Time dynamic topic models

Patrick Jähnichen

UNIVERSITÄT LEIPZIG

Leipzig 2015

Time dynamic topic models

Patrick Jähnichen

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig angenommene

D I S S E R T A T I O N

zur Erlangung des akademischen Grades
Doctor rerum naturalium
(Dr. rer. nat.)

im Fachgebiet Informatik

vorgelegt von
Patrick Jähnichen
geboren am 08.05.1983 in Eilenburg

Die Annahme der Dissertation wurde empfohlen von:
1. Prof. Dr. Gerhard Heyer, Universität Leipzig
2. Prof. Dr. Khurshid Ahmad, Trinity College Dublin

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 22.03.2016 mit dem Gesamtprädikat
magna cum laude

Abstract

Information extraction from large corpora can be a useful tool for many applications in industry and academia. For instance, political communication science has just recently begun to use the opportunities that come with the availability of massive amounts of information available through the Internet and the computational tools that natural language processing can provide. We give a linguistically motivated interpretation of topic modeling, a state-of-the-art algorithm for extracting latent semantic sets of words from large text corpora, and extend this interpretation to cover issues and issue-cycles as theoretical constructs coming from political communication science. We build on a dynamic topic model, a model whose semantic sets of words are allowed to evolve over time governed by a Brownian motion stochastic process and apply a new form of analysis to its result. Generally this analysis is based on the notion of volatility as in the rate of change of stocks or derivatives known from econometrics. We claim that the rate of change of sets of semantically related words can be interpreted as issue-cycles, the word sets as describing the underlying issue. Generalizing over the existing work, we introduce dynamic topic models that are driven by general (Brownian motion is a special case of our model) Gaussian processes, a family of stochastic processes defined by the function that determines their covariance structure. We use the above assumption and apply a certain class of covariance functions to allow for an appropriate rate of change in word sets while preserving the semantic relatedness among words. Applying our findings to a large newspaper data set, the New York Times Annotated corpus (all articles between 1987 and 2007), we are able to identify sub-topics in time, *time-localized topics* and find patterns in their behavior over time. However, we have to drop the assumption of semantic relatedness over all available time for any one topic. Time-localized topics are consistent in themselves but do not necessarily share semantic meaning between each other. They can, however, be interpreted to capture the notion of issues and their behavior that of issue-cycles.

Contents

List of Figures

List of Algorithms

List of Tables

Chapter 1

Introduction

1.1 Information flood and bringing shape to it

Since the beginning of the 1990s and the emergence of the internet, vast numbers of new and old (mostly) textual information became available to the general public. While this is a welcome fact, the problem comes with the sheer amount of information and with the situation that large amounts of the available information are not structured in any way nor have any structured meta-information associated to them. The availability of computers and their growing computational capabilities makes it obvious to utilize them to solve this problem. A specific branch of computer science, *text mining*, is devoted to this task, using computational techniques to gain quantitative and/or qualitative access to large bodies of texts by combining findings from language technology, linguistics, probability theory and computer science (see Heyer et al., 2006; Manning and Schütze, 1999). Text mining primarily uses the raw content of a text as its input data but many solutions are able to incorporate additional knowledge such as authors, publication date (as ours) or citations. How does this form of analysis and automated structuring look like? One of the earliest attempts to automatically capture the contents of an unknown text is that of Luhn (1958) who used statistical properties of word frequency distributions to identify significant portions (words and sentences) of a text to provide an automated abstract of it. In fact, the technique that underlies our research, *topic modeling* (Blei et al., 2003), also makes use of the statistical properties of word frequencies but employs a much more sophisticated probabilistic approach to model these properties and to infer interpretable data from it. It defines an artificial generative process that is assumed to generate the encountered data. Of course this is only a simplified image of reality but is has been shown that the results obtained are indeed interpretable in a qualitative way (e.g. Boyd-Graber et al., 2009).

In this work a specific form of automated quantitative text analysis is examined, providing a way to compress the contents of large text collections into a form easily accessible to humans. We exploit a meta-datum that often is available for text documents: the date of creation or, more often in the type of data that we analyze, the date of publication.

Blei and Lafferty (2006), and extending it, (Wang et al., 2008), have introduced *dynamic* topic models that are able to utilize this additional data. The key concept is to determine thematic structures (as topic models do in general) and then let those structures evolve over time. Here, we are especially interested in the type and behavior of this evolutionary process and use different ideas from linguistics (structural semantics), natural language processing (word co-occurrence analysis and word volatility) and machine learning and probability theory (topic modeling, stochastic processes and time series analysis) to control it.

Structural Semantics

Besides the obvious change of language and points of reference for a specific theme that is modeled by this assumption, a qualitative interpretation of this data is also possible. In particular, the knowledge about the evolution of themes may help to identify specific terms that undergo a change of semantic nature. This idea is mainly based on the doctrine of *structural semantics*. It dates back to Swiss language scientist Ferdinand de Saussure who described the notion of the meaning of a word as something that is not inherently present but is defined by its context[1] (where context can be some unit of analysis, e.g. a document, paragraph or sentence). Topic modeling does just that. It uses the aforementioned statistical properties of word frequencies and builds topics by clustering words that co-occur in documents over a document collection, i.e., here the unit of analysis in the spirit of de Saussure is a document. In consequence, we might interpret clusters of words that co-occur in documents across the collection (the topics) in a semantic way, i.e., we can consider them defining each other's meaning.

Word Volatility

Heyer et al. (2009) have used the above interpretation and examined how word contexts, that is the co-occurrences defining its meaning, change over time. Their approach is based on classical word co-occurrence analysis. For this, the frequency of word pairings in all units of analysis is counted and, using an appropriate measure[2], their statistical significance is measured. The result of this is a list of significant co-occurring words for each term. Heyer et al. (2009) have used this methodology to compute a term's significant co-occurrences for time-sliced subsets of a given text collection, i.e. they divided the collection according to some time-related criterion (this is also sometimes called "binning"). Sorting a term's co-occurrences according to their significance for each time slice results in a time series of ranks for each co-occurring word of a term. Computing the average variance in rank for each of those gives what they call the "word volatility" of the term where they borrowed

[1] Literally, he states that language is a somewhat arbitrary system "in which importance and value of one only emerges from the concurrent existence of the other." (German: "[...] in dem Geltung und Wert des einen nur aus dem gleichzeitigen Vorhandensein des anderen sich ergeben.")(de Saussure, 2001, p. 136).

[2] Log-likelihood ratio test (Dunning, 1993) have proven to be useful for this.

the expression from an econometric context where it (roughly) describes the rate at which stocks vary. Interpreting this in accordance to de Saussure they claim that this volatility can be seen as a measure of change of a term's meaning (due to the more rapid change in a terms context at higher volatility). While their results have proven to be of great value and have been extended(e.g. Heyer et al., 2011; Holz et al., 2010; Holz and Teresniak, 2010), they lack of the usual constraints that come with this classical approach: co-occurrences are by definition semantically related to a term but this relation is (also by definition) limited to direct co-occurrence and can not dissolve e.g. semantic ambiguities. Consider for example the often stressed term "bank". Its significant co-occurrences are (as provided by the Leipzig Wortschatz project[3]) "account", "river", "accounts", "cheque", "Bank", "money", "holiday" etc. While a human may know about the different meanings that these co-occurrences imply, a computer program does not, i.e. the given approach suffers from the inability to semantically separate co-occurrences from each other. This leads to the problem that a resulting high word volatility might be misinterpreted. In the given example one could argue that during the world financial crisis there surely was rapid change in the significant co-occurrences to "bank". The concepts of a river bank and bank holiday however have just as surely stayed the same. The given approach thus has no means of detecting and confining word volatility to a specific semantic aspect of a term. We adopt the basic idea and will develop it in the context of topic modeling to use the semantic resolution that it provides to collectively identify terms that experience a high volatility in a semantic context.

However, before going into detail, there remain some questions to be asked: How does one actually work with textual data?, What is the real-world equivalent of a "set of semantically related terms" or a "topic"?

1.1.1 The Vector Space Model

As already mentioned, Luhn (1958) made an early attempt to automatically identify the significant parts of a text. As said, he proposes to use word frequencies as a measure of significance to a text and, building on that, also measures the significance of a sentence. For this he compiles a dictionary that is ordered by frequency or, as he puts it,

> [i]n other words, an inventory is taken and a word list compiled in descending order of frequency. (Luhn, 1958, p. 160)

Very much along these lines, Salton and McGill (1983) propose a numerical representation of documents. In their approach, documents are transformed into a vector space in which the words of the vocabulary form unit vectors. That is, given a vocabulary of size V, each document will be described by a vector $d \in \mathbb{R}^V$ where components d_i will represent the frequency (or some weight) of the i-th word in the vocabulary in document d. Their idea came from the task of information retrieval in which a query should be answered with the most appropriate documents, i.e. the most relevant ones to the query. Converting both

[3] http://wortschatz.uni-leipzig.de

the query q and the documents $\{d \in D\}$ to vectors opens up for a mathematical treatment of this task in which any measure of similarity between vectors can be used to approximate the similarity of queries to documents. The answer then includes those documents that have the highest similarity to the query. As is obvious, the documents will generally be described by sparse vectors as it can be assumed that the number of nonzero elements in a document will be much smaller than the size of the vocabulary. Hence, they can be stored and retrieved in an efficient way. Topic modeling makes use of this representation. Words and their frequencies are considered to describe a document in a sufficient and adequate way. The assumption is that dropping the internal structure, i.e. the sequence of the words, will of course cause information loss but will not obfuscate the meaning of its content. In fact, the probabilistic model that is applied inherently assumes exchangeability of words in documents (cf. Blei and Lafferty, 2006; Finetti, 1975).

1.2 Topics, Issues and Issue-Cycles

We have above posed the question of what a topic actually is. In the notion of topic modeling, it is a distribution over the vocabulary, in which many words will exhibit low and just a few will have large probability. It emerges as a mixture component that is used (with the other topics) to build up the documents, each with a different mixture of these components. By assuming semantic relatedness among subsets of words in a document it is deduced that the mixture components emerging from the learning process also exhibit semantic relatedness. This can be be backed up both by consulting the theory of structural semantics and of course an intuitive interpretation of words with high probability in a topic by humans.

The idea of finding thematic structures in text streams that are distributed across time is, however, not new. The time-aware analysis of textual data has been an area of active research for almost 20 years now. It gained much popularity through Allan (2002)'s seminal work who tried to apply methods of classical content analysis on a large scale and in an automated manner. Their background in content analysis is reflected in how the authors define a topic. According to them, a topic is

> [...] a seminal event or activity, along with all related events and activities. [...]Allan (2002, p. 19).

We stress here, that Allan (2002)'s topics and topics as produced by a topic model are not synonymous. The definition of a distribution over words is quite straight forward: all word probabilities must be nonnegative and sum up to 1. The assumption that words in this distribution form a semantically coherent set is just an interpretation (although, one might say, a very successful one). Allan (2002)'s definition is more accessible to the human mind but is also more confined. Semantic relatedness among a set of words makes no statement about events of any kind. Consider Table 1.1, the words of highest probability in a sample topic extracted from a large corpus of classic English literature[4]. The words

[4]The Eighteenth Century Collection Online, http://quod.lib.umich.edu/e/ecco/

A "royal" topic
king
earl
lord
duke
great
england
prince
parliament
queen
married

Table 1.1: An example topic extracted from classic English texts.

appearing here are clearly semantically related to each other, they circle around England's King's Court. However, they are just as clearly not related to any particular event as such, they emerge through their document co-occurrence and through their co-occurrence with an event. Political communication science, defines some useful concepts that go beyond the event-centered interpretation of Allan (2002), they are briefly reviewed here. Kantner (2009) describes an *issue* as something that gains attention in the media but is not an event in the sense of "a particular instance of something happening". An issue is described as something broader, a "social problem", that includes related events and their relation. This definition is based on Downs (1996), who coined the notion of issues and *issue-attention cycles* describing how an issue gains and loses attention throughout the media. We feel that this interpretation comes much closer to the means of interpretability of a topic model. Moreover, the evolution of a topic in dynamic topic models certainly is based on the frequency of the terms and thus on the coverage of those terms in a given document collection at any one time. We will consider topics in the topic modeling sense extend their usual interpretation in a time-sensitive model. Using dynamic topic modeling and the idea of word volatility, we aim for a tool aiding exploratory search for events based on the interpretation that the topics we see in a document collection are amalgamations (in time) of issue cycles that are related to each other. Words with high volatility in a topic and the time at which they rapidly change might indicate certain events in an issue. Again, it has to be stressed that this (and in fact all models that, like we, apply the Bayesian paradigm) must undergo a careful inspection and interpretation by domain experts (see Gelman and Shalizi, 2012).

1.3 Data

The remarks made above have already pointed to the type of data we consider. Assuming topics series across time that emerge from a dynamic topic model analysis as unions of

data set	documents	tokens	vocabulary	∅ document length	time points
terror	30852	6163988	8691	199.8	47
riots	13378	1972496	6629	147.4	247
autumn 2001	20099	2503258	4275	124.5	91

Table 1.2: Overview over the data sets.

issue-cycles, inherently includes the supposition that there actually are any issue-cycles in the data to find. The definition of issues and issue-cycles suggests to use mass media content. The basis of our study is the New York Times Annotated corpus (Sandhaus, 2008), consisting of all articles as published by the New York Times (NYT) between 1987 and 2007 and adding up to a total of roughly 1.83 million documents. From this massive data set, different sub-corpora were compiled, a summary of which are given in Table 1.2. The methodology used is quite simple: the "terror" data set consist of all documents containing the string "terror*", i.e. containing the word "terror" and all words starting with it (e.g. "terrorist", "terrorism" etc.), published between January 1st, 2001 and December 31st, 2004; the "riots" data set was compiled by extracting all documents that were tagged by the NYT as "Demonstrations and Riots" throughout the whole data set. The "autumn 2001" data set is simply a collection of all articles as published between September 1st and November 31st, 2001 with no restrictions. All documents underwent a standard preprocessing procedure: punctuation is removed and all words are lowercased. A standard list is then used to delete stop words, additionally we delete all terms occurring less than 15 times. After compiling the sub-corpora, each of them was again sub-divided into separate training, test and validation sets. For each available date, 20% of the documents went into the validation set, the remaining 80% were divided into a training and test by partitioning each document into two parts. Again 20% of a document's content went into the test set whereas the remaining 80% of the data have been used for training.

1.4 Structure of the thesis

This study consists of three main parts. In Part I, we give a specific introduction to methods used for the extraction of semantic concepts from large amounts of text. In chapter 2, we start with a treatment of the (now) classical method of Latent Semantic Indexing (LSI) (Deerwester et al., 1990). In fact this method, although it is mainly motivated by linear algebra and not by probabilistic considerations, can be seen as the ancestor of the models we work with. We describe the main idea behind LSI and its algorithmic structure and then proceed to a probabilistic extension given by Hofmann (1999), probabilistic Latent Semantic Indexing (pLSI). The main ideas remain the same here but Hofmann (1999) gives a probabilistic interpretation of the terms contained in the computational procedure of LSI, alas still not defining a generative procedure for documents and thus not applicable to unknown data. Latent Dirichlet Allocation (Blei and Lafferty, 2006), i.e. topic model-

ing, fills this gap using a fully Bayesian treatment of the problem and we give a precise explanation of the model. Chapter 3 is concerned with approximate inference in Bayesian models. The learning problems we encounter are in general intractable problems and we must appeal to approximations. Several approaches exist for doing that. We briefly introduce sampling and in more detail a variational approach that transforms the problem of finding a proper distribution into that of optimization. Part II deals with the mathematical machinery and the methodology with which topic evolution will be modeled. We give a formal treatment of stochastic processes in general and Gaussian processes in particular in chapter 4, including, but not limited to, Brownian motion which has been used in the model we build upon (Wang et al., 2008). The evolution of random quantities through time is not a novel, not even a young, field of research. Time series modeling consists of a rich family of statistical and probabilistic methods and theories and we give an introduction to it in chapter 5. This includes some classic examples for time series and a selection of the methods usable in our context with an emphasis on those based on Gaussian processes. Chapter 6 gives the two main inference methods for learning in Gaussian process models. Although more sophisticated approaches than the here presented do exist, we decided to keep things simple and present only the analytic and the filtering approach. We demonstrate both learning approaches on the examples introduced in chapter 5. Part III forms the key contribution of this thesis. In chapter 7 we reintroduce the aforementioned basic dynamic topic model as introduced by (Wang et al., 2008). A detailed model description is given, together with a strict re-derivation of the learning procedure used in this model. Chapter 8 introduces Gaussian Process Dynamic Topic Models (GPDTM), a generalization of the above mentioned dynamic topic models, in which we include the results of our studies of stochastic processes and time series analysis. (Wang et al., 2008)'s model is one special case of the GPDTM and we give several other descendant models based on different prior considerations about the behavior of topics in time. This also includes a study with real world data sets (those described in section 1.3).

Summing up our findings and giving possible further research ideas, we conclude in chapter 9.

Part I

Topic Models

Abstract

In the first part of this work we give an (almost historic) overview over the development of approaches to identify semantically meaningful structures in large document collections. Chapter 2 introduces models that identify or can be used to identify clusters of semantically related words. They considerably differ in the approach that is taken but pursue this identical goal. We give an overview over the different techniques, all of which rely on the vector space model and the bag-of-words assumption as described in section 1.1.1. We start with an algebraically motivated approach based on singular value decomposition (Deerwester et al., 1990) and review the historic development of factor analysis and mixture models towards what has been coined as topic modeling. In a nutshell, this covers a technique based on matrix algebra, a probabilistic enhancement of this technique (Hofmann, 1999) and further evolution towards a fully Bayesian interpretation (Blei et al., 2003). Chapter 3 then gives different methodologies to approximate the latter both (singular value decomposition is an exact operation and thus does not need any approximations), including a general overview over Bayesian inference. We give example models and inference algorithms for both sampling methods and a variational approach. We also review the variational inference algorithm for Latent Dirichlet Allocation as given by Blei et al. (2003).

Chapter 2

Topic Models

2.1 Preliminaries

The management and analysis of large text corpora has long been driven by one basic assumption: in natural texts, there exist structures that we as humans would abstractly call themes or topics. In fact, we can motivate this assumption by classical structuralist views. According to de Saussure (2001), terms gain their meaning from their global context, i.e. from other words that appear together with the term in question. Topic structures are just that, sets of words that are helpful for identifying the meaning of their members. Texts are composed of those structures, i.e. each word in the text comes from one of the available themes or topics, common appearance (or co-occurrence) is given by the common document source. Using the same argument, we can also justify usage of the bag-of-words paradigm, in which we neglect the positions of words in a text and make use of the frequency of appearance alone — co-occurrences do not depend on word positions in a document. The difference between the methods we review here is the way of how they arrive at themes or topics. With the advent of modern computers and ongoing research in machine learning, more complex methods became feasible to compute and results became more useful or interpretable or both. However, the main idea of analyzing term-document counts and drawing conclusions about latent, i.e. unobserved, semantic structures that are defined as sets of words, persists.

2.1.1 Latent Semantic Indexing

Latent Semantic Indexing (LSI) is an approach that primarily aims at automatic indexing of documents to answer retrieval queries and to provide documents in the answer that are as closely related to the query as possible. The problem, however, is "(...) that users want to retrieve on the basis of conceptual content, and individual words provide unreliable evidence about the conceptual topic or meaning of a document." (Deerwester et al., 1990) I.e., the key idea was to develop a method that can classify words in the query conceptually and then answer it by providing appropriate documents that contain the desired concepts described by the query words. LSI is a factor model. Its starting point

is a large sparse matrix, the term-document matrix, say X. Its rows are the individual words in the vocabulary, i.e. all of the words that appear across the documents, columns represent the documents. Consider D to be the number of documents and W to be the size of the vocabulary. Consequently, $X \in \mathbb{N}^{W \times D}$, and X_{ij} represents the number of occurrences of term w_i in document d_j. This representation combines document vectors as described in section 1.1.1 to the document-term matrix. Recall the idea of finding concepts in documents. The authors assume an underlying semantic latent structure in the data of which we observe a noisy sample in each document. Using singular-value decomposition (SVD), the authors transform the initial term-document matrix into a space in which documents that share semantic concepts are closer to each other and which they call a "semantic space". Further, words that influence the position of a document in the semantic space only weakly are disregarded. This results in a more dense representation of the underlying semantic space and is effectively a dimensionality reduction technique. A query is then treated as a small document. The documents that are in the neighborhood of the query document in the latent semantic space are returned as the retrieval result.

Formally, SVD factors a matrix into three special matrices: $X = U\Sigma V^*$, where $U \in \mathbb{R}^{W \times W}$ and $V^* \in \mathbb{R}^{D \times D}$ are unitary matrices and $\Sigma \in \mathbb{R}^{W \times D}$ is a diagonal matrix, with elements in its diagonal the singular values of the decomposition in decreasing order. Several interpretations for the resulting matrices do exist but the most useful one for text analysis might be a geometric interpretation. Consider the documents as a cloud of points. When factorizing the term-document matrix, the singular values can be interpreted as the semi-axes of the resulting ellipsoid enclosing the point cloud. Reconstructing the original matrix, i.e. re-multiplying U, Σ and V^* after setting all but the $\hat{K} \ll \text{rank}(X)$ highest singular values in Σ to zero is identical to projecting the point cloud into a lower dimension by collapsing negligible dimensions[1]. According columns in U and rows in V^* are disregarded to form the reduced matrices $\hat{U}$ and $\hat{V}^*$. This is also called rank-reduced SVD. The authors argue, that in this way it is possible to isolate associative structures and get rid off the noise that is introduced by the inherent randomness of word usage. The number of non-zero singular values $\hat{K}$ is a choice of the modeler. Algorithm 2.1 summarizes the technique. The actual index is built from the reduced-rank SVD result. The trimmed matrices $\hat{U}$ and

Algorithm 2.1 Latent Semantic Indexing

Require: $X, \hat{K}$ ▷ the term-document matrix and the required reduced rank
$(U, \Sigma, V^*) \leftarrow \text{svd}(X)$
$\hat{\Sigma} \leftarrow \Sigma_{1:\hat{K},1:\hat{K}}$
$\hat{U} \leftarrow U_{1:W,1:\hat{K}}$
$\hat{V}^* \leftarrow V^*_{1:\hat{K},1:D}$
return $(\hat{U}, \hat{\Sigma}, \hat{V}^*)$ ▷ the rank-reduced matrix factorization

$\hat{V}^*$ describe vector spaces in which the terms and documents live, respectively. Given a

[1] This is identical to doing a Principal Component Analysis without subtracting the means.

query document Q as a term frequency column vector, i.e. $Q \in \mathbb{N}^{W \times 1}$, and rearranging the trimmed matrices, we can determine its position in document space (cf. Algorithm 2.2). To determine similarity between individual terms or documents we can use simple vector similarity measures, e.g. cosine distance (e.g. Manning and Schütze, 1999) that defines similarity as a function of the angle between two vectors. The query is then answered with documents that have a similarity with the query document above some predefined threshold. Further, the trimmed matrices can be used to cluster documents or terms w.r.t. to the underlying concepts using simple clustering techniques such as K-means. Although

Algorithm 2.2 Latent Semantic Indexing query handling

Require: $(\hat{U}, \hat{\Sigma}, \hat{V}^*), Q$ ▷ the factorization as given by LSI and the query document
$\hat{V} = X^* \hat{U} \hat{\Sigma}^{-1}$ ▷ rearrange original factorization
$X \leftarrow Q$
$\tilde{V} = Q^* \hat{U} \hat{\Sigma}^{-1}$
return $\tilde{V}$ ▷ a row vector containing Q's position in the document space

it has proven to be very useful in practical applications, LSI suffers from some serious drawbacks. One (that is shared by all of the techniques discussed here) is the problem of defining the reduced rank $\hat{K}$, i.e. the number of concept dimensions to retain from the dense original SVD. There are recommendations in the literature stating a number 50-1000 concept dimensions according to the number of documents (Landauer and Dumais, 2008), good results are also reported for $\hat{K}$ lying between 50 and 100 (Deerwester et al., 1990). Others recommend e.g. a scree test to only retain dimensions before the "elbow" or retaining the number of dimensions that give rise to a predefined amount of variance in the data (cf. Cangelosi and Goriely, 2007). Another drawback is that previously unseen words in queries are ignored, they have no impact on concept weights for the query document. This may result in empty query answers when all of the query terms are unknown to the index and, as a result, the position of the query in document space cannot be determined. A third major problem is one of interpretation. The concept weights for both documents and terms are defined to be real numbers and as such are also allowed to be negative and unbounded. This makes it impossible to compare the weight values between different sets of documents. In fact, it is impossible to find any interpretation that has a meaning beyond the actual matrix factorization. As another consequence, the reconstructed term-document matrix $\hat{X}$ may have negative entries, again lacking interpretability.

2.1.2 Probabilistic Latent Semantic Indexing

Probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999) tries to overcome some of the drawbacks of LSI by introducing a proper generative model of document construction and a sound statistical basis to the model. The main idea of factorizing the term-document count matrix into three different matrix factors essentially stays the same. However, the procedure of how to arrive at the factors considerably differs. A key component is the

introduction of the Aspect model (Hofmann et al., 1999), a statistical model that associates every observation with a latent class $z \in Z = \{z_1, \ldots, z_K\}$ (the aspects). I.e. every word in every document is assigned to one of K latent classes. The generative procedure for document construction (cf. Hofmann, 1999) is given by

1. picking a document d with probability $p(d)$,
2. selecting a latent class z with probability $p(z|d)$,
3. picking a word w with probability $p(w|z)$.

The resulting probability model defines the joint probability over the documents and words as

$$p(d, w) = p(d) \sum_{z \in Z} p(w|z)p(z|d) \tag{2.1}$$

or, using Bayes' law, identically as

$$p(d, w) = \sum_{z \in Z} p(z)p(w|z)p(d|z). \tag{2.2}$$

The similarity between LSI and pLSI arises through the rearranged joint probability model in Eq. 2.2. When rewriting in matrix notation,

$$\hat{U}_{ik} = p(d_i|z_k), \tag{2.3}$$

$$\hat{V}_{jk} = p(w_j|z_k), \tag{2.4}$$

$$\hat{\Sigma}_{kk} = p(z_k). \tag{2.5}$$

Comparing to LSI, $\hat{U}$ provides a characterization of documents in terms of the set concepts/aspects, whereas $\hat{V}$ provides a characterization of concepts in terms of words. One difference between LSI and pLSI is that in the latter, these characterizations are well-defined probability distributions over the space of concepts and the space of words respectively. Another difference is caused by the different computation of the factors. In LSI, the objective function used to optimize the factorization is the matrix L_2-norm or Frobenius norm (implicitly given using singular-value decomposition). In contrast, pLSI uses a likelihood function

$$\mathcal{L} = \sum_{d \in D} \sum_{w \in W} n(d, w) \log p(d, w), \tag{2.6}$$

with $n(d, w)$ the frequency of term w in document d, as the objective to directly optimize the predictive power of the model. An additional (arbitrary) assumption such as the implicit Gaussian noise on term frequencies that is implied by using the L_2 norm as an objective function is thus avoided. As in LSI, the authors give a geometric interpretation of the factor components used in this model. Starting from the original data where every document lives in a $(W-1)$-dimensional space, consider the K columns of $\hat{V}$ corresponding to the probabilities $p(\cdot|z_k), k \in \{1, \ldots, K\}$. Each of the K columns corresponds to

a point in the $(W-1)$-dimensional word simplex, spanning a $(K-1)$-dimensional sub-simplex. Via the joint probability model, each document's conditional distributions $p(z|d)$ can be approximated by a convex combination of the K distributions over the vocabulary. The components of the mixture distribution $p(z|d)$ thus translate into a point in the $(K-1)$-dimensional sub-simplex. Using the described objective function results in an optimal setting where the projection of the empirical word distribution $\hat{p}(w|d)$ for a document onto the $(K-1)$-dimensional sub-simplex becomes minimal in terms of Kullback-Leibler divergence. When recalling that the factor matrices are defined in terms of probability distributions, the implicit choice of Kullback-Leibler divergence as an objective is quite natural, as it can be seen as a measure (or better a divergence) for the distance between probability distributions. Hofmann (1999) describes an EM-based algorithm for optimizing the objective. We summarize their derivation in Algorithm 2.3. While pLSI does not address the problem of finding the number of latent dimensions, it introduces a joint probability model of word occurrences in documents where the mixture is a well-defined probability distribution and mixture weights can be readily interpreted in a probabilistic fashion. By using the described likelihood function as an optimization objective implicitly uses Kullback-Leibler divergence, resulting in a more realistic optimization procedure. Further, model fitness can be measured using the likelihood function. However, the generative procedure leaves it unclear how documents, latent classes and words are to be selected when constructing a document. The natural next step is to derive a fully Bayesian treatment of the given problem, i.e. to introduce prior distributions over the latent variables and to infer a posterior distribution over the parameters given the data.

Algorithm 2.3 probabilistic Latent Semantic Indexing

Require: $n(d,w) \forall d \in D, w \in W$ ▷ word frequencies
Require: K ▷ dimensionality of the latent space (the sub-simplex)
  $N \leftarrow \sum_{w,d} n(d,w)$
  while $\mathcal{L}$ **not** converged **do**
    $p(z|d,w) \leftarrow \frac{p(z)p(d|z)p(w|z)}{\sum_{z'} p(z')p(d|z')p(w|z')}$ ▷ E-step
    $p(w|z) \leftarrow \frac{\sum_d n(d,w)p(z|d,w)}{\sum_{d,w'} n(d,w')p(z|d,w')}$ ▷ M-step
    $p(d|z) \leftarrow \frac{\sum_w n(d,w)p(z|d,w)}{\sum_{d',w} n(d',w)p(z|d',w)}$
    $p(z) \leftarrow \frac{1}{N} \sum_{d,w} n(d,w)p(z|d,w)$
    $\mathcal{L} \leftarrow$ Eq. 2.6 and 2.2
  return $\hat{U}, \hat{V}$ and $\hat{\Sigma}$ according to Eq. 2.3, 2.4 and 2.5

2.2 Topic Models

Topic models (Blei et al., 2003; Steyvers and Griffiths, 2005; Blei and Lafferty, 2009) are a direct advancement of pLSI as described in the previous section and have become very popular models mostly used for tasks such as semantic clustering or text analysis.

