, 1980, ISBN 91-7372-404-1.

No 55 **Pär Emanuelson:** Performance Enhancement in a Well-Structured Pattern Matcher through Partial Evaluation, 1980, ISBN 91-7372-403-3.

No 58 **Bengt Johnsson, Bertil Andersson:** The Human-Computer Interface in Commercial Systems, 1981, ISBN 91-7372-414-9.

No 69 **H. Jan Komorowski:** A Specification of an Abstract Prolog Machine and its Application to Partial Evaluation, 1981, ISBN 91-7372-479-3.

No 71 **René Reboh:** Knowledge Engineering Techniques and Tools for Expert Systems, 1981, ISBN 91-7372-489-0.

No 77 **Östen Oskarsson:** Mechanisms of Modifiability in large Software Systems, 1982, ISBN 91- 7372-527-7.

No 94 **Hans Lunell:** Code Generator Writing Systems, 1983, ISBN 91-7372-652-4.

No 97 **Andrzej Lingas:** Advances in Minimum Weight Triangulation, 1983, ISBN 91-7372-660-5.

No 109 **Peter Fritzson:** Towards a Distributed Programming Environment based on Incremental Compilation, 1984, ISBN 91-7372-801-2.

No 111 **Erik Tengvald:** The Design of Expert Planning Systems. An Experimental Operations Planning System for Turning, 1984, ISBN 91-7372- 805-5.

No 155 **Christos Levcopoulos:** Heuristics for Minimum Decompositions of Polygons, 1987, ISBN 91-7870-133-3.

No 165 **James W. Goodwin:** A Theory and System for Non-Monotonic Reasoning, 1987, ISBN 91-7870-183-X.

No 170 **Zebo Peng:** A Formal Methodology for Automated Synthesis of VLSI Systems, 1987, ISBN 91-7870-225-9.

No 174 **Johan Fagerström:** A Paradigm and System for Design of Distributed Systems, 1988, ISBN 91-7870-301-8.

No 192 **Dimiter Driankov:** Towards a Many Valued Logic of Quantified Belief, 1988, ISBN 91-7870-374-3.

No 213 **Lin Padgham:** Non-Monotonic Inheritance for an Object Oriented Knowledge Base, 1989, ISBN 91-7870-485-5.

No 214 **Tony Larsson:** A Formal Hardware Description and Verification Method, 1989, ISBN 91-7870-517-7.

No 221 **Michael Reinfrank:** Fundamentals and Logical Foundations of Truth Maintenance, 1989, ISBN 91-7870-546-0.

No 239 **Jonas Löwgren:** Knowledge-Based Design Support and Discourse Management in User Interface Management Systems, 1991, ISBN 91-7870-720-X.

No 244 **Henrik Eriksson:** Meta-Tool Support for Knowledge Acquisition, 1991, ISBN 91-7870-746-3.

No 252 **Peter Eklund:** An Epistemic Approach to Interactive Design in Multiple Inheritance Hierarchies, 1991, ISBN 91-7870-784-6.

No 258 **Patrick Doherty:** NML3 - A Non-Monotonic Formalism with Explicit Defaults, 1991, ISBN 91-7870-816-8.

No 260 **Nahid Shahmehri:** Generalized Algorithmic Debugging, 1991, ISBN 91-7870-828-1.

No 264 **Nils Dahlbäck:** Representation of Discourse-Cognitive and Computational Aspects, 1992, ISBN 91-7870-850-8.

No 265 **Ulf Nilsson:** Abstract Interpretations and Abstract Machines: Contributions to a Methodology for the Implementation of Logic Programs, 1992, ISBN 91-7870-858-3.

No 270 **Ralph Rönnquist:** Theory and Practice of Tense-bound Object References, 1992, ISBN 91-7870-873-7.

No 273 **Björn Fjellborg:** Pipeline Extraction for VLSI Data Path Synthesis, 1992, ISBN 91-7870-880-X.

No 276 **Staffan Bonnier:** A Formal Basis for Horn Clause Logic with External Polymorphic Functions, 1992, ISBN 91-7870-896-6.

No 277 **Kristian Sandahl:** Developing Knowledge Management Systems with an Active Expert Methodology, 1992, ISBN 91-7870-897-4.

No 281 **Christer Bäckström:** Computational Complexity of Reasoning about Plans, 1992, ISBN 91-7870-979-2.

No 292 **Mats Wirén:** Studies in Incremental Natural Language Analysis, 1992, ISBN 91-7871-027-8.

No 297 **Mariam Kamkar:** Interprocedural Dynamic Slicing with Applications to Debugging and Testing, 1993, ISBN 91-7871-065-0.

No 302 **Tingting Zhang:** A Study in Diagnosis Using Classification and Defaults, 1993, ISBN 91-7871-078-2.

No 312 **Arne Jönsson:** Dialogue Management for Natural Language Interfaces - An Empirical Approach, 1993, ISBN 91-7871-110-X.

No 338 **Simin Nadjm-Tehrani:** Reactive Systems in Physical Environments: Compositional Modelling and Framework for Verification, 1994, ISBN 91-7871-237-8.

No 371 **Bengt Savén:** Business Models for Decision Support and Learning. A Study of Discrete-Event Manufacturing Simulation at Asea/ABB 1968-1993, 1995, ISBN 91-7871-494-X.

No 375 **Ulf Söderman:** Conceptual Modelling of Mode Switching Physical Systems, 1995, ISBN 91-7871-516-4.

