Mining Text and Time Series Data with Applications in Finance

Joe Staines

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Computational Finance.

July 2014

UK PhD Centre in Financial Computing & Analytics

University College London

I, Joe Staines, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Contents

List of Figures

List of Tables

Notation

The variable naming convention for this thesis are listed below. Only frequently used variables are included. Where other variables are used they are defined as needed. Interpretations of the common variables are given as well as their domain.

Variable name	Description	Variable domain
D	The number of documents in a corpus of data	$\mathbb{N}$
d	An index refering to a document within a corpus	$[1, \ldots, D]$
N_d	The number of words in document d	$\mathbb{N}$
n	An index refering to a word-position within a document	$[1, \ldots, N_d]$
T	The number of intervals in a time series in the corpus	$\mathbb{N}$
t	An index refering to an interval of a time series	$[1, \ldots, T]$
M	The number of words in a dictionary	$\mathbb{N}$
m	An index refering to the position of a word in a dictionary	$[1, \ldots, M]$
K	The number of factors in a factorization model	$\mathbb{N}$
k	An index refering to a topic	$[1, \ldots, K]$
P	Weight matrix in a factorization model	$\mathbb{R}^{D \times K}$
$\vec{p}_d$	The weight vector for document d in a factorization model, a row of P	$\mathbb{R}^K$
Q	Factor matrix in a factorization model	$\mathbb{R}^{K \times T}$
$\vec{q}_k$	The factor vector for factor k in a factorization model, a column of Q	$\mathbb{R}^T$

θ	The document topic matrix in a topic model (see section 2.3)	$[0,1]^{D\times K}$
$\vec{\theta}_d$	The distribution over topics for document d, a row of θ	$(k-1)$-simplex
β	The topic term matrix in a topic model (see section 2.3)	$[0,1]^{K\times T}$
$\vec{\beta}_k$	The distribution over tokens for topic k, a column of β	$(m-1)$-simplex
z	The set of tokens indicating the topic of a word position in a document (note that this is not a matrix)	$[1,\ldots,K]^{\sum_d N_d}$
$z_{d,n}$	The topic token for word n in document d	$[1,\ldots,K]$
w	The set of words in a corpus in a document (note that this is not a matrix)	$[1,\ldots,M]^{\sum_d N_d}$
$w_{d,n}$	Word n in document d	$[1,\ldots,M]$
X	A matrix of count data, most often the bag-of-words for a corpus defined in equation 2.1	$\mathbb{N}^{D\times M}$
$x_{d,m}$	The count of the occurences of token m in document d	$\mathbb{N}$
R	A matrix of log returns with rows corresponding to documents and columns corresponding to time intervals	$\mathbb{R}^{D\times T}$
$\vec{r}_d$	The column vector of R correspoding to document d	$\mathbb{R}^T$
$\vec{r}_t$	The row vector of R correspoding to time t	$\mathbb{R}^D$
$r_{d,t}$	The log return for document d at time t	$\mathbb{R}$
$s_{d,t}$	The price of asset d at time t	$\mathbb{R}^+$
α	The concentration hyperparameter for the priors over θ in a topic model	$\mathbb{R}^+$
η	The concentration hyperparameter for the priors over β in a topic model	$\mathbb{R}^+$
ρ	Time series balance parameter in TFM (see chapter 3)	$\mathbb{R}^+$
ϕ	Variational parameters for θ in a topic model	${\mathbb{R}^+}^{D\times K}$
γ	Variational parameters for β in a topic model	${\mathbb{R}^+}^{K\times M}$
λ	Variational parameters for z in a topic model	$[0,1]^{\sum_d N_d\times M}$
ζ	Regularization parameters	$\mathbb{R}^+$
ϵ	Independant Gaussian noise	$\mathbb{R}$

μ	The mean of Gaussian distributions, such as the varational distribution over Q in TFM	$\mathbb{R}$
σ	The standard deviation of Gaussian distributions, such as the varational distribution over Q in TFM	$\mathbb{R}^+$
C	covariance matrices of multivariate Gaussian distributions	e.g. $[0,1]^{D^2}$

Table 0: Table of commonly used notation in the thesis with descriptions and domains. Interpretations for and richer descriptions of the variables are given as they are introduced.

To aid comprehension, different parts of the same object are referred to using the same symbol. The indices identify the portion referred to. d always indicate an index over documents, k over topics, n over words in a document, m over words in a dictionary, and t over periods in a time series. Thus $\chi_{d,n}$ would refer to the element of some object χ corresponding to the n-th word in the d-th document. $\vec{\chi}_d$ then refers to the vector of elements $\chi_{d,n}$.

Abstract

Finance is a field extremely rich in data, and has great need of methods for summarizing and understanding these data. Existing methods of multivariate analysis allow the discovery of structure in time series data but can be difficult to interpret. Often there exists a wealth of text data directly related to the time series. In this thesis it is shown that this text can be exploited to aid interpretation of, and even to improve, the structure uncovered. To this end, two approaches are described and tested. Both serve to uncover structure in the relationship between text and time series data, but do so in very different ways.

The first model comes from the field of topic modelling. A novel topic model is developed, closely related to an existing topic model for mixed data. Improved held-out likelihood is demonstrated for this model on a corpus of UK equity market data and the discovered structure is qualitatively examined. To the authors' knowledge this is the first attempt to combine text and time series data in a single generative topic model.

The second method is a simpler, discriminative method based on a low-rank decomposition of time series data with constraints determined by word frequencies in the text data. This is compared to topic modelling using both the equity data and a second corpus comprising foreign exchange rates time series and text describing global macroeconomic sentiments, showing further improvements in held-out likelihood. One example of an application for the inferred structure is also demonstrated: construction of carry trade portfolios. The superior results using this second method serve as a reminder that methodological complexity does not guarantee performance gains.

Chapter 1

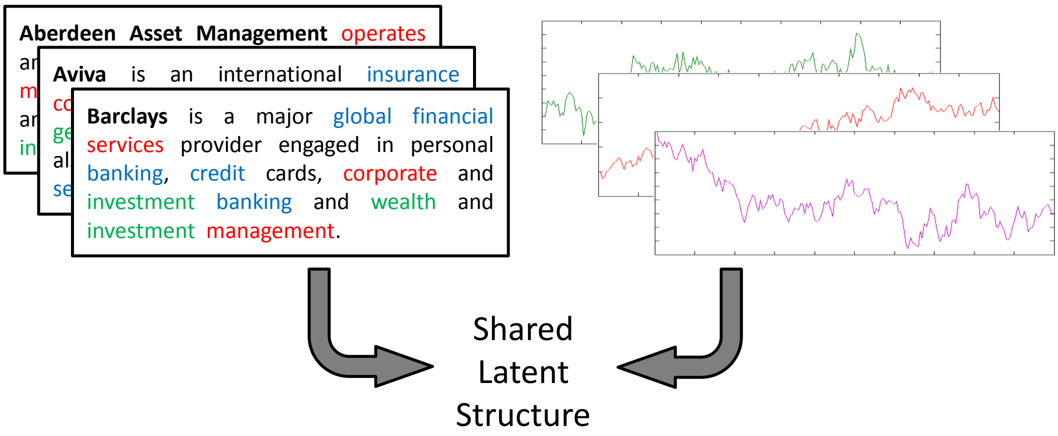
Introduction

This chapter presents the problem that will be solved in the rest of the thesis, outlines the content of each chapter and describes how this work was presented to the community. It also introduces some simple concepts on which later chapters rely.

1.1 Problem statement

Commercial financial institutions, financial regulators, and the research community all contribute to both the production of, and demand for, huge quantities of financial data. Text data are produced by news services, academics and financial analysts. Time series data also receive a great deal of attention, asset prices being the most obvious example. Understanding these data is important for practitioners and researchers alike. The time series data often exhibit high dimensionality and defy explanation of their causes (indeed, they are often viewed simply as random walks). Quantitative analytical methods which aid structure discovery and interpretation can be very useful in financial decision making, and instructive to researchers seeking to understand the markets. With regards to text data, analytical methods can help deal with the sheer volume by providing automated summarization and organization.

One particular area of interest is the relationship between financial assets. Relationships between currency pairs can reveal something of the structure of the global economy, relationships between credit derivatives can help to measure and understand the risk of systemic contagion between markets, and relationships between share prices can aid understanding of the equity markets. In time series data these relationships are represented by the moments of the joint distribution of asset prices returns. In text relationships



Aberdeen Asset Management operates

Aviva is an international insurance

Barclays is a major global financial services provider engaged in personal banking, credit cards, corporate and investment banking and wealth and investment management.

Figure 1.1: The aim of this thesis is to develop methods for finding thematic structure shared between text and time series data.

between assets can be seen in shared semantic content in text written about them. This does not always imply shared words, since the same idea may be expressed in a number of ways (the problem of polysemy).

This thesis deals with the shared thematic structure in corpora containing both text and time series data. Specifically, it aims to address three questions:

- Can text and time series data be combined to uncover shared thematic structure?
- What new methods are required to achieve this?
- Might shared structure be used to add value to financial analysis in industry or academia?

The existence of shared thematic structure would mean that the same themes which drive similarities between time series should also be apparent in the relationships between text documents. For instance, in the financial space, one might expect to be able to find shared structure between financial reports on concerning companies, and the price time series of the shares in those same companies. For instance, if the word “oil” appears repeatedly in descriptions of two companies, one might be able to infer that they are both energy companies and are therefore closely related. The price time series for these two companies would thus be likely affected by many of the same events, and in turn would be more strongly correlated compared to two unrelated companies. The information contained in the text data in these cases could be used to strengthen confidence in any structure found in the time series, and also to aid its interpretation. Likewise the relationships between the time series could improve the discovery of thematic content in the text. This idea, shown in figure 1.1, reflects a popular principle in quantitative finance: ideas are more reliable if they are well supported by data but also have an interpretable, economic explanation (in this case provided by the written text).

This thesis aims to contribute methods and applications for structure discovery in corpora containing both text and time series data. It gives particular attention to correlation structure of asset price time series and the use of text data to augment these. The most obviously useful applications for this are in finance, hence data in this space forms the focus of the examples and experiments in this thesis. In these cases the inputs to the process are price time series R for a set of assets, and a corpus of text w describing these assets. Better understanding of the joint distribution of asset prices, most often considered using correlation, is important since it provides opportunities to improve risk measurement and management, portfolio construction and pricing of derivatives. Some consideration is given in section 8.1 to other potential applications for the methods described.

1.2 List of work presented

The work of this thesis was conducted between September 2012 and June 2014. The initial ideas were first shared in a poster at the 5th York Doctoral Symposium on Computer Science in December 2012. A paper containing the work on FTSE 100 data from chapter 5 was accepted for oral presentation at Business Analytics in Finance and Industry in January 2014 [Staines and Barber, 2014] and is to appear a special issue of Intelligent Data Analysis, due early 2015. A paper applying topic factor modelling to foreign exchange data was accepted for poster presentation at the NIPS workshop, "Topic Models: Computation, Application, and Evaluation" in December 2013 [Staines and Barber, 2013].

1.3 Structure of the thesis

This chapter has introduced the problem addressed in this thesis and described how the work has been shared with the academic community so far. Hereafter it also introduces some important concepts used to describe models through the rest of the thesis. Chapter 2 introduces the ideas of bag-of-words data and data matrix factorization before describing some background literature of structure discovery in data. A particular focus is given to topic modelling, the field into which the first contributed method of this thesis fits.

Chapter 3 introduces a topic modelling approach to structure discovery in text and time series data, called topic factor modelling (TFM). This novel topic model is desribed with reference to latent Dirichlet allocation, a full specification of which is provided in the previous chapter and on which TFM is largely based. A description is given of how inference can be performed, and how the latent parameterization can be interpreted. Some space is also dedicated to evaluation of the model. Difficulties in evaluation are sometimes identified as a weakness of topic modelling.

Chapter 4 shows evidence of the effectiveness of inference in TFM and helps to motivate the choices of inference algorithm and hyperparameter settings. It achieves both of these using experiments with synthetic data, which are important to the conclusions since they provide a controlled example in which the ground truth is known (unlike experiments on financial data).

Chapter 5 shows the effectiveness of TFM on real data: text describing FTSE 100 companies combined with the time series of the returns on their share prices. Better held-out likelihood is seen using TFM than an existing topic model in figure 5.1. The output from TFM is shown to be highly interpretable and even to be related to quantitative economic data separate from the model in figure 6.8. A brief discussion is given of possible applications of TFM to equity data.

Chapter 6 describes a second novel approach to structure discovery in text and time series data, referred to as matrix factorization. This is simpler than topic modelling; a discriminative method based on a low-rank decomposition of time series data with constraints determined by word frequencies in the text data. It is compared to topic modelling using the FTSE 100 data and is shown to improve again on TFM in terms of the likelihood of held-out time series data. It cannot be compared in terms of held-out text since it isn't a generative model for text. Some similarity is found between the output from TFM and matrix factorization. The value of all of the methods in the thesis is highlighted in figure 6.8, where an interpretation of the data by a reader is shown to be strongly related to a time series unseen by the model. This supports the idea that uncovered structure is not reflective of spurious over-fitting, but can in fact give rise to structure with real economic meaning.

Chapter 7 introduces a second corpus comprising foreign exchange rate time series and text describing global macro-economic sentiments. Similar results are seen as for the equity corpus, with matrix factorization giving rise to the greatest likelihood of held-out time series data. It also demonstrates one example of an application for the inferred structure: construction of carry trade portfolios. The use of matrix factorization for correlation prediction is shown to have a similar effect to established robust methods in terms of Sharpe ratio.

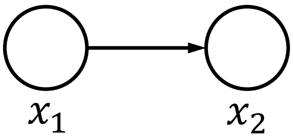


Figure 1.2: The simple Bayesian network whose corresponding distribution factorizes as in equation 1.1.

Chapter 8 describes some possible directions for future work, including new applications and extensions to the model. It also summarizes the conclusions of the thesis. At the end of the thesis are given three appendices which include details of A, the updates for variational inference in TFM; B, a Metropolis-Hastings within Gibbs procedure for inference in TFM; and C, the gradient descent method for the matrix factorization approach.

1.4 Graphical models

To aid the description of structured probabilistic models, graphical depictions of their dependencies are used. Bayesian networks are chosen for their close correspondence to the intuitive meaning of the models used in this thesis. While the description here should permit full understanding of the models used in later chapters, a fuller introduction to probabilistic graphical models can be found in, for instance Barber [2012].

A Bayesian network is a probabilistic graphical model for a set of random variables $X = \{x_1, x_2, \ldots, x_N\}$. The joint distribution of these variables can be factorized into

$$p(X) = \prod_n p\big(x_n|\text{parents}(x_n)\big) \tag{1.1}$$

where parents(x_n) are a subset of X on which x_n directly depends. A factorization of this kind can be depicted graphically by representing each variable by a node and then adding directed arcs pointing to each node from each of the parents of that node. Figure 1.2 shows a trivial example, whose corresponding factorization is written $p(X) = p(x_1)p(x_2|x_1)$. A directed acyclic graph of this kind uniquely specifies a factorization. For full specification of a model, to this must be added the values of each conditional probability table.

The models in this thesis are highly structured. They are thus best understood using graphical representations. To aid understanding three types of variables are used. The first, an observed variable, is represented in the graphical models by a filled circle. These

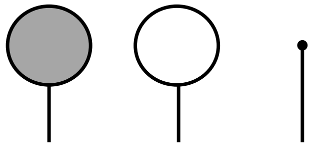


Figure 1.3: Here are shown the three variables types used in this thesis. From left to right these are: an observed variable, a latent variable, and a hyperparameter.

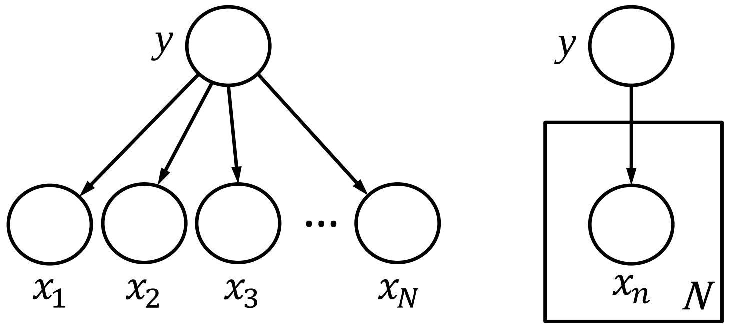


Figure 1.4: On the left is an extended representation of a Bayesian network corresponding to N variables x_n all dependent only on the parent variable y. On the right, the same network is far more compactly represented using plate notation. The repetition over n is represented by the plate.

are the data variables used as input when using the model for inference. When the model is used generatively these are the variables that make up the synthetic corpora (such as the ones used in chapter 4). The second type is a latent variable and is represented by an empty circle. These are the unobserved, structural variables whose values are sought in inference algorithms. True values of these variables are only available in the case of synthetic corpora, and not for real data. The final variable type is a hyperparameter, represented by a smaller, closed circle. These are never treated as stochastic variables and take pre-set values which are chosen for reasons elaborated in section 4.3. Figure 1.3 shows the graphical depiction of each variable type.

Structured graphical models can contain a great number of variables, many of which have may have identical dependencies. If this is the case the representation can be made clearer and more compact using plate notation. This is the use of a box (the eponymous "plate") on the diagram to indicate that everything inside of the box should be repeated a number of times indicated by a number shown in the bottom right of the box. Figure 1.4 shows an example of the use of plate notation. The variables x_n are repeated for n between 1 and N

1.4.1 Generative sampling

Sampling from a model is a process unrelated to any observed data. Sampling means to draw a corpus from the distribution represented by a model. Taking the model shown in figure 1.2 with binary variables, one possible distribution it could represent is given by $p(x_1 = 1) = 0.5$, $p(x_2 = 1|x_1 = 1) = 0.75$, and $p(x_2 = 1|x_1 = 0) = 0.25$. Some samples from this distribution are shown in table 1.1.

Variable	sample 1	sample 2	sample 3	sample 4	sample 5	sample 6
x_1	1	1	0	1	0	1
x_2	1	0	0	1	0	1

Table 1.1: Samples from a distribution over binary random variables, distributed as described above.

Not all statistical models specify a probabilistic distribution over all variables. Those that do are called generative models. The topic modelling approach introduced in chapter 3 is an example of a generative model. Often it suffices to specify a conditional distribution. Such statistical models are known as discriminative, and do not permit generative sampling of corpora. The matrix factorization method introduced in chapter 6 is an example of a discriminative model, since the specified distribution only allows generation of time series variables from given text variables and not sampling from a joint distribution over both, as is possible using the topic models introduced in this thesis. Sampling can be repeated very cheaply for most generative models, making it a useful tool for finding properties of a model.

1.4.2 Inference in Bayesian networks

Bayesian inference is the process of finding a distribution over a subset of variables of a known distribution, or some property of that distribution such as its mode (in this case it is known as maximum likelihood estimation). One important use of this process is to estimate hidden parameters for a model based on some other observed set of variables. While sampling runs the model "forwards" to find corpora that might result from it, inference runs the model "backwards" to find values for hidden variables (assuming that a corpus had been generated by the model). This process relies on the use of Bayes' rule, which relates the prior probability distribution $p(Y)$ of a hidden set of variables Y to the posterior distribution of those variables $p(Y|X)$ given the values of an observed set of variables X.

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \tag{1.2}$$

By way of example, take the model from figure 1.4 with binary random variables, unseen y and observed values $\vec{x} = \{1, 1, 0\}$. Assuming a prior over y of $p(y = 1) = 0.5$, and probabilities $p(x_i = 1|y = 1) = 0.75$ and $p(x_i = 1|y = 0) = 0.25$ for all i, Bayes' rule can be used to infer a posterior distribution over y.

$$\begin{aligned} p(y = 1|\vec{x}) &= \frac{p(y = 1, \vec{x})}{p(\vec{x})} \\ &= \frac{p(y = 1) \displaystyle\prod_i p(x_i|y = 1)}{\displaystyle\sum_y p(y) \prod_i p(x_i|y)} = 0.75 \end{aligned} \tag{1.3}$$

It can thus be inferred that the maximum a posteriori value for y is 1.

In contrast to sampling, inference is often very costly. Efficient methods are thus important across a large number of disciplines. Throughout this thesis the models used are of such complexity that inference would be impossible without the efficient methods discussed in chapters 2 and 3.