Providing a fully Bayesian treatment, they define a family of hierarchical Bayesian models and an artificial generative process for document generation, describing how the actually observable data, the words in the documents get into their place. Their popularity has led to applications in a wide variety of settings and with different types of data, not only text. See e.g. Teh and Jordan (2009); Hoffman et al. (2009) for examples using genetics and music data. For models considering additional data available for documents such as authorship, email recipients or numerical target variables, e.g. movie ratings see Rosen-Zvi et al. (2005); McCallum et al. (2004); Blei and McAuliffe (2008) respectively.

In a simple topic model, document generation is controlled by two latent factors. The topics themselves and the documents' topic proportions. A topic is defined as a probability distribution over the word simplex, i.e. in every topic each word has a certain probability and the probabilities in each individual topic sum to 1. The set of words with highest probability is assumed to describe the individual topics thematically. The second factor, the document topic proportions, is again a set of probability distributions (one for each document), defined over the topic simplex. Every topic gets some probability in a document and the probabilities of topics for a single document sum to 1. Both the topics' distributions over the vocabulary and the documents' distributions over the topics correspond to some extent to $p(w|z)$ and $p(d|z)$ as defined in Eq. 2.1 in the pLSI model. However, being fully Bayesian models, topic models define prior distributions both for the topics and the documents' distributions over topics. Also, the models are statistically motivated. Words, documents and semantic classes are treated as random variables and the most probable setting of the topics is found by using statistical inference techniques. This gives rise to a so called admixture model, in which a document is modeled as a mixture of mixtures (i.e. topics). Another difference between topic models and pLSI is the method of determining parameter settings in the model. pLSI optimizes a likelihood function in terms of the Kullback-Leibler divergence between the empirical word distribution of a document and its approximation given by the multiplication of the appropriate factors given by Eq. 2.3, 2.4 and 2.5. In contrast, topic models define a posterior distribution over the latent variables given the data and try to find an approximation to this true posterior (cf. chapter 3).

2.2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) directly makes use of the generative process that has been introduced in the pLSI model and employs a fully Bayesian treatment by placing prior probability distributions on all latent variables. As before, let $\mathcal{D} = \{d_1, \ldots, d_D\}$ be the set of documents and $\mathcal{W} = \{w_1, \ldots, w_W\}$ the vocabulary. Further, define $\theta_d, d = 1, \ldots, D$ to be the document specific distribution over topics and $\beta_k, k = 1, \ldots, K$ to be the topics, i.e. distributions over the vocabulary. Let θ_{ij} be topic j's probability in document d_i and β_{kn} be word w_n's probability in topic k. Both distributions are multinomial distributions, i.e. distributions over a discrete set (words in the vocabulary and the set of topics respectively). The conjugate prior distribution to the multinomial is the Dirichlet distribution (cf. Kotz et al., 2000). Its governing parameter is called the hyperparameter when the distribution is used as a prior. In the LDA model, symmetric

Dirichlet distributions are placed as priors over each θ_d with hyperparameter α and over each β_k with hyperparameter η. Note that both priors are defined to be distributions on the appropriate simplex, i.e. the prior over each θ_d is a distribution on the topic simplex and the prior over each β_k consequently a distribution on the word simplex. This means that every draw from one of the priors will be a multinomial distribution (a point in the according simplex) as desired. LDA's generative process is given by

1. for all topics k: $\beta_k \sim \mathrm{Dir}_W(\eta)$
2. for all documents d
   (a) $\theta_d \sim \mathrm{Dir}_K(\alpha)$
   (b) for $n = 1, \ldots, N_d$
      i. draw a topic $z_{dn} \sim \mathrm{Mult}(\theta_d)$
      ii. draw a word $w_{dn} \sim \mathrm{Mult}(\beta_{z_{dn}})$

with N_d the length of document d. This gives rise to joint probability model

$$p(\beta, \theta, z, w|\alpha, \eta) = \prod_{k=1}^{K} p(\beta_k|\eta) \prod_{d=1}^{D} \left\{ p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta_{1:K}) \right\}. \tag{2.7}$$

The conditioning of random variable on each other as in Eq. 2.7 can also be visualized graphically using Bayes nets, a special form of probabilistic graphical models. Figure 2.1 shows the LDA model in the so called plate notation. Nodes represent random variables, shaded nodes represent observations and arrows denote conditional dependency between variables. The enclosing rectangles are called plates and symbolize repetition. Note the direct parallelism between Eq. 2.7 and Fig. 2.1.

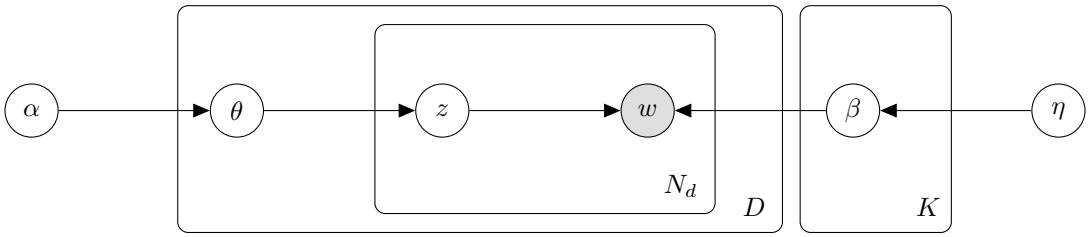


Figure 2.1: LDA model in plate notation.

The benefit from placing prior distributions on the latent variables is threefold. First, introducing prior probabilities on the latent variables provides a more reliable statistic model, prior belief and domain knowledge (or the lack thereof) can be encoded into the prior distribution. This involves the second benefit, direct applicability to unknown data. Hofmann (1999) describes the folding-in method of application of pLSI to unknown documents (queries). However, LDA can be directly applied to new documents to infer the

document specific parameters (i.e. θ_d and $z_{d\cdot}$) given the previously learned model. Third, using the Bayesian paradigm opens up for well-established techniques for statistical inference and model selection. The LDA model serves as a building block for more complex models that incorporate additional data available, see e.g. Dietz et al. (2007); Rosen-Zvi et al. (2005); Blei and Lafferty (2007); Wang and McCallum (2006); Wang et al. (2008).

Constraints

While the LDA model has proven to be extremely helpful in the unsupervised analysis of document collections, it does make assumptions that may be problematic in some settings. As well as the other models reviewed in this chapter, LDA makes use of the bag-of-words paradigm. Essentially, this translates into an assumption of exchangeability from a probability theoretic point of view. Exchangeability, as coined by Finetti (1975), is the assumption that if the joint probability of a collection of random variables is invariant under permutation of the random variables, then this collection can be represented as a mixture distribution (that is generally infinite). This also means that given the parameter of latent shared mixture distribution, the random variables in the collection are conditionally independent of each other. The general applicability to text data now depends on the setting and the type of data. When there is no prior knowledge about the ordering of documents, exchangeability surely is an appropriate assumption, an implicit uninformative prior on the ordering. However, when it comes to document collections that provide meta data such as timestamps, this assumption becomes invalid. Consider a document collection comprising news articles. For a collection of articles that originate from the same day or even month, exchangeability might still be assumed, although sudden events, e.g. natural disasters, terror attacks, elections, may influence this decision. For articles that span whole years, decades or even larger time spans, this assumption is no longer valid. While exchangeability among the words of single documents is akin to the task of finding semantic structures in documents, in document collections the structure to find (the topics) evolves over time. Simultaneously, exchangeability no longer holds when the latent structure across documents evolves and an implicit (time) ordering exists for documents. As time ordered collections of documents and their analysis is the main concern of this theses, we review a specific topic model that captures topics as random variables that are subject to change governed by a stochastic process in chapter 7. A second property that comes by design is the inherent independence of individual topics in documents' topic distributions and of individual words in topics' word distributions. Especially when it comes to topic proportions in documents, intuition forbids the independence of individual topics in that document. If there is a dominant "sports" topic in a document, the probability that other topics that are thematically related (e.g. politics, public relations, events etc.) are present in the document is much higher than that for completely unrelated topics (e.g. military, traffic, communication etc.,). One solution to this problem is to use a prior distribution on the documents' topics distributions that is able to capture this type of correlation. Blei and Lafferty (2007) use a logistic normal distribution, a discrete distribution on the topic simplex. Based on the multivariate normal distribution, it is parameterized by a mean

vector and a covariance matrix, both of which are learned during training. Inspecting the resulting covariance matrix after training reveals positive or negative correlations between topics in a document.

Chapter 3

Approximate inference in Bayesian models

Statistical inference in probabilistic graphical models seeks to find the posterior distribution over unknown variables in the model, given the data. Usually, this posterior distribution is very complicated and hard to compute. Further, in most cases the complexity of its computation increases exponentially with the number of data points used to learn it.

3.1 Foundations

We start with laying the foundations of the procedures used for statistical inference. In particular, we analyze the analytic solution for the posterior distribution and show why this solution becomes intractable for data sets of useful size. We further briefly introduce different techniques for approximating the posterior distribution. In principal these are either sampling based approaches or deterministic approximations. We will give the basic idea of the introduced techniques before concentrating on one specific approach, Variational Bayes, in the next section.

3.1.1 The Model Posterior

Following Hoffman et al. (2013), we will use a running example in our derivations. Consider a simple mixture model as shown in Fig. 3.1. From this, we can read off the joint probability distribution of the model

$$p(\beta, z_{1:n}, x_{1:n}) = p(\beta) \prod_{i=1}^{n} p(z_n|\beta) p(x_n|z_n, \beta) \tag{3.1}$$

Note that we consider potential fixed parameters as part of the latent variables and omit where possible. Further the latent variable β represents the set of possible mixture components and for every observable data point x_i the latent variable $z_i \in \{0,1\}^K, \sum_k z_{ik} = 1$ determines the component from which x_i is drawn, i.e., each z_i is a K-dimensional indicator

vector. Where appropriate, we will suppress the mixture component indices as in Eq. 3.1 although the full joint probability model is given as

$$p(\beta, z_{1:n}, x_{1:n}) = \prod_{k=1}^{K} p(\beta_k) \prod_{i=1}^{n} p(z_n|\beta_{1:K}) p(x_n|z_n, \beta_{1:K}). \tag{3.2}$$

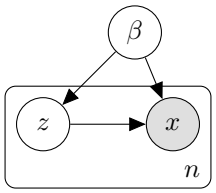


Figure 3.1: A simple graphical model.

Given prior probabilities on the latent variables β and $z_{1:n}$, the goal of statistical inference in Bayesian models is to find the posterior distribution over the latent variables given the data $p(\beta, z_{1:n}|x_{1:n})$. Using the chain and sum rule in probability theory (see e.g. Bishop, 2006), we can rewrite Eq. 3.1 to

$$p(\beta, z_{1:n}, x_{1:n}) = p(\beta, z_{1:n}|x_{1:n}) p(x_{1:n}) \tag{3.3}$$

and

$$p(x_{1:n}) = \int_{\beta} \sum_{z_{1:n}} p(\beta, z_{1:n}, x_{1:n}). \tag{3.4}$$

Combining Eq. 3.3 and 3.4, we arrive at the desired posterior probability

$$p(\beta, z_{1:n}|x_{1:n}) = \frac{p(\beta, z_{1:n}, x_{1:n})}{\int_{\beta} \sum_{z_{1:n}} p(\beta, z_{1:n}, x_{1:n})}. \tag{3.5}$$

Note that when expanding the numerator in Eq. 3.5 as in Eq. 3.3 and collapsing the integral in the denominator as in Eq. 3.4, we can re-derive Bayes' law. While the numerator is easy to compute, the problem comes with the integral in the denominator. Trying to analytically derive a solution we make use of Eq. 3.2 and rewrite Eq. 3.4 to

$$\begin{aligned} p(x_{1:n}) &= \int_{\beta} \sum_{z_{1:n}} p(\beta, z_{1:n}, x_{1:n}) \\ &= \int_{\beta_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\beta_k) \prod_{i=1}^{n} p(z_n|\beta_{1:K}) p(x_n|z_n, \beta_{1:K}) \\ &= \int_{\beta_{1:K}} \prod_{k=1}^{K} p(\beta_k) \prod_{i=1}^{n} \underbrace{\sum_{z_i} p(z_n|\beta_{1:K}) p(x_n|z_n, \beta_{1:K})}_{\text{inner sum}} \end{aligned} \tag{3.6}$$

where we have used the fact that given β, the individual z_is are conditionally independent from one another. While the inner sum is often computable, there are K^n terms to compute. When the number of data points grows to a reasonable size this obviously prohibits exact calculation and we must resort to an approximate solution. This is in fact the main obstacle arising in a vast majority of models that are complex enough to provide interesting insights.

3.1.2 Example model

To tackle this problem, we will consider two different techniques for approximating an intractable integral: sampling and deterministic approximations. For the demonstration of the different approaches we have to further describe our toy model given in Eq. 3.2. Let the model be a mixture model of Gaussians with known variance. The generative model is then

1. $\beta_k \sim \mathcal{N}(\beta_0, \sigma_0^2)$ for $k = 1, \ldots, K$
2. for $i = 1, \ldots, n$
   (a) $z_i \sim \text{Mult}(\pi)$
   (b) $x_i \sim \mathcal{N}(\beta_{z_i}, \sigma^2)$

with $\{\beta_0, \sigma_0^2\}$ the parameters of the prior over the mixture components, π the parameter to a multinomial distribution, i.e. $\pi_i > 0, 1 \leq i \leq K$ and $\sum_{i=1}^{K} \pi_i = 1$, governing which component is chosen and σ^2 the fixed component variance. A slightly extended version of Fig. 3.1 adapted to the mixture of Gaussians is given in Fig. 3.2. Note that we have given all fixed parameters as well for clarity. The joint probability model is given by

$$p(\beta, z_{1:n}, x_{1:n}) = \prod_{k=1}^{K} p(\beta_k|\beta_0, \sigma_0^2) \prod_{i=1}^{n} p(z_i|\pi) p(x_i|z_i, \beta_{1:K}, \sigma^2). \tag{3.7}$$

We note that, if x is a normally distributed (observable) random variable with known

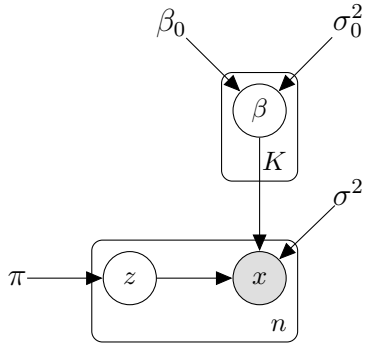


Figure 3.2: The mixture of Gaussians model.

variance, placing a normal prior on its mean (one of the possible β_k) results in a fully conjugate model and all complete conditionals are analytically defined. We give the complete conditionals for both β and $\mathbf{z}$.

Complete conditionals

$$\begin{aligned}
p(\beta_k|\cdot) &\propto p(\beta, z_{1:n}, x_{1:n}) \\
&\propto \prod_{k=1}^{K} p(\beta_k|\beta_0, \sigma_0^2) \prod_{i=1}^{n} p(z_i|\pi) p(x_i|z_i, \beta_{1:K}) \\
&\propto p(\beta_k|\beta_0, \sigma_0^2) \prod_{i=1}^{n} p(x_i|z_i, \beta_{1:K}) \\
&\propto \exp\left\{-\frac{(\beta_k - \beta_0)^2}{2\sigma_0^2}\right\} \prod_{i=1}^{n} \left(\exp\left\{-\frac{(x_i - \beta_k)^2}{2\sigma^2}\right\}\right)^{z_{ik}} \\
&\propto \exp\left\{-\frac{1}{2\sigma_0^2}\beta_k^2\right\} \exp\left\{\frac{\beta_0}{\sigma_0^2}\beta_k\right\} \prod_{i=1}^{n} \exp\left\{-\frac{z_{ik}}{2\sigma^2}\beta_k^2\right\} \exp\left\{\frac{z_{ik}x_i}{\sigma^2}\beta_k\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2} + \frac{\sum_{i=1}^{n} z_{ik}}{\sigma^2}\right)\beta_k^2\right\} \exp\left\{\left(\frac{\beta_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} z_{ik}x_i}{\sigma^2}\right)\beta_k\right\} \qquad (3.8)
\end{aligned}$$

Given that Eq. 3.8 has the form of a normal distribution in canonical form, the full conditional for each β_k is a Gaussian distribution, i.e. $\beta_k|\cdot \sim \mathcal{N}(\beta_k|m, s^2)$ as required by conjugacy with

$$m = \left(\frac{\beta_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} z_{ik}x_i}{\sigma^2}\right)\left(\frac{1}{\sigma_0^2} + \frac{\sum_{i=1}^{n} z_{ik}}{\sigma^2}\right)^{-1} \qquad (3.9)$$

$$s^2 = \left(\frac{1}{\sigma_0^2} + \frac{\sum_{i=1}^{n} z_{ik}}{\sigma^2}\right)^{-1} \qquad (3.10)$$

the mean and variance of the full conditional. Further

$$\begin{aligned}
p(z_i|\cdot) &\propto p(\beta, z_{1:n}, x_{1:n}) \\
&\propto p(z_i|\pi) p(x_i|z_i, \beta_{1:K}) \\
&\propto \prod_{k=1}^{K} \left(\pi_k \exp\left\{-\frac{(x_i - \beta_k)^2}{\sigma^2}\right\}\right)^{z_{ik}} \qquad (3.11) \\
& \qquad (3.12)
\end{aligned}$$

and thus the full conditional for each z_i is a multinomial distribution $z_i|\cdot \sim \text{Mult}(\eta)$ with

$$\eta = \frac{1}{C}\left(\pi \circ \exp\left\{-\frac{(x_i - \beta)^2}{2\sigma^2}\right\}\right) \qquad (3.13)$$

and C the normalization constant, ensuring that $\sum_k \eta_k = 1$ and $\circ : \mathbf{R}^d \times \mathbf{R}^d \to \mathbf{R}^d$ the point-wise multiplication operator, such that for any $a, b \in \mathbf{R}^d$, $a \circ b = (a_1 b_1 \ \ldots \ a_d b_d)^T \in \mathbf{R}^d$.

3.2 Sampling

We briefly review Markov Chain Monte Carlo (MCMC) as a representative of sampling techniques before giving a comprehensive treatment of Variational Bayes or Variational Inference, the latter of which will be used to derive inference algorithms for the models used in Part III.

For the sake of completeness and because a whole range of existing topic models use this technique for statistical inference, we briefly introduce MCMC sampling methods, especially Gibbs sampling. MCMC is an abstract class of algorithms used in statistics and statistical physics to sample from a probability distribution. It is based on the Monte Carlo technique for integration that can be used to compute complex integrals, such as expectations of a random variable under some distribution, via sampling and averaging. Following Gilks et al. (1995)'s general treatment, consider a given posterior distribution $\pi(\cdot)$ and some function $f(\cdot)$ whose expectation we are interested in. Let $X = \{x_1, \ldots, x_m\}$ be a vector of m random variables with distribution $\pi(\cdot)$. Note that naturally[1] all latent variables are encoded in X, i.e. in our example model,

$$X = \{\beta_1, \ldots, \beta_K, z_1, \ldots, z_n\} \tag{3.14}$$

and that $f(\cdot)$ also can be the identity function. If so, the interest lies in the posterior assignment of the latent variables. The expectation

$$\mathbb{E}_{\pi(x)}\left[f(X)\right] = \int f(x)\pi(x)dx \tag{3.15}$$

can be approximated using Monte Carlo integration

$$\mathbb{E}_{\pi(x)}\left[f(X)\right] \approx \frac{1}{n}\sum_{t=1}^{n} f(X_t) \tag{3.16}$$

with $\{X_t, t = 1, \ldots, n\}$ being samples from distribution $\pi(\cdot)$. If the samples $\{X_t\}$ are independent, the law of large numbers applies and the approximation is refined with increasing sample size n. Samples drawn from $\pi(\cdot)$ are in general *not* independent. Fortunately, this requirement can be relaxed. Samples just need to be drawn from $\pi(\cdot)$ with correct proportions throughout its support. As Gilks et al. (1995) states, this can be assured by using a Markov chain whose stationary (equilibrium) distribution is $\pi(\cdot)$. Suppose we draw a sequence of random variables $\{X_1, X_2, \ldots\}$ using a Markov chain. Then for every $t \leq 0$, the next random variable is only dependent on the current state of the chain, i.e.

[1] Recall that the posterior is a distribution over the latent variables given the observable ones.

$p(X_{t+1}|X_{1:t}) = p(X_{t+1}|X_t)$ and X_{t+1} is sampled from $p(X_{t+1}|X_t)$. The crucial requirement now is to have X_t being independent of X_0. If this is the case, the transition kernel $p(\cdot|\cdot)$ becomes invariant and samples from it are samples from the equilibrium distribution of the chain. While this state is theoretically reached when the number of samples n reaches infinity, for practical applications invariance in defined limits is sufficient and is already reached after a manageable number of iterations, the burn-in. When we construct the Markov chain such that its equilibrium distribution is the posterior, we are able to generate samples from this posterior after the burn-in period. Given these samples, we can finally compute any expectation under this distribution using the Monte Carlo technique. This is Markov Chain Monte Carlo for statistical inference in Bayesian models.

3.2.1 Metropolis-Hastings algorithm

The way of how to construct the Markov chain is the difficult part of algorithms that implement this technique. Using the Metropolis-Hastings algorithm, we define a proposal distribution $q(\cdot|X_t)$, i.e. a distribution over the next state given the current one. From this proposal distribution a proposed next state is drawn and then either accepted or rejected according to the following scheme. Let Y be the proposed next state, drawn from $q(\cdot|X_t)$. It is accepted as the next state X_{t+1} with probability

$$\alpha(X_t, Y) = \min\left(1, \frac{\pi(Y)q(X_t|Y)}{\pi(X_t)q(Y|X_t)}\right), \tag{3.17}$$

otherwise the current state persists. The complete algorithm is given in Alg. 3.1.

Algorithm 3.1 Metropolis-Hastings algorithm

Require: $q(\cdot|X_t), X_0, m$ ▷ proposal distribution, initial state, burn-in period
  $t \leftarrow 0$
  repeat
    $Y \sim q(\cdot|X_t)$
    $U \sim \text{Uniform}(0, 1)$
    if $U \leq \alpha(X_t, Y)$ **then**
      $X_{t+1} \leftarrow Y$
    else
      $X_{t+1} \leftarrow X_t$
    $t \leftarrow t + 1$
  until desired number of samples reached
  return $\{X_m, \ldots, X_t\}$ ▷ samples from the stationary distribution $\pi(\cdot)$

Interestingly, any choice of proposal distribution leads to samples from the stationary distribution $\pi(\cdot)$ although the speed of convergence and mixing differs, see again Gilks et al. (1995) for further details on this. The Metropolis-Hastings algorithm is an instance of Random Walk Monte Carlo methods.

3.2.2 Gibbs sampling

A special form of the Metropolis-Hastings algorithm does not update one whole X_t per iteration, but breaks X_t into h components $\{X_{t1}, \ldots, X_{th}\}$. In case of our example, the components typically are the individual latent variables. Components are then updated one by one. Candidate Y_{ti} for state $X_{t+1,i}$ will be dependent on the current setting of all components in X_t, i.e. X_{ti} and $X_{t,-i} = \{X_{t1}, \ldots, X_{t,i-1}, X_{t,i+1}, \ldots, X_{th}\}$. Y_{ti} is thus drawn from $q(Y_{ti}|X_{ti}, X_{t,-i})$. Considering our example again, this translates into iteratively resampling each latent variable, i.e. each of the components of X_t as given in Eq. 3.14. Metropolis et al. (1953) originally introduced this method called Single-Component Metropolis Hastings. When defining $q(Y_{ti}|X_{ti}, X_{t,-i}) = \pi(Y_{ti}|X_{t,-i})$, i.e. letting the proposal distribution be a full conditional, the acceptance probability (Eq. 3.17) of the sample will always be one. This is called Gibbs sampling and also stems from statistical physics where it is called the heat bath algorithm. The motivation is based on the structure of the model. Being fully conjugate, each of the full conditionals does exist in closed form and thus the posterior can be defined in terms of full conditionals. Gibbs sampler proposals are always accepted and finding samples from the equilibrium distribution purely consists of alternately sampling from these full conditional distributions.

3.2.3 Example

Considering our toy model in Fig. 3.2 and given its complete conditionals in section 3.1.2, we can derive the Gibbs sampler for posterior inference. Starting from a random initialization of the latent variables, we resample each z_i-vector according to the multinomial with parameter given by Eq. 3.13 and each β_k from a normal with parameters given by Eq. 3.9 and 3.10 in each iteration. After the burn-in period, we can treat samples obtained in that manner as samples from the posterior and perform the desired actions. The full procedure is given in Alg. 3.2. A Gibbs sampler for approximating the posterior distribution in the LDA model is given by Griffiths and Steyvers (2004).

3.3 Variational Bayes

Compared to the discussed sampling techniques, Variational Bayes (VB) or Variational Inference (VI) is a deterministic approach to approximate the model posterior. The main idea is to define a family of distributions $q(\cdot)$, governed by free *variational* parameters, serving as a proxy distribution to the true posterior. The parameters are then varied to find the member of that family that is closest to the true posterior in terms of Kullback-Leibler divergence. In consequence, the problem of determining the posterior distribution has been reshaped to an optimization problem, namely that of minimizing the KL divergence between the proxy and the true posterior distribution. VI can be much faster than sampling based approaches, producing similar results. Again, consider a posterior distribution $\pi(\cdot)$ that is to be approximated. As before, the posterior is a distribution over the latent variables Θ given the data (i.e. the observable variables) $\mathcal{D}$), i.e. $\pi(\cdot) := p(\Theta|\mathcal{D})$. We

define $q(\cdot)$ to be a distribution over Θ governed by some parameters ν, i.e. $q(\cdot) := q(\Theta|\nu)$. In case of our running example, $\Theta = \{\beta, z_{1:n}\}$. Further, let $\nu = \{\lambda, \phi_{1:n}\}$, defining the proxy $q(\beta, z_{1:n}|\lambda, \phi_{1:n})$. The task in VI is to find the setting of variational parameters that minimizes the KL-divergence of p from q, i.e.

$$\min_{\nu} \mathrm{KL}\left(q(\Theta|\nu)||p(\Theta|\mathcal{D})\right) \tag{3.18}$$

Clearly, we cannot optimize this quantity directly, as we have no means of computing $p(\Theta|\mathcal{D})$. Recall that the problem in computing the true posterior distribution lies in the intractable integral given in Eq. 3.4. In the following we further analyze this term, first by taking the logarithm of both sides and secondly making use of the surrogate distribution q.

$$\begin{aligned}
\log p(\mathcal{D}) &= \log \int p(\Theta, \mathcal{D}) d\Theta \\
&= \log \int p(\Theta, \mathcal{D}) \frac{q(\Theta)}{q(\Theta)} d\Theta \\
&= \log \int q(\Theta) \frac{p(\Theta, \mathcal{D})}{q(\Theta)} d\Theta \\
&= \log \left(\mathbb{E}_{q(\Theta)} \left[\frac{p(\Theta, \mathcal{D})}{q(\Theta)} \right] \right) \\
&\geq \mathbb{E}_{q(\Theta)} \left[\log \left(\frac{p(\Theta, \mathcal{D})}{q(\Theta)} \right) \right] \\
&\geq \mathbb{E}_{q(\Theta)} \left[\log \left(p(\Theta, \mathcal{D}) \right) \right] - \mathbb{E}_{q(\Theta)} \left[\log \left(q(\Theta) \right) \right] \triangleq \mathcal{L}(q)
\end{aligned} \tag{3.19}$$

The inequality is due to Jensen (cf. eg. Perlman, 1974) and is justified by the logarithm being a strictly concave function. The given quantity $\mathcal{L}(q)$ is a lower bound on the log marginal of the data. Because this marginal is also called "the evidence", Eq. 3.19 is often referred to as the Evidence Lower Bound (ELBO). Through the expectations with respect to $q(\Theta)$, the lower bound is a function of the proxy distribution q. By further inspection of the KL-divergence of p from q, we can show that maximizing the ELBO is in fact identical to solving the problem posed in Eq.3.18.

Proof. Consider the KL-divergence of p from q. Then

$$\begin{aligned}
\mathrm{KL}\left(q(\Theta|\nu)||p(\Theta|\mathcal{D})\right) &= \int q(\Theta|\nu) \log \left(\frac{q(\Theta|\nu)}{p(\Theta|\mathcal{D})} \right) \\
&= \mathbb{E}_{q(\Theta|\nu)} \left[\log \left(\frac{q(\Theta|\nu)}{p(\Theta|\mathcal{D})} \right) \right] \\
&= \mathbb{E}_{q(\Theta|\nu)} \left[\log \left(q(\Theta|\nu) \right) \right] - \mathbb{E}_{q(\Theta|\nu)} \left[\log \left(p(\Theta|\mathcal{D}) \right) \right] \\
&= \mathbb{E}_{q(\Theta|\nu)} \left[\log \left(q(\Theta|\nu) \right) \right] - \mathbb{E}_{q(\Theta|\nu)} \left[\log \left(p(\Theta, \mathcal{D}) \right) \right] + \log \left(p(\mathcal{D}) \right) \\
&= -(\mathbb{E}_{q(\Theta|\nu)} \left[\log \left(p(\Theta, \mathcal{D}) \right) \right] - \mathbb{E}_{q(\Theta|\nu)} \left[\log \left(q(\Theta|\nu) \right) \right]) + \log \left(p(\mathcal{D}) \right)
\end{aligned} \tag{3.20}$$

Comparing Eq. 3.19 and 3.20, we see that the KL-divergence is equal to the negative ELBO up to a constant not depending on $q(\Theta)$. Maximizing the ELBO is thus identical to minimizing the KL-divergence of p from q. □

3.3.1 Mean-field Assumption

Given its roots in statistical physics, we adopt a crucial assumption about the structure of the approximating distribution $q(\Theta)$. Recall that $q(\Theta)$ is a distribution over the set of all latent variables Θ. Defining (arbitrary for the moment) subsets of Θ such that $\Theta = \{\theta_1, \ldots, \theta_J\}$, the mean field assumption establishes a factorization of $q(\Theta)$ with all subsets of Θ mutually independent of one another, i.e.

$$q(\Theta) = \prod_{j=1}^{J} q_j(\theta_j). \tag{3.21}$$

The idea is to substitute the dependency on other distributions for each of the subsets with a local-only dependency, rendering all of the subsets conditionally independent. The local dependency is governed by a variational parameter whose state depends on all other subsets but the one it parameterizes. This dependency constitutes a "field" in which the original dependency on the other subset distributions is subsumed. See e.g. Chandler (1987); Baxter (2013) for a more extensive treatment of the mean-field principle. For VI, we employ the most simple factorization possible – each of the latent variables is rendered independent of all others and is equipped with its own variational parameter, i.e. each of the J subsets of Θ corresponds to one latent variable and the J distributions $q_j(\theta_j)$ each are in the same family as the *full conditionals* based on the joint probability model. We point out the methodological similarity to Single-component Metropolis Hastings and the Gibbs sampler: there, each of the proposals for the next state in the Markov chain depends on the data and all other components of the set of latent variables. In VI, the set of other components constitutes the mean field that influences the state of the variational parameters used to approximate the true posterior.