No 383 **Andreas Kågedal:** Exploiting Groundness in Logic Programs, 1995, ISBN 91-7871-538-5.

No 396 **George Fodor:** Ontological Control, Description, Identification and Recovery from Problematic Control Situations, 1995, ISBN 91-7871-603-9.

No 413 **Mikael Pettersson:** Compiling Natural Semantics, 1995, ISBN 91-7871-641-1.

No 414 **Xinli Gu:** RT Level Testability Improvement by Testability Analysis and Transformations, 1996, ISBN 91-7871-654-3.

No 416 **Hua Shu:** Distributed Default Reasoning, 1996, ISBN 91-7871-665-9.

No 429 **Jaime Villegas:** Simulation Supported Industrial Training from an Organisational Learning Perspective - Development and Evaluation of the SSIT Method, 1996, ISBN 91-7871-700-0.

No 431 **Peter Jonsson:** Studies in Action Planning: Algorithms and Complexity, 1996, ISBN 91-7871-704-3.

No 437 **Johan Boye:** Directional Types in Logic Programming, 1996, ISBN 91-7871-725-6.

No 439 **Cecilia Sjöberg:** Activities, Voices and Arenas: Participatory Design in Practice, 1996, ISBN 91-7871-728-0.

No 448 **Patrick Lambrix:** Part-Whole Reasoning in Description Logics, 1996, ISBN 91-7871-820-1.

No 452 **Kjell Orsborn:** On Extensible and Object-Relational Database Technology for Finite Element Analysis Applications, 1996, ISBN 91-7871-827-9.

No 459 **Olof Johansson:** Development Environments for Complex Product Models, 1996, ISBN 91-7871-855-4.

No 461 **Lena Strömbäck:** User-Defined Constructions in Unification-Based Formalisms, 1997, ISBN 91-7871-857-0.

No 462 **Lars Degerstedt:** Tabulation-based Logic Programming: A Multi-Level View of Query Answering, 1996, ISBN 91-7871-858-9.

No 475 **Fredrik Nilsson:** Strategi och ekonomisk styrning - En studie av hur ekonomiska styrsystem utformas och används efter företagsförvärv, 1997, ISBN 91-7871-914-3.

No 480 **Mikael Lindvall:** An Empirical Study of Requirements-Driven Impact Analysis in Object-Oriented Software Evolution, 1997, ISBN 91-7871-927-5.

No 485 **Göran Forslund:** Opinion-Based Systems: The Cooperative Perspective on Knowledge-Based Decision Support, 1997, ISBN 91-7871-938-0.

No 494 **Martin Sköld:** Active Database Management Systems for Monitoring and Control, 1997, ISBN 91-7219-002-7.

No 495 **Hans Olsén:** Automatic Verification of Petri Nets in a CLP framework, 1997, ISBN 91-7219-011-6.

No 498 **Thomas Drakengren:** Algorithms and Complexity for Temporal and Spatial Formalisms, 1997, ISBN 91-7219-019-1.

No 502 **Jakob Axelsson:** Analysis and Synthesis of Heterogeneous Real-Time Systems, 1997, ISBN 91-7219-035-3.

No 503 **Johan Ringström:** Compiler Generation for Data-Parallel Programming Languages from Two-Level Semantics Specifications, 1997, ISBN 91-7219-045-0.

No 512 **Anna Moberg:** Närhet och distans - Studier av kommunikationsmönster i satellitkontor och flexibla kontor, 1997, ISBN 91-7219-119-8.

No 520 **Mikael Ronström:** Design and Modelling of a Parallel Data Server for Telecom Applications, 1998, ISBN 91-7219-169-4.

No 522 **Niclas Ohlsson:** Towards Effective Fault Prevention - An Empirical Study in Software Engineering, 1998, ISBN 91-7219-176-7.

No 526 **Joachim Karlsson:** A Systematic Approach for Prioritizing Software Requirements, 1998, ISBN 91-7219-184-8.

No 530 **Henrik Nilsson:** Declarative Debugging for Lazy Functional Languages, 1998, ISBN 91-7219-197-X.

No 555 **Jonas Hallberg:** Timing Issues in High-Level Synthesis, 1998, ISBN 91-7219-369-7.

No 561 **Ling Lin:** Management of 1-D Sequence Data - From Discrete to Continuous, 1999, ISBN 91-7219-402-2.

No 563 **Eva L Ragnemalm:** Student Modelling based on Collaborative Dialogue with a Learning Companion, 1999, ISBN 91-7219-412-X.

No 567 **Jörgen Lindström:** Does Distance matter? On geographical dispersion in organisations, 1999, ISBN 91-7219-439-1.

No 582 **Vanja Josifovski:** Design, Implementation and Evaluation of a Distributed Mediator System for Data Integration, 1999, ISBN 91-7219-482-0.

No 589 **Rita Kovordányi:** Modeling and Simulating Inhibitory Mechanisms in Mental Image Reinterpretation - Towards Cooperative Human-Computer Creativity, 1999, ISBN 91-7219-506-1.

No 592 **Mikael Ericsson:** Supporting the Use of Design Knowledge - An Assessment of Commenting Agents, 1999, ISBN 91-7219-532-0.

No 593 **Lars Karlsson:** Actions, Interactions and Narratives, 1999, ISBN 91-7219-534-7.

No 594 **C. G. Mikael Johansson:** Social and Organizational Aspects of Requirements Engineering Methods - A practice-oriented approach, 1999, ISBN 91-7219-541-X.