Chapter 2

Background

The essential aim of this thesis is the discovery of structure in data. As background to the contributions made, this chapter details some of the existing methods for structure discovery. It starts by explaining bag-of-words data before reviewing methods for structure discovery based on the factorization of a data matrix. Following that it introduces topic modelling, one approach to the problem of structure discovery in discrete data. The focus is on latent Dirichlet allocation (LDA), a prominent model for topic discovery which has inspired a great deal of subsequent work in the topic modelling community. LDA provides the basis for many newer and more elaborate topic models, including the novel model presented in this thesis.

The final sections of this chapter address areas more directly connected to the contributions of the thesis. Existing methods for topic modelling with mixed data types are discussed and supervised latent Dirichlet allocation (sLDA, which is later used as a benchmark method) is introduced. Existing work is discussed which aims to uncover thematic structure in financial data and finally those few papers which apply topic modelling to financial data are described. This description of the limited existing literature in the area of topic modelling with financial data highlights the potential for developments and provides some motivation for the work that follows.

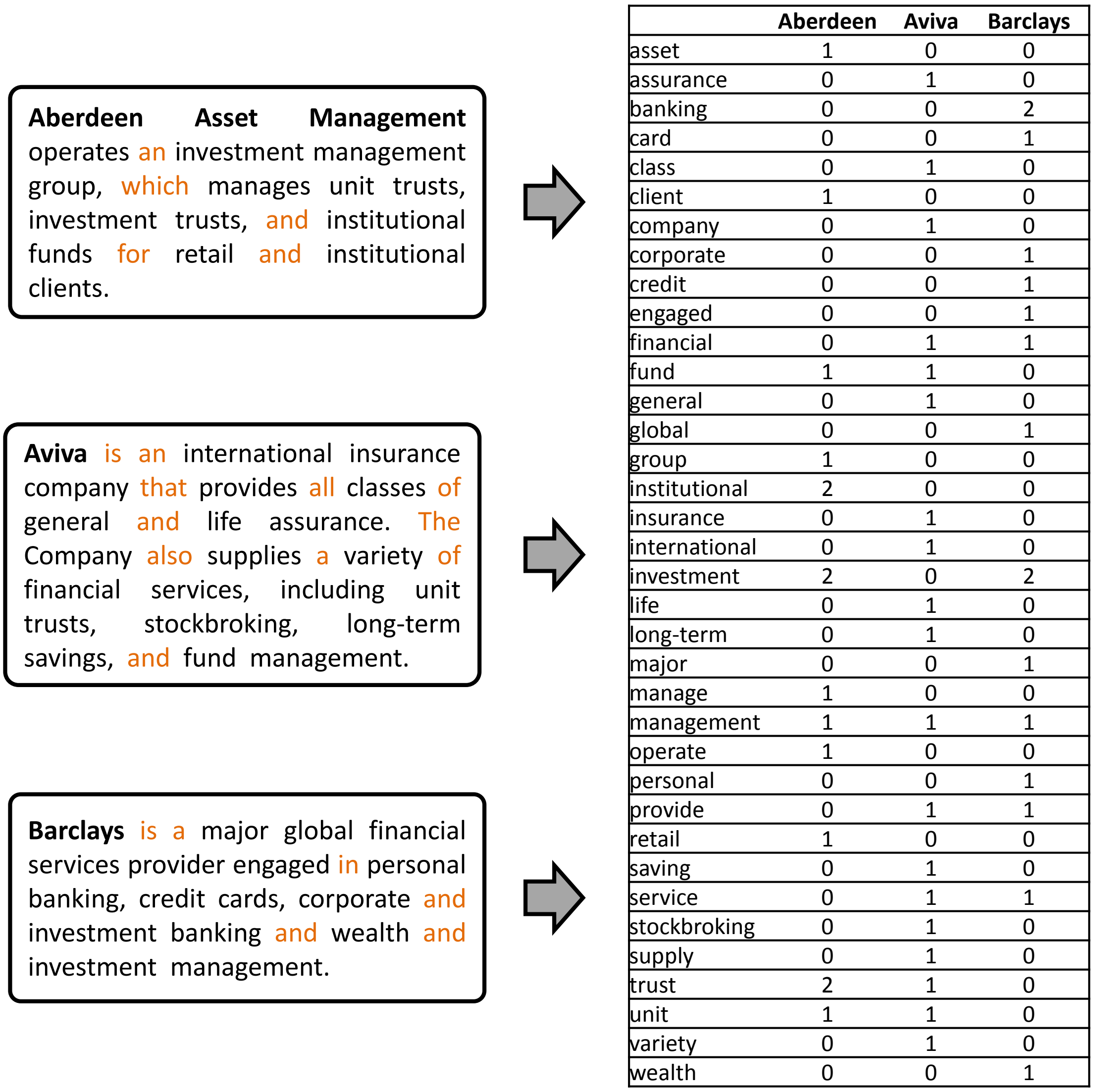
2.1 Bag-of-words data

Text data are crucial to the fields of natural language processing and information retrieval. An efficient representation of this text is needed to make use of it effectively in algorithms. Often the data are collected into distinct documents. In this case the so-called bag-of-words is one possible choice of representation. A document d is represented as a set of words and a counter which describes how many times each word appears in the document. In such a bag-of-words the context in which each word appears is ignored despite being potentially helpful in resolving issues of polysemy (the phenomenon of words with multiple meanings). One way to include context is to count the frequencies of n-grams rather than individual words. An n-gram in this context is a series of n words. For example, the 2-gram "government bond" helps disambiguate the word "bond" which in isolation might mean a connection between things, rather than the intended meaning of a financial security. The bag-of-words disregards this contextual information to obtain a more succinct representation of the data. Models which treat text in this context-free way are exchangeable with respect to word order.

The bag-of-words for a document can be constructed where the set of words includes not only the words contained in the document but all words in a dictionary. The resultant count vector is typically sparse since most documents are composed of only a small subset of all words. For a series of documents these sparse versions can be combined to give the document-term matrix. This is a matrix $X \in \mathbb{N}^{D \times M}$ where D is the number of documents in a corpus and M is the size of the dictionary used. For a corpus in which document d contains N_d words, the words $w_{d,n}$ are used to construct the elements of X:

$$x_{d,m} = \frac{1}{N_d} \sum_{n=1}^{N_d} I[w_{d,n} = m] \tag{2.1}$$

where I is an indicator function and m indexes each unique word's position in the dictionary. In this case the elements have been reweighted by dividing by the number of words in the document so as to correct for document length. Figure 2.1 shows the construction of a document-term matrix without reweighting for documents containing short descriptions of three financial companies. It shows the original documents next to the matrix itself, with each row corresponding to one word used in the documents, and each column to a document. The documents can be processed to ensure that the document-term matrix contains only relevant information. One way to achieve this is by excluding stop words: common, short words which typically don't add meaning to the data. Another possibility is to apply stemming: counting related words, such as declensions and verb conjugations, under the same dictionary entry.



Aberdeen Asset Management operates an investment management group, which manages unit trusts, investment trusts, and institutional funds for retail and institutional clients.

Aviva is an international insurance company that provides all classes of general and life assurance. The Company also supplies a variety of financial services, including unit trusts, stockbroking, long-term savings, and fund management.

Barclays is a major global financial services provider engaged in personal banking, credit cards, corporate and investment banking and wealth and investment management.

	Aberdeen	Aviva	Barclays
asset	1	0	0
assurance	0	1	0
banking	0	0	2
card	0	0	1
class	0	1	0
client	1	0	0
company	0	1	0
corporate	0	0	1
credit	0	0	1
engaged	0	0	1
financial	0	1	1
fund	1	1	0
general	0	1	0
global	0	0	1
group	1	0	0
institutional	2	0	0
insurance	0	1	0
international	0	1	0
investment	2	0	2
life	0	1	0
long-term	0	1	0
major	0	0	1
manage	1	0	0
management	1	1	1
operate	1	0	0
personal	0	0	1
provide	0	1	1
retail	1	0	0
saving	0	1	0
service	0	1	1
stockbroking	0	1	0
supply	0	1	0
trust	2	1	0
unit	1	1	0
variety	0	1	0
wealth	0	0	1

Figure 2.1: The construction of a document-term matrix (right) from three short documents. Stop words are shown in orange in the documents and are excluded from the matrix. Simple stemming has been applied by, for example, taking "fund" in the Aberdeen document and "funds" in the Aviva document to be instances of the same term.

2.2 Data matrix factorization

Across all quantitative disciplines, data frequently take the form of a series of real valued vectors each of equal dimension. In this thesis the d-th data vector is denoted $\vec{r}_d$ and its dimension N. The data matrix is constructed

$$R = [\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_D]^{\mathsf{T}} \tag{2.2}$$

so that the rows of R are the set of D data vectors. The recognition of patterns in data like this is one of the central problems in machine learning. The aim is to find a

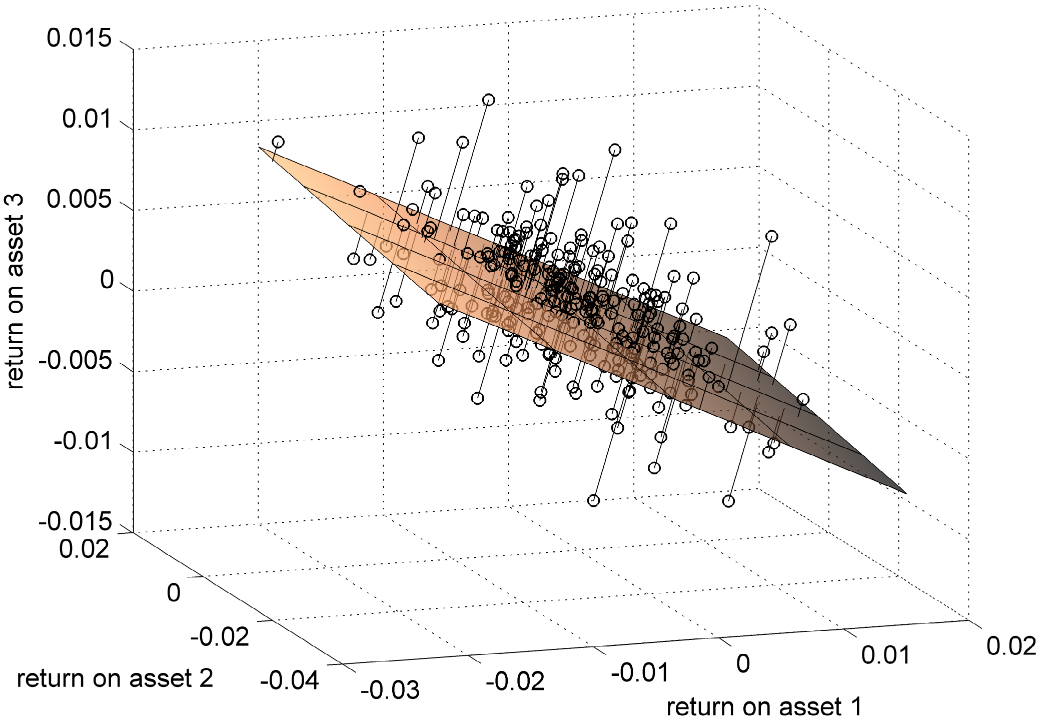


Figure 2.2: A plot of daily returns (see equation 2.20) for three equity assets. Each data vector comprises the return on each asset on a given day. The plane is the two dimensional linear manifold with minimal total perpendicular distance to the data vectors, found using principal component analysis (see section 2.2.1).

representation of the data in a way which reveals its essential structure. One way to go about this is to learn manifolds on which the data approximately lie. In the simplest case one might attempt to find linear manifolds from which the data deviate only slightly.

For continuous, real-valued data, a linear subspace is a set of vectors S in the space of the data, closed under addition and scalar multiplication, and containing the zero vector. A linear manifold is a set of vectors $S + \vec{r}_0$ where $\vec{r}_0$ is an offset vector and S is any linear subspace. Because of the closure property of the subspace, $\vec{r}_0$ can be replaced by any point in the manifold.

Sometimes the real relationships between the data are relatively simple, and can be explained using fewer dimensions than present in the data. The dimension of the data can even be reduced by projecting onto the manifold. In doing this one might hope to capture all the meaningful relationships between the full dimensions of the data. Such an attempt treats the deviations of data from the projection onto the manifold as noise. Figure 2.2 shows the daily returns for three assets plotted with a two dimensional plane fitted to them. In it, the data vectors all appear relatively near to the fitted plane, indicative of some simpler underlying structure. The projections onto the plane can be interpreted as

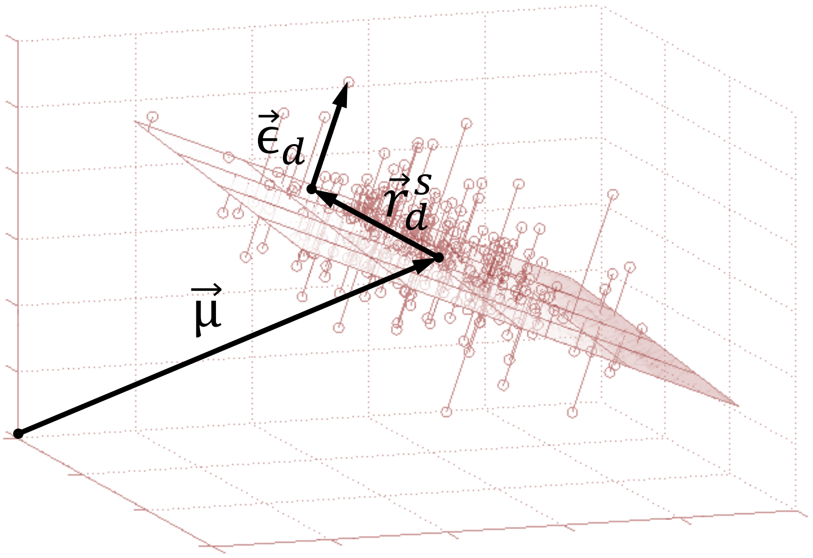


Figure 2.3: A data vector decomposed into the sum of the centre of the data $\vec{\mu}$, reconstruction within the subspace $\vec{r}^s_d$, and a noise vector $\vec{\epsilon}_d$.

reconstructions of each data vector, in which case the perpendicular distance represents one possible choice of reconstruction error. The smaller these distances, the better the reconstruction and the more appropriate the linear manifold.

Under a transformation of the original vector space $\vec{r}' = \vec{r} - \vec{r}_0$, a linear manifold of dimension less than or equal to N forms a subspace whose dimension is the same as that of the original manifold. The problem of learning a manifold can thus be broken down into finding a subspace and finding the offset vector. It is possible to reduce this problem to one of subspace learning by centring the data as a pre-processing step. This can be motivated by finding the offset vector $\vec{r}_0$ which minimizes the total Euclidean distance to all points.

$$\vec{r}_0 = \operatorname*{argmin}_{\vec{r}'} \sum_{d=1}^{D} \sum_{n=1}^{N} \left(r_{d,n} - r'_n\right)^2 = \frac{1}{D} \sum_{d=1}^{D} \vec{r}_d \qquad (2.3)$$

The centred data are given by subtracting the transpose of this, the mean of the data $\vec{r}_0 = \vec{\mu}$, from each row of R. A data vector $\vec{r}_d$ can be decomposed into contributions from the mean of the data $\vec{\mu}$, from a linear combination of the basis vectors $\vec{r}^s_d$, and from additional noise $\vec{\epsilon}_d$. Figure 2.3 shows these three components for one data vector superimposed on figure 2.2.

Given basis vectors $\vec{q}_k$ of the subspace, where $k \in [1, \ldots, K]$ and K is the dimension of the subspace, the reconstruction $\vec{r}^s_d$ of a data vector is then a linear combination of the basis vectors.

$$\vec{r}^s_d = \sum_{k=1}^{K} p_{d,k} \vec{q}_k \qquad (2.4)$$

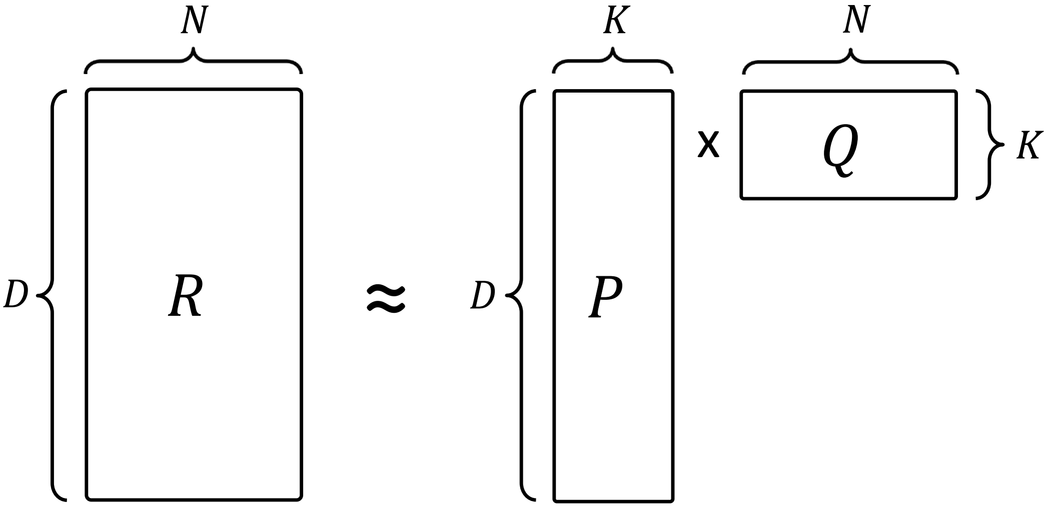


Figure 2.4: A graphical depiction of the approximation of the data matrix by a product of matrices, each with lower rank than the data matrix itself.

Introducing matrices for the weights P whose elements are $p_{d,k}$ and for the basis vectors $Q = [\vec{q}_1, \vec{q}_2, \ldots, \vec{q}_K]^\mathsf{T}$, this can be expressed as a matrix product. The aim of linear subspace learning is to find a reconstructed data matrix as close to the original as possible, sometimes with additional constraints. That is, to tighten the approximation

$$R \approx PQ. \tag{2.5}$$

Figure 2.4 shows this approximation for the case $K < N < D$. That is, where the number of data vectors exceeds the dimension of the data. The product PQ is a low-rank approximation of R so long as $K < \min\{N, D\}$ (assuming R is full row or column rank). If $K > \min\{N, D\}$ then the maximum rank of the product PQ is given not by K but by $\min\{N, D\}$. The patterns in the data represented by the factorization are in that case no simpler than the data itself, and there can exist multiple values of P and Q for which the approximation is exact. Such factorizations are typically not useful. Sometimes the aim is to find some structure in the new basis without dimension reduction. In that case factorizations with $K = \min\{N, D\}$ are sought.

Methods for approximately decomposing a data matrix into a linear combination of factors are widely used in natural and social sciences as well as engineering. Applications for these matrices include feature extraction, the transformation of the data vector into a lower dimensional feature vector to reduce computational cost and eliminate redundancy in the data [Guyon and Elisseeff, 2003]; blind source separation, the separation of mixtures of signals into their components [Comon and Jutten, 2010]; and data analysis and summarization (see for example [Cichocki et al., 2009, chap. 8]). With such a wide variety of fields using them, and such varied motivations, there exists a large number of methods for finding approximate factorizations. Below, some of the most important methods are summarized.

The similarities between these methods are well noted. There have been a number of attempts to give a unified perspective on some subsets of the approximate matrix factorization methods [Singh and Gordon, 2008; Yan et al., 2005; Borga, 1998]. Methods are differentiated firstly by their objective $\mathcal{L}$ which may be, for example, some function of the residual errors.

$$\mathcal{L} = f(R - PQ) \tag{2.6}$$

Other differences come from the constraints placed on P and Q and from the methods' treatment of the data.