3.3.2 General treatment

In mean-field variational inference, the optimization objective (Eq. 3.19) typically is optimized by coordinate ascend in fully conjugate models (cf. Wainwright and Jordan (2007)). This procedure alternately updates each variational parameter while all other parameters are held fixed. In a fully conjugate model in which all distributions are in the exponential family of distributions, these updates are given in closed form. In each iteration we seek

an update for each ν_j, i.e. we optimize $\mathcal{L}(\nu_j)$. Rewriting Eq. 3.19 we arrive at

$$\begin{aligned}\mathcal{L}(\nu_j) &= \mathbb{E}_{q(\Theta|\nu)}\left[\log\left(p(\Theta,\mathcal{D})\right)\right] - \mathbb{E}_{q(\Theta|\nu)}\left[\log\left(q(\Theta|\nu)\right)\right] \\ &= \mathbb{E}_{q(\Theta|\nu)}\left[\log\left(p(\theta_j|\theta_{-j},\mathcal{D})\right) + \log\left(p(\theta_{-j},\mathcal{D})\right)\right] \\ &\quad - \mathbb{E}_{q(\Theta|\nu)}\left[\log\left(q(\theta_j|\nu_j)\right) + \log\left(q(\theta_{-j}|\nu_{-j})\right)\right] \\ &= \mathbb{E}_{q(\Theta|\nu)}\left[\log(p(\theta_j|\theta_{-j},\mathcal{D}))\right] - \mathbb{E}_{q(\Theta|\nu)}\left[\log\left(q(\theta_j|\nu_j)\right)\right] + \text{const.} \end{aligned} \tag{3.22}$$

where we made use of the chain rule and the fact that both $\log\left(p(\theta_{-j},\mathcal{D})\right)$ and $\log\left(q(\theta_{-j}|\nu_{-j})\right)$ are constant w.r.t. ν_j. Adopting the canonical form for exponential family distributions, the general full conditional of latent variable θ_j is given by

$$p(\theta_j|\theta_{-j},\mathcal{D}) = h(\theta_j)\exp\left\{\eta_{\theta_j}(\theta_{-j},\mathcal{D})^T t(\theta_j) - a(\eta_{\theta_j}(\theta_{-j},\mathcal{D}))\right\}. \tag{3.23}$$

The canonical form for the relevant part in the approximating distribution is defined to be in the same family of distributions and thus is given by

$$q(\theta_j|\nu_j) = h(\theta_j)\exp\left\{\nu_j^T t(\theta_j) - a(\nu_j)\right\}. \tag{3.24}$$

Plugging Eq. 3.23 and 3.24 into Eq. 3.22, we obtain

$$\begin{aligned}\mathcal{L}(\nu_j) &= \mathbb{E}_{q(\Theta|\nu)}\left[\log(h(\theta_j)) + \eta_{\theta_j}(\theta_{-j},\mathcal{D})^T t(\theta_j) - a(\eta_{\theta_j}(\theta_{-j},\mathcal{D}))\right] \\ &\quad - \mathbb{E}_{q(\Theta|\nu)}\left[\log\left(h(\theta_j)\right) + \nu_j^T t(\theta_j) - a(\nu_j)\right] + \text{const.} \\ &= \mathbb{E}_{q(\Theta|\nu)}\left[\eta_{\theta_j}(\theta_{-j},\mathcal{D})^T t(\theta_j)\right] - \mathbb{E}_{q(\Theta|\nu)}\left[\nu_j^T t(\theta_j) - a(\nu_j)\right] + \text{const.} \\ &= \mathbb{E}_{q(\Theta|\nu)}\left[\eta_{\theta_j}(\theta_{-j},\mathcal{D})\right]^T \nabla_{\nu_j} a(\nu_j) - \nu_j^T \nabla_{\nu_j} a(\nu_j) - a(\nu_j) + \text{const.} \end{aligned} \tag{3.25}$$

the optimization objective as a function of one variational parameter. The log normalizer $a(\eta_{\theta_j}(\theta_{-j},\mathcal{D}))$ as well es the base measures $h(\cdot)$ have been absorbed into the constant and we made use of the identity $\mathbf{E}[t(x)] = \nabla_\eta a(\eta)$ in exponential family distributions (cf. Bernardo and Smith, 2009), i.e. the expectation of the sufficient statistics under a distribution with natural parameters η is equal to the derivative of its log partition function $a(\cdot)$. Taking the gradient (w.r.t. to the variational parameter to update)

$$\nabla_{\nu_j}\mathcal{L}(\nu_j) = (\mathbb{E}_{q(\Theta|\nu)}\left[\eta_{\theta_j}(\theta_{-j},\mathcal{D})\right] - \nu_j)^T \nabla^2_{\nu_j} a(\nu_j) \tag{3.26}$$

and setting to zero yields the update for η_j

$$\nu_j^* = \mathbb{E}_{q(\Theta|\nu)}\left[\eta_{\theta_j}(\theta_{-j},\mathcal{D})\right], \tag{3.27}$$

i.e. in fully conjugate models where all distributions are in the exponential family, updates for the variational parameters of the approximating distribution q are given by the expectations of the natural parameters of the corresponding full conditional as derived in section 3.1.2 under q.

3.3.3 Example

Consider our toy model in Eq. 3.7. The approximating distribution $q(\Theta)$ is factorized according to the mean-field assumption and is given by

$$q(\Theta) = q(\beta, z_{1:n}) := \prod_{k=1}^{K} q(\beta_k|\lambda_k) \prod_{i=1}^{n} q(z_n|\phi_n), \tag{3.28}$$

where each distribution $q(\beta_k|\lambda_k)$ and $p(z_i|\phi_i)$ is placed in the same exponential family as the corresponding complete conditional such that

$$q(\beta_k|\lambda_k) = h(\beta_k) \exp\left\{\lambda_k^T t(\beta_k) - a_{\beta_k}(\lambda_k)\right\} \tag{3.29}$$

and

$$q(z_i|\phi_i) = h(z_i) \exp\left\{\phi_i^T t(z_i) - a_{z_i}(\phi_i)\right\}. \tag{3.30}$$

Generally, β_k's full conditional in canonical exponential family form is given by

$$p(\beta_k|\cdot) = h(\beta_k) \exp\left\{\eta_{\beta_k}(z_{1:n}, x_{1:n})^T t(\beta_k) - a_{\beta_k}(\eta_{\beta_k}(z_{1:n}, x_{1:n}))\right\}. \tag{3.31}$$

Through inspection of Eq. 3.8 and given that for the normal distribution, $t(\beta) = [\beta \quad \beta^2]^T$, we find that

$$\eta_{\beta_k}(z_{1:n}, x_{1:n}) = \left\langle \frac{\beta_0}{\sigma_0^2} + \frac{\sum_{i=1}^n z_{ik} x_i}{\sigma^2}, -\frac{1}{2}\left(\frac{1}{\sigma_0^2} + \frac{\sum_{i=1}^n z_{ik}}{\sigma^2}\right)\right\rangle. \tag{3.32}$$

Equivalently, from z_i's full conditional in canonical form

$$p(z_i|\cdot) = h(z_i) \exp\left\{\eta_{z_i}(\beta, x_i)^T t(z_i) - a_{z_i}(\eta_{z_i}(\beta, x_i))\right\} \tag{3.33}$$

and the full conditional as given in Eq. 3.11 we find that

$$\eta_{z_i}(\beta, x_i) = \left[\log(\pi_1) - \frac{(x_i - \beta_1)^2}{\sigma^2}, \ldots, \log(\pi_K) - \frac{(x_i - \beta_K)^2}{\sigma^2}\right]^T \tag{3.34}$$

Following the previous general treatment, we can compute the updates for the variational parameters which are

$$\begin{aligned} \lambda_k &= \mathbb{E}_{q(z_{1:m}|\phi_{1:n})}\left[\eta_{\beta_k}(z_{1:n}, x_{1:n})\right] \\ &= \left\langle \frac{\beta_0}{\sigma_0^2} + \frac{\sum_{i=1}^n \phi_{ik} x_i}{\sigma^2}, -\frac{1}{2}\left(\frac{1}{\sigma_0^2} + \frac{\sum_{i=1}^n \phi_{ik}}{\sigma^2}\right)\right\rangle \end{aligned} \tag{3.35}$$

and

$$\begin{aligned} \phi_k &\propto \mathbb{E}_{q(\beta|\lambda)}\left[\eta_{z_i}(\beta, x_i)\right] \\ &\propto \left[\log(\pi_1) - \frac{(x_i - \lambda_1^{(1)})^2 + \lambda_1^{(2)}}{\sigma^2}, \ldots, \log(\pi_K) - \frac{(x_i - \lambda_K^{(1)})^2 + \lambda_K^{(2)}}{\sigma^2}\right]^T \end{aligned} \tag{3.36}$$

where the proportionality stems from the unknown partition constant C introduced in Eq. 3.13. The parameter ϕ_k thus needs to be renormalized after updating each of its components. Alternately recomputing the variational parameters through Eq. 3.35 and 3.36 until convergence of the optimization objective Eq. 3.19 completes the algorithm as shown in Alg. 3.3.

3.4 Variational inference in Topic Models

Following the treatment given in the previous section, we re-derive VI in the LDA model. Blei et al. (2003) derive the updates for variational parameters by conventional methods, i.e. differentiation of the ELBO w.r.t. the different variational parameters, setting the derivative to zero and solving for the parameter. Using the general treatment as in section 3.3.2 we are able to give a more straight forward solution (cf. Hoffman et al., 2013). Recall the joint probability model in Eq. 2.7. We define the variational distribution $q(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N_d})$, give the full conditionals for the latent variables and finally the update equations for the variational parameters to optimize the lower bound on the evidence. As before, we apply the mean-field assumption on the variational distribution,

$$q(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N_d}) = \prod_k q(\beta_k|\lambda_k) \prod_d \left(q(\theta_d|\gamma_d) \prod_n p(z_{dn}|\phi_{dn}) \right) \tag{3.37}$$

defining the variational parameters $\Theta = \{\lambda_{1:K}, \gamma_{1:D}, \phi_{1:D,1:N_d}\}$. We proceed with defining the lower bound on the evidence and the full conditionals of the model. Since every latent variable is equipped with its conjugate prior, all full conditionals exist in closed form.

3.4.1 Full conditionals in LDA

Following our toy example in section 3.1.2 we derive the full conditionals by starting off from the full joint probability model and neglecting terms that do not depend on the latent variable of interest. For the topics β_k we obtain

$$\begin{aligned}
p(\beta_k|\cdot) &\propto p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N_d}, w_{1:D,1:N_d}|\alpha, \eta) \\
&\propto \left(\prod_{k=1}^{K} p(\beta_k|\eta) \right) \left(\prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta_k) \right) \\
&\propto p(\beta_k|\eta) \prod_{d=1}^{D} \prod_{n=1}^{N_d} p(w_{dn}|z_{dn}, \beta_k) \\
&\propto \prod_{w=1}^{W} \beta_{kw}^{\eta-1} \prod_{d=1}^{D} \prod_{n=1}^{N_d} \beta_{kw}^{z_{dnk} w_{dn}} \\
&\propto \prod_{w=1}^{W} \beta_{kw}^{\eta-1} \prod_{w=1}^{W} \prod_{d=1}^{D} \beta_{kw}^{z_{dnk} n_{dw}} \\
&\propto \prod_{w=1}^{W} \beta_{kw}^{(\eta + \sum_{d=1}^{D} z_{dwk} n_{dw}) - 1} \\
&\propto \mathrm{Dir}(\eta + \sum_{d=1}^{D} z_{d \cdot k} n_{d \cdot}).
\end{aligned} \tag{3.38}$$

For the topic proportions θ_d we proceed similarly, i.e.

$$\begin{aligned}
p(\theta_d|\cdot) &\propto p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N_d}, w_{1:D,1:N_d}|\alpha, \eta) \\
&\propto \left(\prod_{k=1}^{K} p(\beta_k|\eta)\right)\left(\prod_{d=1}^{D} p(\theta_d|\alpha)\prod_{n=1}^{N_d} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta_k)\right) \\
&\propto p(\theta_d|\alpha)\prod_{n=1}^{N_d} p(z_{dn}|\theta_d) \\
&\propto \prod_{k=1}^{K} \theta_{dk}^{\alpha-1}\prod_{n=1}^{N_d}\prod_{k=1}^{K}\theta_{dk}^{z_{dnk}} \\
&\propto \prod_{k=1}^{K} \theta_{dk}^{(\alpha+\sum_{n=1}^{N_d} z_{dnk})-1} \\
&\propto \mathrm{Dir}(\alpha + \sum_{n=1}^{N_d} z_{dn}).
\end{aligned} \tag{3.39}$$

Finally, we need to compute the full conditional for the z_{dn}s, the distribution over topics for one word w_{dn} in document d. As before

$$\begin{aligned}
p(z_{dn}|\cdot) &\propto p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N_d}, w_{1:D,1:N_d}|\alpha, \eta) \\
&\propto \left(\prod_{k=1}^{K} p(\beta_k|\eta)\right)\left(\prod_{d=1}^{D} p(\theta_d|\alpha)\prod_{n=1}^{N_d} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta_k)\right) \\
&\propto p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta_{1:K}) \\
&\propto \prod_{k=1}^{K} \theta_{dk}^{z_{dnk}}\prod_{k=1}^{K}(\beta_{kw_{dn}})^{z_{dnk}} \\
&\propto \prod_{k=1}^{K} (\theta_{dk}\beta_{kw_{dn}})^{z_{dnk}} \\
&\propto \mathrm{Mult}\left(\frac{1}{C}\left(\theta_d \circ \beta_{w_{dn}}^T\right)\right)
\end{aligned} \tag{3.40}$$

with C the normalization constant and $\cdot \circ \cdot$ the point-wise multiplication operator similar to Eq. 3.13 in our toy example.

Coordinate ascent updates

Following section 3.3.2, we define the variational distribution over β_k to be a Dirichlet distribution with parameter vector λ_k for all k, the distribution over θ_d also a Dirichlet with parameter γ_d for all d and the distribution over z_{dn} to be a Multinomial with parameter

ϕ_{dn} for all n words and all d, i.e.

$$\begin{aligned}\forall k : q(\beta_k) &\sim \text{Dir}\,(\lambda_k)\\ \forall d : q(\theta_d) &\sim \text{Dir}\,(\gamma_d)\\ \forall d, n : q(z_{dn}) &\sim \text{Mult}\,(\phi_{dn})\end{aligned}$$

Given the general coordinate update in Eq. 3.27 and that all factors of the variational distribution are in the exponential family, we can easily determine the parameter updates for the LDA model. We use standard results for taking expectations of a random variable and of its logarithm and compute the expectations of Eq. 3.38, 3.39 and 3.40 which are given by

$$\begin{aligned}\lambda_k &= \eta + \sum_{d=1}^{D}\prod_{n=1}^{N_d}\phi_{dn}w_{dn}\\ \gamma_d &= \alpha + \sum_{n=1}^{N_d}\phi_{dn}\\ z_{dn} &\propto \psi(\lambda_{\cdot w_{dn}}) - \sum_{w=1}^{W}\psi(\lambda_{\cdot w}) + \psi(\gamma_{d\cdot}) - \sum_{k=1}^{K}\psi(\gamma_{dk})\end{aligned}$$

where the vector z_{dn} has to be renormalized after updating each of its components as before such that $\sum_{k=1}^{K} z_{dnk} = 1$. Following the described update scheme, i.e. alternately updating the variational parameters until convergence of the optimization objective (i.e. the ELBO) constitutes statistical inference in the LDA model.

Algorithm 3.2 Toy model Gibbs sampler.

Require: $\beta_0, \sigma_0^2, \sigma^2, \pi$ ▷ fixed parameters
  $\forall i:$ set $z_{ik} = 1$ for one random k
  $\forall k:$ randomly initialize β_k
  repeat
    for $i = 1$ to n **do**
      $\eta \leftarrow \pi \circ \exp\left\{-\frac{(x_i-\beta)^2}{2\sigma^2}\right\}$
      normalize η
      $z_i \sim \text{Mult}(\eta)$
    for $k = 1$ to K **do**
      $m \leftarrow \left(\frac{\beta_0}{\sigma_0^2} + \frac{\sum_{i=1}^n z_{ik} x_i}{\sigma^2}\right)\left(\frac{1}{\sigma_0^2} + \frac{\sum_{i=1}^n z_{ik}}{\sigma^2}\right)^{-1}$
      $s^2 \leftarrow \left(\frac{1}{\sigma_0^2} + \frac{\sum_{i=1}^n z_{ik}}{\sigma^2}\right)^{-1}$
      $\beta_k \sim \mathcal{N}(m, s^2)$
    if iteration > burn-in **then**
      collect current sample
  until number of iterations reached
  return averaged collected samples

Algorithm 3.3 Toy model variational inference.

Require: $\beta_0, \sigma_0^2, \sigma^2, \pi$ ▷ fixed parameters
  $\forall i:$ randomly initialize z_{ik} and renormalize
  $\forall k:$ randomly initialize β_k
  while converged < convergence criterion **do**
    for $i = 1$ to n **do**
      for $k = 1$ to K **do**
        $\phi_{ik} = \log(\pi_k) - \frac{(x_i-\lambda_k^{(1)})^2+\lambda_k^{(2)}}{\sigma^2}$
      normalize ϕ_i
    for $k = 1$ to K **do**
      $\lambda_k^{(1)} = \frac{\beta_0}{\sigma_0^2} + \frac{\sum_{i=1}^n \phi_{ik} x_i}{\sigma^2}$
      $\lambda_k^{(2)} = -\frac{1}{2}\left(\frac{1}{\sigma_0^2} + \frac{\sum_{i=1}^n \phi_{ik}}{\sigma^2}\right)$
    converged ← relative change of Eq. 3.19
  return $\lambda, \phi_{1:n}$

Part II

Stochastic Processes and Time Series Analysis

Abstract

In this part we will review the mathematical tools that we need for our analysis. We will start with a formal introduction to stochastic processes in chapter 4 and give some basic definitions that are needed to derive dynamic models, i.e. models whose topic evolution is governed by some defined process. We lay emphasis on general Gaussian processes, including one of the most basic stochastic processes, the Wiener process, and the Ornstein-Uhlenbeck process that plays a crucial role in modern econometric stochastic volatility models. We also briefly divert to a general form of representation of stochastic processes in this chapter, although we refer the reader to Çnlar (2011); Øksendal (2003) for an excellent treatment. Some useful general definitions from probability theory are given in appendix A. Chapter 5 introduces time series analysis. Starting out with some introductory examples, we review the formal definition and the different approaches that exist, identifying those techniques that are amenable to the type of analysis that we pursue. It will become apparent that Gaussian processes are a both elegant and powerful tool to aid our needs. In fact, by minor changes to a Gaussian process we can easily modify the assumed dynamics that will drive the dynamics in our approach without having to derive new models. We then give an overview over different methodologies to efficiently compute Gaussian process realizations from a set of noisy observations of that realization in chapter 6. This includes the classical approach motivated by the marginalization property of the Gaussian distribution and the Kalman filter for solving the Wiener problem.

Chapter 4

Stochastic Processes

4.1 Foundations

4.1.1 Definition and Basic Properties

Definition 4.1.1. *Stochastic process.* Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a measurable space $(S, \mathcal{S})$. A **stochastic process** X is a sequence of S-valued random variables $X = \{X_t : t \in T\}$ with $X_t : \Omega \to S, t \in T$, indexed by some totally ordered index set T with $T \subset \mathbb{R}_+$. Then S is called the state space of the stochastic process X and, similarly, X is called S-valued.

In fact, a stochastic process is a function of two parameters, t (the time) and ω (the probability parameter, can be viewed as an experiment or particle). The notation $\{X_t : t \in T\}$ is shorthand for $\{X_t(\omega) : t \in T, \omega \in \Omega\}$. For each fixed $t \in T$, we observe a random variable

$$\omega \to X_t(\omega), \omega \in \Omega$$

and, analogically, for each fixed $\omega \in \Omega$ the function

$$t \to X_t(\omega), t \in T$$

which is called a *path*[1] of X or, identically, a *realization* of X, further on denoted as $\{x_t : t \in T\}$. $X_t(\omega)$ can thus be seen to represent the outcome or position of experiment or particle ω at time t. Consequently, a process can be characterized as

$$(t, \omega) \to X(t, \omega),$$

a function from $T \times \Omega$ into S.

We can distinguish SPs by the characteristics of both their state space and their parameter space (the set T):

[1] If T is discrete, this realization is sometimes called a *sample sequence*, if T is continuous it is also known as a *sample function*.

- When the random variables X_t are discrete, the state space is called a discrete state space. Consequently, if the X_t are continuous, so is the state space.

- When the parameter set T is discrete (i.e. $T \subseteq \mathbb{Z}_+$), the SP is called a discrete parameter process. Again, if the set T is continuous, so is the process.

Definition 4.1.2. *Filtrations.* Given a measurable space $(F, \mathcal{F})$ and a totally ordered index set $T \subset \mathbb{R}_+$, a family of σ-algebras $\{\mathcal{F}_t\}_{t \in T}$ is called a **filtration** if

$$\mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{F}, s < t, s, t \in T.$$

Now, let a set of random variables $X = \{X_t : t \in T\}$ be defined on the measurable space $(F, \mathcal{F})$. Then, $\mathcal{F}_s := \sigma\{X_s : 0 \leq s \leq t\}$ defines the **natural filtration** induced by X.

Definition 4.1.3. *Adaptedness.* Given a filtration $\{\mathcal{F}_t\}_{t \in T}$ and a set of random variables $X = \{X_t : t \in T\}$, X is **adapted** to the filtration $\{\mathcal{F}_t\}_{t \in T}$ if X_t is $\mathcal{F}_t$-measurable for every $t \in T$.

Based on definition 4.1.1, let $\{X_t : t \in T\}$ be a stochastic process.

Definition 4.1.4. *Probability law.* Given the finding that X is a function of two variables (t, the time, ω, the probability parameter), a stochastic process X can be interpreted as a random variable defined on the product space $(E, \mathcal{E}) = (S^T, \mathcal{S}^T)$, with $(S, \mathcal{S})$ its state space. The distribution of X, i.e. the probability measure on $\mathbb{P}X^{-1}$ on $(E, \mathcal{E})$ is called the **probability law** of the process X.

For any finite set of indices $\{t_1, \ldots, t_n\} \triangleq \{t_i\} \in T$, we can construct a (finite-dimensional) joint distribution function and a probability measure on the resulting product space. We make use of the fact that the product σ-algebra $\mathcal{E}$ is generated by (finitely many) measurable rectangles (see Çnlar, 2011, p. 4) and the probability measure on the product space $(E, \mathcal{E})$ is determined by the values it assigns to those rectangles. Thus,

$$\mu_{t_1,\ldots,t_n}(B_1 \times \ldots \times B_n) = \mathbb{P}\{X_{t_1} \in B_1, \ldots, X_{t_n} \in B_n\} \tag{4.1}$$

with $n \in \mathbb{N}^*, t_1, \ldots, t_n \in T, B_1, \ldots, B_n \in \mathcal{S}$ determines the probability law of the process $X = \{X_t : T \in T\}$.

Following the introductions in section A.0.8 we can compute the expectation and variance of a stochastic process, typically as functions of t. These are called the **mean function** and **covariance function** of the process and denoted by

$$m_X(t) = \mathbb{E}[X_t], \tag{4.2}$$

$$k_X(t, \tau) = \mathbb{E}[(X_t - m_X(t))(X_\tau - m_X(\tau))]. \tag{4.3}$$

4.1.2 Markovianity and Stationarity

The Markov Property

Although the Markov property is also known from dealing with probabilistic models and conditional probabilities[2] we primarily describe the implications of the Markov assumption (i.e. assuming that the property holds true) on stochastic processes, first in a general measure-theoretic fashion and second, what implications this has in the context of density or mass functions. Generally, given a stochastic process $X = \{X_t : t \in T\}$, the Markov assumption induces a crucial simplification in the description of the future states given all previous states and the current one. Formally, consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and state space $(S, \mathcal{S})$ as before. Let $\{\mathcal{F}_s\}_{s\in T}$ be a filtration and X a sequence of S-valued random variables defined on Ω and adapted to the filtration. X is a **Markov process**, if for all $A \in \mathcal{S}$ and each $s < t \in T$,

$$\mathbb{P}(X_t \in A|\mathcal{F}_s) = \mathbb{P}(X_t \in A|X_s). \tag{4.4}$$

Recall the definition of the filtration: each $\mathcal{F}_s$ is defined as a σ-algebra generated by all previous states and the present state, i.e. $\mathcal{F}_s = \sigma\{X_u : u \leq s \in T\}$. Thus, the Markov property in the context of stochastic processes tells us, that the probability of X_t being in A, given the σ-algebra generated by all previous states and the current one is just the probability of X_t being in A given the current state X_s alone. Seizing the suggestion about the heuristics of information (see Çnlar, 2011, chapter 2.4), the information encoded in just the current state X_s of the process is the same (under the assumption of Markovianity) as that contained and accumulated in the σ-algebra $\mathcal{F}_s$. In particular, the Markov property states that the future state of a process X_t is independent of the previous states $\{X_u : u < s \in T\}$ given the present state X_s.

Note that we can specify the probability law of the process X by using similar assumptions as above. We consider the joint probability function $p(x_{t_n}, x_{t_{n-1}}, \ldots, x_{t_1})$ (resembling the process' probability law in Eq. 4.1) of a particular (finite-dimensional) realization of the process (cf. section 4.1.1).

Using the chain rule, we rewrite

$$p(x_{t_n}, x_{t_{n-1}}, \ldots, x_{t_1}) = p(x_{t_n}|x_{t_{n-1}}, \ldots, x_{t_1})p(x_{t_{n-1}}, \ldots, x_{t_1})$$

and using the Markov property (Eq. 4.4)

$$p(x_{t_n}, x_{t_{n-1}}, \ldots, x_{t_1}) = p(x_{t_n}|x_{t_{n-1}})p(x_{t_{n-1}}, \ldots, x_{t_1}).$$

Continuing this, we arrive at

$$\begin{aligned} p(x_{t_n}, x_{t_{n-1}}, \ldots, x_{t_1}) &= p(x_{t_n}|x_{t_{n-1}}) \cdot p(x_{t_{n-1}}|x_{t_{n-2}}) \cdots p(x_{t_2}|x_{t_1}) \cdot p(x_{t_1}) \\ &= p(x_{t_1}) \prod_{i=2}^{n} p(x_{t_i}|x_{t_{i-1}}) \end{aligned}$$

[2]Assuming that the Markov property holds in a general probabilistic graphical model induces independence of a random variable from all but its direct neighbors, i.e. it is independent from anything else, given its Markov blanket.

We therefore can define the probability law of a Markov process by specifying the distributions $p(x_t)$ and $p(x_t|x_\tau), t > \tau \in T$, the latter of which are also called the **transition probabilities** of the Markov process.

Returning again to a measure-theoretic character, let $(P_{\tau,t})$ be a family of Markov kernels(i.e. transition probability kernels from $(S, \mathcal{S})$ into $(S, \mathcal{S})$) with $\tau \leq t$. Then this is **Markovian transition function** in $(S, \mathcal{S})$. We now have that $P_{\tau,t}(x_\tau, A) = \{X_t \in A|X_\tau = x_\tau\}$. Clearly, defining $\mathbb{P}\{X_{t_1} = x_{t_1}\}$ (if S is discrete) or $\mathbb{P}\{X_{t_1} \leq x_{t_1}\}$ (if S is continuous) and the Markovian transition function resembles the above description of the probability law of a Markov process.

Stationarity and Nonstationarity

A **strictly stationary** stochastic process is a stochastic process whose joint probability distribution is invariant to translation in time. Let $p(\cdot)$ be the probability distribution function of a realization of a stochastic process $X = \{X_t : t \in T\}$ indexed by some set T as above. If

$$p(x_{t_1+\tau}, x_{t_2+\tau}, \ldots, x_{t_n+\tau}) = p(x_{t_1}, x_{t_2}, \ldots, x_{t_n}), \tau \in T$$

X is a stationary process as its distribution function is *not* a function of time. Another less restrictive notion of stationarity which is important in our context is **weak stationarity** or **covariance stationarity**. It requires that only the first and second moments are not functions of time. For a weakly stationary stochastic process X, we require[3] that $\mathbb{E}[X_t] = m_X(t) = m_X(t + \tau)$ for all $t, \tau \in T$. Further we demand for the covariance, that $k_X(t_1, t_2) = k_X(t_1 - t_2, t_2 - t_2) = k_X(t_1 - t_2, 0)$. Hence, the covariance between two random variables X_{t_1} and X_{t_2} is merely a function of the difference given by their time indices rather than being a function of time itself (recall that $T \subset \mathbb{R}_+$). Given a SP $X = \{X_t, t \in T\}$, if the stochastic process $\{X_{t+s} - X_t, t \in T\}$ is strictly stationary for any $s \in T$, X has **stationary increments**.

Further, if for all finite sets $\{t_i : t_i < t_{i+1}\} \subset T$ the random variables

$$X_{t_2} - X_{t_1}, X_{t_3} - X_{t_2}, \ldots, X_{t_n} - X_{t_{n-1}}$$

are independent of each other, the process has **independent increments**. If in addition stationarity holds also for these increments, i.e., $X_{t+h} - X_{\tau+h}$ has the same distribution as $X_t - X_\tau$ for all $t > \tau \in T$ and all $h > 0$, the process has **stationary independent increments**. To define the probability law of a process with stationary independent increments it suffices to define the distributions of X_t and $X_\tau - X_t, \tau > t \in T$. Note that a stochastic process does not need to be stationary or strictly stationary to have stationary increments (section 4.2.4 describes such a process).