No 595 **Jörgen Hansson:** Value-Driven Multi-Class Overload Management in Real-Time Database Systems, 1999, ISBN 91-7219-542-8.

No 596 **Niklas Hallberg:** Incorporating User Values in the Design of Information Systems and Services in the Public Sector: A Methods Approach, 1999, ISBN 91-7219-543-6.

No 597 **Vivian Vimarlund:** An Economic Perspective on the Analysis of Impacts of Information Technology: From Case Studies in Health-Care towards General Models and Theories, 1999, ISBN 91-7219-544-4.

No 598 **Johan Jenvald:** Methods and Tools in Computer-Supported Taskforce Training, 1999, ISBN 91-7219-547-9.

No 607 **Magnus Merkel:** Understanding and enhancing translation by parallel text processing, 1999, ISBN 91-7219-614-9.

No 611 **Silvia Coradeschi:** Anchoring symbols to sensory data, 1999, ISBN 91-7219-623-8.

No 613 **Man Lin:** Analysis and Synthesis of Reactive Systems: A Generic Layered Architecture Perspective, 1999, ISBN 91-7219-630-0.

No 618 **Jimmy Tjäder:** Systemimplementering i praktiken - En studie av logiker i fyra projekt, 1999, ISBN 91-7219-657-2.

No 627 **Vadim Engelson:** Tools for Design, Interactive Simulation, and Visualization of Object-Oriented Models in Scientific Computing, 2000, ISBN 91-7219-709-9.

No 637 **Esa Falkenroth:** Database Technology for Control and Simulation, 2000, ISBN 91-7219-766-8.

No 639 **Per-Arne Persson:** Bringing Power and Knowledge Together: Information Systems Design for Autonomy and Control in Command Work, 2000, ISBN 91-7219-796-X.

No 660 **Erik Larsson:** An Integrated System-Level Design for Testability Methodology, 2000, ISBN 91-7219-890-7.

No 688 **Marcus Bjäreland:** Model-based Execution Monitoring, 2001, ISBN 91-7373-016-5.

No 689 **Joakim Gustafsson:** Extending Temporal Action Logic, 2001, ISBN 91-7373-017-3.

No 720 **Carl-Johan Petri:** Organizational Information Provision - Managing Mandatory and Discretionary Use of Information Technology, 2001, ISBN 91-7373-126-9.

No 724 **Paul Scerri:** Designing Agents for Systems with Adjustable Autonomy, 2001, ISBN 91-7373-207-9.

No 725 **Tim Heyer:** Semantic Inspection of Software Artifacts: From Theory to Practice, 2001, ISBN 91-7373-208-7.

No 726 **Pär Carlshamre:** A Usability Perspective on Requirements Engineering - From Methodology to Product Development, 2001, ISBN 91-7373-212-5.

No 732 **Juha Takkinen:** From Information Management to Task Management in Electronic Mail, 2002, ISBN 91-7373-258-3.

No 745 **Johan Åberg:** Live Help Systems: An Approach to Intelligent Help for Web Information Systems, 2002, ISBN 91-7373-311-3.

No 746 **Rego Granlund:** Monitoring Distributed Teamwork Training, 2002, ISBN 91-7373-312-1.

No 757 **Henrik André-Jönsson:** Indexing Strategies for Time Series Data, 2002, ISBN 917373-346-6.

No 747 **Anneli Hagdahl:** Development of IT-supported Interorganisational Collaboration - A Case Study in the Swedish Public Sector, 2002, ISBN 91-7373-314-8.

No 749 **Sofie Pilemalm:** Information Technology for Non-Profit Organisations - Extended Participatory Design of an Information System for Trade Union Shop Stewards, 2002, ISBN 91-7373-318-0.

No 765 **Stefan Holmlid:** Adapting users: Towards a theory of use quality, 2002, ISBN 91-7373-397-0.

No 771 **Magnus Morin:** Multimedia Representations of Distributed Tactical Operations, 2002, ISBN 91-7373-421-7.

No 772 **Pawel Pietrzak:** A Type-Based Framework for Locating Errors in Constraint Logic Programs, 2002, ISBN 91-7373-422-5.

No 758 **Erik Berglund:** Library Communication Among Programmers Worldwide, 2002, ISBN 91-7373-349-0.

No 774 **Choong-ho Yi:** Modelling Object-Oriented Dynamic Systems Using a Logic-Based Framework, 2002, ISBN 91-7373-424-1.

No 779 **Mathias Broxvall:** A Study in the Computational Complexity of Temporal Reasoning, 2002, ISBN 91-7373-440-3.

No 793 **Asmus Pandikow:** A Generic Principle for Enabling Interoperability of Structured and Object-Oriented Analysis and Design Tools, 2002, ISBN 91-7373-479-9.

No 785 **Lars Hult:** Publika Informationstjänster. En studie av den Internetbaserade encyklopedins bruksegenskaper, 2003, ISBN 91-7373-461-6.

No 800 **Lars Taxén:** A Framework for the Coordination of Complex Systems´ Development, 2003, ISBN 91-7373-604-X.

No 808 **Klas Gäre:** Tre perspektiv på förväntningar och förändringar i samband med införande av informationssystem, 2003, ISBN 91-7373-618-X.

No 821 **Mikael Kindborg:** Concurrent Comics - programming of social agents by children, 2003, ISBN 91-7373-651-1.

No 823 **Christina Ölvingson:** On Development of Information Systems with GIS Functionality in Public Health Informatics: A Requirements Engineering Approach, 2003, ISBN 91-7373-656-2.