2.2.1 Principal component analysis and singular value decomposition

One established matrix factorization method is principal component analysis, or PCA (see for example [Jolliffe, 2002]). PCA seeks to find the orthonormal basis of a K dimensional subspace which minimizes the mean squared residual error.

$$\min_{P,Q} \quad \frac{1}{DN} \sum_{d,n} \left(r_{d,n} - \sum_{k=1}^{K} p_{d,k} q_{k,n} \right)^2 \qquad \text{subject to } \vec{q}_i^{\,\mathsf{T}} \vec{q}_j = \delta_{i,j} \tag{2.7}$$

where $\delta_{i,j}$ is the Kronecker delta whose value is given by

$$\delta_{i,j} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases} \tag{2.8}$$

Using matrix notation the objective can also be written

$$\min_{P,Q} \quad ||R - PQ||_2^2 \qquad \text{subject to } QQ^{\mathsf{T}} = I \tag{2.9}$$

where I is the identity matrix and $||.||_a^a$ is the a-th power of the entrywise a-norm of a matrix.

$$||M||_a^a = \sum_{i,j} |M_{i,j}|^a \tag{2.10}$$

Differentiating the objective in expression 2.9 with respect to the weight matrix P to find a stationary point gives the weights in terms of the data matrix and the new basis.

$$\frac{\partial}{\partial P}||R - PQ||_2^2 = -2\left(R - PQ\right)Q^\mathsf{T} \tag{2.11}$$

Using the orthonormality of Q, at its stationary points, P is given by

$$P = RQ^\mathsf{T}. \tag{2.12}$$

Substituting this into expression 2.9 gives a new objective in terms of Q alone. This can be expressed as a trace.

$$\min_{Q} \quad \text{trace}\left((R - RQ^\mathsf{T}Q)^\mathsf{T}(R - RQ^\mathsf{T}Q)\right) \qquad \text{subject to } QQ^\mathsf{T} = I \tag{2.13}$$

Using the orthonormality of Q and the invariance of the trace to cyclic permutations this can be simplified to

$$\min_{Q} \quad \text{trace}\left(R^\mathsf{T}R(I - Q^\mathsf{T}Q)\right) \qquad \text{subject to } QQ^\mathsf{T} = I. \tag{2.14}$$

Now applying the method of Lagrange multipliers gives the objective

$$\min_{Q,\Lambda} \quad \text{trace}\left(R^\mathsf{T}R(I - Q^\mathsf{T}Q)\right) - \text{trace}\left(\Lambda(I - QQ^\mathsf{T})\right) \tag{2.15}$$

where Λ is the matrix of Lagrange multipliers. Setting the gradient of this with respect to Q equal to zero gives

$$2QR^\mathsf{T}R - (\Lambda + \Lambda^\mathsf{T})Q = 0. \tag{2.16}$$

Solutions to the eigenvalue equation of the matrix $R^\mathsf{T}R$ are also solutions to equation 2.16, and therefore also stationary points in the objective. In order to minimize expression 2.15, Λ should be chosen to be the diagonal matrix of the K largest eigenvalues and the basis vectors $\vec{q}_k$ should be the corresponding eigenvectors. For zero mean data the matrix $R^\mathsf{T}R$ is proportional to the empirical covariance. For this reason PCA can be computed by finding the eigendecomposition of the covariance matrix. In practice, however, it is preferable to compute the singular value decomposition.

The single value decomposition of the data matrix is a decomposition

$$R = U\Sigma V^\mathsf{T}. \tag{2.17}$$

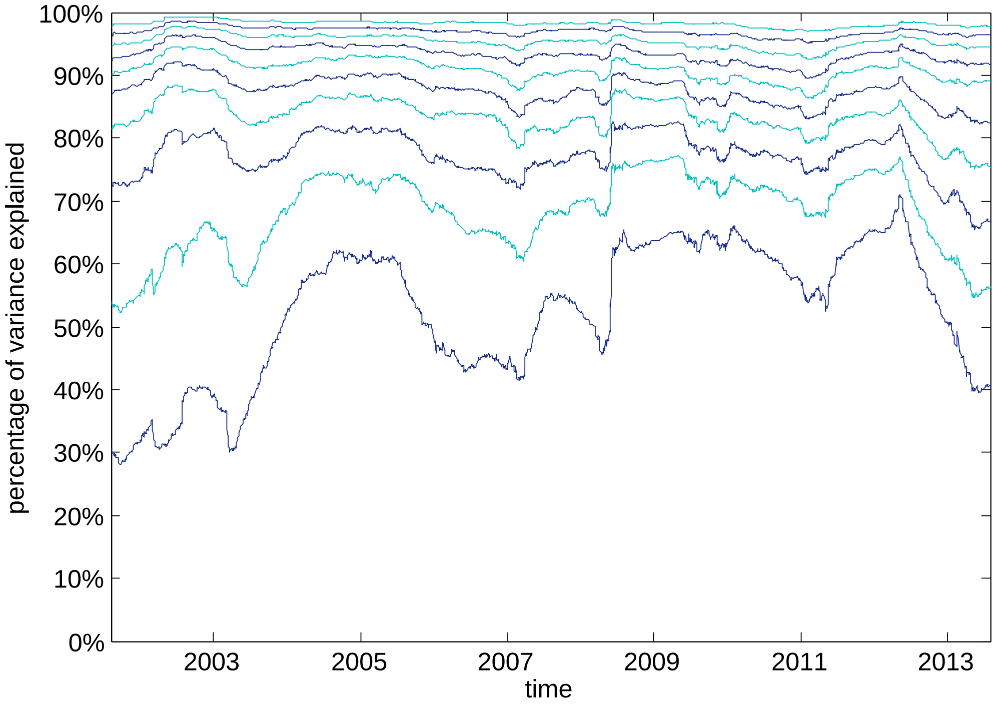


Figure 2.5: A plot showing the proportion of the variance in a rolling window of foreign exchange returns explained by the set of top principal components. The lowest line shows the percentage of variance explained by the first principal component, historically around 50% of the variance. The second line shows the percentage explained by the first two principal components and so forth.

where U and V are orthogonal matrices (i.e. $V^{\mathsf{T}}V = I$ and $U^{\mathsf{T}}U = I$ where I is the identity matrix). These are respectively referred to as the left and right singular vectors of R. Σ is a non-negative, diagonal matrix whose elements are known as the singular values of R. The correspondence to the solution found above can be seen by expressing $R^{\mathsf{T}}R$ in terms of the single value decomposition.

$$\begin{aligned} R^{\mathsf{T}}R &= V\Sigma^{\mathsf{T}}U^{\mathsf{T}}U\Sigma V^{\mathsf{T}} \\ R^{\mathsf{T}}RV &= V\Sigma^{\mathsf{T}}\Sigma \qquad (2.18) \end{aligned}$$

This last equation is the eigenvalue equation of $R^{\mathsf{T}}R$, so it is possible to obtain the eigendecomposition by finding the singular value decomposition. Thus the principal components of R are given by its right singular vectors and the eigenvalues of its covariance matrix are given by the squared singular values. In terms of the value matrices $\Lambda = \Sigma^{\mathsf{T}}\Sigma$.

The full-rank, singular value decomposition can be truncated to construct a low-rank matrix factorization. The low-rank reconstruction is given by

$$R \approx U_K \Sigma_K {V_K}^{\mathsf{T}} \tag{2.19}$$

where U_K, Σ_K, and V_K are the matrices of the K largest singular values and their corresponding singular vectors. This can be related to the approximation 2.5 by recognising the correspondence when $P = U_K \Sigma_K$ and factors $Q = {V_K}^{\mathsf{T}}$.

Principal component analysis can help to explore the sources of covariance in financial data. The logarithmic return on an asset d at time t is given by

$$r_{d,t} = \log(s_{d,t}) - \log(s_{d,t-1}). \tag{2.20}$$

The covariance between these returns for a set of assets is an important consideration when constructing portfolios from those assets. The causes and structure of the covariance are thus of great interest to financial practitioners and researchers alike. For example, Fenn et al. [2011] use the principal components to show that a small number of factors explain a large proportion of the variance in foreign exchange returns. They also relate changes in market structure to changes in the principal components. Figure 2.5 shows the percentage of variance in a rolling window of foreign exchange rates explained by the principal components. The returns are daily log changes in the exchange rates used in chapter 7. The variance (or standard deviation) of returns is commonly used as a proxy for financial risk. The total variance of the data is defined to be the sum of the variances in each exchange rate. For centred data this is proportional to trace$(R^{\mathsf{T}} R)$. The total variance of the PCA low-rank reconstruction of the data is the sum of the K largest eigenvalues.

$$\text{trace}\left(V_K {\Sigma_K}^2 {V_K}^{\mathsf{T}}\right) = \text{trace}\left({V_K}^{\mathsf{T}} V_K {\Sigma_K}^2\right) = \sum_{k=1}^{K} \lambda_k \tag{2.21}$$

The total squared error is proportional to the difference of these two, the sum of the neglected eigenvalues.

$$\begin{aligned}\sum_{d,n} \left(R_{d,n} - [U_K \Sigma_K {V_K}^{\mathsf{T}}]_{d,n}\right) &= \text{trace}\big((U\Sigma V^{\mathsf{T}} - U_K \Sigma_K {V_K}^{\mathsf{T}})^{\mathsf{T}} (U\Sigma V^{\mathsf{T}} - U_K \Sigma_K {V_K}^{\mathsf{T}})\big) \\ &= \text{trace}\big(V^{\mathsf{T}} V \Sigma^{\mathsf{T}} \Sigma - {V_K}^{\mathsf{T}} V \Sigma^{\mathsf{T}} \Sigma_K - V^{\mathsf{T}} V_K {\Sigma_K}^{\mathsf{T}} \Sigma + {V_K}^{\mathsf{T}} V_K {\Sigma_K}^{\mathsf{T}} \Sigma_K\big) \\ &= \sum_{k=K+1}^{\min\{N,D\}} \lambda_k \end{aligned} \tag{2.22}$$

The explained variance is the ratio of the total variance of the reconstruction to the total variance of the data. For accurate reconstructions it is nearer to one. It is clear from figure 2.5 that for foreign exchange data, a small number of principal components can explain a large proportion of the variation in the data. This has implications in terms of diversification of currency portfolios. Note also the sudden increase in the concentration of the variation into fewer principal components at the start of the 2008 financial crisis. This is evidence of significant structural change in the market at that point.

2.2.2 Factor analysis

The descriptions of PCA in terms of squared error minimization and the eigendecomposition of the covariance matrix show the relationship between data matrix factorization and covariance eigendecomposition. Factor analysis is another method with interpretations in terms of both the data and covariance matrices.

In factor analysis the centred data are assumed to be generated by the sum

$$R = PQ + \epsilon \tag{2.23}$$

where the weights in P are Gaussian distributed, with mean zero and covariance equal to the identity matrix, and the Gaussian noise ϵ is uncorrelated with P and has mean zero and diagonal covariance matrix Ψ. The generative marginal probability of $\vec{r}_d$ is Gaussian with mean zero and covariance given by

$$\begin{aligned}\operatorname{cov}(r_{d,n}, r_{d,n'}) &= \mathbb{E}\left[\left(\sum_k p_{d,k} q_{k,n} + \epsilon_{d,n}\right)\left(\sum_k p_{d,k} q_{k,n'} + \epsilon_{d,n'}\right)\right] \\ &= \sum_k q_{k,n} q_{k,n'} + \Psi_{n,n'} \end{aligned} \tag{2.24}$$

where $\mathbb{E}$ is the expectation operator. The factors Q and noise covariance Ψ can be taken to be those values which maximize the likelihood of the data given the generative covariance matrix $Q^{\mathsf{T}}Q + \Psi$.

Solutions for factor analysis are symmetric with respect to rotation of the factors. Note that $Q^{\mathsf{T}}Q = Q^{\mathsf{T}}M^{\mathsf{T}}MQ$ for any unitary matrix M so Q may be replaced by the transformation MQ without altering the generative marginal distribution. The choices of rotation can be made to improve the interpretability of the output.

Factor analysis is very widely used. In finance it can serve a similar purpose to PCA. An example of this type of exploratory analysis can be found in [Hui, 2005]. Factor analysis allows a different noise variance for each variable, as opposed to PCA which corresponds to isotropic noise covariance. This makes it especially appropriate for data sets in which some variables have dramatically differing variance, for instance unstandardized returns data where the assets include both stocks and bonds. When the noise covariance is approximately isotropic, such as standardised returns data, factor analysis behaves much like PCA.

2.2.3 Independent component analysis

Independent component analysis (or ICA, see for example [Comon, 1994]) aims to find a factor matrix Q whose columns are maximally independent of each other. The assumption of zero correlation, made by factor analysis, does not always correspond to statistically independent factors because non-Gaussian data can be dependent without being correlated. In ICA independence is maximized by, for example, maximizing the kurtosis of the columns of Q. This can give rise to dramatically different factorizations from PCA or factor analysis, particularly when the data are strongly non-Gaussian. This is because ICA is able to better capture relationships in these cases. Figure 2.6 gives an illustration of the difference between the principal directions according to these models. In this example the price of an Australian and New Zealand dollar in US dollars are plotted against each other and the principal directions from PCA and ICA superimposed.

ICA typically assumes that the number of factors is equal to the dimension of the data, so that the weight matrix is square. It is therefore not suitable for rank reduction without modification. In the full-rank case it is often convenient to optimize over the so-called "unmixing matrix", the inverse of the weight matrix (P^{-1}). One way to maximize the independence of the components is to maximize their kurtosis.

$$\max_{P^{-1}} ||P^{-1}R||_4^4 \qquad \text{subject to } ||P^{-1}||_2^2 = 1 \tag{2.25}$$

Kurtosis is used as a proxy for non-Gaussianity. Alternatively, minimization of mutual information between the components can be used to justify the minimization of entropy with fixed covariance [Hyvärinen et al., 2001, chap. 10]. This objective is used to derive the popular FastICA algorithm [Hyvärinen and Oja, 2000]. Another, related approach begins from maximum likelihood estimation. Treating the factors as independent random variables and the weights as fixed parameters, the likelihood of the observed data can be factorized

$$p(R) = \det\left(P^{-1}\right) \prod_{t,k} p(q_{k,t}). \tag{2.26}$$

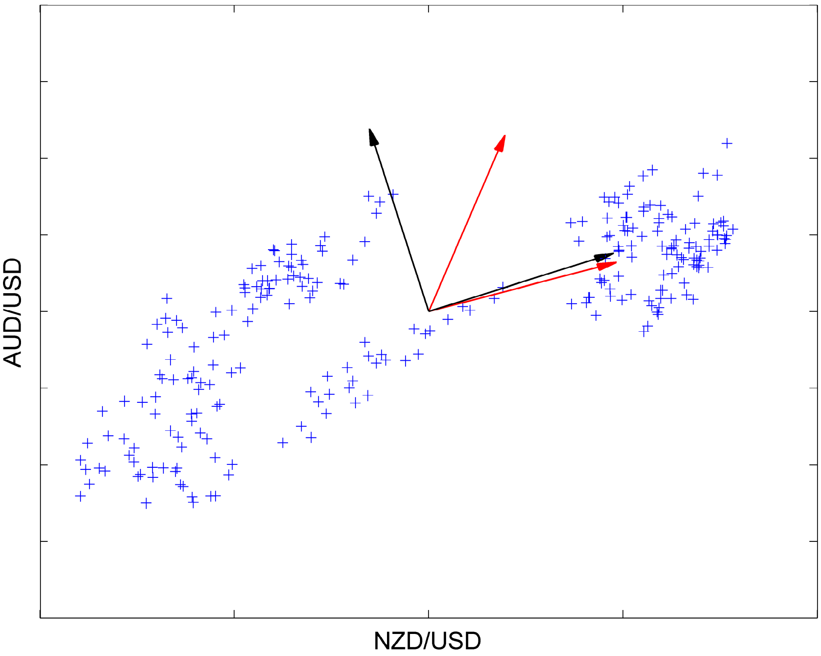


Figure 2.6: The prices of the Australian and New Zealand dollars against the United States dollar during 2012 with the principal directions from two matrix factorization methods. The directions from PCA are shown in black, and ICA in red. Factor analysis gives a result similar to PCA and, since the matrix factorization is not dimension reduced, any pair of linearly independent vectors in the positive quadrant give a solution to NMF.

Using a Gaussian prior on $\vec{q}_t$, this likelihood is proportional to an exponential function which is invariant to orthogonal transformations of the unmixing matrix.

$$p(R) = \det\left(P^{-1}\right) \prod_{t,k} p\left([P^{-1}\vec{r}_t]_k\right) \propto \det\left(P^{-1}\right) \prod_{t} \exp(\vec{r}_t{}^{\mathsf{T}} P^{-1\mathsf{T}} P^{-1}\vec{r}_t) \tag{2.27}$$

Note the invariance of each factor to premultiplication of P^{-1} by some orthogonal matrix M. For this reason the unmixing matrix, and hence the mixing matrix, cannot be uniquely estimated for Gaussian priors. This underlines the importance of non-Gaussianity to ICA.

2.2.4 Non-negative matrix factorization

Non-negative matrix factorization (or NMF, see for example [Seung and Lee, 1999]) aims to minimize some loss function of the reconstruction, constraining the parameter matrices to be non-negative. An example of NMF is the optimization problem

$$\min_{P,Q} \quad ||R - PQ||_2^2 \qquad \text{subject to} \quad p_{d,k} \geq 0, \quad q_{k,n} \geq 0 \quad \forall \;\; d,n,k. \tag{2.28}$$

This can be advantageous in identifying explanatory factorizations when the factors have non-negative interpretations. For instance the data might be thought to be positive mixtures of the underlying factors. It can also be more appropriate for non-negative data such as count data.

For an example with financial data see the work by Drakakis et al. [2008]. They use NMF to find the factors driving heteroskedasticity in asset returns. Volatility is positive by definition, and negative contributions to volatility are hard to interpret. Therefore NMF is more appropriate than other methods for their application.

2.2.5 Discrete component analysis

For discrete data the same kind of approximate factorization can be used, but the interpretation and methodology might be different. For instance, many methods employ the Kullback-Leibler divergence between the data matrix and its reconstruction R^s as the objective.

$$D_{\mathrm{KL}}(R||R^s) = \sum_{d,n} r_{d,n} \log\left(\frac{r_{d,n}}{r^s_{d,n}}\right) \tag{2.29}$$

This use of the Kullback-Leibler divergence on an unnormalized quantity is unfamiliar. It is most often used as an asymmetric measure of the difference between distributions. For continuous variables it is given by

$$D_{KL}\big(p(x)||q(x)\big) = \int_x p(x)\log\left(\frac{p(x)}{q(x)}\right). \tag{2.30}$$

A number of applications for discrete data involve text and, more specifically, the bag-of-words representation of a corpus. Buntine and Jakulin [2006] describe a way to unify NMF, probabilistic latent semantic analysis (pLSA) and latent Dirichlet allocation (LDA) with a number of other methods for structure discovery in text corpora. They define discrete component analysis (DCA) to include any model in which the expectation of the bag-of-words under a generative distribution is a product of two matrices.

$$\mathbb{E}[X] = PQ \tag{2.31}$$

This criterion can be easily satisfied by taking the bag-of-words to be equal to a product of matrices plus some zero mean noise.

$$X = PQ + \epsilon \tag{2.32}$$

Maximizing the likelihood can then be made to correspond to many matrix factorization methods by choosing an appropriate noise model. DCA, however, also stipulates that the generative model be discrete to match the bag-of-words data.

This description of DCA encompasses many possible methods. It can be fully Bayesian, with prior distributions over P and Q so that the parameters must be found using posterior inference. It is also possible to simply maximize a likelihood function with respect to the parameters. The differences between the algorithms which fall under the umbrella of DCA arise from choices of the distribution of X about the mean. For an example of DCA, one could choose Poisson distributed word frequencies with no prior structure on the parameters. The elements of the bag-of-words would then be distributed

$$x_{d,n} \sim \text{Poisson}\left(\sum_k p_{d,k} q_{k,n}\right) \tag{2.33}$$

whose expectation is equal to the matrix PQ as required.