Nonstationarity of a stochastic process is consequently given when the joint probability distribution of a process is not time invariant.

[3]Another requirement is that the process has finite second moments, i.e., $\mathbb{E}[X_t^2] < \infty$. This is implicitly given when both the mean and covariance functions exist and as we only consider such cases we omit it here for clarity.

4.2 Gaussian Processes

4.2.1 Definition

A Gaussian Process (GP) is a stochastic process whose probability law (Definition 4.1.4) is normal, i.e. the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is equipped with the Gaussian measure as a probability measure. In particular, let $X = \{X_t : t \in T\}$ be a stochastic process defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with totally ordered index set T and the X_t taking values in $\mathbb{R}$. Further, consider a finite sub-set of indices $T^* = \{t_i : i = 1, \ldots, n\} \subset T$. Then $X^* = \{X_t : t \in T^*\}$ can be interpreted as a random variable taking values in the product space $(E, \mathcal{E}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ with associated n-dimensional Gaussian measure

$$\gamma^n_{\mu,\sigma^2}(A) = \frac{1}{\sqrt{2\pi\sigma^2}^n} \int_A \exp\left\{-\frac{1}{2\sigma^2}\|x-\mu\|^2_{\mathbb{R}^n}\right\} d\lambda^n(x), A \in \mathcal{E}$$

with λ^n the n-dimensional Lebesgue measure and $\mu \in \mathbb{R}^n$ and $\sigma^2 > 0$ its mean and variance.

The Gaussian measure is absolutely continuous with respect to the Lebesgue measure and so allows for the existence of the well-known Gaussian probability density function (see Johnson et al., 1995, chapter 13). Since this density function is fully defined by its first and second moment, defining its mean and covariance functions fully defines the process' probability law and thus the Gaussian process. Further, this gives GPs the property that whenever a GP is weakly stationary it is also strictly stationary.

Definition 4.2.1. A Gaussian process is a stochastic process whose probability law is normal. It is fully defined by its mean function $m(x)$ and its covariance function (or kernel) $k(x, x')$ and is denoted as $\mathrm{GP}(m, k)$.

4.2.2 Properties

Definition 4.2.1 also allows to transfer a whole set of useful properties of the Gaussian distribution to GPs.

Proposition 4.2.1. *Linear operations on Gaussian random vectors produce Gaussian random vectors. A GP is a set $\{X_t, t \in T\}$ of random variables for which Eq. 4.1.4 is normal, i.e., a GP's realization can be interpreted as a Gaussian random vector with elements x_t. From this follows that linear operations on GPs produce GPs.*

Proposition 4.2.2. *If the random vectors x and y are jointly normally distributed then y's marginal distribution is also normal. We have generally defined the probability law of a SP for any finite set $\hat{T} \subset T$ in Eq. 4.1.4, for a GP the corresponding measure is the Gaussian one. Consequently, the set $\{X_t, t \in T_*\}$ is a Gaussian process for any $T_* \subset \hat{T}$, $T_* \neq \emptyset$.*

Proposition 4.2.3. *If the random vectors x and y are jointly normally distributed then the conditional probability of x given y is multivariate normal, i.e. $x|y \sim \mathcal{N}(\mu, \Sigma)$ with mean*

$$\mu = m_x + K_{xy}K_y^{-1}(y - m_y)$$

and covariance matrix

$$\Sigma = K_x - K_{xy}K_y^{-1}K_{yx}.$$

Here, K_x represents the variance of x and K_{xy} represents the covariance between x and y.

The last property is especially useful in the light of Bayesian inference. Consider y known and x unknown. By assuming a joint normal distribution of x and y, we can deduce mean and variance of x. Clearly, this mirrors the concept of a posterior distribution, the distribution of an unknown quantity, here x, given observed data, here y. For a Gaussian process $X = \{X_t, t \in T\}$, given an observed realization $x^* = \{x_t, t \in T^*\}, T^* \subset T$ we can predict another (unknown) realization $\hat{x} = \{x_t, t \in \hat{T}\}, \hat{T} \subset T$ using this property by again understanding the realizations as normally distributed random vectors. However, this opportunity comes with a trade-off: making predictions about $\hat{x}$ includes inverting x^*'s covariance matrix (that is formed by applying the GP's covariance function on all possible pairings of elements of x^*), an operation of complexity $\mathcal{O}(n^3)$. Clearly, this is only feasible up to moderate sizes of x^* and thus limits our capabilities of predicting the realization $\hat{x}$. The more complex $\hat{x}$ is, the more information about x^* we would need to have. Section 6.1 gives a more thorough description of this dilemma and possible ways to solve it.

4.2.3 Noise

Given an observation, e.g. the average temperature or the closing price of a stock at one particular day, we can conclude the state of the underlying stochastic process modeling the observation's dynamics. Anticipatory of the learning techniques used to approximate stochastic process parameters from data (cf. section 6) however, the measurement we have access to is just a *noisy* realization of the actual stochastic process. This is either because we have reached the level of random errors that come with the measurement equipment or to account for the stochasticity of the process itself due to the lack of more informative data. Another cause can be unpredictable random effects or, as Jazwinski (1970) terms it, "(...) he [the engineer] has reached the "noise level." (...)". For real world examples each model is wrong to some extent[4] and the assumption of noisy measurements accounts for discrepancies between the model and observations. According to their spectral densities, different types of noise have been given color names: white noise (noise with constant power spectral density), pink and blue noise (with power spectral densities inversely proportional and proportional to the frequency f respectively), Brownian/red and violet noise (with power spectral density inversely proportional and proportional to f^2 respectively) and gray noise (with a power spectral density that is psycho-acoustically constant at all frequencies, i.e. produces the perception that all frequencies are equally loud). We restrict further remarks to white noise.

[4]Basically, this is the very definition of a model, to provide a simplified and more usable version of reality.

White noise

Noise sequences and processes are stochastic processes with certain probabilistic properties. A widely applied and useful type of noise is *white noise*. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{G} = \{G_t : t \in T\}$ be a filtration over it. A stochastic process $W = \{W_t : t \in T\}$ that is adapted to $\mathcal{G}$ and is a Markov process with $\mathbb{P}\{W_t \in A|\mathcal{G}_s\} = \mathbb{P}\{W_t \in A\}, A \in \mathcal{F}, s < t \in T$ is called a *white process*. Consequently, a process of this type is completely random and knowledge of the current state gives neither an indication what the next state will be nor what the last state was. Often, the probability law of such a noise sequence is taken to be normal, typically with zero mean and unit variance, in which case it is called a *white Gaussian noise process*. One intuition of its usability is to take a large number of random effects and letting them superimpose each other. This often leads to a superposition having a normal distribution and is in fact one of the main propositions of the *central limit theorem* (cf. Jazwinski, 1970). White Gaussian noise thus can be used to model a combined set of random perturbation of observed measurements.

4.2.4 Wiener process / Brownian motion

Brownian motion is the probably most well-studied stochastic process there is. It underlies many other stochastic processes, in fact all Gaussian process are in some way based on it. We will first make some general remarks on the nature of so-called Lévy processes and then give a formal definition of Brownian motion and Wiener processes. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with a filtration $\mathcal{G} = \{\mathcal{G}_t : t \in \mathbb{R}_+\}$ on it. Further, let $X = \{X_t : t \in \mathbb{R}_+\}$ be a stochastic process with state space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ with $n \leq 1$ the dimension of the space.

Definition 4.2.2. *Lévy processes.* X is called a Lévy process with respect to $\mathcal{G}$ if X is adapted to $\mathcal{G}$ and

a) the path $t \to X_t(\omega)$ is right-continuous and left-limited starting from $X_0(\omega) = 0$ for almost every $\omega \in \Omega$, and

b) the increment $X_{t+s} - X_t$ is independent of $\mathcal{G}_t$ and has the same distribution as X_s for all $t, s \in \mathbb{R}_+$.

Consequently, for X to be a Lévy process, it must have regular paths and its increments have to be stationary and independent.

Definition 4.2.3. *Wiener process.* Now, let $X = \{X_t, t \in \mathbb{R}_+\}$ be a continuous parameter stochastic process. It is called a Brownian motion process if it is Lévy and has state space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. It is a Wiener process if additionally the following conditions hold:

a) $(X_{t+s} - X_t)$ has the Gaussian measure with mean 0 and variance s for every $s, t \in \mathbb{R}$, and

b) $X_0 = 0$.

Brownian motion processes are often denoted as $B = \{B_t : t \in \mathbb{R}_+\}$ and, analogously, Wiener processes as $W = \{W_t : t \in \mathbb{R}_+\}$. We will always refer to Brownian motion or the Wiener process, respectively, in this manner. Fig. 4.1 shows a single realization of a one-dimensional Wiener process.

Note. Formally, white noise is the derivative of the Wiener process (in mean square calculus) (cf. Jazwinski, 1970; Çnlar, 2011). Their paths are nowhere differentiable and they are not stationary. However, their increments are.

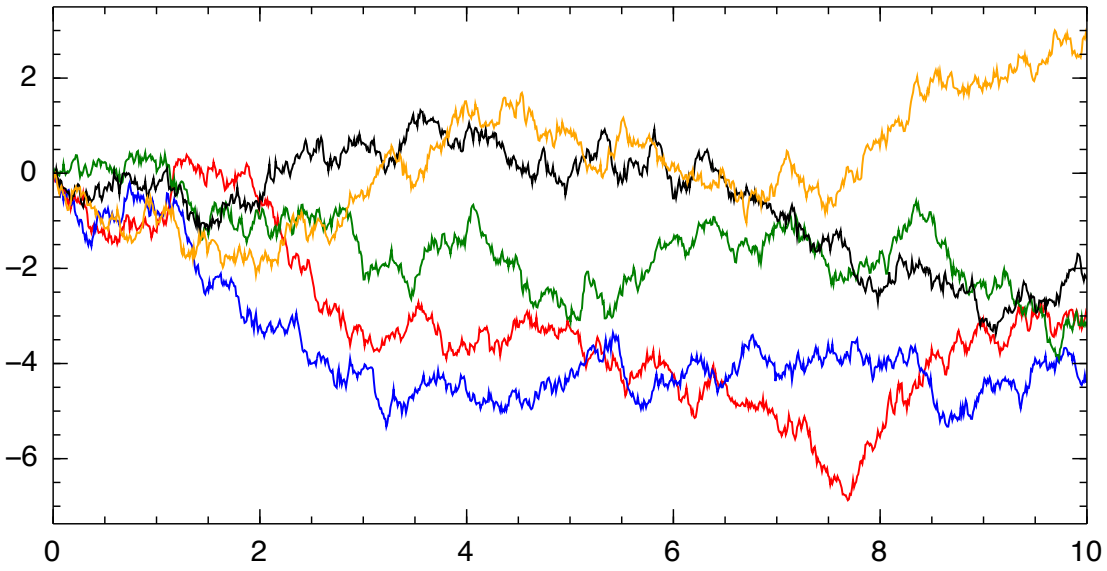


Figure 4.1: Five realizations of the one-dimensional Wiener process with $\sigma^2 = 1$.

4.2.5 Stochastic differential equations

In a deterministic setting we can characterize dynamical phenomena by using information from the derivatives of the function governing the process in question by describing its behavior over time with a differential equation. Following Øksendal (2003), we use a simple population growth model as an example. Let the growth of a population be governed by

$$\frac{dN}{dt} = a(t)N(t), N(0) = N_0 \tag{4.5}$$

with $N(t)$ the size of the population at time t and $a(t)$ the rate of growth at time t. If $a(t)$ is not completely known, i.e. underlies some randomness, the described system is not deterministic any more, thus we assume that $a(t)$ is of the form

$$a(t) = r(t) + \text{"noise"}$$

with $r(t)$ a known, nonrandom function. *Stochastic differential equations* help to describe this problem in a sound mathematical setting and to solve Eq. 4.5. The natural course of action is to model the above "noise" by a white Gaussian noise process as just introduced in section 4.2.3, i.e. define

$$a(t) = r(t) + \alpha w_t$$

with α a constant and w_t white noise[5]. In general, stochastic processes can be interpreted as solutions to stochastic differential equations of the form

$$\frac{dX_t}{dt} = f(t, X_t) + \sigma(t, X_t)w_t, f(t, X_t) \in \mathbb{R}, \sigma(t, X_t) \in \mathbb{R} \tag{4.6}$$

again with w_t white Gaussian noise. Following remark 4.2.4 we put $w_t = \frac{dW_t}{dt}$ in the following. Now, X_t is to be understood to satisfy the following stochastic integral equation in the Itō sense (see Øksendal, 2003, chapter 3)

$$X_t = X_0 + \int_0^t f(s, X_s)ds + \int_0^t \sigma(s, X_s)dW_s, \tag{4.7}$$

where the second term on the right hand side is a conventional Riemann-Stieltjes integral[6] and the third term an Itō integral with respect to the Wiener process. Equivalently reformulated in differential form we proceed to a general stochastic differential equation

$$dX_t = f(t, X_t)dt + \sigma(t, X_t)dW_t. \tag{4.8}$$

Returning to our example above, the stochastic process describing the population size can be described by the following SDE

$$dN_t = r(t)N_t dt + \alpha N_t dW_t \tag{4.9}$$

where we have set $f(t, N_t) = r(t)N_t$ and $\sigma(t, N_t) = \alpha N_t$. The function $f(t, X_t)$ is often called the *drift function*, $\sigma(t, X_t)$ the *diffusion function*. The so described process follows this nomenclature and is called a *diffusion process* and usually has the Markovian property. In our example, the drift function relates to the part of population growth we are sure about. The diffusion function accounts for the "noise" we wanted to be handled as well. By setting $f(t, X_t) = 0$ and $\sigma(t, X_t) = 1$ we arrive at $dX_t = dW_t$, thus we conclude that the Wiener process experiences no drift at all and exhibits a constant diffusion of 1. Consequently, it is governed by the SDE's stochastic part alone and so is completely random as described above, i.e. we cannot be sure about anything concerning its growth and let it be governed by noise alone.

[5]Throughout the literature the terms white noise and white Gaussian noise are often used interchangeably, i.e. white noise (in the context of stochastic processes) is generally Gaussian.

[6]This is only true because we chose the state space to be $\mathbb{R}^n, n \geq 1$, otherwise it would be a Lebesgue integral.

4.2.6 Ornstein-Uhlenbeck process

Having defined the notion of drift and diffusion function, we might also be interested in cases where these are not 0 and 1 respectively. The Ornstein-Uhlenbeck process is one such case. It is a stochastic process that is Gaussian, Markov and stationary and is in fact the only trivial stochastic process fulfilling all of this properties. In differential form, it can be described as

$$dX_t = \theta(\mu - X_t)dt + \sigma dW_t, \tag{4.10}$$

with $\mu \in \mathbb{R}$ the *process mean* and $\theta > 0$ and $\sigma > 0$ its *mean reversion rate* and *volatility*. We observe that the process is driven by a Wiener process W scaled by the volatility parameter σ, i.e. the diffusion function is defined as $\sigma(t, X_t) = \sigma$. The drift function pushes the process towards its mean μ, the higher the mean reversion rate θ, the faster. This property is called the *mean reversion* property and makes the process particularly useful in a variety of applications. Originally, it was designed to model the velocity of a particle undergoing Brownian motion under friction (Uhlenbeck and Ornstein, 1930): the mean is 0 because the particle is slowed down by friction with magnitude given by the mean reversion rate. The random collisions with other particles as described by Brownian motion are modeled by the Wiener process. It has, however, also found considerable interest and extensions in modern econometric and financial mathematics (e.g. (Griffin and Steel, 2006a; Barndorff-Nielsen and Shephard, 2001; Barndorff-Nielsen, 2002)). Figure 4.2 shows three realizations with different mean reversion levels and constant mean and volatility, all three start at $x_0 = 2$. We can clearly see that, the larger θ, the faster the process tends to its mean.

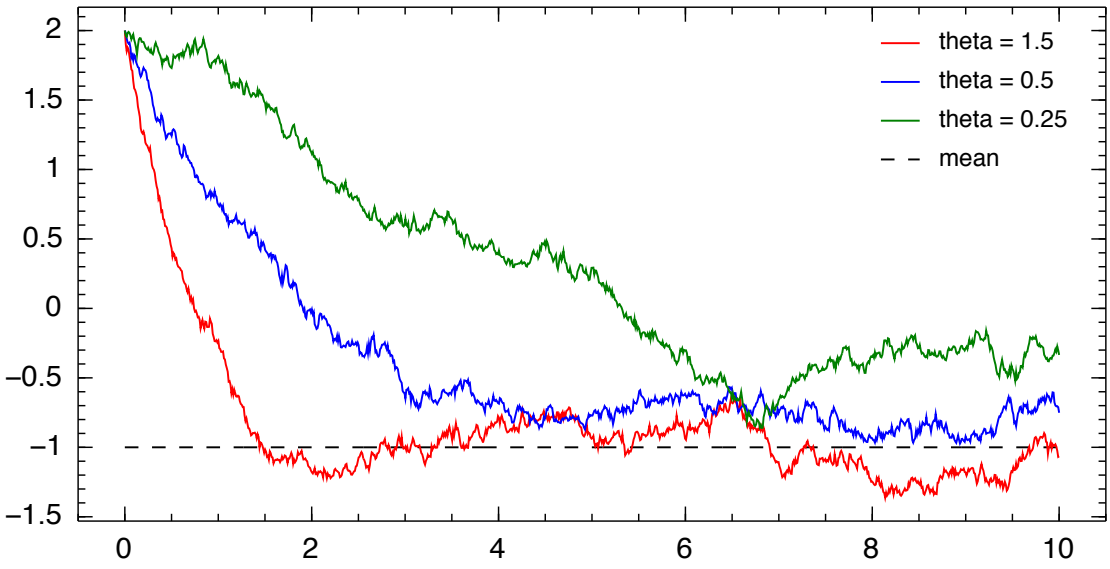


Figure 4.2: Three realizations of the one-dimensional Ornstein-Uhlenbeck process. Other parameters are $\mu = -1$, $\sigma = .25$ and $x_0 = 2$.

Chapter 5

Time Series Analysis

5.1 Some introductory Examples

When analyzing large amounts of text, we need to use any information provided by the data that helps to bring an order to the informational clutter that Big Data actually is. One of the most basic types of meta data that often is provided is a time-stamp for documents, e.g. the time it was collected into a corpus or (preferably) the date it was published. As has been described in chapter 2, we are concerned with modeling the content of documents through topics as semantic aggregations. Our ultimate goal is to model the behavior of topics (in the sense as described in the previous part) over the period of time, for which we have data available. These kinds of analyses are called Time Series Analyses (TSA). Different types of TSA exist, ranging from relatively simple analyses of frequencies over time to complex stochastic models mainly used in econometrics or weather forecasting; examples 5.1.1 and 5.1.2 present example data amenable to a time series analysis for both those research fields.

Given that topics are latent variables in probabilistic graphical models, they are random per definitionem. Any time series of random variables is a *stochastic process* (cf. Taylor, 2011) and stochastic processes can be used to mathematically model properties of dynamic systems over time. In fact, more sophisticated econometric models in (e.g. Barndorff-Nielsen, 2002; Barndorff-Nielsen and Shephard, 2001) or weather prediction (e.g. Archambeau et al., 2007) use stochastic processes to model the behavior of highly dynamic systems such as stock or option prices and weather forecasts.

We base our research on existing work that models topic evolution as Brownian motion, both in the discrete (see Blei and Lafferty, 2006) and continuous setting (see Wang et al., 2008), and aim to develop a more general approach in part III. In order to fully understand the machinery underlying this, however, we have to lay down the mathematical and methodological bases of the analyses we want to perform. In the current part we will first introduce the foundations of stochastic processes and how *stochastic differential equations* can be used to describe them. This includes a formal introduction to stochastic processes and *Gaussian processes* in particular. We further give a brief overview over the

ideas behind *time series analysis* and quickly divert to stochastic methods rather than the classical statistical understanding of time series analysis. Concluding, we describe different approaches for *learning dynamic models* based on stochastic processes using *filtering*, *Gaussian process regression* and a *variational inference algorithm* for general dynamic models defined by stochastic differential equations.

Examples

Example 5.1.1. *Stock and exchange rates* The classical example for TSA is that of stock and exchange rates. Economy is concerned with maximizing the profit and minimizing the risk and various techniques and theories have been introduced to achieve either one or the other (see Taylor (2011) for a thorough introduction or e.g. Drapeau (2010) for a recent work on risks). The data we are dealing with in this context typically include dates (time) and the corresponding value of the stock at that particular time. Obviously, this constitutes a time series as we have a sequence of values together with appropriate meta data. The target of TSA is now to analyze the values of the stock in question, identify long term and short term behavior of the time series (i.e. shocks, long-term and short-term trends, cyclical behavior etc.) and to help making predictions of future values. We show two different examples of stock price time series of car manufacturers in Figs. 5.1 and 5.2 at the New York and Frankfurt stock exchange respectively.

Example 5.1.2. *Weather and climate data* Another obvious example is that of weather and climate data. Again, we have functional values (e.g. temperature, rainfall, CO_2-emissions etc.) together with the time of their individual measurement. As before, the aim lies in analyzing the historical behavior of values over time and the prediction of future data. An example of temperature and precipitation in Leipzig over the last 40 years is shown in Fig. 5.3.

5.2 Definition

Box et al. (2013) start with a basic definition of what a time series is:

> A *time series* is a sequence of observations taken sequentially in time.

In other words, everything that we observe, measure or deduct[1] can be seen as a time series, as long as it comes with the appropriate meta data (i.e. a time stamp). Further they state that

> *Time series analysis* is concerned with the techniques for the analysis of (...) dependence [...]

[1]For the sake of clarity we will refer to the term *data point* irrespective of the actual type and origin of the data, i.e. be it observed or latent.

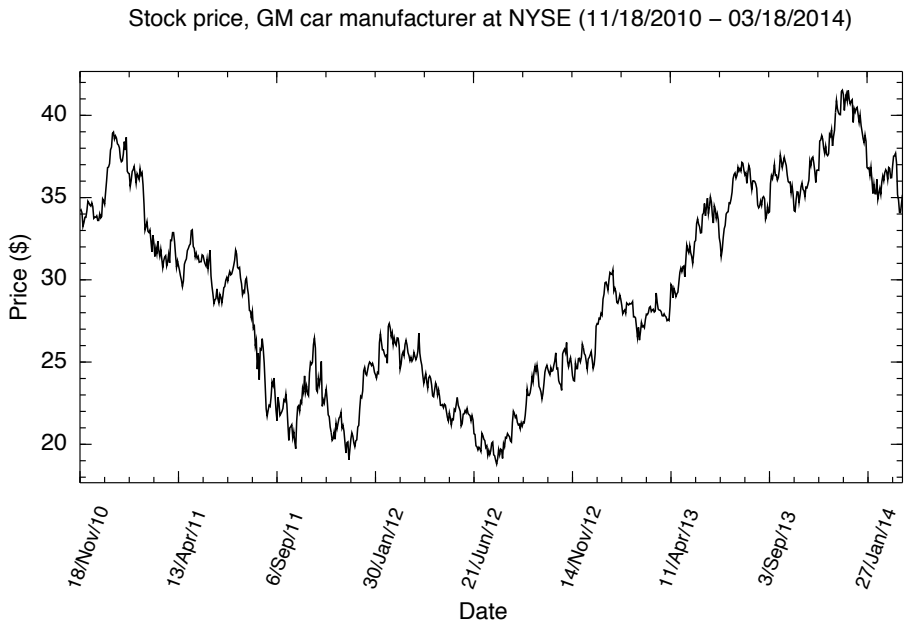


Figure 5.1: GM stock price time series, data obtained via Google Finance (`http://www.google.com/finance`)

Figure 5.2: Porsche stock price time series, data obtained via Yahoo! Finance (`http://finance.yahoo.com`)

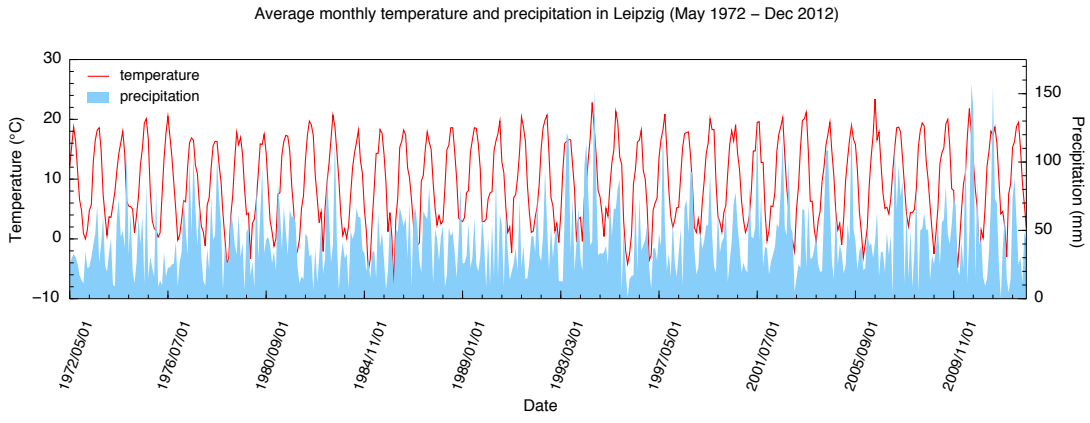


Figure 5.3: Average monthly temperature and average monthly precipitation in Leipzig from May 1972 to December 2012, data obtained via the Weather Request and Distribution System of Deutscher Wetterdienst (`https://werdis.dwd.de`)

between such data points. Typically, data points that are close to each other in time exhibit higher dependency on each other than data points that appear far from each other.

Being more specific, Brockwell and Davis (2009) discriminate TSA in their definition:

> A time series is a set of observations x_t, each one being recorded at a specified time t. A discrete-time series (...) is one in which the set T_0 of times at which observations are made is a discrete set (...). Continuous-time series are obtained when observations are recorded continuously over some time interval, e.g. when $T_0 = [0, 1]$. [...]

Note the strong resemblance to the definitions of stochastic processes from the previous section. A set of such observations is always a realization of the random variables modeled by a stochastic process. Brockwell and Davis (2009) state that "(...) frequently (...) the term time series (is used) to mean both the data and the process of which it is a realization", i.e. often the term "time series" is used interchangeably for the observations and the driving stochastic process in the background. In our particular case, data points are topics (i.e. distributions over the vocabulary) and we expect them to change through time. In this context, the set of timestamps that are associated with observations, i.e. documents, is a subset of T_0. It is of course possible to consider T_0 to be either discrete, consisting of predefined points in time, or continuous. The usage of discrete-time or continuous-time series therefore is a matter of how the model is designed, the structure of the data we have available and what questions we are trying to answer.

Definition 5.2.1. In the context of probabilistic topic models, TSA is concerned with *describing* the evolution of distributions over the vocabulary and *predicting* their state at points in time where it is unknown. Special interest lies in the identification of points in time when topics change considerably and/or unexpectedly.

Box et al. (2013) give the following list of five important applications of TSA:

1. The *forecasting* of future values of a time series from current and past values
2. The determination of the *transfer function* of a system to inertia–the determination of a dynamic input-output model that can show the effect on the output of a system of any given series of inputs
3. The use of indicator input variables in transfer function models to represent and assess the effects of unusual *intervention* events on the behavior of a time series
4. The examination of interrelationships among several related time series variables of interest and determination of appropriate *multivariate* dynamic models to represent these joint relationships among variables over time

5. The design of simple *control schemes* by means of which potential deviations of the system output from a desired target may, so far as possible, be compensated by adjustment of the input series values

The nature of our data limits the fields of possible applications to the first and fourth point mentioned above. This is because the topics (whose dynamics we are trying to model) are *latent* random variables, i.e. they cannot be observed directly. This consequently forbids attempts of willingly provoking or even just influencing the outcome of such a system by simply altering the input. In our case, the input is dependent on the composition of the individual documents and underlies random perturbations itself (as introduced by the probabilistic model). It is thus highly arguable that a willing manipulation of topic evolution by indirectly changing topics through altering document structure is at all possible. Another exemplary use of TSA is the analysis and forecasting of stock and option markets as described by e.g. Taylor (2011) or Øksendal (2003). In fact, the idea to interpret stochastic volatility in our context (discussed later) stems from here.

Corollary 1. *In probabilistic topic modeling, TSA is limited to the prediction of unknown values from available data and the study of interrelationships of different time series, i.e. multivariate time series.*

5.3 Gaussian Processes for Time Series Analysis

According to Roberts et al. (2012), the problem of time series analysis can be recast to that of regression of a form $x_t = f(t) + \nu$ with ν an additive white noise process, often called the *process noise* and accounting for the randomness of x_t. They argue that, having defined the problem, we can pursue two different possible goals:

1. Find the presumed form of $f(\cdot)$

2. Find the distribution over x^* given unknown input points t_*, i.e. evaluate $p(x_*|t_*)$

that translate into **function mapping** and **curve fitting** as approaches to tackle these, respectively. The first of the two is not of interest in our setting. We do not want to fix $f(\cdot)$ to an explicit form but rather allow the relationship between t and x_t to be conditioned on observational data in a Bayesian sense. Curve fitting on the contrary completely adheres to our definition: the idea is to fit a curve to a set of known (t, x_t) points and to predict unknown data by extrapolating the curve fitted to the observed "parts of the path". In a sense, we are not fixing $f(\cdot)$ but place a prior distribution on it that can be refined in the light of data (cf. section 5.3.1). Using this approach, we can tackle one of the main applications of TSA to topic modeling, the prediction of unknown values from available data, i.e. given input points x_t at certain (time) dates t, we want to fit a curve/function in a high dimensional space that predicts the (previously unknown) state x_* of a system at times t_*.