No 828 **Tobias Ritzau:** Memory Efficient Hard Real-Time Garbage Collection, 2003, ISBN 91-7373-666-X.

No 833 **Paul Pop:** Analysis and Synthesis of Communication-Intensive Heterogeneous Real-Time Systems, 2003, ISBN 91-7373-683-X.

No 852 **Johan Moe:** Observing the Dynamic Behaviour of Large Distributed Systems to Improve Development and Testing – An Empirical Study in Software Engineering, 2003, ISBN 91-7373-779-8.

No 867 **Erik Herzog:** An Approach to Systems Engineering Tool Data Representation and Exchange, 2004, ISBN 91-7373-929-4.

No 872 **Aseel Berglund:** Augmenting the Remote Control: Studies in Complex Information Navigation for Digital TV, 2004, ISBN 91-7373-940-5.

No 869 **Jo Skåmedal:** Telecommuting's Implications on Travel and Travel Patterns, 2004, ISBN 91-7373-935-9.

No 870 **Linda Askenäs:** The Roles of IT - Studies of Organising when Implementing and Using Enterprise Systems, 2004, ISBN 91-7373-936-7.

No 874 **Annika Flycht-Eriksson:** Design and Use of Ontologies in Information-Providing Dialogue Systems, 2004, ISBN 91-7373-947-2.

No 873 **Peter Bunus:** Debugging Techniques for Equation-Based Languages, 2004, ISBN 91-7373-941-3.

No 876 **Jonas Mellin:** Resource-Predictable and Efficient Monitoring of Events, 2004, ISBN 91-7373-956-1.

No 883 **Magnus Bång:** Computing at the Speed of Paper: Ubiquitous Computing Environments for Healthcare Professionals, 2004, ISBN 91-7373-971-5.

No 882 **Robert Eklund:** Disfluency in Swedish human-human and human-machine travel booking dialogues, 2004, ISBN 91-7373-966-9.

No 887 **Anders Lindström:** English and other Foreign Linguistic Elements in Spoken Swedish. Studies of Productive Processes and their Modelling using Finite-State Tools, 2004, ISBN 91-7373-981-2.

No 889 **Zhiping Wang:** Capacity-Constrained Production-inventory systems - Modelling and Analysis in both a traditional and an e-business context, 2004, ISBN 91-85295-08-6.

No 893 **Pernilla Qvarfordt:** Eyes on Multimodal Interaction, 2004, ISBN 91-85295-30-2.

No 910 **Magnus Kald:** In the Borderland between Strategy and Management Control - Theoretical Framework and Empirical Evidence, 2004, ISBN 91-85295-82-5.

No 918 **Jonas Lundberg:** Shaping Electronic News: Genre Perspectives on Interaction Design, 2004, ISBN 91-85297-14-3.

No 900 **Mattias Arvola:** Shades of use: The dynamics of interaction design for sociable use, 2004, ISBN 91-85295-42-6.

No 920 **Luis Alejandro Cortés:** Verification and Scheduling Techniques for Real-Time Embedded Systems, 2004, ISBN 91-85297-21-6.

No 929 **Diana Szentivanyi:** Performance Studies of Fault-Tolerant Middleware, 2005, ISBN 91-85297-58-5.

No 933 **Mikael Cäker:** Management Accounting as Constructing and Opposing Customer Focus: Three Case Studies on Management Accounting and Customer Relations, 2005, ISBN 91-85297-64-X.

No 937 **Jonas Kvarnström:** TALplanner and Other Extensions to Temporal Action Logic, 2005, ISBN 91-85297-75-5.

No 938 **Bourhane Kadmiry:** Fuzzy Gain-Scheduled Visual Servoing for Unmanned Helicopter, 2005, ISBN 91-85297-76-3.

No 945 **Gert Jervan:** Hybrid Built-In Self-Test and Test Generation Techniques for Digital Systems, 2005, ISBN 91-85297-97-6.

No 946 **Anders Arpteg:** Intelligent Semi-Structured Information Extraction, 2005, ISBN 91-85297-98-4.

No 947 **Ola Angelsmark:** Constructing Algorithms for Constraint Satisfaction and Related Problems - Methods and Applications, 2005, ISBN 91-85297-99-2.

No 963 **Calin Curescu:** Utility-based Optimisation of Resource Allocation for Wireless Networks, 2005, ISBN 91-85457-07-8.

No 972 **Björn Johansson:** Joint Control in Dynamic Situations, 2005, ISBN 91-85457-31-0.

No 974 **Dan Lawesson:** An Approach to Diagnosability Analysis for Interacting Finite State Systems, 2005, ISBN 91-85457-39-6.

No 979 **Claudiu Duma:** Security and Trust Mechanisms for Groups in Distributed Services, 2005, ISBN 91-85457-54-X.

No 983 **Sorin Manolache:** Analysis and Optimisation of Real-Time Systems with Stochastic Behaviour, 2005, ISBN 91-85457-60-4.

No 986 **Yuxiao Zhao:** Standards-Based Application Integration for Business-to-Business Communications, 2005, ISBN 91-85457-66-3.

No 1004 **Patrik Haslum:** Admissible Heuristics for Automated Planning, 2006, ISBN 91-85497-28-2.

No 1005 **Aleksandra Tešanovic:** Developing Reusable and Reconfigurable Real-Time Software using Aspects and Components, 2006, ISBN 91-85497-29-0.

No 1008 **David Dinka:** Role, Identity and Work: Extending the design and development agenda, 2006, ISBN 91-85497-42-8.