2.3 Topic modelling

In recent years a need has emerged for automated identification of thematic structure in discrete corpora such as text or digital images. This has been driven by the growth in digitization and storage of data and the desire to navigate, organise and understand large data sets. This need has been met in part by so-called topic models. These are generative models which have latent topic variables upon which the observed data are conditioned. The states of these latent variables are then linked to a greater or lesser extent with certain values of the observed data, so that their posterior likelihood reveals something of the thematic structure of the corpus. This structure takes the form of relationships between the topics and the documents, and between the topics and the observed data tokens, and is typically interpretable to a human reader. Figure 2.7 illustrates the generative process for topic modelling, using the example of a document describing a company. Topic tokens are drawn from document topic distributions and then based on these word are drawn from topic word distributions. In this thesis all the topic models simultaneously learn both the word content of the topics and the topic proportions in the documents. The titles identifying the documents are in that case not given, but can be assigned by the user of the model. In this case a topic containing the words {“banking”,“services”,“clients”,“management”,“savings”} has been assigned the title “Banking”. It is this intuitive interpretation of the latent variables as a summary

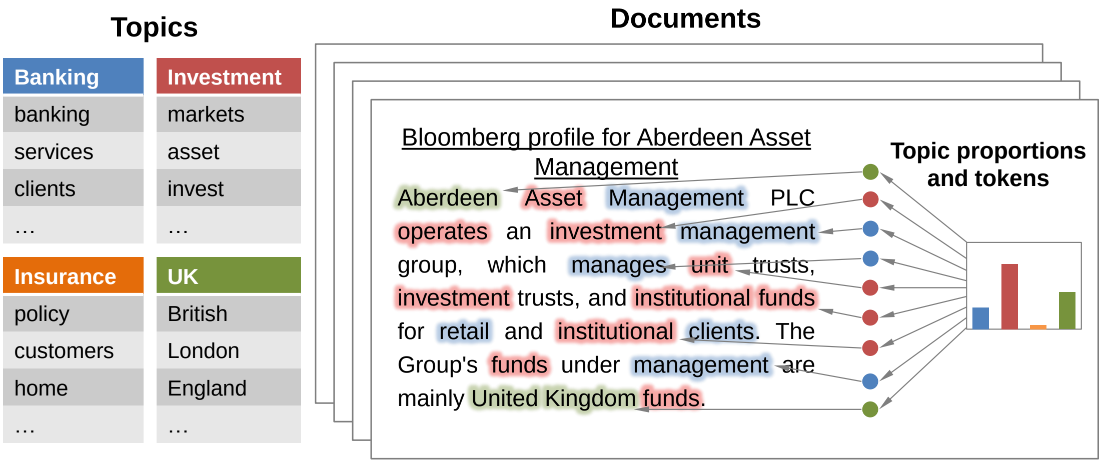

Topics
Documents
Banking
banking
services
clients
…
Investment
markets
asset
invest
…
Insurance
policy
customers
home
…
UK
British
London
England
…
Bloomberg profile for Aberdeen Asset Management
Aberdeen Asset Management PLC operates an investment management group, which manages unit trusts, investment trusts, and institutional funds for retail and institutional clients. The Group's funds under management are mainly United Kingdom funds.
Topic proportions and tokens

Figure 2.7: This figure illustrates the generative process behind topic modelling. On the left are shown a set of topics and the words they contain. On the right is shown one of the documents from a corpus. The document has associated with it a histogram of topic proportions, from which the topic tokens (the coloured circles) are drawn. The words are drawn from the topics indicated by the topic tokens. The words, apart from stop words, are coloured according to the topic with which they are most strongly associated. The topics, though inferred from a probabilistic model, are interpretable to the human reader. In this case three of them appear to represent sub-categories of the financial sector and a fourth represents operations in the UK. These are the titles given to them at the top of each column.

of content of a document that gives topic modelling its name. As this thesis will go on to show, these models can also be described in terms of data matrix factorization.

Early applications for topic modelling came from natural language processing and information retrieval. In natural language processing a corpus comprising text documents might be summarized in terms of the topics represented in each document and the words represented in each topic. In information retrieval this description of a document in terms of topics helps deal with vocabulary mismatch; the similarity between queries and documents can be reflected in similar topic content even where the vocabulary used is different. The text corpora to which topic modelling has been successfully applied include scientific journals [Grifiths and Steyvers, 2004], news articles [Wang et al., 2008], and emails [Joty et al., 2009]. Using efficient methods, truly huge corpora can be approached with topic modelling. Some more ambitious projects have worked with corpora of 19th-century literature [Jockers, 2013] and millions of Wikipedia articles [Hoffmann et al., 2010].

One important precursor to topic modelling was latent semantic analysis, or LSA [Deerwester et al., 1990]. This involves taking the singular value decomposition of the document-term matrix (X, whose elements are given by equation 2.1). The left singular vectors then give the relevance of a topic to a document and the right singular vectors

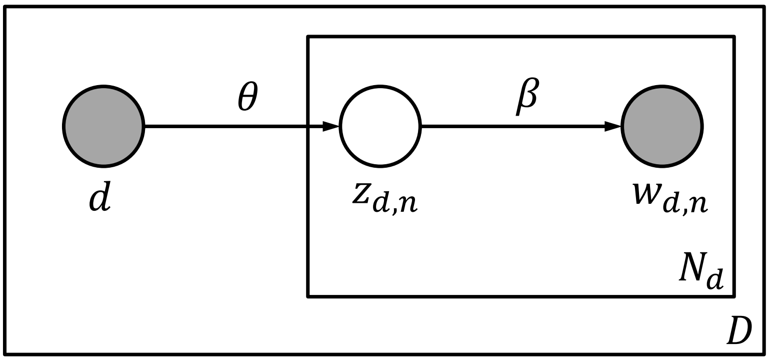


Figure 2.8: The Bayesian network for probabilistic latent semantic analysis, with plates representing repetition over the documents d and words n. $z_{d,n}$ are the latent variables which describe the topic to which each word $w_{d,n}$ is attributed. The shaded nodes are the observable document index d and words $w_{d,n}$. For an explanation of graphical models see section 1.4.

the relevance of a word to a topic. The singular values indicate the significance of the topic within the corpus. Truncating the decomposition to K topics allows retention of significant thematic structure while reducing noise.

$$X \approx U_K \Sigma_K {V_K}^{\intercal} \tag{2.34}$$

X might first be reweighted to mitigate the impact of non-thematically relevant features. A popular choice is TF-IDF reweighting [Baeza-Yates and Ribeiro-Neto, 1999] which helps to reduce the impact of common words. TF-IDF covers a number of weighting schemes but all are some ratio wherein the numerator give some description of the frequency of a word in a document (term frequency) and the denominator increases with the frequency of a word in the corpus as a whole (inverse document frequency). For example, the TF-IDF reweighted document term matrix element corresponding to document d and word m can be given by

$$\mathrm{TF}-\mathrm{IDF}(d,m) = \log\left(\frac{x_{d,m}+1}{\frac{1}{D}\sum_d I[x_{d,m} \neq 0]}\right) \tag{2.35}$$

where the document frequency is taken to be the log of the fraction of documents in which word m occurs and the term frequency is simply log reweighted. The addition of one to the numerator avoids undefined values when $x_{d,m} = 0$. Under this definition, if a word occurs often in a document but in few other documents at all it is taken to be more discriminative and given a higher value in the TF-IDF reweighted matrix.

LSA contains the key assumptions that would motivate the later development of topic modelling: that documents can be described in terms of their relationship to a set of themes and that those themes are reflected in the occurrence of words in the document. The use of the singular value decomposition, however, is suggestive of a Gaussian generative model. This doesn't match the discrete text data. Probabilistic latent semantic analysis (pLSA) explicitly models the generative process for each word as a mixture of categorical distributions [Hofmann, 1999].

$$p(w_{d,n} = m) = \sum_{k=1}^{K} \theta_{d,k}\beta_{k,m} \qquad \forall\ n \in [1, \ldots, N_d] \tag{2.36}$$

This distribution can be thought of in terms of a latent topic token $z_{d,n}$ per word position. That makes $\theta_{d,k}$ the probability of topic k being generated for each word position in document d, $p(z_{d,n}|d)$. $\beta_{k,m}$ is then the probability of generating the word m given that the topic token k has been generated, $p(w_{d,n} = m|z_{d,n} = k)$. The generative process for pLSA thus proceeds as follows, where θ and β are parameters of the distribution.

1. For each word position $n \in [1, \ldots, N_d]$ in each document $d \in [1, \ldots, D]$ draw a token $z_{d,n}$ independently from the categorical distribution $p(w_{d,n} = m) = \theta_{d,m}$
2. For each word in each document draw the word $w_{d,n}$ independently for each word position from the categorical distribution $p(w_{d,n} = m) = \beta_{z_{d,n},m}$

Imagine a corpus of news articles containing an article d whose subject is the economy. The topic distribution $\vec{\theta}_d$ for that document might be concentrated on a topic k pertaining to the economy whose word distribution $\vec{\beta}_k$ would in turn be concentrated on words pertaining to the economy. The most likely words for article d then would be related to the economy. Typically the parameters are found using a corpus rather than specified. When maximum likelihood methods are used there is no guarantee that the structure represented by the parameters will correspond to semantic meaning, but they are often highly interpretable.

The graphical model for this distribution is shown in figure 2.8. It is easy to show that pLSA belongs to the category of discrete component analysis, as defined by Buntine and Jakulin [2004].

$$\mathbb{E}[x_{d,m}] = \mathbb{E}\left[\frac{1}{N_d}\sum_{n=1}^{N_d} I[w_{d,n} = m]\right] = \sum_k \theta_{d,k}\beta_{k,m} \tag{2.37}$$

The factor matrix is given by $Q = \beta$ and the weight matrix by $P = \theta$, neither with any prior weighting. These parameters can be found using the expectation-maximization algorithm [Dempster et al., 1977]. The objective is the log likelihood of a corpus given the parameters.

$$\mathcal{L}(w|\theta,\beta) = \sum_{d=1}^{D}\sum_{n=1}^{N_d} \log\left(\sum_k \theta_{d,k}\beta_{k,w_{d,n}}\right) \tag{2.38}$$

Optimizing this directly is challenging because of the high combined dimensionality of the latent variables z. Instead, the Kullback-Leibler divergence between a variational distribution over the latent topics and their conditional probability gives a lower bound on the objective.

$$\begin{aligned} D_{\mathrm{KL}}\big(q(z)||p(z|w,\theta,\beta)\big) &= \mathbb{E}_q\Big[\log\big(q(z)\big) - \log\big(p(z|w,\theta,\beta)\big)\Big] \geq 0 \\ \log\big(p(w|\theta,\beta)\big) &\geq \mathbb{E}_q\Big[\log\big(p(w,z|\theta,\beta)\big)\Big] - \mathbb{E}_q\Big[\log\big(q(z)\big)\Big] \end{aligned} \tag{2.39}$$

This bound is tightened with respect to the variational distribution by setting it equal to the conditional probability of the latent topics.

$$\begin{aligned} q(z_{d,n}) &= p(z_{d,n}|w_{d,n},\theta,\beta) \\ &\propto p(w_{d,n},z_{d,n}|\theta,\beta) = \theta_{d,z_{d,n}}\beta_{z_{d,n},w_{d,n}} \end{aligned} \tag{2.40}$$

This is the expectation step. The maximization step is then to optimize the lower bound 2.39 with respect to θ and β for fixed $q(z)$, adding appropriate Lagrange multipliers to account for the normalization constraints on θ and β. This gives update equations

$$\begin{aligned} \theta_{d,k} &\propto \sum_{n=1}^{N_d} q(z_{d,n}=k) \\ \beta_{k,m} &\propto \sum_{d=1}^{D}\sum_{n=1}^{N_d} q(z_{d,n}=k)I[w_{d,n}=m]. \end{aligned} \tag{2.41}$$

While pLSA is a full generative model for the corpus, it cannot be generalized to new documents. Furthermore, great care must be taken to avoid overfitting. A generative model for the document topic proportions was required to address these issues. This gave rise to latent Dirichlet allocation (LDA), which is described fully below. As well as its prior structure, another benefit of LDA is its modular nature. This makes it easy to extend to build topic models with more complex structure. Indeed, this is how the new models in the following chapter were conceived.

Latent semantic analysis, probabilistic latent semantic analysis, and latent Dirichlet allocation all treat each document in a corpus as a bag-of-words, neglecting the context of words. While this is adequate for many applications, it falls short of the best models for natural language processing. Work has been done on integrating semantic features of natural language models into topic modelling [Wallach, 2006]. N-gram topic modelling is one example of what is possible. Such work highlights the flexibility of the modular nature of topic models. Since the focus of this thesis does not lie in natural language processing only bag-of-words models are considered.

2.3.1 Latent Dirichlet allocation

Latent Dirichlet allocation is arguably the prototypical topic model. Extensive descriptions may be found in, among others, [Blei et al., 2003; Blei and Lafferty, 2009]. In LDA, each topic is represented by a distribution over the dictionary and each document has a corresponding distribution over topics representing its thematic content. LDA is thus a type of mixed membership model Erosheva et al. [2004]. Each document belongs not to a single cluster, but to a mixture of topics. The weights allocated to each topic have a prior given by the Dirichlet distribution.

In the earliest work, the Dirichlet prior was used for only the document-topic matrix θ and the topic-term matrix β was left as a parameter. This was later improved upon by also applying a similar prior to the topic-term matrix. This allowed automated topic discovery rather than merely attribution of documents to predetermined topics. It is this version, sometimes called smoothed LDA, which is now described.

Topics and document distributions are drawn from Dirichlet priors, and each word in a document is sampled by first drawing a topic and then drawing a word from that topic. Each of these steps is independent of each other, each topic independent of each other topic, each word independent of each other word (conditional on te topic). This leads to a comparatively simple distribution over the bag-of-words. LDA is, in essence, an extension of pLSA with priors on the parameters. It has been argued that this prior structure regularizes the model and encourages sparsity in the output [Steyvers and Griffiths, 2006], but it also complicates the parameter estimation algorithms required. By using the conjugacy of the Dirichlet distribution with the categorical distribution, LDA adds prior structure while also being easily scalable. This scalability has improved steadily as new developments have been made in approximate methods for inference.

The generative process for LDA is now described in detail. The notation used is drawn from the literature. For all later models in this thesis, variables are named for consistency with this section. Consider a corpus with themes drawn from K different topics. For each topic a categorical distribution over all M words in a dictionary is

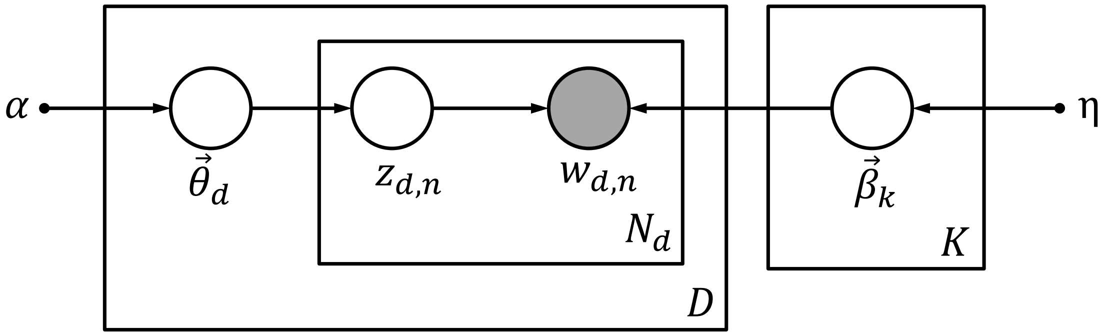


Figure 2.9: The Bayesian network for latent Dirichlet allocation, with plates representing repetition over the documents d, words n and topics k. The shaded node represents words $w_{d,n}$. $z_{d,n}$ are the latent variables which describe the topic to which each word $w_{d,n}$ is attributed. $\vec{\theta}_d$ are the topic distributions for document d and $\vec{\beta}_k$ the word distributions for topic k. For an explanation of graphical models see section 1.4.

sampled from a symmetric Dirichlet distribution with concentration parameter η.

$$p(\vec{\beta}_k|\eta) = \frac{\Gamma(M\eta)}{\Gamma(\eta)^M} \prod_{m=1}^{M} \beta_{k,m}{}^{\eta-1} \tag{2.42}$$

The resulting vector $\vec{\beta}_k$ defines the probability of picking a word from topic k. Independent of this, for each of the D documents in the corpus a vector $\vec{\theta}_d$ of dimension K is sampled from a symmetric Dirichlet distribution with concentration parameter α.

$$p(\vec{\theta}_d|\alpha) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \theta_{d,k}{}^{\alpha-1} \tag{2.43}$$

The vector $\vec{\theta}_d$ defines the probability of drawing the token $z_{d,n}$ which determines the topic from which the word $w_{d,n}$ will subsequently be drawn, exactly as in pLSA.

$$p(z_{d,n} = k|\theta) = \theta_{d,k} \tag{2.44}$$

$$p(w_{d,n} = m|z_{d,n}, \beta) = \beta_{z_{d,n},m} \tag{2.45}$$

The graphical model for LDA is shown in figure 2.9. Just like pLSA, LDA can be viewed as a factorization of the distribution over words. The conditional expectation of bag-of-words is proportional to the product of the two parameter matrices.

$$\mathbb{E}[x_{d,m}|\theta, \beta] = \mathbb{E}\left[\frac{1}{N_d}\sum_{n=1}^{N_d} I[w_{d,n} = m]\middle|\theta, \beta\right] = \sum_k \theta_{d,k}\beta_{k,m} \tag{2.46}$$

Using matrix notation this can be written as

$$\mathbb{E}[X|\theta, \beta] = \theta\beta. \tag{2.47}$$

In this way LDA may be and is included under the umbrella of discrete component analysis. The simplicity with which LDA can be understood (given the inherent complexity of simultaneous learning of topics and document membership), and the efficient inference methods which have allowed it to be applied to large corpora, have given it a central importance within the topic modelling community.

2.3.2 Inference for LDA

The quantities of interest for LDA are the most likely posterior settings of the latent variables for a given data set. The value of $\vec{\theta}_d$ can be interpreted as a summary of the thematic content of that document and $\vec{\beta}_k$ can be interpreted as a description of the theme of topic k. The posterior

$$p(\theta, \beta, z|w) = \frac{p(\theta, \beta, w, z)}{p(w)} \tag{2.48}$$

is intractable because the denominator requires marginalization over all settings of z. A number of efficient methods exist for finding approximate maximum a posteriori settings for θ and β. These include mean field variational inference [Blei et al., 2003], collapsed variational inference [Teh et al., 2007], expectation propagation [Minka and Lafferty, 2002], and Gibbs sampling [Grifiths and Steyvers, 2004]. Gibbs sampling and variational inference for LDA are described below, taken from these references. The focus here is on these two methods since they are the ones applied in the following chapter. For smoothed LDA they are not necessarily the preferred methods. The optimal choice of inference algorithm for topic modelling will depend on the application in question. The quality of solution and speed of inference change for both different corpora and different algorithms.