5.3.1 GP as a prior over functions

Let us now examine how we can use Gaussian processes for this approach. Consider a Gaussian process $X = \{X_t : t \in T\}$ as defined in section 4.2 with $T \subseteq \mathbb{R}_+$ and T totally ordered. Then, for all $\omega \in \Omega$, let $x_t = X_t(\omega)$ and each path $\{x_t : t \in T\}$ be a function

$$\begin{aligned} f : \mathbb{R}_+ &\to \mathbb{R} \\ t &\mapsto x_t. \end{aligned}$$

Now, fix ω such that for every finite subset $x = \{x_t : t \in \hat{T}\}$ of size $n = |\hat{T}|$ of a path, we can define a joint probability distribution $p(x_{t_1}, \ldots, x_{t_n})$, which is of course an n-dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$ with parameters defined by the underlying process' mean and covariance function, i.e. $\mu_i = m_x(t_i)$ and $\Sigma_{ij} = k_x(t_i, t_j)$. Effectively, this results in a prior distribution over possible paths $p(x) = p(x_{t_1}, \ldots, x_{t_n})$ and thus over $f(\cdot)$.

Rasmussen and Williams (2006) develop a framework to extend this definition, fitting the GP into the framework of Bayesian inference and reasoning. It may be understood as a prior over an infinite-dimensional object (the function space), effectively rendering it a nonparametric one (cf. Orbanz and Teh, 2010). As we have seen before, a Gaussian process is defined by its mean and covariance functions. Without loss of generality, we assume a zero mean function, i.e. $m_x(t) = 0$ for all $t \in T$. For the covariance function, we have several candidates that determine the nature of the Gaussian process. Note that the Gaussian process is not a specific stochastic process but rather a family of processes that differ in their path covariance structure. However, the covariance function will be a positive semi-definite, finite function of tuples (t_i, t_j) by definition. Specifying this function implies a distribution over the function space from which we can draw random samples (which are functions). We can condition this distribution on observations to effectively restrict the space to functions that agree with the observations.

5.3.2 Covariance Functions

As has been described, a finite, semi-definite covariance function fully specifies a Gaussian stochastic process (considering a zero mean function) in the sense above. For the purpose of modeling topic evolution over time with stochastic processes, we introduce some covariance function that might be of interest and which we will use in our models in part III. We include the basic fully random Wiener process, the mean-reverting Ornstein-Uhlenbeck process and two other processes that are (to our knowledge) as yet only defined in terms of their covariance structure: the squared exponential and the periodic covariance function.

Wiener Covariance

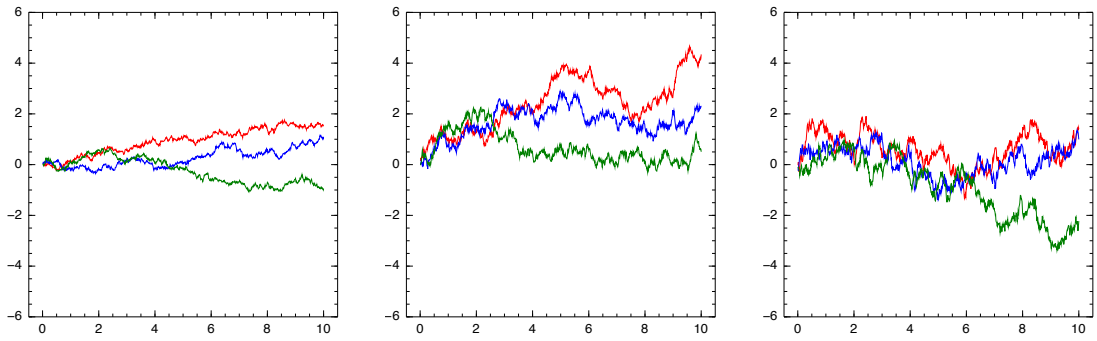
We have introduced the Wiener process before in section 4.2.4 as a Gaussian process. We have, however, not yet described an explicit form (except the trivial SDE $dX_t = dW_t$) but rather defined the process by its properties. Defining a GP by its covariance structure gives us the opportunity to do so.

Definition 5.3.1. *Wiener covariance.* A Gaussian process with covariance function

$$k_x^{\text{Wiener}}(t_i, t_j) = \sigma^2 min(t_i, t_j) \tag{5.1}$$

is a Wiener process with variance σ^2.

Each of the examples in Figure 5.4 shows three random draws from the GP with Wiener covariance (with differing variance parameter σ^2), i.e. three sample function from the function space confined by that covariance structure.



(a) $\sigma^2 = 0.25$ (b) $\sigma^2 = 0.5$ (c) $\sigma^2 = 1$

Figure 5.4: Three realizations of random functions (paths) from a GP with Wiener covariance function.

Ornstein-Uhlenbeck Covariance

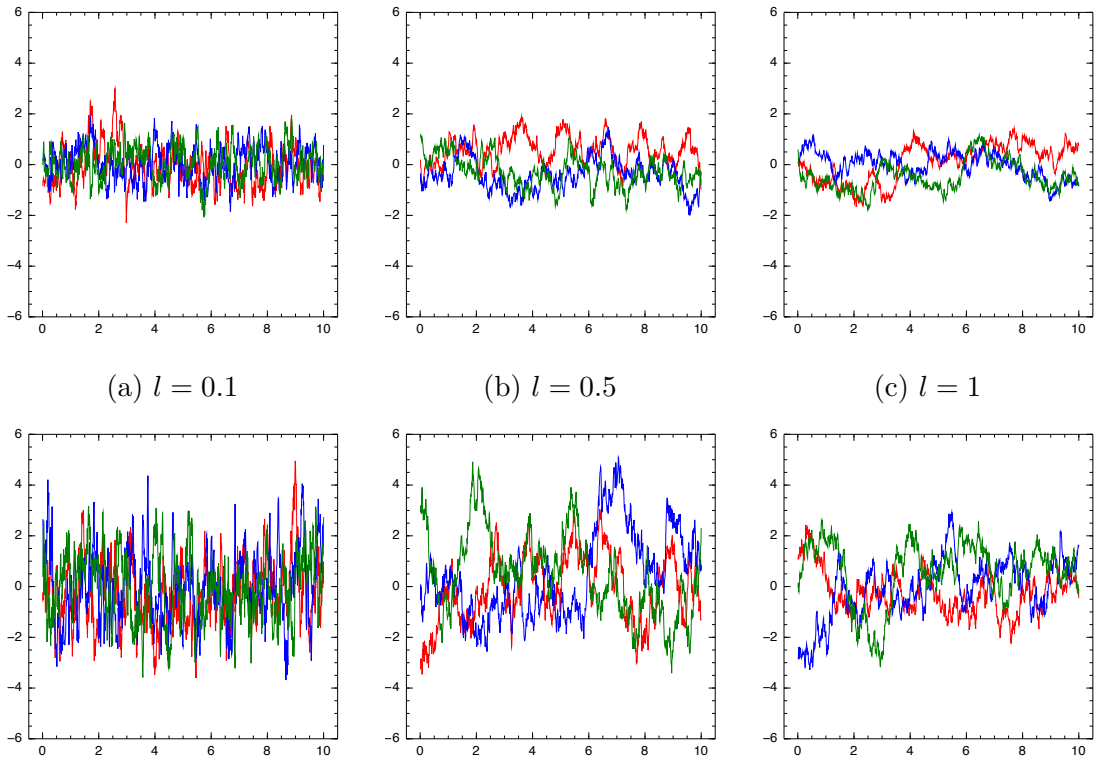
Defining Gaussian processes by their covariance structure alone lets us easily switch between different stochastic processes and thus between different subspaces of the function space. We review and redefine the Ornstein-Uhlenbeck process as described in section 4.2.6. This process is (strictly) stationary and stationarity implies that the process' covariance between two random variables must be invariant to time transformation (cf. section 4.1.2) and thus a function of their difference in time. We will define this difference as $\delta = t_i - t_j$ for $t_i > t_j$ and their covariance as $k(t_i, t_j) = k(\delta)$. In contrast to the Wiener process this allows for a bounded variance, caused by the process' mean reversion property.

Definition 5.3.2. *Ornstein-Uhlenbeck covariance.* A Gaussian process with covariance function

$$k^{\text{OU}}(\delta) = \sigma^2 \exp\left\{-\frac{\delta}{l}\right\}, \tag{5.2}$$

is an Ornstein-Uhlenbeck process with l the characteristic length scale and σ^2 the process variance.

Inspecting the function, l governs the covariance in the following way: the larger l, the larger the covariance for a given time difference. As a consequence a lower l allows a function from the function space to change more rapidly as two adjacent time locations have lower covariance and hence can divert more quickly. The process variance σ^2 controls the amplitude of these changes. We show sample paths of this process in Figure 5.5 for different process variance and characteristic length scale values. Note also the difference to Figure 4.2 where we have chosen $x_{t_0} = 2$ to be fixed. Here, x_{t_0} is chosen from the Gaussian distribution $\mathcal{N}(0, \sigma^2)$, effectively rendering the process probability measure invariant. The Ornstein-Uhlenbeck covariance function is a member of the larger class of Matérn covariance functions.



(a) $l = 0.1$ (b) $l = 0.5$ (c) $l = 1$

(d) $l = 0.1$ (e) $l = 0.5$ (f) $l = 1$

Figure 5.5: Realizations of random functions (paths) from a GP with Ornstein-Uhlenbeck covariance function. Upper row with $\sigma^2 = 0.5$, lower row with $\sigma^2 = 2$.

Squared Exponential Covariance

As has been suggested, the squared exponential covariance function is probably the most widely-used example of GP covariance function. It is infinitely differentiable causing the resulting process to be very smooth. Rasmussen and Williams (2006) point out that there is dissent about whether

> "(...) that such strong smoothness assumptions are unrealistic for modeling many physical processes (...)"

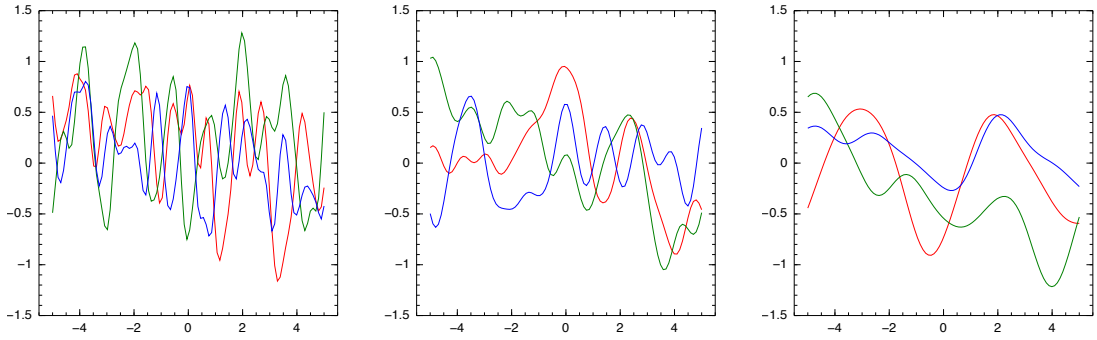
but nevertheless, the resulting process is used excessively throughout the literature (see e.g. Titsias (2009); Titsias and Lawrence (2010); Roberts et al. (2012); Hensman et al. (2013) all of which use the squared exponential as their running example). This is because the squared exponential gives rise to a feature space defined by Gaussian shaped basis functions, to which the input is being transformed, see MacKay (1998) and Rasmussen and Williams (2006, chapter 4). The squared exponential covariance function is stationary, we give it again as a function of time difference δ.

Definition 5.3.3. *Squared exponential covariance.* The covariance function

$$k_x^{\mathrm{SE}}(\delta) = \sigma^2 \exp\left\{-\frac{\delta^2}{2l^2}\right\} \tag{5.3}$$

is called the squared exponential covariance function with variance σ^2 and characteristic length scale l.

Both variance and length scale fulfill the same function as before. Figure 5.6 shows three random functions for differing length scale respectively. We skip inspection of different process variances as this only affects the process' amplitude as before.



(a) $l = 0.25$ (b) $l = 0.5$ (c) $l = 1$

Figure 5.6: Realizations of random functions from a GP with squared exponential covariance function, constant signal noise $\sigma^2 = 0.2$ and differing length scales l.

Periodic Covariances

Considering our goal of describing large collections of time stamped data and the nature of the dynamic variable being topics, it naturally occurs to seek out for periodic repetition in high probability of a topic. For instance, topics with high probability words associated with sports events in general are assumed to periodically rank words higher that are associated

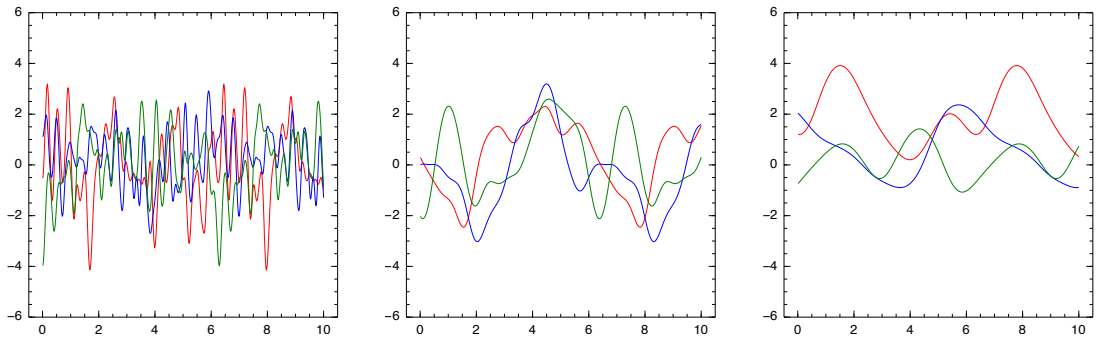
with recurring events (e.g. the Olympic Games or Soccer World Championships). For this purpose, we introduce periodic covariance functions which are also stationary functions. However, they are formed by translating the input (i.e. the time locations) to another space using a non-stationary transformation and then use a stationary covariance function on that space. The following covariance function is based on the squared exponential covariance, applied to transformed inputs $\mathbf{t}_i = (\cos(t_i), \sin(t_i))$ and $\mathbf{t}_j = (\cos(t_j), \sin(t_j))$. Rearranging terms instantly leads to:

Definition 5.3.4. *Periodic covariance.* A covariance function of the form

$$k_x^{\text{periodic}}(\delta) = \sigma^2 \exp\left\{-\frac{2\sin^2(\frac{\delta}{2})}{l^2}\right\}, \tag{5.4}$$

is called periodic with period 2π and process variance and characteristic length scale as before.

Figure 5.7 show random realizations with this covariance, again for constant variance and differing length scale.



(a) $l = 0.25$ (b) $l = 0.5$ (c) $l = 1$

Figure 5.7: Realizations of random functions from a GP with periodic covariance function based on the squared exponential function. Signal noise $\sigma^2 = 0.2$ is constant and length scales l differ as described.

Chapter 6

Bayesian Inference for Stochastic Processes

In this chapter we introduce techniques for learning the parameters of stochastic processes in the Bayesian setting, i.e. under the influence of data. Our aim is to develop a general framework of inference tools for SPs that is usable in the context of dynamic mixture models such as topic models. We first turn our attention to the previously described problem of curve fitting. Our goal is to find the "best" setting of the parameters to a stochastic process in the light of observations. As we have concerned ourselves mainly with Gaussian processes we shall start with the exact but approach based on the marginalization property of the Gaussian distribution. We proceed with an approximation - the Kalman filter, which is more efficient than the exact approach but limits itself to linear state transitions.

6.1 Gaussian Process Regression

In the last chapter we have shown that we are able to sample random functions from the part of the function space defined by a certain covariance function. We will now see, how we can incorporate observed data and draw conclusions about the structure of the underlying process. For this, let $x = \{x_t : t \in T\}$ be a path of the underlying Gaussian process $X = \{X_t : t \in T\}$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with totally ordered index set T and $\mathbb{P}$ the appropriate Gaussian measure. Further let $y = \{y_t : t \in \hat{T}\}$ with $\hat{T}$ a finite subset of T be a set of noisy observations of x available to us. Following section 4.2.3, we assume $y_t = x_t + \epsilon, t \in \hat{T}$ to have *observation noise* $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$, representing the uncertainty in each individual observation y_t and being a zero-mean, constant variance additive white noise process. Often, an additional mapping function $h : \mathbb{R} \to \mathbb{R}$ from the process' state space into its observation space is defined, i.e. $y_t = h(x_t) + \epsilon$, however we do not do this here. We assume h to be the identity function and let $h(x_t) = x_t$. The covariance between individual observations y_{t_i} and y_{t_j} is then given by the same covariance function k_x that defines the GP (cf. section 4.2) for all $i \neq j$ and $k_x(t_i, t_j) + \epsilon$ for $i = j$,

i.e.

$$k_y(t_i, t_j) = k_x(t_i, t_j) + \epsilon\delta_{ij} \tag{6.1}$$

with δ the Kronecker delta.

6.1.1 Exact inference

Given a set of observed data points y, the predictive distribution of a (discretized and unknown) path x derives from the assumption that both the observed data points and the path components stem from the same distribution, in our case indicating their derivation from the same stochastic process. In applying the marginalization property (cf. proposition 4.2.3), we arrive at the joint posterior distribution over the function space, defined by the posterior GP[1]. From it, we can again sample random functions that now conform to the observed function values at the training point locations. Graphically speaking, we consistently sample random paths x from the GP and reject all those that do not agree with our observations. Clearly this would be a laborious approach but it fortunately corresponds to simple conditioning of the joint prior distribution over paths on the observed data points y. For this, let T^* be another finite subset of T representing a discretization of T, with T^* and $\hat{T}$ not necessarily disjunct. Using this assumption, we can augment the observed distribution over the points y with the unknown path x:

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{xx} & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{yx} & \mathbf{\Sigma}_{yy} \end{bmatrix} \right),$$

with covariance and cross-covariance matrices given by

$$\begin{aligned} (\mathbf{\Sigma}_{xx})_{ij} &= k_x(t_i, t_j), \forall t_i, t_j \in T^*, \\ (\mathbf{\Sigma}_{xy})_{ij} &= k_y(t_i, t_j) = (\mathbf{\Sigma}_{yx})_{ji}, \forall t_i \in T^*, t_j \in \hat{T}, \\ (\mathbf{\Sigma}_{yy})_{ij} &= k_y(t_i, t_j), \forall t_i, t_j \in \hat{T}. \end{aligned}$$

Now the predictive distribution of x is given by

$$x \sim \mathcal{N}(m, \mathbf{S})$$

with

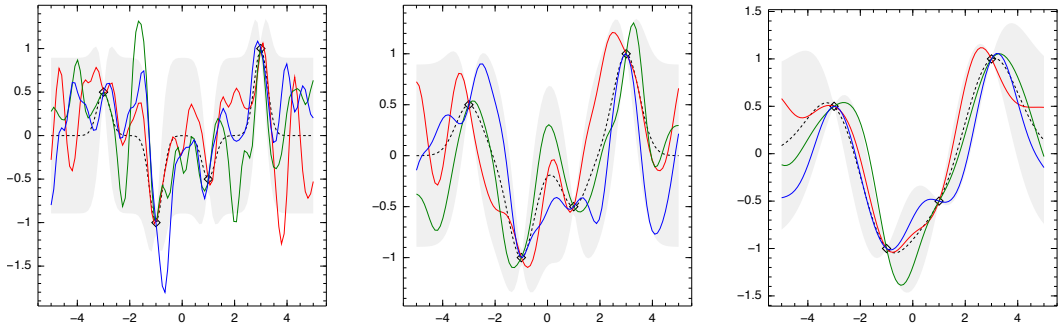
$$\begin{aligned} m &= \mu_x + \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}(y - \mu_y), \\ \mathbf{S} &= \mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx}. \end{aligned}$$

This computation is dominated by the inversion of the matrix $\mathbf{\Sigma}_{yy}$ which has complexity $\mathcal{O}(n^3)$ where n is the number of observed data points, i.e. $n = |\hat{T}|$. Hence, this approach

[1]Recall that the posterior is the "refined" prior distribution in the light of data. In our case, the GP with its covariance function is the prior over paths and we refine it with observations to obtain the posterior over paths.

to inference in Gaussian processes is only suitable for up to medium sized problems (with observation set sizes of a few hundred to a few thousand points) and looses applicability when approaching larger data set complexity or when frequent re-computation of the GP posterior is needed. We do not treat this problem here, however, it can be relaxed by making use of faster learning algorithms for Gaussian processes such as sparse approximations to the full GP (e.g. Snelson and Ghahramani, 2005) and variational approaches (e.g. Titsias, 2009; Hensman et al., 2013).

Example 6.1.1. Examples corresponding to conditioning the sample functions from the prior GP (Figure 5.6) to observed data are given in Figure 6.1. This time we sampled three functions from the posterior GP respectively. Each of the random function is fixed at the training points and is thus a member of the subset of functions in the function space that agree with the training data. We additionally show the posterior process mean $m_x(t), t \in T^*$ (dashed curve) and the 95% confidence interval $m_x(t) \pm 2\sqrt{k_x(t,t)}, t \in T^*$ (gray shaded area).

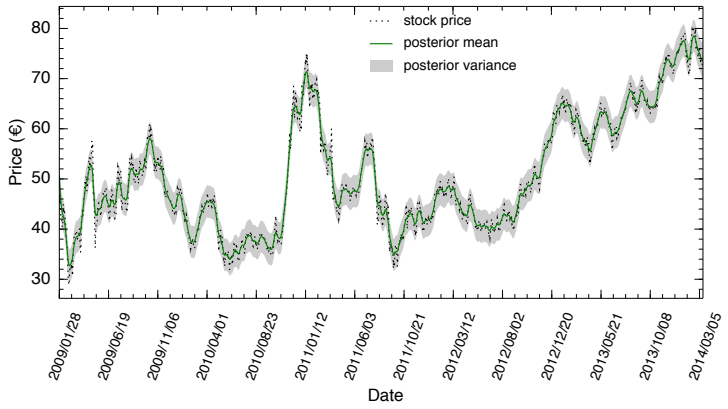


(a) $l = 0.25$ (b) $l = 0.5$ (c) $l = 1$

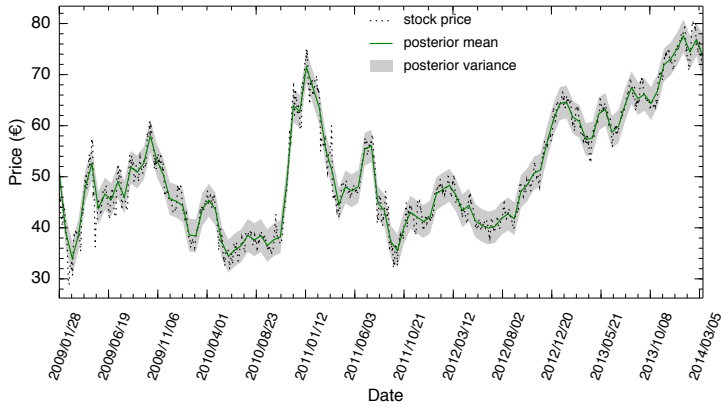
Figure 6.1: Realizations of random functions drawn from a GP prior with constant signal noise $\sigma^2 = 0.2$ and differing length scales.

Note how this takes on our understanding in section 5.3.1 of a Gaussian process being a prior over the function space again.

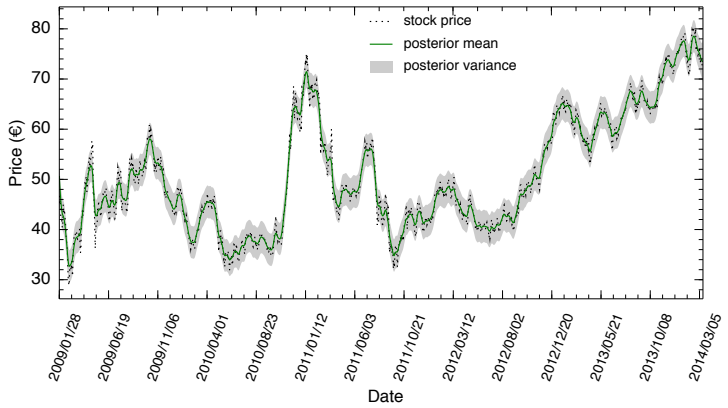
Example 6.1.2. Coming back to our motivating example of stock prices, we show a Gaussian process with Wiener covariance function conditioned on observations taken from Example 5.1.1 (the Porsche stock price time series). We selected a uniformly distributed subset of the observations and conditioned unknown locations on these observations. Again, we used a discretized surrogate of the true path of the underlying stochastic process for that. For comparison we show a path discretization of size 1000, 100 and number of observations respectively in Figure 6.2. For each figure, we show the resulting mean (green curve) and and two standard deviations as given by the diagonal of the posterior covariance matrix. As is obvious, the complexity of path discretization scarcely affects the result, i.e. a reasonable approximation of the realization of a stochastic process is possible by estimating the process state at only a limited number of locations.



(a) Path discretization with 1000 locations.



(b) Path discretization with 100 locations.



(c) Path discretization with same number of locations as there are observations (53 in this case).

Figure 6.2: Posterior Gaussian process approximating the distribution over paths. We show the actual data (dotted curve), mean (green curve) and two standard deviations (shaded region) of the posterior process.

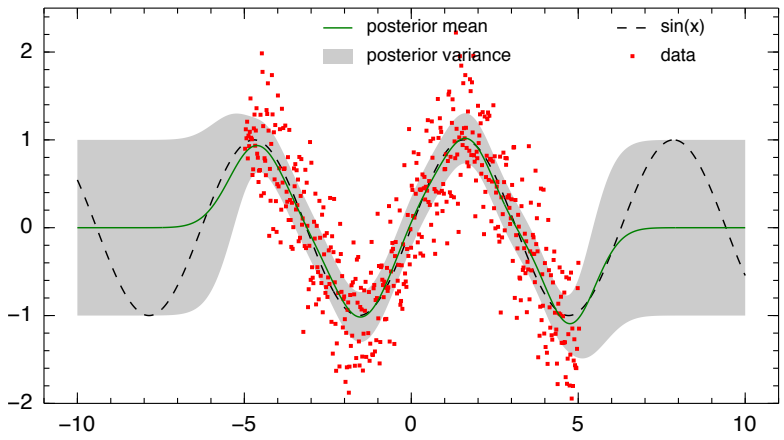


Figure 6.3: A fitted Gaussian Process. Data points are taken from $\sin(t)$ and independent noise has been added. We show the predictive mean two standard deviations of the predicted variance and the original sine function together with the observed data.

Example 6.1.3. Another example is given in Fig. 6.3 where we have fitted a Gaussian process with squared exponential covariance function to noisy observations of a sine wave. We generated observations y by taking the sine value $\sin(\cdot)$ at locations $t \in \hat{T}$ and adding a noise term ϵ with $\epsilon \sim \mathcal{N}(0, .5)$. After that we took the interval $[-10, 10]$ and extracted 1000 uniformly distributed location points that act again as the set of discretized locations of path components T^*. Fig. 6.3 shows the data points as red dots and the posterior mean as green curve together with two standard deviations derived from the posterior covariance shaded in gray. Figure 6.4 shows three sampled paths from the posterior GP. For comparison we also show the original sine function (dashed curve) in both figures.

6.2 Filtering

Filtering in general is part of a larger field of research known as *control engineering* or *control theory* where it is used to approximate the state of some dynamic system given a limited number of uncertain measurements of this state. Given a time series of data, the task of predicting the process at time t_i, given noisy observations for times $t_0, \ldots, t_i$ is called *filtering*. Estimating the state of the process at time t_l is called *prediction* if $l > i$ or *smoothing* if $l < i$. The resemblance to our previous remarks on the usage of stochastic processes for time series analysis is obvious. As a matter of fact, Kalman (1960) developed a filter as an efficient way to solve what is known as the *Wiener problem* - to estimate the state of a dynamic system given noisy observations under the assumption that its state evolution is governed by a Wiener process. We will first introduce this *Kalman filter* and then give some remarks on state space models, a methodology to model dynamic models under the influence of noise. Based on that, we review a method to transform a certain class of Gaussian process covariance functions in such a way that we can derive algorithms

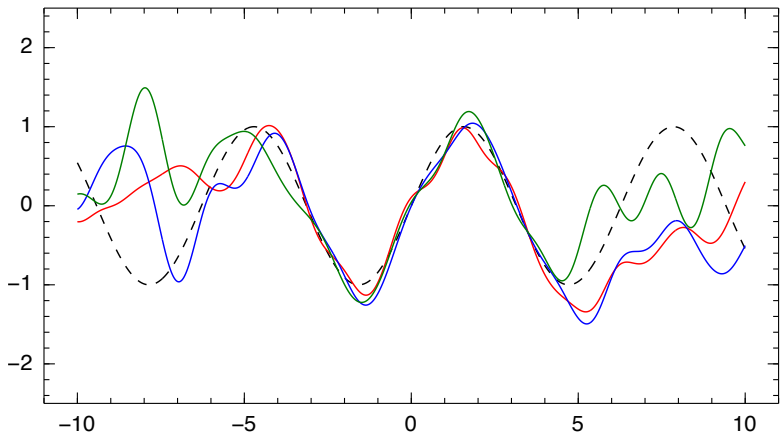


Figure 6.4: Three sample paths of the posterior Gaussian process in Figure 6.3. We again show the original sine function (the dashed curve) for comparison.

resembling the main idea of filtering, making them much more efficient than the exact solution (of complexity $\mathcal{O}(n^3)$) described in the previous section.

6.2.1 The Kalman filter

The Kalman filter was introduced by Kalman (1960) as a recursive estimator to solve the *Wiener problem*. It is essentially a recursive and more effective inference algorithm for a GP that is Wiener. The actual task of filtering consists of two steps: a) an update step, dependent on the previous state, called the *time update* and b) an update step, dependent on the current measurement, called the *measurement update*. Jointly, a) and b) are sometimes called the forward step. To refine the result, we can reestimate the state after we have learned about all measurements available, computing a smoothing distribution for the state of the system on the way. This is the backward step. Effectively, we refine the state estimate for a given time t_i that has been derived from measurements at times t_s, $s \leq i$ with information from the future, i.e. given at times t_r, $i < r \leq |T|$.
Let $x = \{x_t : t \in T\}$ be a path of a stochastic process and $y = \{y_t : t \in \hat{T}\}$ be observations of this process, where $\hat{T} \subseteq T$ and T be ordered as before. The Kalman filter model assumes the state to recursively evolve from previous states via the following recursion (the time update):

$$x_t = \Phi_t x_{t-1} + w_t, \tag{6.2}$$

where Φ_t is a linear operation called the *state transition* and w_t is Gaussian white noise with $w_t \sim \mathcal{N}(0, Q_t)$ and Q_t the noise covariance. w_t is called the *process noise* (cf. 5.3). The model of observations (the measurement update) assumes that

$$y_t = H_t x_t + v_t, \tag{6.3}$$

where H_t is the observation model that maps the state space to the observation space and v_t is the *measurement noise* or *observation noise* (cf. 6.1). Again, the noise is white, zero-mean Gaussian with measurement noise covariance R_t. Following section 6.1, from now on we assume H_t to be the identity matrix that leaves x_t unaltered, i.e. $H_t = \mathbb{I}$.