No 1009 **Iakov Nakhimovski:** Contributions to the Modeling and Simulation of Mechanical Systems with Detailed Contact Analysis, 2006, ISBN 91-85497-43-X.

No 1013 **Wilhelm Dahllöf:** Exact Algorithms for Exact Satisfiability Problems, 2006, ISBN 91-85523-97-6.

No 1016 **Levon Saldamli:** PDEModelica - A High-Level Language for Modeling with Partial Differential Equations, 2006, ISBN 91-85523-84-4.

No 1017 **Daniel Karlsson:** Verification of Component-based Embedded System Designs, 2006, ISBN 91-85523-79-8

No 1018 **Ioan Chisalita:** Communication and Networking Techniques for Traffic Safety Systems, 2006, ISBN 91-85523-77-1.

No 1019 **Tarja Susi:** The Puzzle of Social Activity - The Significance of Tools in Cognition and Cooperation, 2006, ISBN 91-85523-71-2.

No 1021 **Andrzej Bednarski:** Integrated Optimal Code Generation for Digital Signal Processors, 2006, ISBN 91-85523-69-0.

No 1022 **Peter Aronsson:** Automatic Parallelization of Equation-Based Simulation Programs, 2006, ISBN 91-85523-68-2.

No 1030 **Robert Nilsson:** A Mutation-based Framework for Automated Testing of Timeliness, 2006, ISBN 91-85523-35-6.

No 1034 **Jon Edvardsson:** Techniques for Automatic Generation of Tests from Programs and Specifications, 2006, ISBN 91-85523-31-3.

No 1035 **Vaida Jakoniene:** Integration of Biological Data, 2006, ISBN 91-85523-28-3.

No 1045 **Genevieve Gorrell:** Generalized Hebbian Algorithms for Dimensionality Reduction in Natural Language Processing, 2006, ISBN 91-85643-88-2.

No 1051 **Yu-Hsing Huang:** Having a New Pair of Glasses - Applying Systemic Accident Models on Road Safety, 2006, ISBN 91-85643-64-5.

No 1054 **Åsa Hedenskog:** Perceive those things which cannot be seen - A Cognitive Systems Engineering perspective on requirements management, 2006, ISBN 91-85643-57-2.

No 1061 **Cécile Åberg:** An Evaluation Platform for Semantic Web Technology, 2007, ISBN 91-85643-31-9.

No 1073 **Mats Grindal:** Handling Combinatorial Explosion in Software Testing, 2007, ISBN 978-91-85715-74-9.

No 1075 **Almut Herzog:** Usable Security Policies for Runtime Environments, 2007, ISBN 978-91-85715-65-7.

No 1079 **Magnus Wahlström:** Algorithms, measures, and upper bounds for Satisfiability and related problems, 2007, ISBN 978-91-85715-55-8.

No 1083 **Jesper Andersson:** Dynamic Software Architectures, 2007, ISBN 978-91-85715-46-6.

No 1086 **Ulf Johansson:** Obtaining Accurate and Comprehensible Data Mining Models - An Evolutionary Approach, 2007, ISBN 978-91-85715-34-3.

No 1089 **Traian Pop:** Analysis and Optimisation of Distributed Embedded Systems with Heterogeneous Scheduling Policies, 2007, ISBN 978-91-85715-27-5.

No 1091 **Gustav Nordh:** Complexity Dichotomies for CSP-related Problems, 2007, ISBN 978-91-85715-20-6.

No 1106 **Per Ola Kristensson:** Discrete and Continuous Shape Writing for Text Entry and Control, 2007, ISBN 978-91-85831-77-7.

No 1110 **He Tan:** Aligning Biomedical Ontologies, 2007, ISBN 978-91-85831-56-2.

No 1112 **Jessica Lindblom:** Minding the body - Interacting socially through embodied action, 2007, ISBN 978-91-85831-48-7.

No 1113 **Pontus Wärnestål:** Dialogue Behavior Management in Conversational Recommender Systems, 2007, ISBN 978-91-85831-47-0.

No 1120 **Thomas Gustafsson:** Management of Real-Time Data Consistency and Transient Overloads in Embedded Systems, 2007, ISBN 978-91-85831-33-3.

No 1127 **Alexandru Andrei:** Energy Efficient and Predictable Design of Real-time Embedded Systems, 2007, ISBN 978-91-85831-06-7.

No 1139 **Per Wikberg:** Eliciting Knowledge from Experts in Modeling of Complex Systems: Managing Variation and Interactions, 2007, ISBN 978-91-85895-66-3.

No 1143 **Mehdi Amirijoo:** QoS Control of Real-Time Data Services under Uncertain Workload, 2007, ISBN 978-91-85895-49-6.

No 1150 **Sanny Syberfeldt:** Optimistic Replication with Forward Conflict Resolution in Distributed Real-Time Databases, 2007, ISBN 978-91-85895-27-4.

No 1155 **Beatrice Alenljung:** Envisioning a Future Decision Support System for Requirements Engineering - A Holistic and Human-centred Perspective, 2008, ISBN 978-91-85895-11-3.

No 1156 **Artur Wilk:** Types for XML with Application to Xcerpt, 2008, ISBN 978-91-85895-08-3.

No 1183 **Adrian Pop:** Integrated Model-Driven Development Environments for Equation-Based Object-Oriented Languages, 2008, ISBN 978-91-7393-895-2.

No 1185 **Jörgen Skågeby:** Gifting Technologies - Ethnographic Studies of End-users and Social Media Sharing, 2008, ISBN 978-91-7393-892-1.