Mean field variational inference

Since finding maximum a posteriori settings is intractable, one possible approach is to approximate the posterior $p(\theta, \beta, z|w)$ using a simpler, variational distribution $q(\theta, \beta, z)$. The settings of θ, z and β which maximize q can then be taken as an approximation to the maximum a posteriori settings. The variational distribution should be as close to the true posterior as possible, by Kullback-Leibler divergence.

$$
\begin{aligned}
q^*(\theta, \beta) &= \underset{q}{\operatorname{argmin}}\, D_{\mathrm{KL}}\big(q(\theta, \beta, z)||p(\theta, \beta, z|w)\big) && (2.49)\\
&= \underset{q}{\operatorname{argmin}}\ \mathbb{E}_q\big[\log\big(q(\theta, \beta, z)\big)\big] - \mathbb{E}_q\big[\log\big(p(\theta, \beta, z|w)\big)\big] && (2.50)
\end{aligned}
$$

Then, using Bayes' rule and the fact that $p(w)$ is a constant, this becomes

$$
q^*(\theta, \beta) = \underset{q}{\operatorname{argmin}}\ \mathbb{E}_q\big[\log\big(q(\theta, \beta, z)\big)\big] - \mathbb{E}_q\big[\log\big(p(\theta, \beta, z, w)\big)\big]. \tag{2.51}
$$

The simplifying assumption in mean field variational inference is to make all latent variables independent.

$$
q(\theta, \beta, z) = \prod_d q(\vec{\theta}_d) \prod_{d,n} q(z_{d,n}) \prod_k q(\vec{\beta}_k) \tag{2.52}
$$

The variational factors should also be of the same form as the complete conditionals of the true posterior. The appropriate choice for LDA is thus

$$
\begin{gathered}
\vec{\theta}_d \overset{q}{\sim} \text{Dirichlet}(\vec{\phi}_d) \qquad\qquad \vec{\beta}_k \overset{q}{\sim} \text{Dirichlet}(\vec{\gamma}_k)\\
q(z_{d,n} = k) = \lambda_{d,n,k}
\end{gathered} \tag{2.53}
$$

where $\overset{q}{\sim}$ denotes the distribution of the given variable in q. The parameter vectors $\vec{\phi}_d$ and $\vec{\gamma}_k$ must be positive, and λ appropriately constrained.

$$
\begin{aligned}
\lambda_{d,n,k} &\geq 0 \quad \forall\ \ d, n, k\\
\sum_k \lambda_{d,n,k} &= 1
\end{aligned} \tag{2.54}
$$

Given this choice of distribution, the variational objective is tractable and differentiable with respect to each of the variational parameters.

$$
\begin{aligned}
\mathcal{L} &\equiv D_{\mathrm{KL}}\big(q(\theta, \beta, z)||p(\theta, \beta, z|w)\big)\\
&= \mathbb{E}_q\left[\sum_d \log q(\vec{\theta}_d|\vec{\phi}_d) + \sum_{d,n} \log q(z_{d,n}|\lambda) + \sum_k \log q(\vec{\beta}_k|\vec{\gamma}_k)\right]\\
&- \mathbb{E}_q\left[\sum_d \log p(\vec{\theta}_d) + \sum_{d,n}\big(\log p(z_{d,n}|\theta) + \log p(w_{d,n}|z_{d,n}, \beta)\big) \sum_k \log p(\vec{\beta}_k)\right]\\
&+ \text{constant}
\end{aligned} \tag{2.55}
$$

Differentiating (using the method of Lagrange multipliers in the case of the constrained λ) gives rise to closed form updates. The following updates can be repeated until convergence to optimize $\mathcal{L}$.

$$\phi_{d,k} = \alpha + \sum_n \lambda_{d,n,k} \qquad\qquad \gamma_{k,m} = \eta + \sum_{d,n} \lambda_{d,n,k}\delta_{w_{d,n},m}$$
$$\lambda_{d,n,k} \propto \exp\bigl(\psi_k(\vec{\phi}_d) + \psi_{w_{d,n}}(\vec{\gamma}_k)\bigr) \tag{2.56}$$

$\vec{\psi}$, used in the update for λ, is the expectation of the log of a Dirichlet random vector. Its elements can be expressed in terms of the digamma function Ψ and the parameter vector $\vec{\alpha}$ of the Dirichlet distribution.

$$\psi_i(\vec{\alpha}) \equiv \Psi(\alpha_i) - \Psi\left(\sum_j \alpha_j\right) \tag{2.57}$$

The values of θ and β to output can be chosen in a number of ways. They can be taken to be those values which maximize the optimized variational distribution. They can also be taken to be the means of the variational distribution, that is

$$\vec{\theta}^*_d = \frac{\vec{\phi}_d}{\sum_k \phi_{d,k}} \qquad\qquad \vec{\beta}^*_k = \frac{\vec{\gamma}_k}{\sum_m \gamma_{k,m}}. \tag{2.58}$$

Collapsed Gibbs sampling

Rather than explicitly approximating the posterior, Gibbs sampling forms a Markov chain which has the posterior as a stationary distribution. This is achieved by resampling each variable in turn, conditioned on the state of all other variables. After a sufficient number of iterations this corresponds to sampling from the posterior. A significant difficulty with this method is deciding the necessary length of the burn-in period before the chain has converged.

Since under LDA one can marginalize over θ and β, it would be counterproductive to resample these variables. A more efficient method is to resample each element of z from the conditional with θ and β marginalized. This is known as collapsed Gibbs sampling. It has improved convergence properties relative to full Gibbs sampling because it deals with the conditional distribution over two sets of latent variables exactly rather than by sampling. The resampling distribution for word n in document d is

$$p^{\text{Gibbs}}(z_{d,n}) = p(z_{d,n}|z_{\backslash d,n}, w) \propto p(z_{d,n}, z_{\backslash d,n}|w) \tag{2.59}$$

where $z_{\setminus d,n}$ are the settings of the z tokens excluding the one for the resampled word. The marginal posterior $p(z_{d,n}, z_{\setminus d,n}|w)$ is thus stationary with respect to resampling a single variable (and hence also to a series of such resamples). The individual updates themselves are drawn from

$$p(z_{d,n}|z_{\setminus d,n}, w) \propto \frac{\big(\alpha + \#\{d,k\}\big)\big(\eta + \#\{k, w_{d,n}\}\big)}{M\eta + \#\{k\}}. \tag{2.60}$$

The functions # refer to counts of the occurrence of certain combinations of states in the set of all words in the corpus excluding the one corresponding to the token being resampled. $\#\{d,k\}$ refers to occurrences of topic k in document d, $\#\{k\}$ to occurrences of topic k in the whole corpus, and $\#\{k, w_{d,n}\}$ to occurrences of topic k associated with the word observed at position n in document d.

Once the burn-in period is complete (as mentioned it can be difficult to reliably determine when this might be) one would normally take a number of samples and use this empirical distribution as an estimate of the posterior. However, it is generally θ and β which are of interest rather than the states of z. Since the maximum a posteriori states of the other latent variables are not very sensitive to individual tokens in z, the last state of z can be taken and the most likely settings of θ and β given that state calculated. The parameters conditioned on z have a posterior distribution

$$\begin{aligned} p(\theta, \beta|z, w) &\propto p(\theta)p(\beta)p(z|\theta)p(w|z,\beta) \\ &\propto \prod_{d,k} \theta_{d,k}^{\alpha-1} \prod_{k,m} \beta_{k,m}^{\eta-1} \prod_{d,n} \theta_{d,z_{d,n}} \beta_{z_{d,n},w_{d,n}} \\ &\propto \prod_{d,k} \theta_{d,k}^{\alpha-1+\#\{d,k\}} \prod_{k,m} \beta_{k,m}^{\eta-1+\#\{k,m\}} \end{aligned} \tag{2.61}$$

where the counts # are those present in the final sample of z. Applying the method of Lagrange multipliers to constrain the parameters to the probability simplex and differentiating gives rise to optimal parameter settings from Gibbs sampling:

$$\begin{aligned} \theta^*_{d,k}|z &\propto \alpha - 1 + \#\{d,k\} \\ \beta^*_{k,m}|z &\propto \eta - 1 + \#\{k,m\}. \end{aligned} \tag{2.62}$$

Alternatively, notice that the posterior is Dirichlet with parameter vector given by the sum of the relevant symmetric parameter and the vector of the counts in z. The optimal values can again be chosen to be, for example, the mean or mode of that distribution.

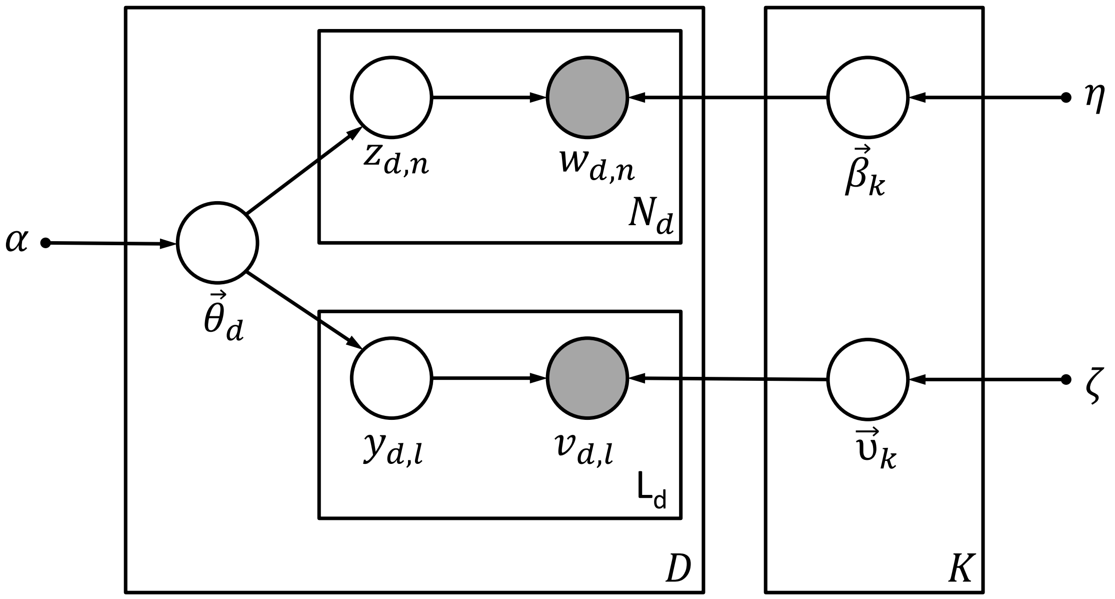


Figure 2.10: The Bayesian network corresponding to a topic model in the style of LDA with two discrete data types. In addition to a generative distribution identical to LDA, a second observed variable $v_{d,l}$ is emitted with latent topic tokens $y_{d,l}$ in exactly the same way. A topic consists of distributions over the dictionaries of the two data types $\vec{\beta}_k$ and $\vec{\upsilon}_k$. For an explanation of graphical models see section 1.4.

2.4 Topic modelling with mixed data

While early work focused on text data, there is no need to restrict topic modelling to this domain. LDA and other topic models can be applied to other discrete data types by direct analogy. LDA can also be easily extended to mixed data types. Two type, discrete topic models could easily be constructed by simply having topics defined as a pair of distributions over each data type, as in figure 2.10. In this figure two parallel topic models share the same document topic distributions θ. One could alternatively use a single distribution over both data types, but a more structured model better reflects the data, aiding interpretability, and will increase the likelihood of data. For the same reasons a number of further structures for topic modelling with mixed data have been developed.

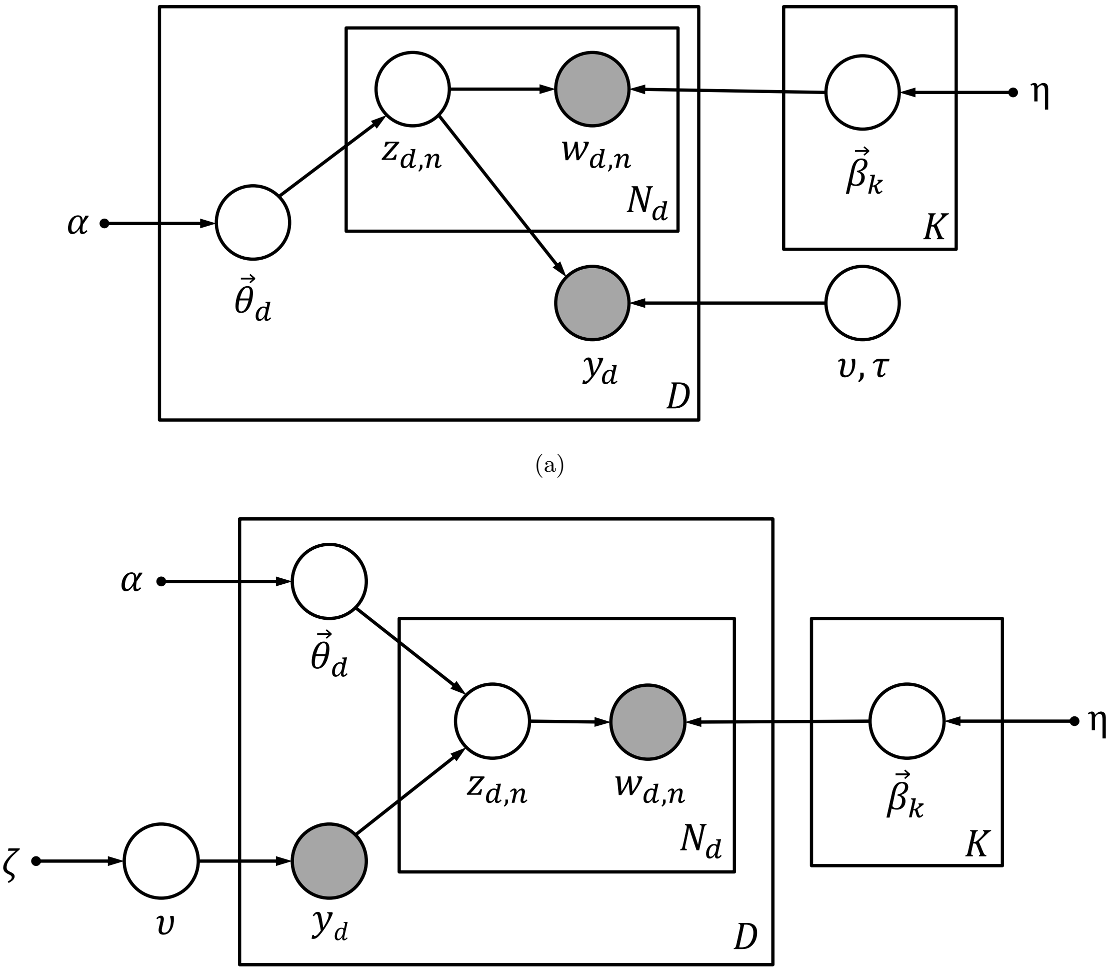
Topic models with additional data types are often motivated by corpora where documents are annotated by metadata. In this context metadata can refer to both the relationships between documents in a corpus or to simpler additional information associated with documents. The former can be modelled by so-called relational topic models which use text and the hyperlinks between the documents, assuming that the probability of a hyperlink between two documents depends on the similarity between their topic proportions [Chang and Blei, 2009]. Das et al. [2011] compare various possible joint and conditional models for text data with part-of-speech tags. The probability of a word in this case typically depends on both the topic token and the part-of-speech tag associated with it and topics are a set of unigrams. Other metadata used include associated locations [Wang et al., 2007] and the document authorship information [Rosen-Zvi et al., 2004].

Another motivation comes from annotated data. This is in some sense the opposite of the metadata in that text is treated as an annotation to the other data type rather than the metadata annotating text. Blei and Jordan [2003] use a good example of this type of corpus: images with captions. Their model is an example of a downstream topic model. That is, the topic tokens of the text variable are ancestral to a response variable. By comparison, an upstream model is one in which the second data type is ancestral to the text. Figure 2.11 illustrates the difference in ancestral order between upstream and downstream models. When the second data type is annotated by text, an upstream topic model gives the intuitively appropriate dependence structure. One might expect metadata inspired models to be upstream, with the metadata ancestral to the text. In fact both upstream and downstream structures are used.

One prominent upstream model is Dirichlet-multinomial regression [Mimno and McCallum, 2008]. The generative model looks like LDA with the exception that the parameter vector of each document's Dirichlet distribution is dependent on further variables. Specifically, the Dirichlet parameter vector is constructed by exponentiating a linear combination of features. Those features can take any form, making this a flexible model suited to many of the tasks that have been addressed with bespoke models. However, it does not generalize to new features. While it is easy to propose an appropriate generative process for features (none is given in the original paper) calculating dependency structure between features is difficult. Downstream models lead much more naturally to such a generative structure. For this reason Dirichlet-multinomial regression is not considered in this thesis, with supervised latent Dirichlet allocation preferred as a benchmark.

Supervised latent Dirichlet allocation

Supervised latent Dirichlet allocation is a general method for incorporating additional data into topic modelling [Blei and McAuliffe, 2008]. This additional data takes the form of a response variable distributed with a generalized linear model (or GLM, see for example [McCullagh and Nelder, 1989]), dependent on the topic tokens from LDA. The link functions used in the GLM are canonical so that the response variable is distributed



(a)

(b)

Figure 2.11: Example Bayesian networks for (a) downstream and (b) upstream topic modelling. Note the altered position of the text variables w in the ancestral order. Both are based on LDA, with an additional observed variable y_d per document. In the first case the additional observed data type is dependent on the latent topic tokens. In the second the tokens are dependent on the additional data. (a) corresponds to supervised latent Dirichlet allocation. For an explanation of graphical models see section 1.4.

with

$$p(y_d|z, \vec{v}, \tau) = f(y_d, \tau) \exp\left(\frac{\vec{v}^\top \vec{z}_d y_d - g(\vec{v}^\top \vec{z}_d)}{\tau}\right) \tag{2.63}$$

where $\vec{z}_d$ is the normalized vector of the frequencies of each topic in $\vec{z}_d$, τ is a dispersion parameter, and f and g are known functions. This form permits many common distributions to be applied to a response variable. One can use a categorical response, unconstrained real response or constrained real response. The generative model for the documents is identical to LDA. The graphical model for sLDA is shown in figure 2.11(a).

Exact inference is impractical for the same reasons as for LDA. Both mean field variational inference and collapsed Gibbs sampling can be applied to sLDA, though some complications may arise depending on the choice of GLM. Blei and McAuliffe [2008] use mean field variational inference for sLDA with a Gaussian response variable. In this case the update to each variable's variational distribution is tractable. In section 3.4.2 this form of inference is set out in detail as well as collapsed Gibbs sampling for a specific form of sLDA.

2.5 Exploring thematic structure in financial data

In finance, decomposition of time series into the key driving forces is commonplace. Most commonly, returns are attributed to some set of economic variables by regression. The resulting, ubiquitous, probabilistic models of returns are called factor models (see, for example [Fama and French, 1996]). More recently, PCA and factor analysis have been used to decompose price return time series into contributions from factors [Conlon et al., 2009; Shen and Zheng, 2009]. These two approaches can even be combined; Driessen et al. [2003] for example, correct for a reference rate of interest before fitting a factor model to bond returns.

Factors inferred from data don't necessarily correspond to financial fundamentals, and may not be economically interpretable at all. This makes their practical value somewhat limited; in a dynamic environment it is difficult to have confidence in the continued existence of a pattern unless one can also explain why the pattern is there. For instance if two stocks happen to have had a high correlation over the past year one should be more confident that their relationship is not spurious if they come from the same industry. Adding financial text to a model can help to incorporate this kind of economic support to patterns found in financial time series.

Combining financial text and time series data

Though they both use text and time series data, the aim of this thesis is entirely distinct from the wide literature on stock recommendation and stock return prediction (see for example [Geva and Zahavi, 2014; Mittermayer, 2004; Gidofalvi and Elkan, 2003]). Similar methods can often be applied, but the challenges and aims are very different. This thesis thus contains no comment on any of these works, instead focussing on more relevant literature that attempts discovery of shared structure.

The potential for uncovering shared thematic structure in both text and returns data together is increasingly being recognised in a variety of academic communities. Yogatama et al. [2014] use context variables from financial time series to inform a text model. They leave the extension to topic modelling to future work, and consider the online rather than batch learning context, but their work is in a very similar spirit to that contained in this thesis. Figure 1 in [Yogatama et al., 2014] is closely related to the temporal topic factor model proposed in figure 8.1 of this thesis (but with unigrams rather than topic modelling, and using an upstream context variable).

Another example of combining text and time series data to find themes is given in [Hisano et al., 2013]. They use the LASSO [Tibshirani, 1996] method to find the topics (derived using LDA) whose prevalence in news flow best explains trading volume (that is, the rate of transactions in dollars per unit time). These topics are then shown to explain some of the contemporaneous normalized daily volume. The relational nature of the structure discovered is quantified using a distance metric over topics, and visualized using networks. While this work does explore the links between financial time series, it makes no attempt to model them. Indeed the majority of the time series are ignored in evaluation, since they assert that the highest 5% by transaction volume are the most significant. Furthermore, there is no attempt to benchmark their efforts, so is unclear how much can be gained over naive methods such as simple markets news volume. It should also be noted that, while this work does tackle the problem of relational structure, trading volume structure is a fundamentally easier problem than price correlation since its nature includes only magnitude comovement in markets, and not directional comovement.

Similar, ad-hoc, connections of time series models to text models can be frequently found. Lavrenko et al. [2000] use a great deal of domain specific knowledge and ad-hoc choices to associate news stories with trends in price to enable prediction of future trends. While impressive, this work is heavily reliant on the input of an analyst. For example, they take the Yahoo recommended articles for a stock rather than using information retrieval techniques to align stories with price surges (though they do claim that an automated technique would be equally effective). A generative model of both text and time series offers a chance to avoid many of the challenges in implementation encountered by Lavrenko et al. [2000] and to exploit any benefits of combining text and time series. Topic modelling appears to be an ideal tool for this.