Forward step The forward step consisting of the time and measurement updates computes an a-posteriori state $\{\hat{x}_t : t \in T\}$ of the system, i.e. its state after incorporating knowledge from observations up to time t. A prior state of the system is computed by the time update:

$$\hat{x}_{t|t-1} = \Phi_t \hat{x}_{t-1|t-1}, \tag{6.4}$$

where $\hat{x}_{t-1|t-1}$ is the previous posterior state estimation, emphasizing the recursive nature of the algorithm. In case of a driving canonical Wiener process with drift zero, $\Phi_t = \Phi = \mathbb{I}$. The prior covariance $P_{t|t-1}$ is also computed in this step:

$$P_{t|t-1} = \Phi_t P_{t-1|t-1} \Phi_t^T + Q_t, \tag{6.5}$$

where $P_{t-1|t-1}$ is the previous posterior covariance estimation and Q_t is the process noise covariance.

The measurement update then computes the posterior state of the system after incorporating the measurement at time t. First, we compute the measurement residual, the difference between the prior state estimate and the current measurement

$$\begin{aligned} r_t &= y_t - H_t \hat{x}_{t|t-1} \\ &= y_t - \hat{x}_{t|t-1}, \end{aligned}$$

together with the residual covariance, the prior covariance plus the measurement noise

$$\begin{aligned} S_t &= H_t P_{t|t-1} H_t^T + R_t \\ &= P_{t|t-1} + R_t. \end{aligned}$$

From this, the *optimal Kalman gain* is obtained:

$$\begin{aligned} K_t &= P_{t|t-1} H_t^T S_t^{-1} \\ &= P_{t|t-1} S_t^{-1}. \end{aligned}$$

Using the Kalman gain, we can update the state and covariance prior estimates to arrive at the state and covariance posterior estimate:

$$\begin{aligned} \hat{x}_{t|t} &= \hat{x}_{t|t-1} + K_t r_t \\ &= \hat{x}_{t|t-1} + P_{t|t-1}(P_{t|t-1} + R_t)^{-1}(y_t - \hat{x}_{t|t-1}) && (6.6) \\ P_{t|t} &= (\mathbb{I} - K_t H_t) P_{t|t-1} \\ &= (\mathbb{I} - P_{t|t-1}(P_{t|t-1} + R_t)^{-1} P_{t|t-1} && (6.7) \end{aligned}$$

Backward step The backward step (also called the smoothing step) computes a refined a posteriori estimate of the system state after incorporating all available observations $\{y_1, \ldots, y_T\}$. Smoothing corresponds to a backward recursion to refine the posterior estimations from the forward step (the filter estimation), initial settings are given by the last forward recursion step. There exist several approaches to smoothing, we concentrate on the popular Rauch-Tung-Striebel smoother here (Rauch et al., 1965). The outcomes of the backward step are the posterior state and covariance estimates $\{\hat{x}_{t|T} : t \in T\}$ and $\{P_{t|T} : t \in T\}$ conditioned on all observations, given by

$$\hat{x}_{t|T} = x_{t|t} + C_t(\hat{x}_{t+1|T} - \hat{x}_{t+1|t}) \tag{6.8}$$

$$P_{t|T} = P_{t|t} + C_t(P_{t+1|T} - P_{t+1|t})C_t^T \tag{6.9}$$

with $C_t = P_{t|t}\Phi_t^T P_{t+1|t}^{-1}$.

Solving the Wiener Problem

Recall the Lévy property of the Wiener process: given a Wiener process W_t, its increments $W_t - W_s \sim \mathcal{N}(0, t-s)$ are independent. Consider (ordered) noisy measurements y of a path x. For the sake of a more intuitive understanding and referring to chapter 5, we call y a time series of measurements, i.e., $y = \{y_t : t \in T^*\}$ with t representing time and T^* totally ordered. Assuming that the measurements are of a dynamic model driven by a canonical Wiener process we set the linear state transformation Φ in Eq. 6.2 to $\mathbb{I}$, i.e., there is no drift. Following previous sections, we further define the difference in time between t_i and t_{i-1} as δ_t. The process noise thus becomes $w_t \sim \mathcal{N}(0, \delta_t)$, i.e., the diffusion function is constant 1. We can now proceed to solve the Wiener problem and make predictions of the system state at unknown time locations, i.e., approximate the path of the underlying stochastic process that gave rise to the measurements. Our running example will again be that of stock prices introduced in Example 5.1.1. In Figure 6.5 we show the outcome of a forward-backward sweep of a Kalman filter (i.e. after filtering and smoothing) using again every 25th observation as in Example 6.1.2. We note that, as expected, the outcome of this algorithm is identical to that of a GP with Wiener covariance function and predicted path state at observation time locations as shown in Figure 6.2 (c).

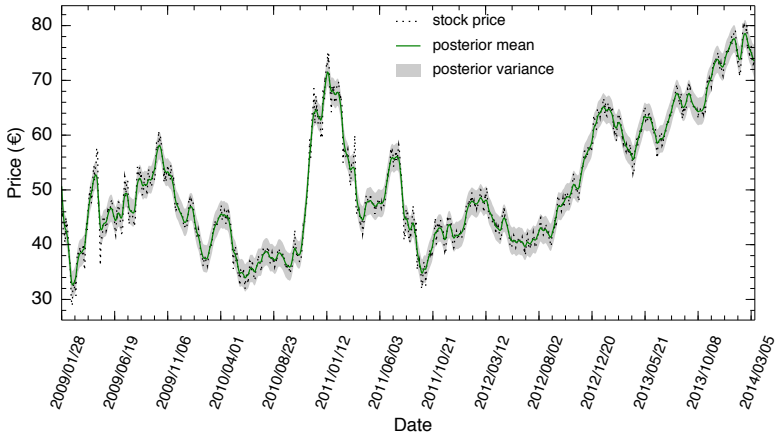


Figure 6.5: Using data from Example 5.1.1 we computed the backward (smoothed) system state of a Kalman filter. We show the resulting smoothed state (green curve) and the two standard deviation tube derived from the smoothed variance (gray shaded area) together with the original data (dotted curve).

Part III

Dynamic Topic Models

Abstract

In the following chapters we will bring together the approaches and mathematical machinery described in parts I and II. We start with a basic definition of what dynamic topic models are and in which way we understand dynamics in this setting by reviewing related work, including both parametric and nonparameteric models. Note that all previous work concentrates on the model structure as such, the type and behavior of motion in time is described only vaguely. We then examine an existing topic model that assumes topics to follow a Brownian motion through time and handles inference using a variational Kalman filter approach (Blei and Lafferty, 2006; Wang et al., 2008). We study this inference approach and then show how we can extend it to use general Gaussian processes to model the topics' motions on the simplex. Again, we lay particular emphasis on the inference procedure. We will report experiments on the data sets described in section 1.3. In particular, we will test the predictive abilities of the model in terms of held-out likelihood, predictive likelihood and on predicting timestamps given the document content and a learned model. We further add a new measure of how to identify particularly interesting terms in a learned topic model, based on previous findings concerning word volatility (see Heyer et al., 2009) and present our findings for the given data sets.

Finally, we conclude our findings.

Chapter 7

Continuous-Time Dynamic Topic Model

There are different approaches to modeling the behavior of topics over time. Loosely, we can differentiate between two main ideas here: (a) models that assume topics to be static semantic concepts (event-like) that are used uniquely (or at least irregularly) over the period of analysis and (b) models that allow for a dynamic change of topics by modeling a "movement" on the word simplex over time. Before going into detail, we give a brief summary of related work that uses topic modeling to analyze time-stamped text data and point out how it relates to ours.

7.1 Related Work

Srebro and Roweis (2005)'s approach is probably the first to appear throughout the literature. They make use of a dependent Dirichlet process where the dependency is modeled by a Gaussian process or is order based (see Griffin and Steel, 2006b). Topics in their model are found using the time marginal Dirichlet process. While this approach uses Gaussian processes for modeling topic covariance in time, it relies on topics as stable concepts, rather specific events, and thus models the probability of their concurrent occurrence. Another approach is that of (Zhu et al., 2005) who introduce a simple (asymmetric) time kernel to model influence of documents on a topic as a decaying function in time. In their model, document clusters from the past have influence on the decision to which cluster a document is assigned at present. They use an exponential decay function, not unlike an asymmetric Ornstein-Uhlenbeck covariance kernel (with no covariance into the past). Wang and McCallum (2006)'s approach uses a parametric model to also find event-like topics whose usage in time depends on the additional data of the time-stamp. In particular, the time-stamp influences a document's distribution over topics and finally leads to a posterior distribution over a topic's usage in time. They show that the topics found in this way describe more event-specific details than vanilla LDA (i.e. the model in section 2.2.1) and that their usage in time can be reasonably modeled with a Beta distribution. However,

they assume this distribution to be unimodal, which is often but not necessarily true, especially when the extracted topics cannot be related to a specific event and describe a more general semantic concept. Continuing their work, Walker et al. (2012) introduce Topic over Nonparametric Time, a model that replaces the topics' Beta distributions over time by a Dirichlet process mixture of normals, leading to a much richer class of densities over time that can be described by this approach. Again, they treat topics as stable concepts over time and do not allow for a change or shift of topics as time moves on.

The Dynamic Topic Model (DTM) introduced in Blei and Lafferty (2006) is the seminal work on which we base our research. Here, topics cease to explicitly represent event-like stable concepts but rather generalize to semantic concepts that may be interpreted but naturally undergo some change, be it because the whole concept's semantics, or the vocabulary used to refer to one and the same concept changes. Other related work that is based on this assumption was done by Caron et al. (2007) who make use of Blei and Lafferty (2006)'s methodology in a nonparametric setting and pick up Srebro and Roweis (2005)'s idea of correlating Dirichlet process mixture models in time. They use the Markov assumption on Dirichlet process mixtures (DPM) and thus model their covariance structure using a Gauss-Markov process. Wang et al. (2008) extend Blei and Lafferty (2006)'s model to a continuous time setting and call it the Continuous-Time Dynamic Topic Model (cDTM). It is this model that we extend and generalize in the following. Ahmed and Xing (2012) introduce the infinite dynamic topic model, a nonparametric model with unbounded number of topics per epoch and topic evolution based on Brownian motion. This model is based on the Hierarchical Dirichlet process model (Teh et al., 2006) which effectively is a DPM of Dirichlet processes. (Zhang et al., 2010) extend this approach to multiple data sources, finding time correlations between topics in different corpora.

As becomes clear, there exist numerous different models that take time information into account. We do not intend to explicitly assume event-like topics that have individual distributions over time indicating their usage but rather concentrate on the notion of dynamic topics that are allowed to develop over time (model type (b) above). Further, we are generally more interested in the *type* and *behavior* of topic evolution through time rather than on the specific model structure. In the following we will make use of the techniques described in part II and examine the information about topic evolution that we are able to extract from time-stamped document collections.

7.2 The Model

For the sake of clarity, and as it serves as a starting point for our investigation, we first review the cDTM introduced in Wang et al. (2008). The cDTM is a dynamic document model that assumes a constant topic drift, only depending on the difference between two points in time (i.e. the drift function of the associated SDE is constant one). The authors model that drift by introducing normally distributed increments for topics whose variance depends on their time difference. At each point in time, documents are then generated using the current state of the topic at that time. The generative process for documents at

time t in this model is as follows:

1. for all $k = 1, \ldots, K$ draw topics $\beta_{k,t} \sim \mathcal{N}_V(\beta_k^{t-1}, v\Delta_{s_t,s_{t-1}}\mathbb{I})$
2. for all documents $d^t \in D_t$
   (a) draw topic proportions $\theta_d^t \sim Dir(\alpha)$
   (b) for all words $w_{d,n}^t \in N_d^t$ in the document
      i. draw an assignment $z_{d,n}^t \sim Mult(\theta_d^t)$
      ii. draw a word $w_{d,n}^t \sim Mult(\pi(\beta_z^t))$

As the time marginal topic distribution β_k^t is normal, it is not conjugate to the multinomial[1] used to model the word-to-topic assignments in documents. Blei and Lafferty (2006) use a mean parameterization of the natural parameters to the multinomial. That is, the normal distribution is mapped onto the word simplex by a function $\pi(\cdot)$, where $\pi(\beta_k^t)_w = \frac{\exp\{\beta_{k,w}^t\}}{\sum_{w'}\exp\{\beta_{k,w'}^t\}}$, i.e. the prior on the topics is effectively a logistic normal (see Aitchison and Shen, 1980).

The resulting topic increments are thus $\beta_{1:K}^{t-1} - \beta_{1:K}^t \sim \mathcal{N}_V(0, v\Delta_{s_t,s_{t-1}}\mathbb{I})$ which shows that the underlying dynamics are in fact driven by a standard Brownian motion (cf. section 4.2.4).

Blei and Lafferty (2006) suggest to use a Kalman filtering scheme to facilitate inference in a model using these dynamics. The specific forward and backward step equations for the state space model are given in appendix B. Brownian motion, being a Gauss-Markov process, induces a multivariate normal distribution over time for all topics:

$$p(\beta_{k,1:T}|\mu_0, \sigma_0^2) = \mathcal{N}(\beta_{k,0}|\mu_0, \sigma_0^2)\prod_{t=1}^{T}\mathcal{N}(\beta_{k,t}|\beta_{k,t-1}, v\Delta_{s_t,s_{t-1}}) \tag{7.1}$$

with $k = 1, \ldots, K$. Implicitly, the authors kept the assumption of near independence as proposed by the Dirichlet prior on the topic distributions in the LDA model by modeling Eq. 7.1 for each word in each topic. A graphical representation as a plate model is given in Figure 7.1. We have omitted all indices besides time (t_-, t, t_+) for clarity.

7.2.1 Inference

Wang et al. (2008) use a variational scheme for inference in this model, turning inference into an optimization problem. As usual (cf. section 3.3), it consists of two steps: optimizing local variational parameters (those that are document-specific) and global variational parameters (the corpus-wide variational parameters, i.e. the topics). Iterating between

[1] A conjugate prior distribution has the property to produce a posterior distribution in the same family as itself when multiplied by the likelihood distribution. The conjugate distribution to the multinomial is the Dirichlet distribution as is exploited in the LDA model. For more on conjugacy see Bernardo and Smith (2009, chapter 5.2)

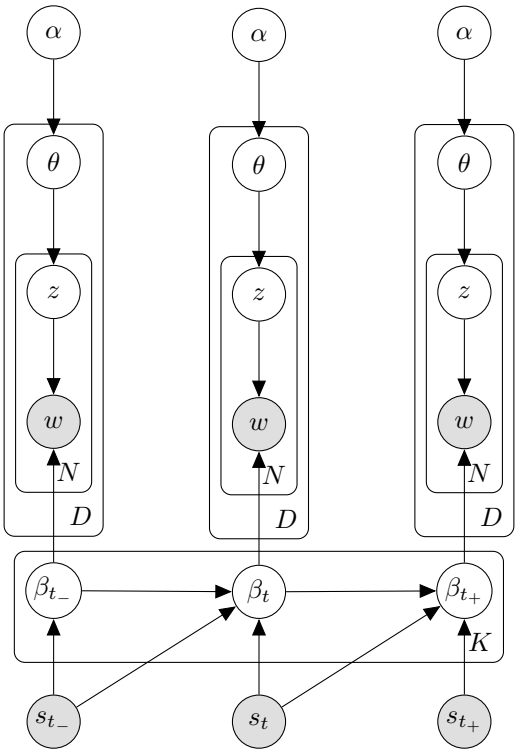


Figure 7.1: Continuous time dynamic topic model in plate notation.

these until convergence of the objective completes the algorithm. As mentioned above, the multinomial and the logistic normal are not a conjugate pair and thus not all full conditionals exist. The authors circumvent this by applying appropriate approximations. The first step is to formulate the optimization objective, i.e. the lower bound on the marginal distribution over the data (the ELBO)

$$\log p(w|\beta,\theta,z) \geq \mathcal{L} = \mathbb{E}_q\left[\log p(\beta,\theta,z,w)\right] - \mathbb{E}_q\left[\log q(\beta,\theta,z)\right], \tag{7.2}$$

with

$$\begin{aligned}\log p(\beta,\theta,z,w|\mu_0,\sigma_0^2,\alpha) = &\sum_{k=1}^{K}\sum_{w=1}^{V} p(\beta_{k,w}^{1:T}|\mu_0,\sigma_0^2) \\ &+ \sum_{t=1}^{T}\sum_{d=1}^{D_t}\log p(\theta_d^t|\alpha) + \sum_{n=1}^{N_d^t}\log p(z_{d,n}^t|\theta_d^t) + \log p(w_{d,n}^t|\pi(\beta_{z_{d,n}^t}^t))\end{aligned} \tag{7.3}$$

the joint model probability. The variational distribution is defined as

$$\log q(\beta, \theta, z) = \sum_{k=1}^{K}\sum_{w=1}^{V} \log q(\beta_{k,w}^{1:T}|\hat{\beta}_{k,w}^{1:T}) + \sum_{t=1}^{T}\sum_{d=1}^{D_t}\sum_{k=1}^{K}\left\{ \log q(\theta_{d,k}^{t}|\lambda_{d,k}^{t}) + \sum_{n=1}^{N_{t,d}} \log q(z_{d,n,k}^{t}|\phi_{d,n,k}^{t}) \right\} \tag{7.4}$$

with variational parameters $\hat{\beta}$, λ and ϕ to fit. Note that the usual assumption of complete independence among parameters in the variational distribution is not favorable in this case(this would render topics at different times independent). In the model this is reflected by the chaining of topics through employing a similar coupling as in the original βs on the variational $\hat{\beta}$s, i.e. to establish a drift on the variational parameters instead. Consequently, the parameters' distribution is similar to Eq. 7.1 and given by

$$\begin{aligned} q(\hat{\beta}_{k,w}^{1:T}) &= q(\hat{\beta}_{k,w}^{0}) \prod_{t=1}^{T} q(\hat{\beta}_{k,w}^{t}|\hat{\beta}_{k,w}^{t-1}, v\Delta_{s_t,s_{t-1}}) \\ &= \mathcal{N}(\hat{\beta}_{k,w}^{0}|m_{k,w}^{0}) \prod_{t=1}^{T} q(\hat{\beta}_{k,w}^{t}|\hat{\beta}_{k,w}^{t-1}, v\Delta_{s_t,s_{t-1}}). \end{aligned} \tag{7.5}$$

Following Blei and Lafferty (2006), we can interpret each $\hat{\beta}_{k,w}^{t}$ as a *noisy observation* at time t of the underlying stochastic process governing the drift of $\beta_{k,w}^{1:T}$. The tricky part is thus to determine the variational distribution $q(\beta_{k,w}^{1:T}|\hat{\beta}_{k,w}^{1:T})$ (for all k and w). From the Kalman filter model it is known that a noisy observation is given by the process state at the time of observation obfuscated by a zero mean white Gaussian measurement noise with some variance $\hat{\nu}^2$, i.e. the distribution of $\hat{\beta}_{k,w}^{t}$ given the true state $\beta_{k,w}^{t}$ is

$$\hat{\beta}_{k,w}^{t}|\beta_{k,w}^{t} \sim \mathcal{N}(\hat{\beta}_{k,w}^{t}|\beta_{k,w}^{t}, \hat{\nu}^2). \tag{7.6}$$

Using Bayes' law we can now easily derive the variational distribution which is given by

$$q(\beta_{k,w}^{1:T}|\hat{\beta}_{k,w}^{1:T}) = \frac{p(\hat{\beta}_{k,w}^{1:T}|\beta_{k,w}^{1:T})p(\beta_{k,w}^{1:T})}{p(\hat{\beta}_{k,w}^{1:T})} \tag{7.7}$$

$$= \frac{\prod_{t=1}^{T} p(\hat{\beta}_{k,w}^{t}|\beta_{k,w}^{t})p(\beta_{k,w}^{1:T})}{p(\hat{\beta}_{k,w}^{1:T})}, \tag{7.8}$$

and, as all distributions involved are normal, will be denoted by

$$q(\beta_{k,w}^{1:T}|\hat{\beta}_{k,w}^{1:T}) \triangleq \mathcal{N}(\tilde{m}_{k,w}^{1:T}, \tilde{V}_{k,w}). \tag{7.9}$$

The variational posterior is then treated as the outcome of the Kalman filter's backward step. Plugging Eq. 7.8 and Eq. 7.4 into Eq. 7.2 and rearranging terms leads to a reformu-

lated ELBO as given by

$$\begin{aligned}\mathcal{L} =&\mathbb{E}_q\left[\sum_{t=1}^{T}\sum_{d=1}^{D_t}\log p(\theta_d^t|\alpha)+\sum_{n=1}^{N_d^t}\log p(z_{d,n}^t|\theta_d^t)+\log p(w_{d,n}^t|\pi(\beta_{z_{d,n}^t,w_{d,n}^t}^t))\right]\\ &-\mathbb{E}_q\left[\sum_{k=1}^{K}\sum_{w=1}^{V}\sum_{t=1}^{T}\log p(\hat{\beta}_{k,w}^t|\beta_{k,w}^t)\right]+\mathbb{E}_q\left[\sum_{k=1}^{K}\sum_{w=1}^{V}\log p(\hat{\beta}_{k,w}^{1:T})\right]\end{aligned} \tag{7.10}$$

which we seek to optimize w.r.t. the variational parameters. The updates of the document level parameters are easily derived and resemble the updates as found in the inference procedure for LDA. In particular, the closed form update is given by

$$\begin{aligned}\lambda_{d,k}^t &= \sum_{w=1}^{V} n_{d,w}^t \mathbb{E}_q\left[z_{d,w,k}^t\right]\\ &= \sum_{w=1}^{V} n_{d,w}^t \phi_{d,w,k}^t.\end{aligned} \tag{7.11}$$

For the nonconjugate pair, the update is given by

$$\begin{aligned}\phi_{d,w,k}^t &\propto \exp\left\{\mathbb{E}_q\left[\log\theta_{d,k}^t\right]+\mathbb{E}_q\left[\log\pi\left(\beta_{k,w}^t\right)\right]\right\}\\ &\propto \exp\left\{\psi\left(\lambda_{d,k}^t\right)-\psi\left(\sum_{k'}\lambda_{d,k'}^t\right)+\mathbb{E}_q\left[\log\pi\left(\beta_{k,w}^t\right)\right]\right\}\end{aligned} \tag{7.12}$$

where the remaining expectation in Eq. 7.12,

$$\mathbb{E}_q\left[\log\pi\left(\beta_{k,w}^t\right)\right] = m_{k,w}^t - \mathbb{E}_q\left[\log\sum_{w'}\exp\left\{\beta_{k,w'}^t\right\}\right], \tag{7.13}$$

needs to be approximated as it does not exist in closed form. We could either apply a Taylor expansion around some additional variational parameter ζ_k^t on the second term as suggested by Blei and Lafferty (2006) or simply bound the expression from below using Jensen's inequality as has been done by Wang et al. (2008), completing the document level updates. The corresponding sub-procedure is also sometimes called the variational E-step (in reminiscence to the EM-algorithm (Dempster et al., 1977)) and is given in Algorithm 7.1. The returned values $\Xi_{t,k}$ and $\Upsilon_{t,k,w}$ are also called the sufficient statistics to the global topic distribution. We will need them when optimizing the global variational parameters.

For the corpus level parameters $\hat{\beta}$ a numerical optimization procedure is applied. In order to do so, we have to define an objective function, i.e. isolate those parts depending

Algorithm 7.1 Variational E-step in the cDTM model.

Require: β ▷ the current topic state
  $\forall k:$ compute logistic projection π_{β_k}
  for $t = 1$ to T **do**
    for $d = 1$ to D_t **do**
      while converged $<$ convergence criterion **do**
        $\forall k:$ set $\lambda_{d,k}^t$ to prior
        for $n = 1$ to N_d^t **do**
          for $k = 1$ to K **do**
            $\phi_{d,n,k}^t \leftarrow \exp\left\{\mathbb{E}[\log \theta_{d,k}^t] + \mathbb{E}[\log \pi(\beta_{k,w}^t)]\right\}$ ▷ Eqs. 7.12, 7.13
          normalize $\phi_{d,n,\cdot}^t$ across topics
          for $k = 1$ to K **do**
            update expectation: $\mathbb{E}[z_{d,n,k}^t] \leftarrow \phi_{d,n,k}^t$
        $\lambda_{d,k}^t \leftarrow \sum_{w=1}^V n_{d,w}^t \mathbb{E}[z_{d,w,k}^t$ ▷ Eq. 7.11
        converged $\leftarrow$ relative change of Eq. 7.10

  return sufficient statistics$\Xi_{t,k} = \sum_{d=1}^{D_t} \sum_{w=1}^V n_{d,w}^t \phi_{d,w,k}^t$, $\Upsilon_{t,k,w} = \sum_{d=1}^{D_t} n_{d,w}^t \phi_{d,w,k}^t$

on $\hat{\beta}$ from Eq. 7.2. When fixing k and w, the objective

$$\begin{aligned}
\mathcal{L}(\hat{\beta}_{k,w}) &= \mathbb{E}_q \left[\sum_{t=1}^{T} \sum_{d=1}^{D_t} \sum_{n=1}^{N_d^t} \log p(w_{d,n}^t | \pi(\beta_{z_{d,n}^t, w_{d,n}^t}^t)) \right] \\
&\quad - E_q \left[\sum_{t=1}^{T} \log p(\hat{\beta}_{k,w}^t | \beta_{k,w}^t) \right] + E_q \left[\log q(\hat{\beta}_{k,w}^{1:T}) \right] \\
&= \sum_{t=1}^{T} \sum_{d=1}^{D_t} \sum_{w=1}^{V} \sum_{k=1}^{K} n_{d,w}^t \phi_{d,w,k}^t \mathbb{E}_q \left[\pi(\beta_{k,w_{d,n}^t}^t)) \right] \\
&\quad - \sum_{t=1}^{T} \sum_{w=1}^{V} \sum_{k=1}^{K} -\frac{1}{2\hat{\nu}^2} E_q \left[(\hat{\beta}_{k,w}^t - \beta_{k,w}^t)^2 \right] \\
&\quad + \sum_{w=1}^{V} \sum_{k=1}^{K} \sum_{t=1}^{T} -\frac{1}{2v\Delta_{s_t, s_{t-1}}} (\hat{\beta}_{k,w}^t - \hat{\beta}_{k,w}^{t-1})^2 \\
&\quad + \text{constant}
\end{aligned} \tag{7.14}$$

where all terms that are independent of $\hat{\beta}_{k,w}$ have been absorbed into the constant. Note that we can again make use of the approximation as given in Eq. 7.13 in the first term on the right-hand side. By taking the gradient of Eq. 7.14, a gradient based numerical optimization algorithm can be used and completes the update of the global variational parameters. The key ingredient here is to treat the "latent observation" $\hat{\beta}_{k,w}^t$ as a function of

data set	model log likelihood	test log likelihood	∅ prediction error
terror (5 topics)	-70345762.7	-8.17	9.41
terror (10 topics)	-69861394.4	-8.19	9.43
riots (5 topics)	-22330251.6	-8.26	47.34
riots (10 topics)	-22214626.9	-8.32	46.47
autumn 2001 (10 topics)	-28194572.8	-8.06	0.96
autumn 2001 (15 topics)	-28334530.4	-8.14	1.00

Table 7.1: Results for the cDTM.

$\beta_{k,w}^t$ and to construct the gradient accordingly so that the objective is effectively optimized w.r.t. to the posterior state directly. Blei and Lafferty (2006) give an explicit derivation of the gradient used and we do not replicate it here, however we will derive a similar algorithm when working with general Gaussian processes. Iterating between updating the local and global parameters eventually converges to a local optimum of the ELBO and we can use the variational parameter settings to approximate the true posterior and work with it.

7.3 Experiments

For computation we have used the publicly available program code that is kindly provided by the author[2]. Their algorithm produces predictive distributions for all documents in the test set and reports the per-word likelihood, i.e. the test set performance. It further conducts a time-stamp prediction on these documents. We report this outcome as generated by the model for the datasets described in section 1.3 in Table 7.1. The "terror" and "riots" data set are both measured in months, i.e. there is a set of documents aggregated for each month. The "autumn 2001" data set aggregates documents per day. The average time-stamp prediction error is of the same scale as the data set. Note that the longer the time series, the higher the error becomes. Considering that the cDTM tries to find a set of semantically related words for each topic that is consistent over time, it naturally becomes harder to predict the time of a specific document given the topics. We give further example topics as generated by different runs using a 5 and 10-topic cDTM respectively. Note that we can only show topics at certain points in time, however, the distribution over words is known for all times (and will be of interest in the next chapter).

7.3.1 "Autumn 2001" Dataset

Table 7.2 shows words that are clearly connected to sports news coverage. Terms like "play", "game", "team", "season" etc. are the semantic "anchors" of the topics and words like "mets", "yankees" or "series" rise and fall. Intuitively this is perfectly reasonable

[2]http://www.cs.princeton.edu/ chongw/resource.html

9/1/2001	9/15/2001	9/30/2001	10/16/2001	10/31/2001	11/15/2001	11/30/2001
years	years	years	game	game	years	years
play	game	game	years	years	game	game
game	play	season	play	series	season	season
team	games	play	season	yankees	team	team
season	team	team	series	season	play	play
league	season	games	yankees	play	games	games
games	league	league	team	team	points	football
football	mets	yankees	games	games	playing	points
yankees	football	series	league	league	league	playing
playing	playing	football	manager	playing	football	league

Table 7.2: Top probability words for the sports topic.