No 1187 **Imad-Eldin Ali Abugessaisa:** Analytical tools and information-sharing methods supporting road safety organizations, 2008, ISBN 978-91-7393-887-7.

No 1204 **H. Joe Steinhauer:** A Representation Scheme for Description and Reconstruction of Object Configurations Based on Qualitative Relations, 2008, ISBN 978-91-7393-823-5.

No 1222 **Anders Larsson:** Test Optimization for Core-based System-on-Chip, 2008, ISBN 978-91-7393-768-9.

No 1238 **Andreas Borg:** Processes and Models for Capacity Requirements in Telecommunication Systems, 2009, ISBN 978-91-7393-700-9.

No 1240 **Fredrik Heintz:** DyKnow: A Stream-Based Knowledge Processing Middleware Framework, 2009, ISBN 978-91-7393-696-5.

No 1241 **Birgitta Lindström:** Testability of Dynamic Real-Time Systems, 2009, ISBN 978-91-7393-695-8.

No 1244 **Eva Blomqvist:** Semi-automatic Ontology Construction based on Patterns, 2009, ISBN 978-91-7393-683-5.

No 1249 **Rogier Woltjer:** Functional Modeling of Constraint Management in Aviation Safety and Command and Control, 2009, ISBN 978-91-7393-659-0.

No 1260 **Gianpaolo Conte:** Vision-Based Localization and Guidance for Unmanned Aerial Vehicles, 2009, ISBN 978-91-7393-603-3.

No 1262 **AnnMarie Ericsson:** Enabling Tool Support for Formal Analysis of ECA Rules, 2009, ISBN 978-91-7393-598-2.

No 1266 **Jiri Trnka:** Exploring Tactical Command and Control: A Role-Playing Simulation Approach, 2009, ISBN 978-91-7393-571-5.

No 1268 **Bahlol Rahimi:** Supporting Collaborative Work through ICT - How End-users Think of and Adopt Integrated Health Information Systems, 2009, ISBN 978-91-7393-550-0.

No 1274 **Fredrik Kuivinen:** Algorithms and Hardness Results for Some Valued CSPs, 2009, ISBN 978-91-7393-525-8.

No 1281 **Gunnar Mathiason:** Virtual Full Replication for Scalable Distributed Real-Time Databases, 2009, ISBN 978-91-7393-503-6.

No 1290 **Viacheslav Izosimov:** Scheduling and Optimization of Fault-Tolerant Distributed Embedded Systems, 2009, ISBN 978-91-7393-482-4.

No 1294 **Johan Thapper:** Aspects of a Constraint Optimisation Problem, 2010, ISBN 978-91-7393-464-0.

No 1306 **Susanna Nilsson:** Augmentation in the Wild: User Centered Development and Evaluation of Augmented Reality Applications, 2010, ISBN 978-91-7393-416-9.

No 1313 **Christer Thörn:** On the Quality of Feature Models, 2010, ISBN 978-91-7393-394-0.

No 1321 **Zhiyuan He:** Temperature Aware and Defect-Probability Driven Test Scheduling for System-on-Chip, 2010, ISBN 978-91-7393-378-0.

No 1333 **David Broman:** Meta-Languages and Semantics for Equation-Based Modeling and Simulation, 2010, ISBN 978-91-7393-335-3.

No 1337 **Alexander Siemers:** Contributions to Modelling and Visualisation of Multibody Systems Simulations with Detailed Contact Analysis, 2010, ISBN 978-91-7393-317-9.

No 1354 **Mikael Asplund:** Disconnected Discoveries: Availability Studies in Partitioned Networks, 2010, ISBN 978-91-7393-278-3.

No 1359 **Jana Rambusch**: Mind Games Extended: Understanding Gameplay as Situated Activity, 2010, ISBN 978-91-7393-252-3.

No 1373 **Sonia Sangari**: Head Movement Correlates to Focus Assignment in Swedish, 2011, ISBN 978-91-7393-154-0.

No 1374 **Jan-Erik Källhammer**: Using False Alarms when Developing Automotive Active Safety Systems, 2011, ISBN 978-91-7393-153-3.

No 1375 **Mattias Eriksson**: Integrated Code Generation, 2011, ISBN 978-91-7393-147-2.

No 1381 **Ola Leifler**: Affordances and Constraints of Intelligent Decision Support for Military Command and Control – Three Case Studies of Support Systems, 2011, ISBN 978-91-7393-133-5.

No 1386 **Soheil Samii**: Quality-Driven Synthesis and Optimization of Embedded Control Systems, 2011, ISBN 978-91-7393-102-1.

No 1419 **Erik Kuiper**: Geographic Routing in Intermittently-connected Mobile Ad Hoc Networks: Algorithms and Performance Models, 2012, ISBN 978-91-7519-981-8.

No 1451 **Sara Stymne**: Text Harmonization Strategies for Phrase-Based Statistical Machine Translation, 2012, ISBN 978-91-7519-887-3.

No 1455 **Alberto Montebelli**: Modeling the Role of Energy Management in Embodied Cognition, 2012, ISBN 978-91-7519-882-8.

No 1465 **Mohammad Saifullah**: Biologically-Based Interactive Neural Network Models for Visual Attention and Object Recognition, 2012, ISBN 978-91-7519-838-5.

No 1490 **Tomas Bengtsson**: Testing and Logic Optimization Techniques for Systems on Chip, 2012, ISBN 978-91-7519-742-5.