Topic modelling with financial data

To the authors' knowledge no attempt has previously been made to use topic modelling with emission variables for both text and time series. There have, however, been attempts to apply topic models to financial data. For example, Doyle and Elkan [2009a] discretize returns data so LDA can be applied. In their case documents correspond to days and the ticker symbol for a stock is added with a "+" appended once per 1% increase in the stock price and with a "−" appended once per 1% fall in the price. For example the document { "ABC+", "ABC+", "ABC+", "DEF−", "DEF−"} describes a day where shares in company ABC rose by 3% and those in DEF fell by 2%. The principal conclusion of the work is that more specialized topic models are needed for financial data. They highlight the need for improvements in terms of automatically finding an appropriate number of topics, changing topic content through time, and handling burstiness in the data. And indeed they then follow up with a bursty topic model in [Doyle and Elkan, 2009b]. Perhaps a more pressing concern is the awkward treatment of the data to fit existing methods. Nonetheless their work demonstrates that there is interest in financial topic modelling in the machine learning community.

One area of interest is finding topics which have some sort of causal relationship with financial time series. In contrast to this thesis, which focuses on topics which determine correlation between time series, Kim et al. [2013] attempt to find topics which cause outright changes in individual time series. They extract topics, using pLSA, from news items which have a time stamp, then aim to find those topics which have an impact on the time series throughout the time covered by their data. They determine causality using Granger tests [Granger, 1969]. The most likely causal topics are then used to reconstruct the topics. These two steps are iterated. They call this iterative topic modelling with time series feedback (ITMTF). The aim is to produce topics which have both internal coherence and a causal relationship with the time series. This they apply to finding topics which drive changes in the share price of two companies (American Airlines and Apple). This work does relate time series and topic structure in text but gives no evidence of performance out of sample. The authors note in their conclusion that a method truly integrating text and time series would be beneficial. That is exactly what this thesis achieves.

Shah and Smith [2010] discuss using a supervised topic model with financial data to predict risk. In this they built on earlier work attempting the same thing using regression [Kogan et al., 2009]. They collate text data from financial filings and take for their response variable the volatility of the share price. In doing so they aim to find topics which predict risk (by, for example, expressing the language of financial distress). This has issues from both a modelling and financial point of view. One problem with their model is the assumption that thematic similarities in text reflect degrees of financial instability. Companies may actually express uncertainty in different ways. For instance, an investment firm might talk about "high drawdowns" while a high street retailer might mention "decreasing margins". Similarities in language are more likely to reflect similarities in the type of company than a similar level of financial risk. A second issue comes arises from financial considerations. Contrary to Shah's introduction, volatility prediction is subject to market efficiency considerations because volatility is implied by the fairly liquid options markets in any major company. In contrast, correlation (while it is to some extent tradable in the prices of some more exotic securities) is less reliably implied by the market. The pairing of the thematic content of text with price comovement, the subject of this thesis, is therefore both more pertinent to topic modelling and more likely to provide useful insight into the sources of risk in financial assets.

2.6 Summary of the background and its relevance to the thesis

In this chapter, a broad background was presented to the work that follows. Since the aim of this thesis is to find structure in the relationship between text and time series data it is important to first understand independent structure discovery in either text or time series. Section 2.1 introduced a representation of text data often used in this type of work and section 2.3 described the most relevant models of thematic structure in text. Section 2.2 described some broader methods which can be applied to structure discovery as well as examples of how some of them are applied to financial time series. The start of the following chapter goes into greater detail about how financial time series can be viewed in terms of matrix factorization.

The rest of the chapter dealt with work more closely related to that contained in the rest of the thesis. Section 2.4 describes how the topic modelling community has dealt with joint data. That work is a direct inspiration for the work of this thesis and provides the competing methodologies (in particular sLDA). Finally in section 2.5, the most similar prior work is detailed. The significant differences between this and the work of this thesis, in both method and motivation, reveal an opportunity to develop new methods as well as to highlight a useful application for structure discovery. This thesis goes some way towards exploiting that opportunity.

Chapter 3

Topic Factor Models

A review of the literature reveals no previous attempt to apply topic modelling with joint emission of text and time series variables. This chapter describes one approach to achieve just that. Later chapters will demonstrate the importance of the slight differences of this approach from pre-existing topic models. After a definition of the model, its interpretation for financial data is explained. Inference algorithms using mean field variational inference and Gibbs sampling are then described. The chapter ends with a discussion on evaluation techniques for topic modelling and the choices of evaluation methods made in this thesis.

3.1 Factorizing matrices of financial returns data

One type of data set to which the matrix factorization methods of the previous chapter can be applied is the matrix of financial returns. The log return of an asset is given by

$$r_{d,t} = \log(s_{d,t}) - \log(s_{d,t-1}) \tag{3.1}$$

where $s_{d,t}$ is the price at time t of the asset d. The matrix $R \in \mathbb{R}^{D \times T}$, in that case, contains the return on a set of D assets over the time periods $[1, \ldots, T]$. Factorizing this is useful because it can help analysts to construct a simpler explanation for the returns and perhaps to propose underlying causes. The log return is used over the simple return $(\frac{s_{d,t}-s_{d,t-1}}{s_{d,t-1}})$ because it more naturally reflects compounding of returns, though both may be treated in much the same way.

The approximate factorization $R \approx PQ$ comprises the matrix of basis vectors $Q \in$

$\mathbb{R}^{K\times T}$ and a matrix of weights $P \in \mathbb{R}^{D\times K}$. It is most natural to interpret the returns for a given asset as being composed of a linear combination of factor time series (see figure 3.1). The approximate reconstruction of a data vector is in that case given by

$$\vec{r}_d^{\,s} = \sum_{k=1}^{K} p_{d,k}\vec{q}_k. \tag{3.2}$$

However, since the elements of the data matrix are approximated by

$$r_{d,t} \approx \sum_{k=1}^{K} p_{d,k}q_{k,t} \tag{3.3}$$

one can equally view each day's returns as a linear combination of vectors of the asset weights. Vectors of the returns at a given time t are in that case approximately reconstructed instead.

$$\vec{r}_t^{\,s} = \sum_{k=1}^{K} q_{t,k}\vec{p}_k \tag{3.4}$$

In essence, this switches the interpretations of the weights and factors. Or equally, represents an approximate factorization of $R^\intercal$ rather than R.

$$R^\intercal \approx (PQ)^\intercal = Q^\intercal P^\intercal \tag{3.5}$$

Note the weight matrix in this factorization is now $Q^\intercal$ and the factor matrix $P^\intercal$. For consistency, the data matrix factorization is denoted $R = PQ$, with $R \in \mathbb{R}^{D\times T}$, throughout this thesis.

The basis vectors $\vec{q}_k$ can be interpreted as the returns attributable to some underlying risk factor at each time period in the data set. The weight $p_{d,k}$ then defines the exposure of asset d to risk factor k. The vector $\vec{p}_d$ is a summary of the risk exposure of asset d; comparing two assets' vectors gives some information about the sources of correlation between those assets. Topic factor modelling is motivated by the idea that these sources of correlation should also be reflected in qualitative analysis written about the assets. If the same words appear frequently in documents describing two assets it may be because they are similar businesses or are exposed to similar risk factors. For instance, if the word "oil" appears repeatedly in descriptions of two companies, one might be able to infer that they are both energy companies and will have higher correlation because of their shared dependence on the price of oil. Topic modelling with text and time series data should be able to uncover these types of relationships.

3.2 Topic factor modelling - adding structured, continuous variables to LDA

This thesis aims to model the relationship between text and time series data. A topic modelling approach requires emission of text and time series data. These data, namely a corpus of text w and set of time series R, are inputs into the process. The aim is to discover a topic description θ of the documents and corresponding descriptions of the topics themselves. This is to be achieved using a generative model. Then using inference, the latent structure (made up of θ, β, z and Q) can be found from the observed variables w and R. The additional variables ρ, α and η are treated as hyperparameters and set as required to give satisfactory output (see section 4.3). The size of the data (D, M and N_d) is given by the corpus but the number of topics K is another fixed value which must be chosen.

In order to construct a generative model of text and time series, this thesis proposes adding a model of structured, continuous data to LDA Blei et al. [2003]. The generative process for a corpus of mixed data then follows the sequence below.

1. Generate $\vec{\theta}_d$ independently for each of D documents, $\vec{\theta}_d \sim \text{Dirichlet}(\alpha)$
2. Generate $\vec{\beta}_k$ for each of K topics, $\vec{\beta}_k \sim \text{Dirichlet}(\eta)$, dependant of each other and of θ
3. Draw the token $z_{d,n}$ independently for each word position from the categorical distribution $p(z_{d,n} = k) = \theta_{d,k}$
4. Draw the word $w_{d,n}$ independently for each word position from the categorical distribution $p(w_{d,n} = m) = \beta_{z_{d,n},m}$
5. Generate latent time series parameters independently for each topic at each time series interval
6. Generate a time series $\vec{r}_d$ for each document given the topic distribution $\vec{\theta}_d$ and the latent time series parameters

For the model of continuous time series data an approximate low-rank matrix factorization is used

$$R \approx PQ \tag{3.6}$$

whose factors are associated with the topics in the text model. In terms of the data vectors

$$r_{d,t} \approx \sum_{k=1}^{K} p_{d,k}(\vec{\theta}_d) q_{k,t}. \tag{3.7}$$

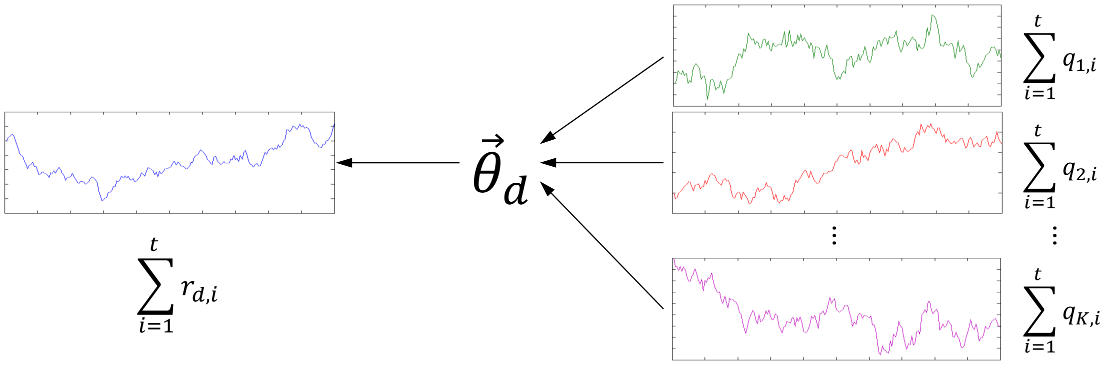


Figure 3.1: This figure shows schematic of the generative process for time series data in TFM. The variable $\vec{\theta}_d$ from LDA is reinterpreted as a weight vector for the linear combination of topic factors to construct a document time series.

This is referred to as topic factor modelling, which indicates its relationships to both topic modelling and matrix factorization. In topic factor modelling then, each topic consists of not only a distribution over words but also a time series $\vec{q}_k$ of length T. The topic time series is chosen to be independent of the other topic time series and to have a Gaussian prior distribution at each time interval t.

$$p(\vec{q}_k) = \prod_t \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-{q_{k,t}}^2}{2}\right) \tag{3.8}$$

It should be noted that this generative model is exchangeable with respect to time series interval ordering. What matters is merely that the corresponding time intervals for each document occur concurrently. While this is not a sequentially causal structure, the observed data used in this thesis are exclusively time series. For this reason this part of the model is referred to as a time series.

The document time series are constructed by combining the time series of the topics which make up the model's description of that document. Figure 3.1 shows the combination of the K topic time series to give each document time series. The cumulative topic time series returns combined linearly (with weights given by $\vec{\theta}_d$) give rise to the document time series. The dependence of both data variable types on θ allows inference of latent structure from a joint data set. $\vec{\theta}_d$ now plays the part of both the distribution over topics in the text model and the parameter in a factor model for time series.

It is simplest to take a linear combination of the topic time series at each time interval to construct time series returns for each document d. Adding a term $\epsilon_{d,t} \sim \mathcal{N}(0,1)$ specific to each document ensures that the latent structure can always fit to the data. The significance of contributions from the topics and from this document specific component is controlled by a parameter ρ. The returns are then given by

$$r_{d,t} = \frac{\rho}{\sqrt{v(\alpha)}} \sum_k \theta_{d,k} q_{k,t} + \sqrt{1-\rho^2}\, \epsilon_{d,t} \tag{3.9}$$

where the scale factor v is added so that the expectation of the variance of the returns is one. Taking expectations with respect to the prior on Q and the Dirichlet prior on θ, the mean of the observed returns will thus be zero and the variance given by

$$\mathbb{E}[r_{d,t}^2] = \frac{\rho^2}{v(\alpha)} \mathbb{E}\left[\sum_k \theta_{d,k}^2 \middle| \alpha \right] + 1 - \rho^2. \tag{3.10}$$

The concentration hyperparameter α from LDA has a critical impact on this: as the expected concentration of θ increases so does the variance of the returns. In terms of interpretation, the sources of price risk are less diversified. For the returns to have unit variance then the scale factor must be equal to

$$v(\alpha) = \frac{\alpha + 1}{K\alpha + 1}. \tag{3.11}$$

A generative model whose returns have unit variance is desirable so that standardized time series data can be used. Otherwise it is necessary to simultaneously learn the variance of each document so that inference isn't biased towards documents with higher variance time series. There are three reasons why treatment in this way is preferable to fitting a variance in the model. The first is that it lowers the cost of inference; inference is already time consuming without adding an additional variable. The second is that future variance is implied by options markets, making correlation prediction the more important application. Fitting historical variance is made less practically useful by the ready availability of strong indicators of expectations of future variance.

This standardization of data involves look-ahead adjustments to the returns in a corpus (i.e. the value $r_{d,\tau}$ depends on information from $t > \tau$). This is permissible here because the entire corpus is analysed post-hoc without time labels on the text. Finally, this treatment helps to keep the model comparatively simple. Standardization of returns as a pre-processing step, though unusual in finance, is easier to understand and interpret than adding additional variables to the model.

An alternative to this model is to take the square root of θ before the linear combi-

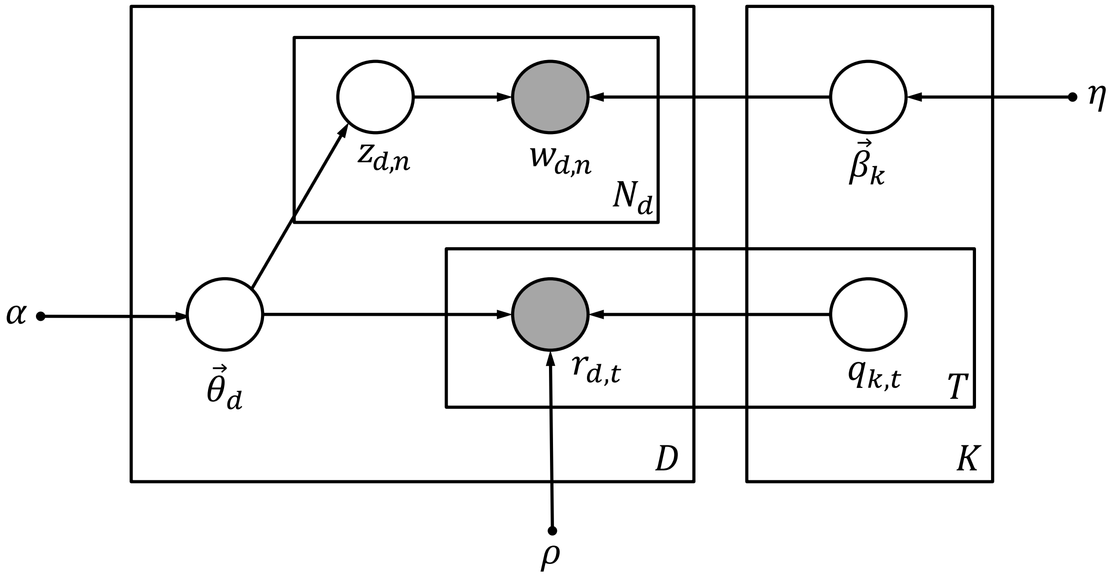


Figure 3.2: The Bayesian network for a topic factor model. The shaded nodes are the observed data: words $w_{d,n}$ and time series intervals $r_{d,t}$. Each topic comprises a distribution over words $\vec{\beta}_k$ and a time series $\vec{q}_k$. The latent topic content of each document is described by the vectors $\vec{\theta}_d$ which determine how the topics are combined in the generative model of the data. For an explanation of graphical models see section 1.4.

nation. The t-th time series element for document d would then be:

$$r_{d,t} = \rho \sum_k \sqrt{\theta_{d,k}}\; q_{k,t} + \sqrt{1-\rho^2}\; \epsilon_{d,t}. \tag{3.12}$$

This has unit variance without the need for a scale factor. The principal benefit to generating time series intervals in this way is that the conditional variance is a constant with respect to $\vec{\theta}_d$. The conditional variance of $r_{d,t}$ as generated using equation 3.9 is given by

$$\mathbb{E}\left[r_{d,t}^2 | \theta\right] = \frac{\rho^2}{v(\alpha)} \sum_k \theta_{d,k}{}^2 + 1 - \rho^2. \tag{3.13}$$

As $\vec{\theta}_d$ becomes more concentrated on a small number of topics, this variance increases, which may not be desirable. Using the alternative generative process does, however, add difficulty to the already costly inference problem. Empirically there turns out to be little difference between the results for both models. No space is therefore dedicated to the model in equation 3.12, other than to note that it is an acceptable alternative.

Putting together the model described above, the graphical model representing a TFM is shown in Figure 3.2. This corresponds to a factorization

$$p(w, R, z, \theta, \beta, Q) = \prod_d \left(p(\vec{\theta}_d) \prod_n \Big(p(z_{d,n}|\vec{\theta}_d) p(w_{d,n}|z_{d,n}, \beta) \Big) \prod_t p(r_{d,t}|\vec{\theta}_d, Q) \right) \\ \times \prod_k \left(p(\vec{\beta}_k) \prod_t p(q_{k,t}) \right) . \tag{3.14}$$

With empty text this looks similar to factor analysis with the weights constrained to be non-negative and with Dirichlet priors. If the time series is empty it reduces to LDA. The parameter ρ plays the role of tuning the significance of the time series in the model. At $\rho = 0$ the time series has no dependence on θ and the model corresponds to LDA. In contrast, for ρ near 1, the probability of the time series will be vanishingly small, and fitting θ to the time series will overwhelm the influence of the text.

The choice of ρ can also help to address the issue of combining discrete and continuous data. Since the generative model contains both types of variable, the likelihood function is a product of both probability mass functions and probability density functions. The value of a probability density function can be very large, if it is largely confined to a small range of values, or very small, if it is more diffuse. A model combining this function with a probability mass function can thus be biased toward fitting either the discrete or continuous data. By choosing an appropriate value of ρ, the hope is that solutions which merely find the best fit to one data type (which might easily be achieved with a simpler model) can be avoided. In section 4.3 the impact of the choice of ρ on the empirical success of inference is discussed. It would also be possible to have a different value of ρ for each document. For instance, had the documents different length one might wish to adjust ρ to counteract the differing relationships between the text and time series. For simplicity as single, homogenous value of ρ is used throughout this thesis.

3.3 Interpreting topic factor modelling for financial data

A data corpus for TFM needs a set of text documents each with an associated time series. The financial corpora proposed in this thesis contain text made up of corporate material or analysis relevant to one asset per document. These could be equities, with each document referring to one company and its share price. The time series should be made up of the changes in the log price of the asset. If the returns are Gaussian distributed, as in TFM, the resultant prices are log normal. This fits with a widely used paradigm of stock prices. Log normal prices are popular because they are non-negative.

This is fitting for equities, which confer rights to cash flows with limited liability, but may not always be appropriate. Furthermore, because returns on financial assets are well known to be heavy-tailed (see figure 5.2), the tail risk implied by a log normal price model is understated.

Each topic then contributes to each asset's price according to the weight allocated to it. $\vec{\theta}_d$ thus has a dual interpretation: it is the topic distribution for the text written about an asset and it is the topic factor exposure of the price. In the case of equity data, the text could for example pertain to the likely determinants of success for the company (for example if it discusses the sectors and geographies that the company operates in, the strategy it employs, or the principal sources of expense and risk). In theory then, the two interpretations of θ should be closely related for real data. This being the case, topics could be interpreted as the risk factors affecting the corpus. $q_{k,t}$ should then be interpreted as the price pressure of a risk factor corresponding to topic k at time t. Of course the real number of risk factors with price impact is in general far greater than the number of assets, whereas the aim in TFM is to produce a dimension reduced description of the space. $\epsilon_{d,t}$ refers to the contribution to the change in price from an asset's idiosyncratic risk (the portion of the price risk unique to that asset).