9/1/2001	9/15/2001	9/30/2001	10/16/2001	10/31/2001	11/15/2001	11/30/2001
year	attacks	attacks	american	year	american	year
percent	american	united	united	american	year	united
states	united	american	states	officials	united	american
state	states	states	year	states	states	states
united	president	year	officials	united	taliban	percent
american	officials	percent	attacks	percent	president	officials
president	year	president	percent	president	percent	today
officials	today	officials	government	state	officials	state
million	percent	state	state	attacks	state	afghanistan
today	security	terrorist	president	today	afghanistan	taliban

Table 7.3: Top probability words for the President of the United States topic.

as on September, 21st the New York Mets played the first baseball game after the 9/11 attacks and in late October and early November the New York Yankees played the 2001 World Series finals (but lost against Arizona Diamondbacks). The words in Table 7.3 are about the 9/11 terror attacks on the World Trade Center. The words "attacks", "security", "terrorist" or "taliban" rise and fall whereas the terms "american", "united", "states" or "president" remain stable. Looking at both these topics reveals the twofold interpretation that is possible. For instance, the second example can be either viewed as a general topic about the President of the United States or as a topic describing the event 9/11 attacks.

7.3.2 "Terror" Dataset

As has been indicated by a larger average error in time-stamp prediction for this data set, the topics produced are less descriptive for a certain point in time. Handling longer time spans, makes it harder to assign words to topics and at the same time to preserve the semantic meaning of the overall topic. Table 7.4 shows a topic that is roughly centered around security. Its most probable words are covering the Israeli-Palestinian conflict in the beginning and are then influenced by the 9/11 attacks to cover air-travel related security, a

12/31/2000	9/1/2001	5/1/2002	1/1/2003	8/31/2003	5/1/2004
israeli	security	security	north	american	american
palestinian	airlines	school	united	military	iraq
sharon	officials	york	american	united	iraqi
minister	flight	students	nuclear	muslim	military
israel	airport	sept	states	americans	soldiers
peace	federal	coast	korea	soldiers	killed
palestinians	attacks	schools	trade	forces	saudi
prime	people	guard	world	army	people
today	planes	nuclear	international	officials	army
bank	passengers	system	countries	islamic	forces

Table 7.4: Top probability words for a topic roughly related to security.

12/31/2000	9/1/2001	5/1/2002	1/1/2003	8/31/2003	5/1/2004
american	street	street	street	city	intelligence
united	show	show	york	people	officials
officials	york	york	show	security	information
court	theater	film	theater	york	report
laden	music	west	west	department	prisoners
states	museum	directed	city	officials	qaeda
trial	center	life	tickets	attack	attacks
case	tickets	time	center	passengers	united
judge	broadway	world	music	sept	military
prosecutors	avenue	play	work	manhattan	states

Table 7.5: Top probability words for a switching topic.

possible nuclear threat and the relationship to North Korea and finally handle the beginning war in Iraq. Although a certain development is observable it does not resemble the usual notion of a topic in the topic modeling sense. The effect is even more extreme in the example shown in Table 7.5. Here, the key terms circle around "officials", "court", bin "laden", then switch to a Broadway related topic and then back to security and the beginning war in Iraq. We will come back to this example in section 8.1.

7.3.3 "Riots" Dataset

The "riots" data set is defined over the longest time span, covering 247 months of data. Looking at the previous example, one would expect the changing of semantic context in time to be even more severe. Indeed this is the case. Table 7.6 shows an example topic that starts off with high probability terms that describe problems with the Ku-Klux-Klan in Atlanta and then goes on with terms connected to the anti-abortion movement and after that violence in school. It stays stable for some time, covering the Israeli-Palestinian conflict and then turns to violence in general. Clearly, the terms that appear over time have no semantic relation (besides that they are all related to demonstrations and riots)

12/1/1986	4/1/1990	7/31/1993	12/1/1996	4/1/2000	8/1/2003	12/1/2006
county	abortion	mayor	city	people	israeli	people
atlanta	square	dinkins	people	house	palestinian	group
people	people	city	local	violence	israel	country
king	movement	crown	palestinian	israeli	people	local
group	court	people	israeli	palestinian	hamas	city
house	israeli	abortion	economic	land	violence	sunday
american	group	violence	made	local	women	time
klux	prime	school	violence	small	time	violence
georgia	house	jewish	palestinians	million	killed	prime
weeks	today	group	streets	town	center	made

Table 7.6: Top probability words for a rapidly switching topic.

but point to certain sub-topics that were formed over sub-sets of time. We will expand this finding and give a qualitative interpretation of it in the next chapter.

Chapter 8

Gaussian Process Dynamic Topic Models

8.1 Word Volatility in Dynamic Topic Models

Recall Heyer et al. (2009)'s definition of word volatility as introduced in section 1.1. Highly volatile terms, i.e. those that rapidly change their co-occurrence context, are considered "hotly discussed" and can to some extent be compared to issues as something that gets attention by the media[1]. This means that their approach can be used to point out key terms related to them, but is limited to this raw statement. No conclusions on the precise semantic nature of the hypothetical issue nor on the "attention-cycle" it might be connected to can be drawn. We now develop this idea in the context of topic models. Consider a dynamic topic model as in the previous chapter. Assuming that topics that are evolving over time are affected by certain events[2] (reflected by the data) we utilize the interpretation of topics as semantic clusters of terms and deduce that the events that contribute to a topic's current state are also to some extent related to each other. This nicely resembles the definition of an issue as given in section 1.2 as the result is a stream of latent word clusters that are formed by analyzing documents that comprise events *and* language describing the bigger picture (what Kantner (2009) calls a "social problem"). Recall for instance the sports topic as given in Table 7.2. Although we have given a reasonable explanation as for why terms like "mets" or "yankees" do appear, this required some research about the baseball series in 2001 (a rather laborious task for a non-American to be honest). Considering the notion of word volatility this goal is reached much faster. The approach for a specific topic k is as follows: for every word w, compute the variance of the expected probability of w in topic k, the *topic word volatility*

$$v_{w,k} = \mathrm{var}(\mathbb{E}_q[\pi\left(\beta_k\right)_w^{1:T}]) \tag{8.1}$$

[1] A single volatile term does not form an issue, of course. However, hotly discussed terms can give valuable hints to what the actual issue might be.

[2] Events are not defined in a particular way here, in this case we understand an event simply as something that has happened and is covered by observed documents.

and its average over topics

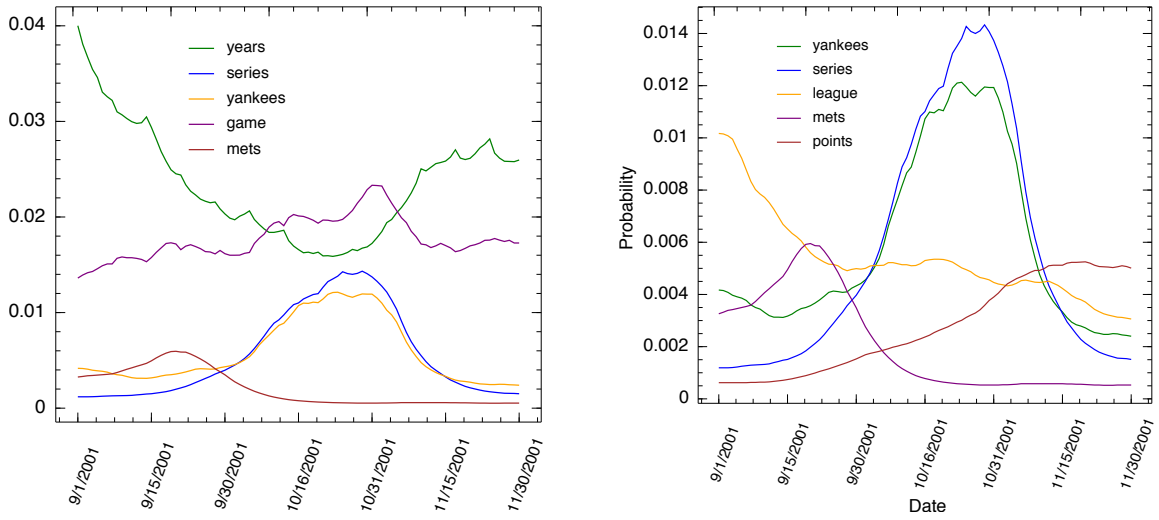
$$\bar{v}_{k,w} = \frac{\mathrm{var}(\mathbb{E}_q[\pi\,(\beta_k)_w^{1:T}])}{\sum_{k'} \mathrm{var}(\mathbb{E}_q[\pi\,(\beta_{k'})_w^{1:T}])]}. \tag{8.2}$$

Plain topic word volatility can be used to find the key terms that relate to a hypothesized issue, its average over topics is usable for identifying the events that contribute to the current state. In other words, we find terms whose expected probability varies most over time, both in a specific topic alone and in comparison to all other topics.

high volatility	high averaged volatility
years	yankees
series	series
yankees	league
game	mets
mets	points
points	team
points	manager
league	game
team	season
season	play

Table 8.1: Highly volatile terms in the sports topic.

For the topic we referred to above, the terms with highest topic word volatility are given in Table 8.1. *Words that describe the overall theme have higher volatility, those that are connected with specific events a higher averaged volatility.* A closer look at their expected probability value over time as in Fig. 8.1 then reveals the approximate time of the event. For this particular topic, the result is not too different from the topics shown in Table 7.2 as the covered time span is rather small. This is, however, not always the case. Turning back to the example topic from the "terror" data set given in Table 7.5, we are able to uncover the events that led to the development of the topic. Table 8.2 shows highly volatile terms in this topic. Figure 8.2 shows the probability paths of the top five words on the lists. The high volatility terms again describe the hypothetical issue, high average volatility terms identify events that contributed to the topic's state. The term "blackout" refers to the Northeast blackout on August 14th, 2004 and "madrid" to the Madrid bombing on March 11th, 2004, both are events that are security related.



(a) High topic word volatility.

(b) High averaged topic word volatility.

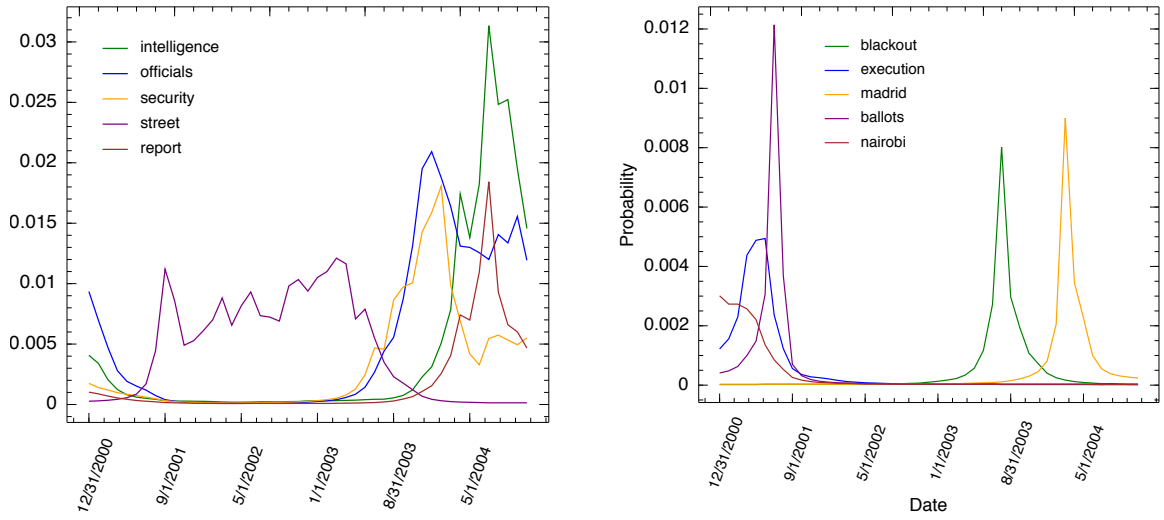
Figure 8.1: Probabilities of highly volatile terms in the example sports topic.

high volatility	high averaged volatility
intelligence	blackout
officials	execution
security	madrid
street	ballots
report	nairobi
house	tanzania
city	sand
american	mohamed
trial	grid
united	spanish

Table 8.2: Highly volatile terms in the switching topic.

8.2 Modifying the Time Dynamics

In the cDTM, the topic drift is modeled by a Brownian motion, i.e. it is completely random and its variance is unbounded (covariance between two points in time linearly grows with their distance). In a sense, this can be seen as an almost non-informative prior on the expected topic drift that is completely overwritten by the observed data. A process with bounded variance that is used in stochastic volatility models in econometrics is the previously introduced Ornstein-Uhlenbeck (OU) process (cf. section 4.2.6). The process'



(a) High topic word volatility.

(b) High averaged topic word volatility.

Figure 8.2: Probabilities of highly volatile terms in the example switching topic.

property of "drawing" its state towards a long-term mean and the ability to parameterize this behavior are of main interest to us here. The idea is to allow topics to change considerably in a short amount of time but only temporarily and to reflect that in the stochastic process used as a prior. Probabilities are rapidly reverted to their mean value thus favoring terms that describe a sudden change in the reception or treatment of a semantic concept without changing the topic's original meaning. However, using specific stochastic processes it is possible to encode certain assumptions directly into the prior on the topic drift process. Akin to expecting only a few words to characterize a topic and thus using a Dirichlet prior on it (besides the mathematical simplifications that come with this), we expect words to considerably change their probability in a topic (e.g. an event-specific term) from low to high and back to low (when the event is over). Assuming a mean value of zero, an Ornstein-Uhlenbeck process describes such a behavior. Our first intuition is thus to model term covariance in time by the appropriate covariance matrix (cf. section 5.3.2). However, the exponential nature of the corresponding covariance function creates a rapid decline in covariance which quickly approaches zero (depending on the parameters of the function). This implicates the risk of losing semantic coherence. When the covariance for all words drops too fast, the behavior of words in a topic might become erratic, destroying the semantic clustering. To relax the strong assumption of the OU process we can resort to custom covariance computations (see Rasmussen and Williams, 2006, chapter 4). For instance, given two covariance matrices K_1 and K_2, their sum $K_1 + K_2$ (including weighted summing (e.g. Plate, 1999)) is again a valid (i.e. positive semi-definite) covariance matrix. Enriching a basic Brownian motion with the OU process may produce better results by allowing stable terms in a topic while preserving the intended volatile behavior. Additionally

we take other covariances into account. We examine whether a smoother function such as the squared exponential covariance function is also amenable to this problem setting and if periodic covariance functions can help to better find recurring events or even issues in the data. Before reporting our findings we extend the variational algorithm from section 7.2.1 to the more general Gaussian process setting. Given the algorithm, changing the underlying stochastic process (or combinations thereof) by modifying the respective covariance functions is trivial.

8.3 The Model

We develop a dynamic model based on the general definition of a Gaussian process, i.e. the ability to describe a realization of a GP as a multivariate normal distribution and to work with that instead. Doing so lets us easily switch between different processes. Recall that the Gaussian process is fully defined by its covariance function. Modifying it thus automatically changes the underlying process. For this study, we consider the same document model as the cDTM, i.e. we assume that topic proportions θ_d^t for document d at time t are draws from a Dirichlet prior, word assignment $z_{d,n}^t$ of word n in document d is drawn from a multinomial distribution over the K topics parameterized by θ_d^t, and word $w_{d,n}^t$ in document d at position n is drawn from the logistic projection $\pi(\cdot)$ of the normally distributed topic variable $\beta_{z_{d,n}^t}^t$ onto the word simplex. Note however that the general procedure is not bound to this model, any document model that is able to work with (or transform) the normally distributed topic variable may as well be used. The generalization affects the objective function in Eq. 7.14 that needs to be modified in order to incorporate a Gaussian process into the model. We define a generative process for the Gaussian Process Dynamic Topic Model (GPDTM) and then proceed to a strict description of the inference algorithm used for learning in the model. Note that we follow Blei and Lafferty (2006) in their assumption of independence between individual words in topics. As they have done, we treat each word in every topic separately and place a GP prior on its evolution over time.

1. for all $k = 1, \ldots, K$ and $w = 1, \ldots, V$ draw probability path $\beta_{k,w}^{1:T} \sim \mathrm{GP}(0, K)$
2. for all times t in T
   (a) for all documents $d^t \in D_t$
      i. draw topic proportions $\theta_d^t \sim Dir(\alpha)$
      ii. for all tokens $n = 1, \ldots, N_d^t$ in the document
         A. draw an assignment $z_{d,n}^t \sim Mult(\theta_d^t)$
         B. draw a word $w_{d,n}^t \sim Mult(\pi(\beta_{z_{d,n}^t, w_{d,n}^t}))$

Without loss of generality (Rasmussen and Williams, 2006, cf.), we assume a zero mean Gaussian process. The covariance matrix K is derived from a covariance function of time

$k(\cdot,\cdot)$ as described in section 5.3.2 such that $K_{i,j} = k(t_i, t_j), t_i, t_j \in T$. Besides Brownian motion, all induced stochastic processes will be stationary.

8.4 Inference

Inference for the global topic variables includes again numerical optimization of the objective function in Eq. 7.14, where the distribution of each $\hat{\beta}_{k,w}^{1:T}$ is now given by

$$\log q(\hat{\beta}_{k,w}^{1:T}) = -\frac{T}{2}\log(2\pi) - \log|K + \hat{\nu}\mathbb{I}| - \frac{1}{2}(\hat{\beta}_{k,w}^{1:T})^T(K + \hat{\nu}\mathbb{I})^{-1}(\hat{\beta}_{k,w}^{1:T}) \tag{8.3}$$

instead of the Markov chain representation as in Eq. 7.9. Note that adding noise $\hat{\nu}^2$ to the covariance matrix corresponds to assuming noisy observations $\hat{\beta}$ (as also defined in the state space model in Eq. 7.6). The optimization objective for the global variational parameters $\hat{\beta}$ is

$$\begin{aligned}\mathcal{L}(\hat{\beta}_{k,w}) = \mathbb{E}_q &\left[\sum_{t=1}^{T}\sum_{d=1}^{D_t}\sum_{n=1}^{N_d^t}\log p(w_{d,n}^t|\pi(\beta_{z_{d,n}^t,w_{d,n}^t}^t))\right] \\ &\underbrace{- E_q\left[\sum_{t=1}^{T}\log p(\hat{\beta}_{k,w}^t|\beta_{k,w}^t)\right]}_{\text{term 1}} + \underbrace{E_q\left[\log q(\hat{\beta}_{k,w}^{1:T})\right]}_{\text{term 2}}.\end{aligned} \tag{8.4}$$

We will concentrate on deriving the marked terms which are different from the objective function given for the cDTM. Term 1 can be expanded to

$$E_q\left[\sum_{t=1}^{T}\log p(\hat{\beta}_{k,w}^t|\beta_{k,w}^t)\right] = -\frac{1}{2\hat{\nu}^2}E_q\left[(\hat{\beta}_{k,w} - \beta_{k,w})^T(\hat{\beta}_{k,w} - \beta_{k,w})\right] + \text{constant}$$

where we have transformed it to a multivariate normal distribution and absorbed all terms independent of β into a constant. From the marginalization property (cf. 4.2.3) and assuming the variational parameters $\hat{\beta}$ to be noisy measurements to a Gaussian process, we can again treat the $\hat{\beta}$ as functions of β. Optimizing w.r.t. $\hat{\beta}$ is thus again equal to optimizing w.r.t. β directly. The measurements as derived from the GP posterior are given by

$$\hat{\beta}_{k,w} = (K + \hat{\nu}^2\mathbb{I})K^{-1}(\beta_{k,w} - \mu_{k,w}) \tag{8.5}$$

where $\mu_{k,w}$ is simply the mean of the current posterior state transformed to the appropriate column vector. Note that in this case both covariance matrices K are identical, as we are obviously deriving measurements only at those points in time where a certain term w appears (resembling the sparse nature of the cDTM). For notational convenience, we introduce the auxiliary variable $\Sigma_0 = (K + \hat{\nu}^2\mathbb{I})K^{-1}$ and proceed with expanding term 1

to

$$\begin{aligned} E_q\left[\sum_{t=1}^{T}\log p(\hat{\beta}_{k,w}^t|\beta_{k,w}^t)\right] &= -\frac{1}{2\hat{\nu}^2}E_q\left[(\Sigma_0(\beta_{k,w}-\mu_{k,w})-\beta_{k,w})^T((\Sigma_0(\beta_{k,w}-\mu_{k,w})-\beta_{k,w})\right] \\ &= -\frac{1}{2\hat{\nu}^2}E_q\left[((\Sigma_0-\mathbb{I})\beta_{k,w}-\Sigma_0\mu_{k,w})^T((\Sigma_0-\mathbb{I})\beta_{k,w}-\Sigma_0\mu_{k,w})\right] \\ &= -\frac{1}{2\hat{\nu}^2}((\Sigma_0-\mathbb{I})m_{k,w}-\Sigma_0\mu_{k,w})^T((\Sigma_0-\mathbb{I})m_{k,w}-\Sigma_0\mu_{k,w})+\text{constant} \end{aligned} \tag{8.6}$$

where we have again absorbed terms independent of β into the constant. The corresponding gradient is now easily found to be

$$\frac{\partial}{\partial m_{k,w}}E_q\left[\sum_{t=1}^{T}\log p(\hat{\beta}_{k,w}^t|\beta_{k,w}^t)\right] = -\frac{1}{\hat{\nu}^2}(\Sigma_0-\mathbb{I})^T((\Sigma_0-\mathbb{I})m_{k,w}-\Sigma_0\mu_{kw}). \tag{8.7}$$

We now turn to the second term of interest in the objective function. Term 2 can be expanded using the standard multivariate normal form for the realization of a GP and yields

$$E_q\left[\log q(\hat{\beta}_{k,w}^{1:T})\right] = -\frac{1}{2}E_q\left[\hat{\beta}_{k,w}^T K^{-1}\hat{\beta}_{k,w}\right]+\text{constant}.$$

Using identity 8.5 we can retrieve $\hat{\beta}$ as above and term 2 yields

$$\begin{aligned} E_q\left[\log q(\hat{\beta}_{k,w}^{1:T})\right] &= -\frac{1}{2}E_q\left[(\Sigma_0(\beta_{k,w}-\mu_{k,w}))^T K^{-1}(\Sigma_0(\beta_{k,w}-\mu_{k,w}))\right] \\ &= -\frac{1}{2}(\Sigma_0(\beta_{k,w}-\mu_{k,w}))^T K^{-1}(\Sigma_0(\beta_{k,w}-\mu_{k,w}))+\text{constant}. \end{aligned} \tag{8.8}$$

The corresponding gradient is then

$$\frac{\partial}{\partial m_{k,w}}E_q\left[\log q(\hat{\beta}_{k,w}^{1:T})\right] = -\Sigma_0^T K^{-1}\Sigma_0(m_{k,w}-\mu_{k,w}) \tag{8.9}$$

which enables us to use a gradient based numerical optimizer to maximize the ELBO[3]. As has been briefly mentioned in section 6.1, the computations associated with learning in Gaussian processes are rather costly. As the covariance matrices are not dependent on β in our case (as they only depend on the known observation times) they can be precomputed, greatly improving performance. Further, both the variational e-step (Algorithm 7.1) and the global optimization can be parallelized when appropriate resources are available. The computation for the former distributes across documents in each time step and the latter across words in each topic.

[3]We have tested both the Fletcher-Reeves conjugate gradient method (see e.g. Wright and Nocedal, 2006) and the RPROP algorithm (Riedmiller and Braun, 1993). Both have shown similar results whereas RPROP is slightly faster.

data set	best model	log likelihood	∅ ts error	∅ ts error cDTM
terror (5 topics)	bm+periodic	-3093063.1	9.14	9.41
terror (10 topics)	ou	-3127147.2	9.32	9.43
riots (5 topics)	ou	-268447.0	39.22	47.34
riots (10 topics)	ou	-227285.0	34.65	46.47
autumn 2001 (10 topics)	ou	-1960723.2	0.63	0.96
autumn 2001 (15 topics)	periodic	-1928392.0	0.58	1.00

Table 8.3: Predictive likelihood on the validation set for GPDTM. ou stands for the Ornstein-Uhlenbeck, bm+periodic for the sum of a Brownian motion and periodic covariance function.

8.5 Experiments

Following the methodology described in the previous chapter, different models have been computed for each data set. We report the predictive likelihood and the average time-stamp prediction error for the best model together with the outcome of the cDTM for comparison in Table 8.3. Note that values cannot be compared across rows but only between models run on the same data set. Further inspection of the topics produced revealed that often the semantic relatedness is *not* given over all time steps for a given topic. However, we can report semantically stable topics over subsections of the time line. To some degree this behavior must be expected due to the exponential form of the covariance functions used. Also, the large decrease in time-stamp prediction error can be explained in that way. If the high probability words in a topic are not bound by an overall semantic theme they will be optimized to be more concentrated in time, as this explains the data best. The results for the "terror" data set were not distinguishable from those obtained by the cDTM. The most probable explanation for this is that the "terror" data set comprises a (considerably) smaller time span than the other data sets (47 vs. 91 and 247). As we have used identical length scales for all exponential covariance function in our study, this suggests that the effect of modifying the time dynamic does not emerge on a smaller scale. We exclude this data set from further analysis.

In the following we present example topics (as produced by the corresponding best performing model) and charts of words with high topic word volatility that back our assumption of the presence of issue-like term clusters in the data. Again, this is our interpretation, there is no guarantee nor any claim that the found structures actually are what a political communication scientist calls issue.

8.5.1 "Autumn 2001" Dataset

Tables 8.4 and 8.5 show the top probability words and terms with highest volatility from a topic as generated by the best performing model. Here, highly volatile terms reasonably describe the key terms in the topic and are also reflected by the top probability words in

9/1/2001	9/15/2001	9/29/2001	10/13/2001	10/27/2001	11/10/2001	11/24/2001
people	life	people	american	york	back	music
president	people	world	officials	officials	people	work
world	president	school	united	anthrax	good	president
life	family	years	states	people	president	york
long	american	attacks	people	center	american	american
dont	home	told	anthrax	health	made	world
american	didnt	attack	laden	university	make	life
story	dont	president	world	department	told	film
book	young	american	afghanistan	public	store	story
made	found	back	saudi	state	dont	people

Table 8.4: Top probability words for the "anthrax" topic.

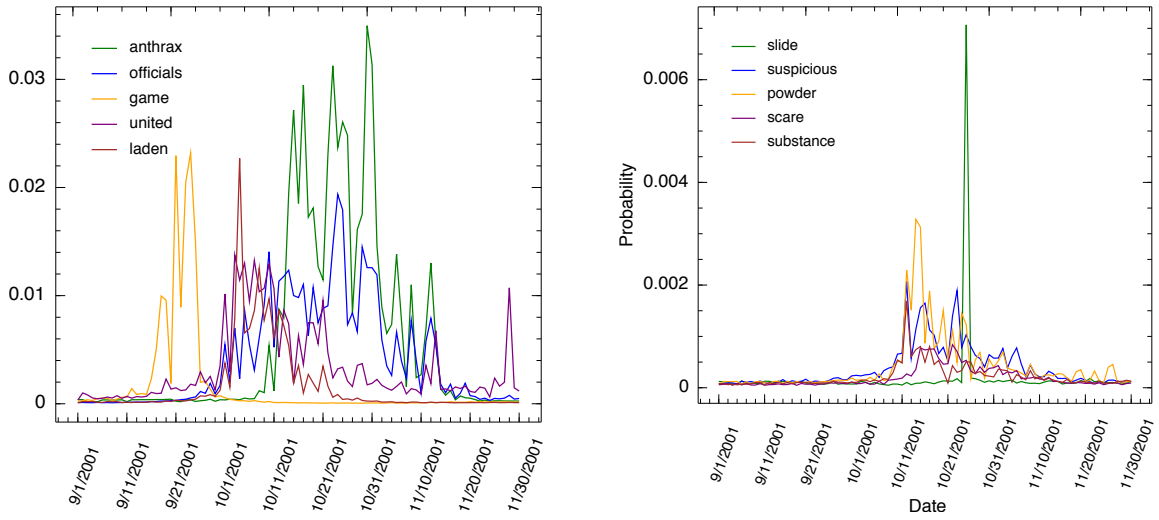
it. Terms with high average volatility describe a specific event, in this case the anthrax attacks shortly after the 9/11 attacks. High average volatility terms describe the event even further, using very rare words (for instance, the frequency of "scare" in the whole corpus is only 722, compared to 8460 occurrences of "anthrax").

high volatility	high averaged volatility
anthrax	slide
officials	suspicious
game	powder
united	scare
laden	substance
states	bioterrorism
american	contained
attacks	harry
street	facility
film	antibiotics

Table 8.5: Highly volatile terms in the "anthrax" topic.

8.5.2 "Riots" Dataset

Table 8.6 gives an example topic that is again composed of several time constrained subtopics. The most prominent one is again the Israeli-Palestinian conflict. Looking at the expected probability time series of highly volatile terms in this topic, we can give a qualitative interpretation of the analysis. As seen in Fig. 8.4, the terms "israeli", "palestinian" and "palestinians" experience a sharp rise in the beginning of 1987 and again in 2000. This coincides with the media coverage of the first and second Palestinian intifada and thus clearly resembles an issue-cycle. High average volatility terms do not provide



(a) High topic word volatility.

(b) High averaged topic word volatility.

Figure 8.3: Probabilities of highly volatile terms in the example "anthrax" topic.

any further insight in this case. Another example is given in Table 8.8. Here the terms "students" and "university" have high expected probability in 1997. Looking at the high average volatility terms for this topic, we can assume that this refers to the June Democratic Uprising in Seoul. The surge in "police" and "officers" can again be explained by consulting the high average volatility terms. It refers to the uprising in Indonesia that occurred in 1998.

12/1/1986	9/1/1990	6/1/1994	3/1/1998	12/1/2001	9/1/2005
today	military	people	people	palestinian	china
people	soviet	today	police	israeli	government
lead	congress	united	black	arafat	people
government	palestinians	government	city	palestinians	bush
violence	government	political	march	israel	united
years	today	black	mayor	israelis	president
protests	united	president	rally	people	states
week	states	states	officers	bank	political
time	indians	american	white	violence	american
political	minister	military	giuliani	today	officials

Table 8.6: Top probability words for the "intifada" topic.