No 1481 **David Byers**: Improving Software Security by Preventing Known Vulnerabilities, 2012, ISBN 978-91-7519-784-5.

No 1496 **Tommy Färnqvist**: Exploiting Structure in CSP-related Problems, 2013, ISBN 978-91-7519-711-1.

No 1503 **John Wilander**: Contributions to Specification, Implementation, and Execution of Secure Software, 2013, ISBN 978-91-7519-681-7.

No 1506 **Magnus Ingmarsson**: Creating and Enabling the Useful Service Discovery Experience, 2013, ISBN 978-91-7519-662-6.

No 1547 **Wladimir Schamai**: Model-Based Verification of Dynamic System Behavior against Requirements: Method, Language, and Tool, 2013, ISBN 978-91-7519-505-6.

No 1551 **Henrik Svensson**: Simulations, 2013, ISBN 978-91-7519-491-2.

No 1559 **Sergiu Rafiliu**: Stability of Adaptive Distributed Real-Time Systems with Dynamic Resource Management, 2013, ISBN 978-91-7519-471-4.

No 1581 **Usman Dastgeer**: Performance-aware Component Composition for GPU-based Systems, 2014, ISBN 978-91-7519-383-0.

No 1602 **Cai Li**: Reinforcement Learning of Locomotion based on Central Pattern Generators, 2014, ISBN 978-91-7519-313-7.

No 1652 **Roland Samlaus**: An Integrated Development Environment with Enhanced Domain-Specific Interactive Model Validation, 2015, ISBN 978-91-7519-090-7.

No 1663 **Hannes Uppman**: On Some Combinatorial Optimization Problems: Algorithms and Complexity, 2015, ISBN 978-91-7519-072-3.

No 1664 **Martin Sjölund**: Tools and Methods for Analysis, Debugging, and Performance Improvement of Equation-Based Models, 2015, ISBN 978-91-7519-071-6.

No 1666 **Kristian Stavåker**: Contributions to Simulation of Modelica Models on Data-Parallel Multi-Core Architectures, 2015, ISBN 978-91-7519-068-6.

No 1680 **Adrian Lifa**: Hardware/Software Codesign of Embedded Systems with Reconfigurable and Heterogeneous Platforms, 2015, ISBN 978-91-7519-040-2.

No 1685 **Bogdan Tanasa**: Timing Analysis of Distributed Embedded Systems with Stochastic Workload and Reliability Constraints, 2015, ISBN 978-91-7519-022-8.

No 1691 **Håkan Warnquist**: Troubleshooting Trucks – Automated Planning and Diagnosis, 2015, ISBN 978-91-7685-993-3.

No 1702 **Nima Aghaee**: Thermal Issues in Testing of Advanced Systems on Chip, 2015, ISBN 978-91-7685-949-0.

No 1715 **Maria Vasilevskaya**: Security in Embedded Systems: A Model-Based Approach with Risk Metrics, 2015, ISBN 978-91-7685-917-9.

No 1729 **Ke Jiang**: Security-Driven Design of Real-Time Embedded System, 2016, ISBN 978-91-7685-884-4.

No 1733 **Victor Lagerkvist**: Strong Partial Clones and the Complexity of Constraint Satisfaction Problems: Limitations and Applications, 2016, ISBN 978-91-7685-856-1.

No 1734 **Chandan Roy**: An Informed System Development Approach to Tropical Cyclone Track and Intensity Forecasting, 2016, ISBN 978-91-7685-854-7.

No 1746 **Amir Aminifar**: Analysis, Design, and Optimization of Embedded Control Systems, 2016, ISBN 978-91-7685-826-4.

No 1747 **Ekhiotz Vergara**: Energy Modelling and Fairness for Efficient Mobile Communication, 2016, ISBN 978-91-7685-822-6.

No 1748 **Dag Sonntag**: Chain Graphs – Interpretations, Expressiveness and Learning Algorithms, 2016, ISBN 978-91-7685-818-9.

No 1768 **Anna Vapen**: Web Authentication using Third-Parties in Untrusted Environments, 2016, ISBN 978-91-7685-753-3.

No 1778 **Magnus Jandinger**: On a Need to Know Basis: A Conceptual and Methodological Framework for Modelling and Analysis of Information Demand in an Enterprise Context, 2016, ISBN 978-91-7685-713-7.

No 1798 **Rahul Hiran**: Collaborative Network Security: Targeting Wide-area Routing and Edge-network Attacks, 2016, ISBN 978-91-7685-662-8.

No 1813 **Nicolas Melot**: Algorithms and Framework for Energy Efficient Parallel Stream Computing on Many-Core Architectures, 2016, ISBN 978-91-7685-623-9.

No 1823 **Amy Rankin**: Making Sense of Adaptations: Resilience in High-Risk Work, 2017, ISBN 978-91-7685-596-6.

No 1831 **Lisa Malmberg**: Building Design Capability in the Public Sector: Expanding the Horizons of Development, 2017, ISBN 978-91-7685-585-0.

No 1851 **Marcus Bendtsen**: Gated Bayesian Networks, 2017, ISBN 978-91-7685-525-6.

No 1852 **Zlatan Dragisic**: Completion of Ontologies and Ontology Networks, 2017, ISBN 978-91-7685-522-5.

No 1854 **Meysam Aghighi**: Computational Complexity of some Optimization Problems in Planning, 2017, ISBN 978-91-7685-519-5.

No 1863 **Simon Ståhlberg**: Methods for Detecting Unsolvable Planning Instances using Variable Projection, 2017, ISBN 978-91-7685-498-3.