The topics identified are not guaranteed to correspond to real economic variables. Non-economic topics might result from, for example, inhomogeneity in the vocabulary used in the corpus. In topic modelling there can also exist "junk topics" which do not make up a large part of any document. TFM provides a description of the space of equity price time series which should be richer than that gleaned from simply using correlation data, since it uses human analysis as a further input. It is a numerical realization of the adage of quantitative finance that patterns in data are more trustworthy if there is an economic explanation for them. In this case, one might expect that equities which have correlated in the past are more likely to continue to do so if text written about them has more language in common.

3.4 Inference with topic factor models

The quantity of interest under this model is the posterior of the latent variables given a corpus. The posterior

$$p(\theta, \beta, z, Q|w, R) = \frac{p(\theta, \beta, w, R, z, Q)}{p(w, R)} \tag{3.15}$$

is intractable because of the high dimensionality of z and Q. In the case of vanilla LDA, a number of efficient methods exist for finding approximate maximum a posteriori settings for θ and β, including mean field variational inference [Blei et al., 2003], collapsed variational inference [Teh et al., 2007], and Gibbs sampling [Grifiths and Steyvers, 2004].

Collapsed Gibbs sampling relies on being able to marginalize out θ to calculate the complete conditional distribution for each $z_{d,n}$ (from which Gibbs samples would then be taken). In the case of TFM there is no closed form solution for the complete conditional so the efficient inference can be done either by finding another way to sample from the conditional or by using variational inference.

3.4.1 Variational inference

Since finding maximum a posteriori settings is intractable, one possible approach is to approximate the posterior $p(\theta, \beta, z, Q|w, R)$ using a simpler, variational distribution $q(\theta, \beta, z, Q)$. The settings of θ, β, z and Q which maximize q can then be interpreted as an approximation to the maximum a posteriori settings. The optimal variational distribution $q^*(\theta, \beta, z, Q)$ should be as close to the true posterior as possible, by KL divergence:

$$\begin{aligned} & \operatorname*{argmin}_q D_{\mathrm{KL}}\big(q(\theta,\beta,z,Q)||p(\theta,\beta,z,Q|w,R)\big) \\ = & \operatorname*{argmin}_q \ \mathbb{E}_q\big[\log\big(q(\theta,\beta,z,Q)\big)\big] - \mathbb{E}_q\big[\log\big(p(\theta,\beta,z,Q|w,R)\big)\big]. \end{aligned} \tag{3.16}$$

Then using Bayes' rule and the fact that $p(w, R)$ is a constant

$$q^*(\theta,\beta,z,Q) = \operatorname*{argmin}_q \ \mathbb{E}_q\big[\log\big(q(\theta,\beta,z,Q)\big)\big] - \mathbb{E}_q\big[\log\big(p(\theta,\beta,z,Q,w,R)\big)\big]. \tag{3.17}$$

In the case of LDA one would then take the variational factors for each component to be of the same (exponential family) form as the corresponding complete conditional in the true posterior. Update rules tighten the bound with respect to each variational parameter in turn. This yields an iterative algorithm to find local optima of the posterior distribution of parameters θ and β. Unfortunately this approach cannot be followed in the case of TFM because of the form of the complete conditional in $\vec{\theta}_d$. $p(\theta|z, Q, R)$ is still exponential family, but the required expectations for update rules cannot be expressed in closed form.

It is possible, however, to choose a simpler variational distribution such that the objective $\mathcal{L} = \mathbb{E}_q\big[\log\big(q(\theta,\beta,z,Q)\big)\big] - \mathbb{E}_q\big[\log\big(p(\theta,\beta,z,Q,w,R)\big)\big]$ and its gradients are

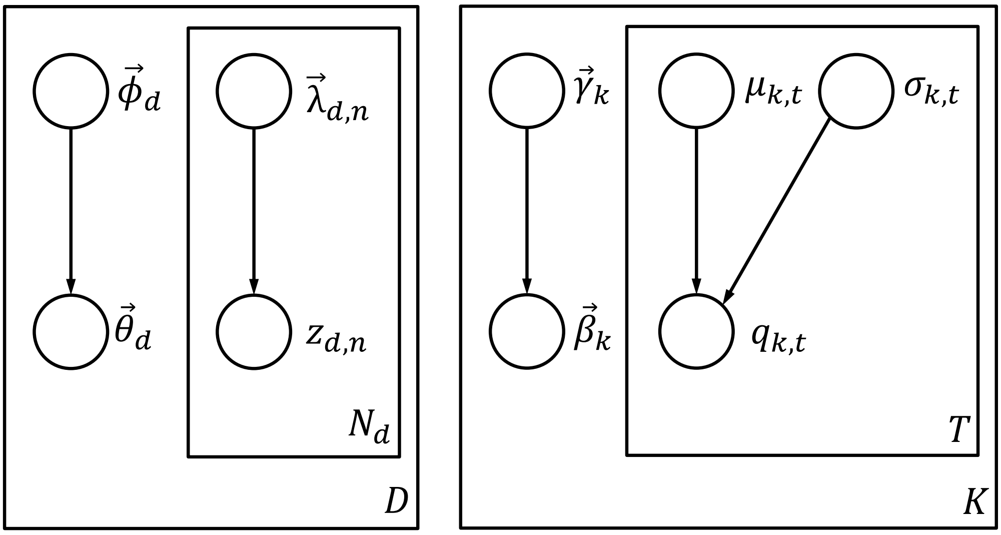


Figure 3.3: The Bayesian network for the variational distribution to approximate the posterior on the latent variables. Variational parameters ϕ, λ, γ, μ, and σ describe a distribution over the latent variables in TFM.

tractable. In that case, gradient descent can be applied in the variational parameter space. The variational distributions of $\vec{\theta}_d$ and $\vec{\beta}_d$ should be taken to be Dirichlet distributions with parameter vectors $\vec{\phi}_d$ and $\vec{\gamma}_d$ respectively, $z_{d,n}$ to be categorical $q(z_{d,n} = k|\lambda) = \lambda_{d,n,k}$, and $q_{k,t}$ to be independently Gaussian distributed with mean $\mu_{k,t}$ and standard deviation $\sigma_{k,t}$. The complete variational distribution is shown in figure 3.3 and can be written:

$$\begin{aligned} q(\theta, \beta, z, Q) = &\prod_d \left(q(\vec{\theta}_d|\vec{\phi}_d) \prod_n q(z_{d,n}|\lambda) \right) \\ &\times \prod_k \left(q(\vec{\beta}_k|\vec{\gamma}_k) \prod_t q(q_{k,t}|\mu_{k,t}, \sigma_{k,t}) \right) . \end{aligned} \tag{3.18}$$

Then the gradients of the objective with respect to each parameter can be calculated. For λ, γ and $\{\mu, \sigma\}$, update rules allow us to find the minima with respect to each parameter explicitly. These updates can be found in appendix A. In the case of ϕ however, no such simple update exists, and a gradient following method is needed.

One might think that a sensible start point for inference would be a uniform variational distribution. However, since both the model and variational distributions are symmetric with respect to topic order, the gradients for each topic are identical when the topic parameters are the same. It is thus extremely important to break this symmetry before conducting updates. This is achieved by adding a small noise component to the uniform parameters.

3.4.2 Inference for sLDA and TFM using Gibbs sampling

The reason that Gibbs sampling is so compelling for LDA (and indeed for sLDA) is that so-called collapsed Gibbs sampling is possible. For sLDA or TFM this requires that, given a set of work topic tokens z, the optimal settings of all other latent variables can be found analytically. This means that all other variables can be marginalized out and the inference algorithm reduces to resampling z iteratively until convergence.

Before describing a Gibbs sampling method for sLDA, the exact form it should take to generate text and time series in the same way as TFM is given. The response variable for document d, $\vec{r}_d$, is drawn from the same distribution as used for TFM, replacing θ in the probability density function with $\overline{z}$ where

$$\overline{z}_{d,k} = \frac{1}{N_d} \sum_{n=1}^{N_d} z_{d,n,k}. \tag{3.19}$$

The time series elements are then given by

$$r_{d,t} = \frac{\rho}{\sqrt{v(\alpha)}} \sum_k \overline{z}_{d,k} q_{k,t} + \sqrt{1-\rho^2}\ \epsilon_{d,t}. \tag{3.20}$$

Collapsed Gibbs sampling (see page 43) for sLDA requires resampling from the complete conditional of a single $z_{d,n}$. If $z_{\backslash d,n}$ represents z excluding the word being resampled, the distribution from which $z_{d,n}$ is resampled is given by

$$\begin{aligned} p\left(z_{d,n} = k \mid z_{\backslash d,n}, w, r\right) &\propto p\left(z_{d,n} = k, z_{\backslash d,n}, w, r\right) \\ &\propto \frac{\big(\alpha + \#\{d,k\}\big)\big(\eta + \#\{k, w_{d,n}\}\big)}{M\eta + \#\{k\}} \int_Q p(R|z,Q)p(Q) \\ &\propto \frac{\big(\alpha + \#\{d,k\}\big)\big(\eta + \#\{k, w_{d,n}\}\big)}{M\eta + \#\{k\}} \times \frac{\exp\left(-\frac{1}{4}\vec{b}_k{}^{\mathsf{T}}(A_k)^{-1}\vec{b}_k\right)}{\sqrt{\det(A_k)}} \end{aligned} \tag{3.21}$$

where $\#\{d,k\}$ refers to occurrences in $z_{\backslash d,n}$ of topic k in document d, $\#\{k, w_{d,n}\}$ to occurrences in $z_{\backslash d,n}$ of topic k associated with the word $w_{d,n}$, and where the matrices A_k and vectors $\vec{b}_k$ for each k are given by the expression below.

$$A_k = \frac{1}{2}\left(I + \frac{\rho^2\, \vec{z}_d(k)\vec{z}_d(k)^{\mathsf{T}}}{v(1-\rho^2)}\right) \tag{3.22}$$

$$\vec{b}_k = \frac{\rho\, \vec{z}_d(k)}{2\sqrt{v}(1-\rho^2)} \sum_t r_{j,t} \tag{3.23}$$

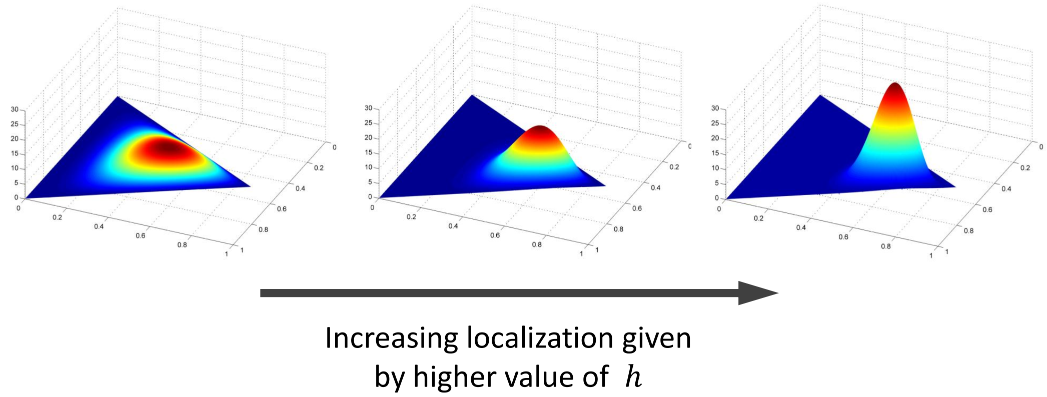


Figure 3.4: A toy example of changing h to localize the transition distribution. The 3-dimensional simplex is mapped onto an equilateral triangle, and the transition probability density function from expression 3.27 plotted for 3 values of h. Higher h gives rise to a concentration of probability mass near the current value of θ.

$\vec{z}_d(k)$ is the vector of topic frequencies in document d if the token $z_{d,n} = k$. It has elements

$$[\vec{z}_d(k)]_{k'} = \frac{1}{N_d}\left(\sum_{n' \neq n} z_{d,n',k'} + \delta_{k,k'}\right). \tag{3.24}$$

The method described above requires matrix inversion for each possible k at each word position and so has time complexity of $O\left(\sum_d N_d K^4\right)$. This can be reduced to $O\left(\sum_d N_d K^3\right)$ by noting that the matrices A_k can be efficiently computed from A_{k-1} using the Sherman–Morrison formula.

$$(A_{\setminus k})^{-1} = (A_{k-1})^{-1} + \frac{(A_{k-1})^{-1}\vec{z}_d(k-1)\vec{z}_d(k-1)^{\mathsf{T}}(A_{k-1})^{-1}}{\frac{\rho^2}{v(1-\rho^2)} + \vec{z}_d(k-1)^{\mathsf{T}}(A_{k-1})^{-1}\vec{z}_d(k-1)} \tag{3.25}$$

$$(A_k)^{-1} = (A_{\setminus k})^{-1} - \frac{(A_{\setminus k})^{-1}\vec{z}_d(k)\vec{z}_d(k)^{\mathsf{T}}(A_{\setminus k})^{-1}}{\frac{\rho^2}{v(1-\rho^2)} - \vec{z}_d(k)^{\mathsf{T}}(A_{\setminus k})^{-1}\vec{z}_d(k)} \tag{3.26}$$

z can thus be iteratively resampled relatively efficiently. After a sufficient number of iterations, the settings of z can be found in one of two ways. Either the settings of z for a single iteration can be taken as a point estimate, and the optimal settings of the other parameters computed, or a series of samples of z can be used to approximate the marginal posterior and corresponding posterior estimates for other parameters computed.

For TFM the complete conditional $p\left(z_{d,n} = k \mid z_{\setminus d,n}, w, R\right)$ is not tractable (since the time series factors in that case interfere with the integral over θ). It is possible, however, to collapse β when resampling z, and to sample from the complete conditional over Q. The key challenge to Gibbs sampling for TFM is then resampling θ. Rejection sampling is intractable because the probability density changes very quickly with respect to θ so samples drawn from distributions other that the true conditional are rejected with very high probability. It is thus necessary to use a Metropolis-Hastings sampling method. The effectiveness of this is dependent on being able to tune the variance of the transition density. This is achieved using a Dirichlet transition density with parameters given by a linear combination of the LDA conditional density and the previous sample for θ. The resampled variable is distributed

$$\vec{\theta}'_d \sim \text{Dirichlet}(\vec{v}_d) \tag{3.27}$$

where the vector $\vec{v}_d$ has elements

$$[\vec{v}_d]_k = \alpha + \#\{d, k\} + h\,\theta_{d,k}. \tag{3.28}$$

Figure 3.4 shows the effect of changing the concentration parameter h on this transition density. In section 4.2 it is shown that for an appropriately chosen value of h this Metropolis-Hastings sampling method can traverse the probability simplex quickly by maintaining a relatively high acceptance rate.

Inference using a Gibbs method thus comprises resampling first z, collapsed on β, then approximately resampling θ using a Metropolis-Hastings step, and alternating until convergence. The net result is a partially collapsed Metropolis-Hastings within Gibbs inference algorithm which is referred to in this thesis as MHWG. Full details of this resampling procedure are presented in appendix B.

3.5 Evaluating topic models

A common criticism of topic models is the difficulty of assessing the effectiveness of an inferred parameterization. Once a set of parameters are found which are locally optimal, there is no way of efficiently determining how good a solution has been found. One way to measure success is to hold out a portion of the corpus and find the probability of generating it with latent parameters inferred from the rest of the data. The Bayesian nature of topic modelling gives a choice of two measures of success depending on the context of the work. Most simply, inference can be used to find a point estimate of the parameters and the model evaluated on the basis of the likelihood of held-out data given this estimate. Alternatively, one can compute the expectation of this likelihood with respect to the posterior distribution over the parameters inferred from the in-sample data. In intractable cases the expectation can be approximated by drawing samples from the posterior.

The held-out portion of the corpus can be a whole document, in which case the likelihood assesses only the quality of the topics (and not the quality of the document-topic distributions) in the inferred solution. To test the document-topic distributions one would have to use so-called document completion methods, where the held-out data are individual words from each document. Success measurement based on held-out data requires inference to be performed repeatedly (once for each held-out portion of data) which is obviously extremely costly. For most applications it will be necessary to apply efficient methods.

For held-out documents the problem of repeated inference is not the only issue. Finding the probability of a document $\prod_n p(w_{d,n}|\beta, \alpha)$ given only the topic-word distributions β requires marginalizing over the topic assignments $z_{d,n}$ at each word position. This is intractable so must be approximated. A number of methods exist for estimating this probability:

- importance sampling using the prior on z or some more sophisticated proposal distribution (as in Li and McCallum [2006])
- annealed importance sampling
- the harmonic mean method, which uses the fact that the harmonic mean of $\prod_n p(w_{d,n}|z, \beta)$ where z are sampled from $p(z|w)$ is an unbiased estimator of the document likelihood [Grifiths and Steyvers, 2004]
- an estimator based on Chib's approximation [Chib, 1995]
- the left-to-right algorithm as introduced in Wallach [2008, p.65].

All of the above methods are described and compared in terms of likelihood, sensitivity to parameters and computation cost in Wallach et al. [2009a].

Document completion is an alternative where the probability of only a held-out portion of a document is used as the evaluation measure $p(w_{d,1:n}|w_{d,n:N}, \beta, \alpha)$. Since the training data contains a number of words in document d, $\vec{\theta}_d$ can be estimated. This gives the very simple option of using the estimated theta to find the probability of the held-out document. It was, however, shown by Wallach et al. [2009a] that this gives rise to a lower likelihood on held-out data than either annealed importance sampling or the left-to-right method.

Assessment using only held-out log likelihood is often criticised as contrary to the reasons for using topic modelling in the first place. That is, it ignores the coherency and interpretability of topics. Indeed, Chang et al. [2009] found negative correlation between likelihood and topic coherency as measured by human interpreters as the number of topics is increased. This is likely to be true for topic factor models in just the same way. Topic coherency measures are based on setting tasks to humans. In so-called word intrusion tests, an impostor word is added to words sampled from a topic to see if a human interpreter can identify the odd one out (see table 3.1). Alternatively in topic intrusion, an extract from a document is presented alongside a representation of topics sampled from the document's topic distribution and an impostor topic. The success rates of users identifying the impostors in these tests can provide excellent evidence for claims that a topic model is producing semantically meaningful structure.

home	steel	aerospace	customers
retail	bank	systems	management
medicine	banking	defence	businesses
brand	security	mining	investment
stores	services	aircraft	department

Table 3.1: A series of examples of word intrusion tests with words sampled from the most likely topic in a document and the intruder word sampled from another topic chosen at random. Intruder words are highlighted. Note that in the final test it is not clear which word is intruding. Human subjects will likely misidentify it and thus this topic will have a low score for semantic coherency from the word intrusion test.

Finally, relative performance with respect to a benchmark method can be assessed by using human judgement of coherency. The top words for a topic from each model can be placed side by side to allow human evaluation of their relative merit. The topic which receives the most votes should be considered to have superior thematic coherency. This was used to argue the superiority of pachinko allocation over latent Dirichlet allocation by Li and McCallum [2006].

3.6 Evaluating topic factor models

While semantic validity is the sole focus of most topic model applications, in topic factor modelling with financial data the inferred structure has a more nuanced meaning. The aim is to reflect underlying economic structure. For that reason tests in terms of semantic divisions wouldn't be satisfactory. Moreover, the structure in question is (at least in part) only familiar to experts, making recruiting test subjects difficult. Then there is the issue of subjectivity and competing interpretations of economic reality. Because of these issues this thesis focuses on likelihood based evaluation measures despite their known drawbacks. This is also more appropriate for modelling the time series portion of the data, which gives rise to the most interesting applications of topic factor models.