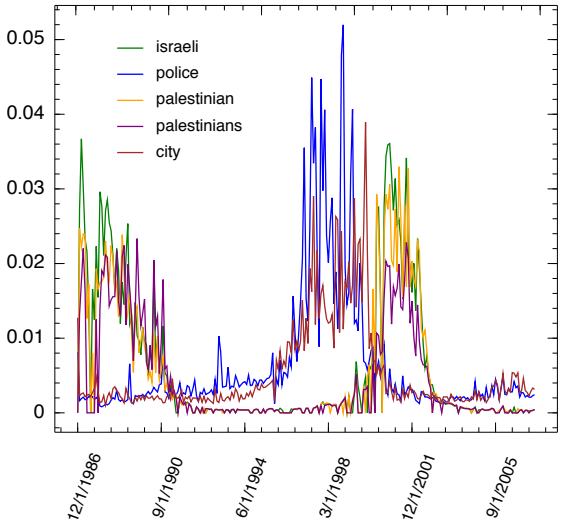


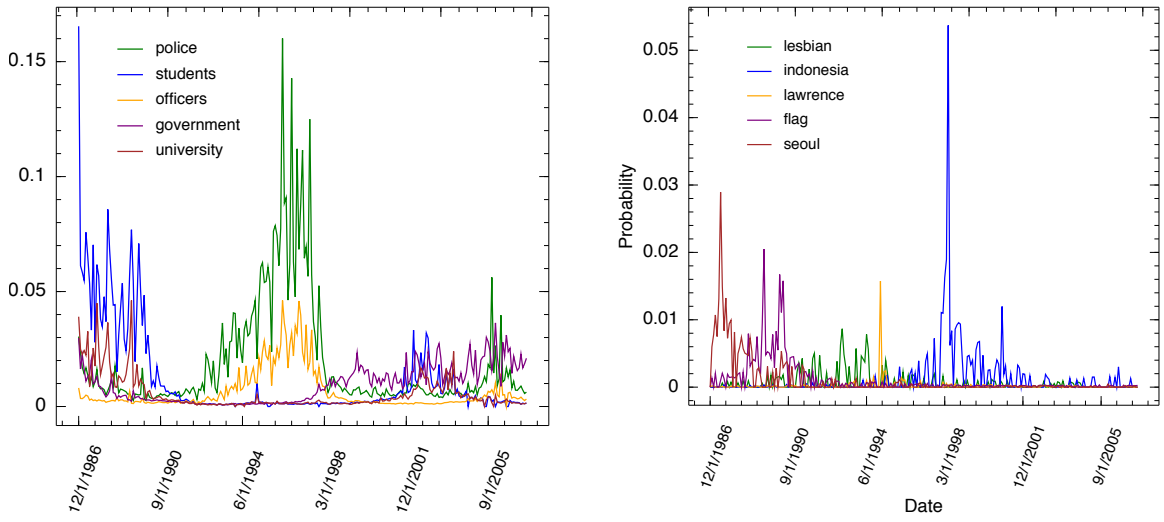
Figure 8.4: Probabilities of highly volatile terms in the example "intifada" topic.

12/1/1986	9/1/1990	6/1/1994	3/1/1998	12/1/2001	9/1/2005
students	president	march	police	government	government
student	bush	police	indonesia	economic	people
university	people	city	china	people	opposition
police	rights	people	people	students	police
government	students	games	chinese	groups	china
south	human	avenue	officials	protests	news
protest	house	park	human	china	movement
president	today	york	economic	years	political
today	political	lesbian	officers	movement	rights
protests	states	officers	government	rights	president

Table 8.7: Top probability words for the "indonesia" topic.

high volatility	high averaged volatility
police	lesbian
students	indonesia
officers	lawrence
government	flag
university	seoul
chinese	cuomo
china	activists
bush	followers
student	response
president	games

Table 8.8: Highly volatile terms in the "indonesia" topic.



(a) High topic word volatility.

(b) High averaged topic word volatility.

Figure 8.5: Probabilities of highly volatile terms in the example "indonesia" topic.

Chapter 9

Conclusion and Future Research

We give a profound introduction into the field of semantic analysis of textual data and review both classical and state-of-the-art approaches. Our main focus lies on topic modeling where topics, i.e. sets of words that share latent semantic meaning, are extracted from documents based on their document co-occurrence frequencies. We further give a linguistic interpretation of topics that was also previously used to motivate classical co-occurrence analysis. We, however, concentrate on a specific type of topic models, dynamic ones, that allow topics to evolve over time.

As Heyer et al. (2009) have shown, the change of word context can be an indicator of changing media coverage and thus a change of meaning. They call the average rank variation of a term's significant co-occurrences the word-volatility. While adopting the general idea, we redefine word volatility as the variation of expected word probability over time in a specified topic of a dynamic topic model. This novel type of posterior analysis is able to both uncover key terms of the topics as well as rare words referring events that affect topic evolution.

As seen in the previous chapter, a topic may well exhibit high probability words across time that are not necessarily semantically similar. One conclusion is, that in the specific context of newspaper data, the topics we extract over time cease to be what traditionally is coined a topic in the topic modeling sense. At certain points in time, the semantic relatedness between words in a topic is guaranteed but on the whole may change (e.g. when an issue-cycle is over). Blei and Lafferty (2006) use the content 30,000 articles of *Science* magazine from 1881 to 1999 and fit a 20-topic model to the data. Their reported results show stable and interpretable topics in the conventional sense, bearing a clear semantic meaning throughout time. (Wang et al., 2008) in contrast use data very similar to ours. They report results on a news corpus (as we do) and on a set of top website-articles that are concerned with an already defined issue, the U.S. presidential elections in 2008. While we can assume stable topics inside an issue, they naturally evolve (by definition) but do not only stay in direct vicinity in the semantic space defined by the distribution over the vocabulary. They may as well "jump" to other issues. Furthermore the type of data we analyzed lacks of the typical stable semantic content. Newspaper texts are naturally subject to massive, and sometimes erratic, variation. Introducing stochastic

processes more complex than standard Brownian motion in fact encourages this "jumpy" behavior. We show that this can indeed be beneficial to do. Using more appropriate priors on the diffusion of topics in time, we are able to identify patterns in the data that might be interpreted as either events in an issue or even issue-cycles (in the political communication sense) by using a new interpretation of word volatility. Using this new definition, we extract words from topics that exhibit a) a high variance and b) a high averaged (across topics) variance in their expected probability to identify key terms for the present issues and occurring events in a topic stream respectively.

This raises the argument whether the generated structures are still topics (in the topic modeling sense). We call them *time-localized topics* as they exhibit the typical structure of a topic, a set of semantically related terms, but are confined to a certain time-frame. Continuing our interpretation, we can call them *topical issues*. Words that build up the time-localized topics have the typical properties of issues and we are able to identify references to events that have contributed to the state of the topic. We stress here, that this assumption does not entirely break the definition of a topic in a dynamic topic model as given in section 7.1. Time-localized topics behave in the same manner as before *during* the topical issue-cycle. They only change indefinitely when topical issues are over.

In future we plan to investigate whether change point detection (Roberts et al., 2012) of highly volatile terms could be a useful approach to automatically separate different issues in topics. Another way to reach this goal is to use another document model, for instance a nonparametric model such as the ones described by Ahmed and Xing (2012) and (Zhang et al., 2010). These are based on the idea that topics emerge and die. They model this behavior by using a Dirichlet process model as a prior on the number of topics in the model. As we encounter similar behavior here, these models seem beneficial for separating topics in time and thus also thematically.

The second opportunity for improvement is that of posterior inference. By now we have used a rather expensive approach in computing Gaussian processes analytically. The integration of sparse approximation as suggested by Snelson and Ghahramani (2005) and variational approaches such as introduced by (Titsias, 2009; Hensman et al., 2013) can considerably speed up the inference process, thus allowing to process much larger data sets. However, the next step in further developing the model is to involve domain experts from political communication science to test our method against theoretical assumption made there and to either conform or falsify our assumption about the similarity of time-localized topics and issues.

Appendix A

Foundations of probability theory

We give some basic definitions from probability theory her that might be useful, including the basic notions of measurable and probability spaces, measurable functions and random variables. The following remarks are loosely based on Çnlar (2011, chapters 1 and 2) and Øksendal (2003, chapter 2) to which we refer the reader for a deeper treatment.

Definition A.0.1. *Measurable spaces and sets.* Given a set Ω, a family $\mathcal{F}$ of subsets of Ω with properties

(i) $\emptyset \in \mathcal{F}$

(ii) $F \in \mathcal{F} \Rightarrow F^C \in \mathcal{F}$, with $F^C = \Omega \setminus F$ the complement of F in Ω

(iii) $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow A := \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

is called a **σ-algebra** on Ω and the pair $(\Omega, \mathcal{F})$ is called a **measurable space**. The subsets F of Ω that belong to $\mathcal{F}$ are called $\mathcal{F}$-measurable sets.

Given any set $\mathcal{G}$ of subsets of Ω, there exists a smallest σ-algebra $\mathcal{H}_{\mathcal{G}}$ that contains $\mathcal{G}$, i.e. $\mathcal{H}_{\mathcal{G}} = \bigcap \{\mathcal{H} : \mathcal{H} \text{ is a } \sigma\text{-algebra of } \Omega, \mathcal{G} \subset \mathcal{H}\}$, $\mathcal{H}_{\mathcal{G}}$ is called the σ-algebra *generated by* $\mathcal{G}$. If Ω is a topological space (such as e.g. $\mathbb{R}^n$), the σ-algebra generated by the collection of all open subsets of Ω is called the *Borel σ-algebra* on Ω denoted $\mathcal{B}(\Omega)$ with its elements called *Borel sets*.

Definition A.0.2. *Measurable functions.* Let $(\Omega, \mathcal{F})$ and $(\Sigma, \mathcal{S})$ be measurable spaces. A function $g : \Omega \to \Sigma$ is a **measurable function** relative to $\mathcal{F}$ and $\mathcal{S}$ if $g^{-1}B \in \mathcal{F}$ for every $B \in \mathcal{S}$.

Definition A.0.3. *Probability measure and space.* Given a measurable space $(\Omega, \mathcal{F})$, a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ with properties

1. $\mathbb{P}(\emptyset) = 0, \mathbb{P}(\Omega) = 1$ and

2. given $A_1, A_2, \ldots \in \mathcal{F}$ and $A_i \cap A_j = \emptyset$ if $i \neq j$, i.e. $\{A_i\}_{i=1}^{\infty}$ is disjoint,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

is called a **probability measure** on $(\Omega, \mathcal{F})$ and the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**. Considering a probability space, the elements of its σ-algebra are often called **events**.

The necessary properties of the function P in definition A.0.3 are also called the *probability* or *Kolmogorov axioms*.

Definition A.0.4. *Random variables.* Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(S, \mathcal{S})$ a measurable space. A function $X : \Omega \to S$ is a **random variable** taking values in S or, simply, an **S-valued random variable**, if X is measurable relative to both $\mathcal{F}$ and $\mathcal{S}$. I.e. for every $B \in \mathcal{S}$,

$$X^{-1}B = \{X \in B\} = \{\omega \in \Omega : X(\omega) \in B\}$$

is an event. We will generally be dealing with topological spaces, i.e. S is $\mathbb{R}$ or $\mathbb{R}^n$ and $\mathcal{S}$ is the Borel σ-algebra on S, $\mathcal{B}(S)$.

A random variable is *simple* if it takes only finitely many values in $\mathbb{R}$ and *discrete* if it takes only countably many values. Every S-valued random variable X induces a probability measure μ_X on $\mathcal{S}$ with μ_X being the image of $\mathbb{P}$ under X, i.e.

$$\mu_X(B) = \mathbb{P}(X^{-1}B) = \mathbb{P}\{X \in B\}, B \in \mathcal{S}.$$

μ_X is thus a probability measure on $(S, \mathcal{S})$ and is called the *distribution* of X.

Definition A.0.5. *Cumulative distribution function.* Given the distribution of X and assuming that $S = \mathbb{R}$, there exists a function F, the *cumulative distribution function* that is defined by

$$F(x) = \mathbb{P}\{X \leq x\}. \tag{A.1}$$

This intuitively makes sense when considering the Borel σ-algebra generated by $\mathbb{R}$ which includes all half-open intervals on the real line, i.e. $(X \leq x) \in \mathcal{S} = \mathcal{B}(S)$ is an event for all $x \in \mathbb{R}$.

Definition A.0.6. *Probability density and mass function.* Let a function p satisfy

$$\mathbb{P}\{X \in B\} = \int_B p(x) d\mu(x) \tag{A.2}$$

with $\mu(x)$ being the flat and neutral Lebesgue measure on Ω. If again $S = \mathbb{R}$ and F as defined above is absolutely continuous with respect to the Lebesgue measure, p is called the **probability density function** and additionally satisfies

$$\mathbb{P}\{X \leq x\} = F(x) = \int_{-\infty}^{x} p(s) ds,$$

otherwise, i.e. when there are countably many possibly events, it is called a *probability mass function* and satisfies

$$\mathbb{P}\{X \in B\} = \sum_{x \in B} p(x).$$

Now, given a second random variable Y taking values in the measurable space $(T, \mathcal{T})$, we can also draw conclusions about the pair $Z = (X, Y) : \omega \mapsto Z(\omega) = (X(\omega), Y(\omega))$.

Definition A.0.7. *Joint probability.* As before, let $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space. Further, let X and Y be random variables, taking values in the measurable spaces $(S, \mathcal{S})$ and $(T, \mathcal{T})$ respectively. The pair $Z = (X, Y) : \omega \mapsto Z(\omega) = (X(\omega), Y(\omega))$ is measurable relative to $\mathcal{F}$ and the **product σ-algebra** $(\mathcal{S} \otimes \mathcal{T})$. In other words, Z is a random variable taking values in the space $(S \times T, \mathcal{S} \otimes \mathcal{T})$. The distribution of Z is called the **joint probability of X and Y** and is a probability measure π on the product space. The product σ-algebra $(\mathcal{S} \otimes \mathcal{T})$ is generated by a system of measurable rectangles (see Çnlar, 2011, chapter 1) and it thus suffices to define

$$\pi(A, B) = \mathbb{P}\{X \in A, Y \in B\}, A \in \mathcal{S}, B \in \mathcal{T},$$

the probability that X is in A *and* Y is in B.

Using the above definition, we are also able to infer the *marginal probabilities* from a given joint. Given the joint distribution π, we derive the marginal probability measure μ on X and ν on Y as

$$\mu(X) = \mathbb{P}\{X \in A\} = \pi(A \times T), \qquad \nu(Y) = \mathbb{P}\{Y \in B\} = \pi(S \times B).$$

Definition A.0.8. *Expectations.* Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X a random variable defined on Ω and taking values in $\mathbb{R}$[1]. X is measurable relative to $\mathcal{F}$, so its integral with respect to $\mathbb{P}$ exists and is meaningful. It is called the **expected value** of X and denoted by

$$\mathbb{E}[X] = \int_\Omega X(\omega)\mathbb{P}(d\omega) = \int_\Omega X d\mathbb{P} = \mathbb{P}\{X\},$$

and thus consequently (cf. Definition A.2) by

$$\mathbb{E}[X] = \int_{\mathbb{R}} xp(x)d\mu(x)$$

in the context of probability density functions.

The expected value of the n-th power of X is called the n-th moment of X, in particular, the first moment $\mathbb{E}[X]$ is called the *mean* of X. Given that the first moment is finite, the n-th moment of $(X - \mathbb{E}[X])$ is called the n-th centered moment. Of special interest is the second centered moment, i.e. $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$ which is called the *variance* of X and is denoted by $\mathbb{V}[X]$.

[1]Implicitly, X takes values in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, see definition A.0.4.

Appendix B

Variational Kalman Filtering

For learning in the cDTM, Wang et al. (2008) use a variational interpretation of the Kalman filter introduced in section 6.2.1. We here give a derivation of the forward and backward equations they use in their variational algorithm. Recall, that one basic property of the Wiener process is the independence of its increments. Let B_t be a standard Brownian motion, then $B_t - B_s \sim \mathcal{N}(0, t-s)$. We have seen in section 7.2, that this describes topic evolution in the cDTM model. Given the generative story of the model, the filter time update is governed by $\beta_t \sim \mathcal{N}(\beta_{t-1}, v\Delta_{s_t})$, where Δ_{s_t} denotes the time difference between time t and $t-1$ and v is the process noise. Further, let the measurement update be governed by $\hat{\beta}_t \sim \mathcal{N}(\beta_t, \hat{v}_t)$ with $\hat{v}_t$ the measurement noise at time t.
In terms of the Kalman filter, the recursive time and measurement updates (Eq. 6.2 and 6.3) are thus

$$\begin{aligned} \beta_t &= \beta_{t-1} + w_t \\ w_t &\sim \mathcal{N}(0, v\Delta_{s_t}) \end{aligned} \tag{B.1}$$

$$\begin{aligned} \hat{\beta}_t &= \beta_t + v_t \\ v_t &\sim \mathcal{N}(0, \hat{v}_t) \end{aligned} \tag{B.2}$$

Note that in our case of standard Brownian motion, $\Phi_t = H_t = I$. The forward step is then given by[1]

$$m_t = m_{t-1} \tag{B.3}$$

$$P_t = V_{t-1} + v\Delta_{s_t} \tag{B.4}$$

[1]To relate to Wang et al. (2008), we change notation in the following way: $\hat{x}_{t|t-1} = m_t$ if there is no observation, otherwise $\hat{x}_{t|t} = m_t$, $P_{t|t-1} = P_t$, $P_{t|t} = V_t$, $\hat{x}_{t|T} = \tilde{m}_t$, $P_{t|T} = \tilde{V}_t$.

for the priors. In case there is no observation, priors are used as posterior estimates in the forward step, i.e. $V_t = P_t$, m_t stays unchanged. Otherwise, the update proceeds as

$$\begin{aligned} r_t &= \hat{\beta}_t - m_{t-1} \\ S_t &= P_t + \hat{v}_t \\ K_t &= P_t S_t^{-1} = \frac{P_t}{P_t + \hat{v}_t} \end{aligned}$$

$$\begin{aligned} m_t &= m_{t-1} + K_t r_t = m_{t-1} + \frac{P_t}{P_t + \hat{v}_t}(\hat{\beta}_t - m_{t-1}) \\ &= \hat{\beta}_t \frac{P_t}{P_t + \hat{v}_t} + m_{t-1}(1 - \frac{P_t}{P_t + \hat{v}_t}) = \hat{\beta}_t \frac{P_t}{P_t + \hat{v}_t} + m_{t-1}\frac{\hat{v}_t}{P_t + \hat{v}_t} \\ &= \frac{\hat{\beta}_t P_t + m_{t-1}\hat{v}_t}{P_t + \hat{v}_t} \end{aligned} \tag{B.5}$$

$$\begin{aligned} V_t &= (I - K_t)P_t = (I - \frac{P_t}{P_t + \hat{v}_t})P_t \\ &= \frac{\hat{v}_t}{P_t + \hat{v}_t} P_t = \hat{v}_t \frac{P_t}{P_t + \hat{v}_t}. \end{aligned} \tag{B.6}$$

For the backward step Wang et al. (2008) employ the RTS smoother (see section 6.2.1) yielding

$$\begin{aligned} C_t &= \frac{V_t}{P_{t+1}} \\ \tilde{m}_t &= m_t + \frac{V_t}{P_{t+1}}(\tilde{m}_{t+1} - m_{t+1}) \\ \tilde{V}_t &= V_t + \frac{V_t}{P_{t+1}}(\tilde{V}_{t+1} - P_{t+1})\frac{V_t}{P_{t+1}}. \end{aligned}$$

Now, setting $t = t - 1$ and noting that in the backward step, we are not incorporating

any observations, i.e. Equations B.3 and B.4 hold, this transforms into

$$\begin{aligned}
\tilde{m}_{t-1} &= m_{t-1} + \frac{V_{t-1}}{P_t}(\tilde{m}_t - m_t) \\
&\overset{(B.3)}{=} m_{t-1}(1 - \frac{V_{t-1}}{P_t}) + \tilde{m}_t \frac{V_{t-1}}{P_t} \\
&\overset{(B.4)}{=} m_{t-1}(1 - \frac{V_{t-1}}{V_{t-1} + v\Delta_{s_t}}) + \tilde{m}_t \frac{V_{t-1}}{P_t} \\
&= m_{t-1}\frac{v\Delta_{s_t}}{P_t} + \tilde{m}_t \frac{V_{t-1}}{P_t} \qquad \text{(B.7)}
\end{aligned}$$

$$\begin{aligned}
\tilde{V}_{t-1} &= V_{t-1} + \frac{V_{t-1}}{P_t}(\tilde{V}_t - P_t)\frac{V_{t-1}}{P_t} \\
&= V_{t-1} + \frac{V_{t-1}^2}{P_t^2}(\tilde{V}_t - P_t). \qquad \text{(B.8)}
\end{aligned}$$

Equations B.5, B.6, B.7 and B.8 together replicate the forward and backward equations as found in Wang et al. (2008).

Bibliography

Amr Ahmed and Eric P Xing. Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. *arXiv preprint arXiv:1203.3463*, 2012.

John Aitchison and S M Shen. Logistic-Normal Distributions - Some Properties and Uses. *Biometrika*, 67(2):261–272, 1980.

James Allan. *Topic Detection and Tracking*. Event-Based Information Organization. Springer Science & Business Media, February 2002.

Cédric Archambeau, Dan Cornford, Manfred Opper, and John Shawe-Taylor. Gaussian process approximations of stochastic differential equations. *The Journal of Machine Learning Research*, 1:1–16, 2007.

Ole E Barndorff-Nielsen. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):253–280, 2002.

Ole E Barndorff-Nielsen and Neil Shephard. NonGaussian Ornstein–Uhlenbeckbased models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):167–241, 2001.

Rodney J Baxter. *Exactly Solved Models in Statistical Mechanics*. Courier Dover Publications, July 2013.

José M Bernardo and Adrian F M Smith. *Bayesian Theory*. John Wiley & Sons, September 2009.

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

David M Blei and John D Lafferty. Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

David M Blei and John D Lafferty. A Correlated Topic Model of Science. *The Annals of Applied Statistics*, 2007.

David M Blei and John D Lafferty. Topic models. In Ashok N Srivastava and Mehran Sahami, editors, *Text Mining: Classification, Clustering, and Applications*, page 71. CRC Press, 2009.

David M Blei and John D McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems*, pages 121–128, 2008.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, March 2003.

George E P Box, Gwilym M Jenkins, and Gregory C Reinsel. *Time Series Analysis. Forecasting and Control*. John Wiley & Sons, May 2013.

Jordan Boyd-Graber, Jonathan Chang, Sean Gerrish, Chong Wang, and David M Blei. Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 31, 2009.

Peter J Brockwell and Richard A Davis. *Time Series: Theory and Methods*. Springer, April 2009.

Richard Cangelosi and Alain Goriely. Component retention in principal component analysis with application to cDNA microarray data. *Biology Direct*, 2(1):2, 2007.

François Caron, Manuel Davy, and Arnaud Doucet. Generalized Polya urn for time-varying Dirichlet process mixtures. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence (UAI 2007)*, pages 33–40, 2007.

David Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, 1987.

Erhan Çnlar. *Probability and Stochastics*. Springer Science & Business Media, February 2011.

Ferdinand de Saussure. *Grundfragen der allgemeinen Sprachwissenschaft*. de Gruyter, 3 edition, 2001.

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

Arthur Dempster, Nan Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

L Dietz, S Bickel, and T Scheffer. Unsupervised prediction of citation influences. *Proceedings of the 24th International Conference on Machine Learning*, pages 233–240, 2007.

Anthony Downs. Up and Down with Ecology: The" Issue-Attention Cycle. *The Politics of American Economic Policy Making*, 1996.

Samuel Drapeau. *Risk Preferences and Their Robust Representation*. PhD thesis, Humboldt-Universität Berlin, April 2010.

Ted Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. New Mexico State University. Computing Research Laboratory, 1993.

Bruno de Finetti. *Theory of Probability: A critical introductory treatment*, volume 2 of *Wiley series in probability and mathematical statistics*. Wiley, 1975.

Andrew Gelman and Cosma Shalizi. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 2012.

Walter R Gilks, S Richardson, and David J Spiegelhalter. *Markov Chain Monte Carlo in Practice*. CRC Press, December 1995.

Jim E Griffin and Mark F J Steel. Inference with non-Gaussian Ornstein–Uhlenbeck processes for stochastic volatility. *Journal of Econometrics*, 134(2):605–644, 2006a.

Jim E Griffin and Mark F J Steel. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101(473):179–194, 2006b.

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

James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian Processes for Big Data. *arXiv.org*, September 2013.

Gerhard Heyer, Uwe Quasthoff, and Thomas Wittig. Text Mining: Wissensrohstoff Text, 2006.

Gerhard Heyer, Florian Holz, and Sven Teresniak. Change of Topics over Time-Tracking Topics by their Change of Meaning. *KDIR*, 9:223–228, 2009.

Gerhard Heyer, Daniel Keim, Sven Teresniak, and Daniela Oelke. Interaktive explorative Suche in großen Dokumentbeständen. *Datenbank-Spektrum*, 11(3):195–206, 2011.

Matt Hoffman, David M Blei, and P R Cook. Finding latent sources in recorded music with a shift-invariant HDP. *International Conference on Digital Audio Effects (DAFx)(under review)*, 2009.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Thomas Hofmann. Probabilistic latent semantic indexing. *Proceedings of the Twenty-Second Annual International SIGIR Conference*, 1999.

Thomas Hofmann, Jan Puzicha, and Michael I Jordan. Learning from Dyadic Data. *Advances in Neural Information Processing Systems*, pages 466–472, 1999.

Florian Holz and Sven Teresniak. Towards automatic detection and tracking of topic change. *Computational Linguistics and Intelligent Text Processing*, pages 327–339, 2010.

Florian Holz, Sven Teresniak, Gerhard Heyer, and Gerik Scheuermann. Generating a Visual Overview of Large Diachronic Document Collections based on the Detection of Topic Change. In *Proc. IVAPP 2010: International Conference on Information Visualization Theory and Applications*, 2010.

Andrew H Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, January 1970.

Norman Lloyd Johnson, Samuel Kotz, and N Balakrishnan. *Continuous univariate distributions Vol. 1*, volume 1. John Wiley & Sons, 2 edition, 1995.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45, 1960.

Cathleen Kantner. *Transnational Identity Discourse in the Mass Media. Humanitarian Military Interventions and the Emergence of a European Identity (1990-2006)*. Unveröffentlichte Habilitation, 2009.

Samuel Kotz, N Balakrishnan, and Norman Lloyd Johnson. *Continuous Multivariate Distributions: Models and Applications*, volume 1. John Wiley & Sons, 2 edition, 2000.

Thomas Landauer and Susan Dumais. Latent semantic analysis. *Scholarpedia*, 3(11):4356, 2008.

Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958.

David J C MacKay. Introduction to Gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168:133–166, 1998.

Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT Press, January 1999.

Andrew McCallum, Andrés Corrada-Emmanuel, and Xuerui Wang. The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email. *NIPS'04 Workshop on Structured Data and Representations in Probabilistic Models for Categorization*, 2004.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of State Calculation by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953.

Bernt Øksendal. *Stochastic Differential Equations*. An Introduction with Applications. Springer Science & Business Media, January 2003.

Peter Orbanz and Yee-Whye Teh. Bayesian nonparametric models. In *Encyclopedia of Machine Learning*, pages 81–89. Springer, 2010.

Michael D Perlman. Jensen's inequality for a convex vector-valued function on an infinite-dimensional space. *Journal of Multivariate Analysis*, 4(1):52–65, 1974.

T A Plate. Accuracy versus interpretability in flexible modeling: Implementing a tradeoff using gaussian process models. *Behaviormetrika*, 1999.

Carl Edward Rasmussen and Christopher K I Williams. *Gaussian Processes for Machine Learning*. Mit Press, January 2006.

Herbert E Rauch, C T Striebel, and F Tung. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450, 1965.

Martin Riedmiller and Heinrich Braun. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. pages 586–591, 1993.

Stephen Roberts, M Osborne, M Ebden, Steven Reece, N Gibson, and S Aigrain. Gaussian Processes for Timeseries Modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371, 2012.

Michal Rosen-Zvi, Thomas L Griffiths, and Mark Steyvers. Learning Author Topic Models from Text Corpora. *The Journal of Machine Learning Research*, 2005.

Gerard Salton and Michael J McGill. *Introduction to modern information retrieval*. McGraw-Hill computer science series. McGraw-Hill, 1983.

Evan Sandhaus. The New York Times Annotated Corpus. Linguistic Data Consortium, January 2008.

Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, pages 1257–1264, 2005.

Nathan Srebro and Sam Roweis. Time-varying topic models using dependent Dirichlet processes. *UTML, TR# 2005*, 3, 2005.

Mark Steyvers and Thomas L Griffiths. Probabilistic topic models. In Thomas Landauer, D McNamara, and W Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2005.

Stephen J Taylor. *Asset Price Dynamics, Volatility, and Prediction*. Princeton University Press, February 2011.

Yee-Whye Teh and Michael I Jordan. Hierarchical Bayesian nonparametric models with applications. *Bayesian Nonparametrics*, page 158, 2009.

Yee-Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 2006.

Michalis K Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 567–574, 2009.

Michalis K Titsias and Neil D Lawrence. Bayesian Gaussian process latent variable model. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2010.

George Eugene Uhlenbeck and Leonard Salomon Ornstein. On the theory of the Brownian motion. *Physical Review*, 36:823–841, September 1930.

Martin J Wainwright and Michael I Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2007.

Daniel D Walker, Kevin Seppi, and Erik K Ringger. Topics Over Nonparametric Time: A Supervised Topic Model Using Bayesian Nonparametric Density Estimation. In *Citeseer*, 2012.

Chong Wang, David M Blei, and D Heckerman. Continuous time dynamic topic models. *Proc. of UAI*, 2008.

Xuerui Wang and Andrew McCallum. Topics over time: a non-Markov continuous-time model of topical trends. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, 2006.

Stephen J Wright and Jorge Nocedal. *Numerical Optimization*. Springer, New York, 2^{nd} edition, 2006.

J Zhang, Yangqiu Song, C Zhang, and Shixia Liu. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1079–1088, 2010.

Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Time-Sensitive Dirichlet Process Mixture Models. Technical Report CMU-CALD-05-104, Carnegie Mellon University, Pittsburgh, May 2005.

Statement of Authorship

I hereby declare that the dissertation at hand is solely my own work and that I utilized no illegitimate further help. I did not use any but the specified sources and auxiliary means and I marked all passages in this text that have been cited directly or indirectly from published or unpublished sources and all other statements based on personal communication accordingly. Furthermore, I indicated all materials or services that have been provided by others.

Leipzig, 11.05.2015