No 1879 **Karl Hammar**: Content Ontology Design Patterns: Qualities, Methods, and Tools, 2017, ISBN 978-91-7685-454-9.

No 1887 **Ivan Ukhov**: System-Level Analysis and Design under Uncertainty, 2017, ISBN 978-91-7685-426-6.

No 1891 **Valentina Ivanova**: Fostering User Involvement in Ontology Alignment and Alignment Evaluation, 2017, ISBN 978-91-7685-403-7.

No 1902 **Vengatanathan Krishnamoorthi**: Efficient HTTP-based Adaptive Streaming of Linear and Interactive Videos, 2018, ISBN 978-91-7685-371-9.

No 1903 **Lu Li**: Programming Abstractions and Optimization Techniques for GPU-based Heterogeneous Systems, 2018, ISBN 978-91-7685-370-2.

No 1913 **Jonas Rybing**: Studying Simulations with Distributed Cognition, 2018, ISBN 978-91-7685-348-1.

No 1936 **Leif Jonsson**: Machine Learning-Based Bug Handling in Large-Scale Software Development, 2018, ISBN 978-91-7685-306-1.

Linköping Studies in Arts and Science

No 504 **Ing-Marie Jonsson:** Social and Emotional Characteristics of Speech-based In-Vehicle Information Systems: Impact on Attitude and Driving Behaviour, 2009, ISBN 978-91-7393-478-7.

No 586 **Fabian Segelström:** Stakeholder Engagement for Service Design: How service designers identify and communicate insights, 2013, ISBN 978-91-7519-554-4.

No 618 **Johan Blomkvist:** Representing Future Situations of Service: Prototyping in Service Design, 2014, ISBN 978-91-7519-343-4.

No 620 **Marcus Mast:** Human-Robot Interaction for Semi-Autonomous Assistive Robots, 2014, ISBN 978-91-7519-319-9.

No 677 **Peter Berggren:** Assessing Shared Strategic Understanding, 2016, ISBN 978-91-7685-786-1.

No 695 **Mattias Forsblad:** Distributed cognition in home environments: The prospective memory and cognitive practices of older adults, 2016, ISBN 978-91-7685-686-4.

Linköping Studies in Statistics

No 9 **Davood Shahsavani:** Computer Experiments Designed to Explore and Approximate Complex Deterministic Models, 2008, ISBN 978-91-7393-976-8.

No 10 **Karl Wahlin:** Roadmap for Trend Detection and Assessment of Data Quality, 2008, ISBN 978-91-7393-792-4.

No 11 **Oleg Sysoev:** Monotonic regression for large multivariate datasets, 2010, ISBN 978-91-7393-412-1.

No 13 **Agné Burauskaite-Harju:** Characterizing Temporal Change and Inter-Site Correlations in Daily and Subdaily Precipitation Extremes, 2011, ISBN 978-91-7393-110-6.

No 14 **Måns Magnusson:** Scalable and Efficient Probabilistic Topic Model Inference for Textual Data, 2018, ISBN 978-91-7685-288-0.

Linköping Studies in Information Science

No 1 **Karin Axelsson:** Metodisk systemstrukturering- att skapa samstämmighet mellan informationssystemarkitektur och verksamhet, 1998. ISBN 9172-19-296-8.

No 2 **Stefan Cronholm:** Metodverktyg och användbarhet - en studie av datorstödd metodbaserad systemutveckling, 1998, ISBN 9172-19-299-2.

No 3 **Anders Avdic:** Användare och utvecklare - om anveckling med kalkylprogram, 1999. ISBN 91-7219-606-8.

No 4 **Owen Eriksson:** Kommunikationskvalitet hos informationssystem och affärsprocesser, 2000, ISBN 91-7219-811-7.

No 5 **Mikael Lind:** Från system till process - kriterier för processbestämning vid verksamhetsanalys, 2001, ISBN 91-7373-067-X.

No 6 **Ulf Melin:** Koordination och informationssystem i företag och nätverk, 2002, ISBN 91-7373-278-8.

No 7 **Pär J. Ågerfalk:** Information Systems Actability - Understanding Information Technology as a Tool for Business Action and Communication, 2003, ISBN 91-7373-628-7.

No 8 **Ulf Seigerroth:** Att förstå och förändra systemutvecklingsverksamheter - en taxonomi för metautveckling, 2003, ISBN 91-7373-736-4.

No 9 **Karin Hedström:** Spår av datoriseringens värden – Effekter av IT i äldreomsorg, 2004, ISBN 91-7373-963-4.

No 10 **Ewa Braf:** Knowledge Demanded for Action - Studies on Knowledge Mediation in Organisations, 2004, ISBN 91-85295-47-7.

No 11 **Fredrik Karlsson:** Method Configuration method and computerized tool support, 2005, ISBN 91-85297-48-8.

No 12 **Malin Nordström:** Styrbar systemförvaltning - Att organisera systemförvaltningsverksamhet med hjälp av effektiva förvaltningsobjekt, 2005, ISBN 91-85297-60-7.

No 13 **Stefan Holgersson:** Yrke: POLIS - Yrkeskunskap, motivation, IT-system och andra förutsättningar för polisarbete, 2005, ISBN 91-85299-43-X.

No 14 **Benneth Christiansson, Marie-Therese Christiansson:** Mötet mellan process och komponent - mot ett ramverk för en verksamhetsnära kravspecifikation vid anskaffning av komponentbaserade informationssystem, 2006, ISBN 91-85643-22-X.