Typically in topic modelling, the measure of success is perplexity, the exponentiated negative cross entropy of the model distribution (given latent parameters) and empirical distribution of observed data. Log likelihood is preferred in this thesis, since it gives rise to the same ordering of parameterizations but has a clearer interpretation. The log likelihood per word

$$P_{\text{word}}(\theta, \beta, w) = \frac{1}{\sum\limits_{d} N_d} \sum_{d,N_d} \log\big(p(w_{d,n}|\theta, \beta)\big) \tag{3.29}$$

and log likelihood per time interval

$$P_{\text{time}}(\theta, Q, R) = \frac{1}{T} \sum_{d,t} \log\big(p(r_{d,t}|\theta, Q)\big) \tag{3.30}$$

are compared to evaluate models.

For out-of-sample testing of text, the estimated theta method described above (and in Wallach et al. [2009a]) is used. This is directly applicable to TFM. Because of the cost of repeated inference it is sometimes necessary to sample a set of words to exclude uniformly, rather than leave out each in turn. If v is a set of N_v held-out words $v_{d,n}$ taken from any number of documents, denoting the inferred parameters without the excluded set by $\{\theta_{\backslash v}, \beta_{\backslash v}\}$ the log likelihood of held-out text is given by

$$\hat{P}_{\text{word}}(v) = \frac{1}{N_v} \sum_{\nu} \log\big(p(v_{d,n}|\beta_{\backslash v}, \theta_{\backslash v})\big). \tag{3.31}$$

The likelihood of held-out time series intervals is more challenging. In all applications, time series intervals for assets arrive simultaneously. The appropriate portion of data to hold out then seems to be the time series intervals for each asset at a given time. In that case there is no posterior over $q_{k,t}$ to give the conditional log likelihood. This is a problem because it will later become apparent that there is a great deal of positive correlation between the factors underlying real data (the mean correlation of μ typically exceeds 0.2 for an equity corpus). Ignoring correlation in Q thus underestimates the correlation of the observed series. To combat this, the likelihood of held-out returns with marginalizing over the prior

$$p(R|\theta) = \int_Q p(R|\theta, Q)p(Q) \tag{3.32}$$

is not used. Rather, the likelihood of held out data is taken to be the log likelihood of a Gaussian with covariance given by the expectation, under the variational posterior, of the covariance of the training data. That covariance matrix V has elements given by

$$\begin{aligned}[V]_{d,d'} &= \mathbb{E}_q\left[\frac{\rho^2}{v(\alpha)}\vec{\theta}_d^{\mathsf{T}}\left(\frac{1}{T}QQ^{\mathsf{T}}\right)\vec{\theta}_{d'}\right] + (1-\rho^2)\delta_{dd'} \\ &= \frac{\rho^2\vec{\phi}_d^{\mathsf{T}}C\vec{\phi}_{d'}}{v(\alpha)\Phi_d\Phi'_d} + \left(\frac{-\rho^2\vec{\phi}_d^{\mathsf{T}}C\vec{\phi}_d}{v(\alpha)\Phi_d^2(\Phi_d+1)} + \frac{\rho^2\sum_k\phi_{d,k}C_{kk}}{v(\alpha)\Phi_d(\Phi_d+1)} + 1 - \rho^2\right)\delta_{dd'} \qquad (3.33)\end{aligned}$$

where $\Phi_d = \sum_k \phi_{d,k}$ and C is given by

$$C_{j,k} = \frac{1}{T}\sum_t \left(\mu_{j,t}\mu_{k,t} + \sigma_{k,t}^2\delta_{jk}\right). \qquad (3.34)$$

The measure of success for a held-out vector of time series data, $\vec{r}_\tau = [r_{1,\tau}, r_{2,\tau}, \ldots, r_{D,\tau}]^{\mathsf{T}}$, is then the log probability of the held-out data given this covariance. Note that the returns for the whole corpus are standardized before the held-out portion is removed.

$$\hat{P}_{\text{time}}(r_\tau|\phi,\mu,\sigma) = -\frac{D}{2}\log\bigl(2\pi\det(V)\bigr) - \vec{r}_\tau^{\mathsf{T}}V^{-1}\vec{r}_\tau \qquad (3.35)$$

This is referred to throughout this thesis as the log likelihood of held-out time series data. Holding out whole intervals of the time series also helps to highlight that it is the shape of the joint distributions of the asset returns $p(r_{d,t}, r_{d',t}|\theta_d, \theta'_d)$ which are most important. Typically when working on supervised topic models one would focus on conditional distributions of individual response variables.

In the next chapter, rather than an empirical corpus, data from known generative models are considered. In this case it is tempting to propose some sort of distance measure from the true parameters. A major problem with this stems from the symmetry of parameterization under permutations of topics. The distance from $\vec{\theta}_{d'}$ to $\vec{\theta}_d$ will change when two non-identical topics are permuted, while of course the order in which topics are presented has no impact on their semantic validity or explanatory power over the corpus. The document word distribution doesn't change with permutations of topics. One valid choice of evaluation measure for generative experiments would be the Kullback-Leibler divergence between the inferred and generative distributions. In the next chapter, the log likelihood of the data given the true generative parameters (θ^g, β^g, and Q^g) is chosen instead to provide a point of reference for the held-out log likelihood in empirical experiments. The log likelihood per word is given by

$$P^g_{\mathrm{word}}(\theta, \beta, w) = \frac{1}{\sum\limits_d N_d} \sum_{d,N_d} \log\bigl(p(w_{d,n}|\beta^g, \theta^g)\bigr) \tag{3.36}$$

and the log likelihood per time interval by

$$P^g_{\mathrm{time}}(\theta, Q, R) = \frac{1}{T} \sum_{d,t} \log\bigl(p(r_{d,t}|\theta^g, Q^g)\bigr). \tag{3.37}$$

3.7 Summary of topic factor models

This section described the use of topic modelling for text and time series data. No previous attempt has been made by the topic modelling community to model both text and time series data. The community has, however, produced a model capable of being used for that purpose. In section 3.4.2 the required generative form of the response variable was described. This was chosen to correspond to the closely related topic model which constituted the main contribution of this chapter. TFM, the first contributed model of this thesis, was described in full in section 3.2. The chapter also described how topic models for financial text and time series data can be interpreted and evaluated.

These models represent one of the two approaches to mining text and time series data provided in this thesis. They are evaluated using synthetic data in chapter 4 and using a corpus of equity data in chapter 5. The second, simpler approach is described and contrasted to this topic modelling approach in chapter 6.

Chapter 4

Validating Topic Factor Modelling with Synthetic Data

In this chapter synthetic data are used to validate the methods later used for analysis of financial data. Data corpora are constructed using the generative models described earlier. These artificial corpora are then used to test the importance of model specification, to compare the qualities of inference methods, and to examine the impact of changing the hyperparameters on the model. The results indicate the sensitivity of modelling to small changes in specification. They also show relative robustness to hyperparameter settings. Variational methods appear superior to Gibbs sampling based methods for both TFM and sLDA, in contrast to typical findings for LDA [Welling et al., 2008].

4.1 Comparing generative models

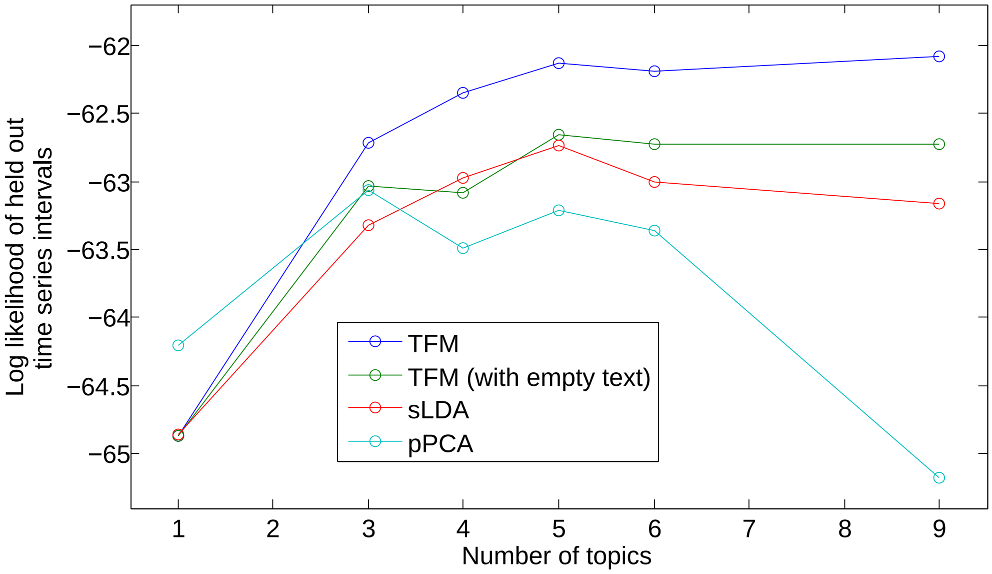
In order to justify the use of TFM in place of sLDA it is necessary to show that the small difference between the two models does result in material difference in inference. The features of the corpora generated by the two models are also important. It is possible to argue that the flexibility of TFM to allocate mass to topics for the time series without then generating text from that topic is critical and of particular importance in the case of financial data. This section seeks to validate that by comparing the parameters inferred from a synthetic corpus by the two models.

A corpus of 50 documents is generated using TFM as described in section 3.2 with 5 topics, a dictionary of 500 words, 200 time series intervals, and the number of words per document Poisson distributed with intensity 100. Inference was performed on corpora, holding out each time interval (for all documents at once) and subsequently holding out one word at a time from each document (sampled without replacement). Using these corpora the values of the latent parameters are inferred for both TFM and sLDA, using variational inference methods for both. The results presented are the mean log likelihood of these held-out data given parameters inferred from the rest of the corpus. A fuller discussion of the evaluation method is given in section 3.6. Results are divided into held-out text and held-out time series intervals.

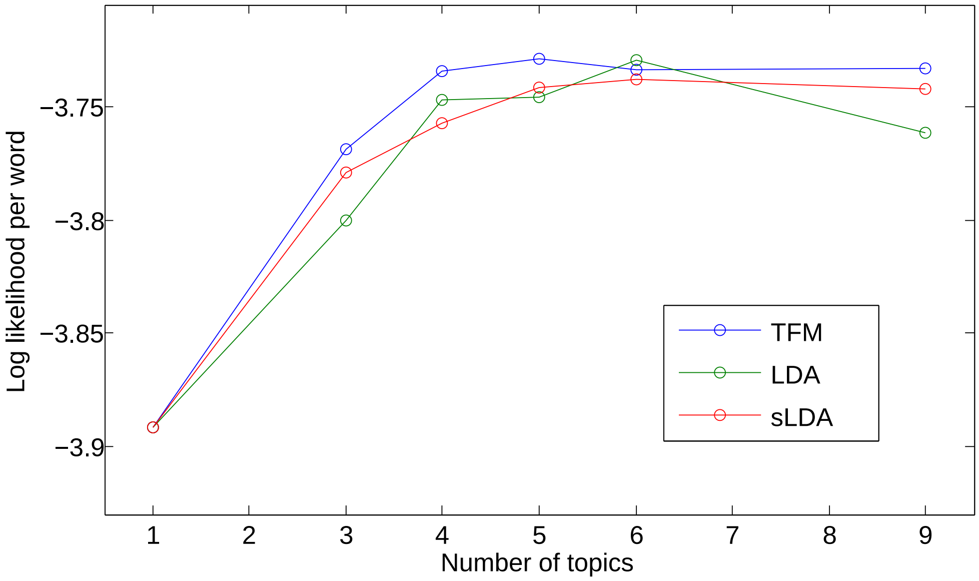
Figure 4.1(a) shows TFM outperforming sLDA in terms of the likelihood of time series data held-out from a synthetic corpus. Using sLDA on the joint corpus gives superior performance to training probabilistic PCA on only the time series data. However, sLDA proves to be not much better than using TFM with empty text data. Given that sLDA has the extra information of the text data it would outperform the time series only model if it were well specified (as indeed TFM does). This is a clear indication of the issue of model misspecification. When the model used for inference is different to the structure underlying the data, adding a second data type is no help in improving the parameter estimates versus using the first data type in isolation. This experiment represents a sanity check on the inference process, demonstrating in a controlled case that everything works as expected. When data are generated using TFM then, as expected, TFM is the best model with which to uncover the generative structure. It remains to be shown for any given application that TFM is more appropriate than sLDA.

Figure 4.1(b) shows the log likelihood of held-out text data. This case is different from the time series results. While TFM does outperform sLDA the difference is much smaller. The improvement over LDA (i.e. using just text) is marginal. It seems that it is easier to improve inference of time series parameters using text than to improve the text model using the time series.

Differences in time taken to perform inference under the two models are also important. Using similar, mean field, variational inference methods, sLDA is faster since it has closed form updates for each variable. On the above corpus, inferring 5 topics takes 89 seconds on average for TFM versus just 0.14 seconds for sLDA. The scaling properties in time are similar for both methods. The time complexity of a single variational update for each variable for sLDA is $O\left(K^2\left(K+T+\sum_d N_d\right)+DKT\right)$. For TFM the same complexity is $O\left(K\left(K^2+KT+KD+\sum_d N_d\right)+DT\right)$. These are both linear with respect to number of documents and number of time series intervals and cubic with respect to number of topics. The large timing differences can be attributed not to the complexity



(a)



(b)

Figure 4.1: Figures showing comparisons of TFM and sLDA on a synthetic corpus in terms of held-out text likelihood and held-out time series likelihood. Each is shown with an appropriate benchmark using only one data type.

of the updates but to the number of updates required for convergence. In particular, since the update for the document topic distribution is gradient following in TFM, it takes many updates to converge with all other variational distributions held equal.

Memory scaling is a problem for TFM inference unless limited memory methods are employed for the update in ϕ (see appendix A.5). For some applications with particularly large corpora the time cost for TFM might be unacceptable and any model misspecification would be a necessary sacrifice to perform inference in reasonable time.

4.2 Comparing inference methods

The choice of inference method can be very important both in terms of speed and avoiding local optima. In section 3.4 two methods are outlined for inference in topic factor modelling. For LDA collapsed Gibbs sampling generally proves to be the fastest method, and is also often able to find superior solutions [Teh et al., 2007; Welling et al., 2008]. To compare these methods in this section representative problems are generated and each method applied to them.

With the use of synthetic data, one has access to the true settings of the latent variables in the generative model. Because of this, it is tempting to measure the success of inference by some distance measure between the inferred variable values and the actual values generated. However, TFM is (like many other topic models) symmetric with respect to permutations of topic order. That is, the distribution of the observed variables is unchanged by permutations of the rows of θ so long as they are matched by permutations of the columns of β and Q. As a result for any value of the generative parameters there are $K!$ values for the inferred parameters all having equal quality, only one of which has minimum distance between itself and the generative parameters.

Thus comparisons are made between inference methods by assessing the log likelihood of the training data given a point estimate of the parameters from each inference method alongside the time taken to reach that point estimate. These log likelihoods are rescaled so that the figure for the generative parameters occurs at -1 (or 1 for positive generative log likelihood; note that there are continuous variables so the likelihood function may be greater than 1). This rescaled measure of success is given by the ratio

$$\frac{\log p(r, w, \theta, \beta, Q)}{\left|\log p(r, w, \theta^g, \beta^g, Q^g)\right|} \tag{4.1}$$

where θ^g, β^g and Q^g are the generated instances of the latent variables.

First, to demonstrate the performance of the two approaches on a smaller model a set of ten corpora are generated using the following set of parameters.

- $\alpha = 0.5$, the parameter for the Dirichlet prior over document topic distributions
- $\eta = 0.1$, the parameter for the Dirichlet prior over topic word distributions
- $\rho = 0.7$, the TFM parameter
- $D = 10$, the numbers of documents
- $N_d \sim \text{Poisson}(200)$, the number of words in document d

- $K = 3$, the number of topics
- $M = 2000$, the size of the dictionary
- $T = 250$, the number of time periods

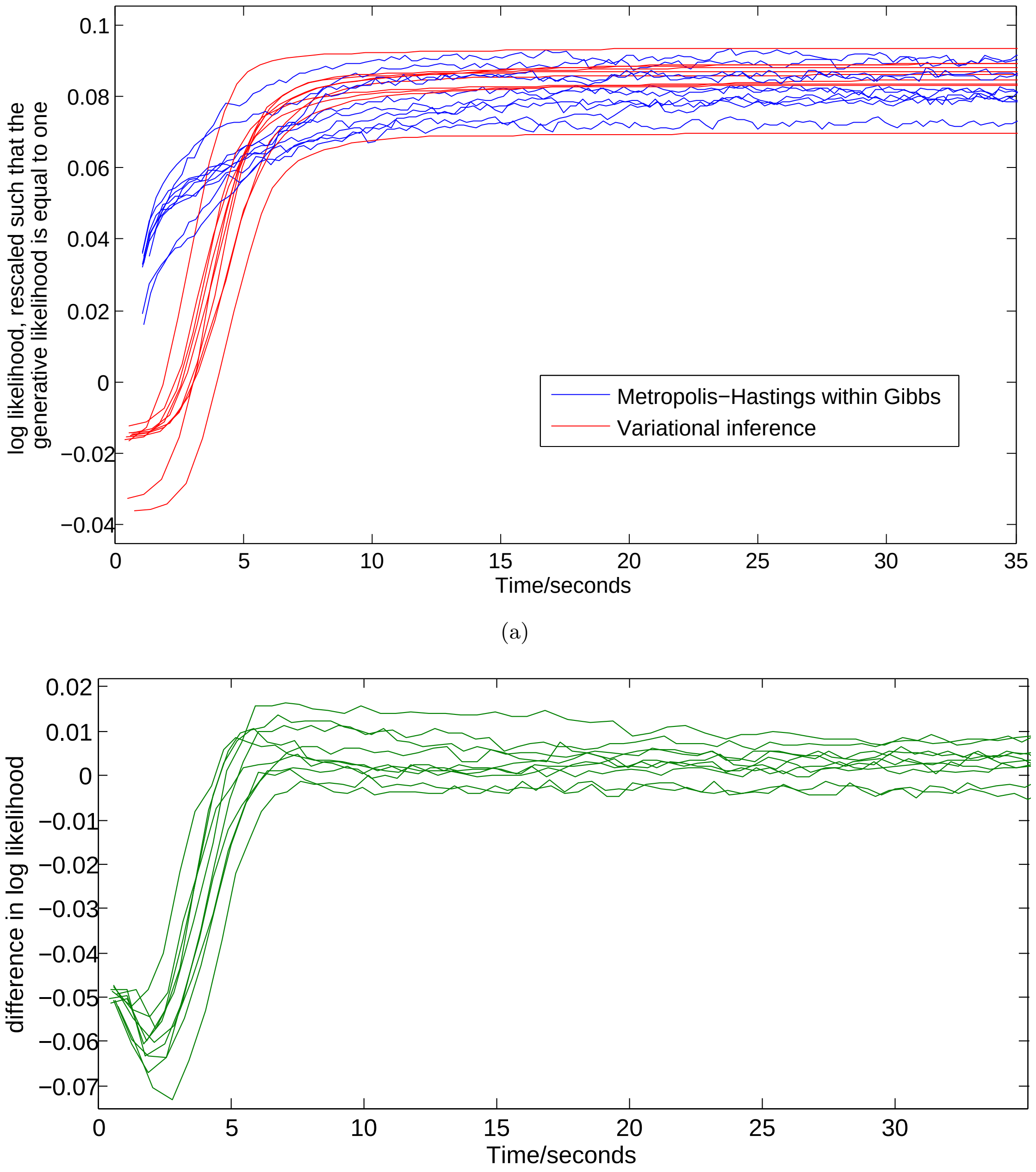
Figure 4.2(a) shows the performance of variational inference and the partially collapsed Metropolis-Hastings within Gibbs method (MHWG) described in section 3.4.2. The plots shown demonstrate effective convergence of both methods in a matter of seconds for this small problem. It also shows that both methods fail to reach a log likelihood comparable to that of the generative parameters. That variational methods should be stuck in local optima is not surprising, since they directly optimize a non-convex function. This work shows that MHWG is prone to getting stuck in local optima too in this case, and at similar likelihood levels. Since it is not clear from figure 4.2(a) which plots relate to which test corpus the differences between the log likelihood of point estimates from the variational method and MHWG are also plotted for each corpus. This more clearly shows that variational inference gives superior results in the majority of cases.

The next important quality is the scaling of the method with respect to problem size. There are a wide variety of variables which impact the size of the problem. For the variational method, optimization over θ consumes the vast majority of the time taken. D and K are therefore the most important considerations in computational cost. Within MHWG, each sample is made quicker by using a Metropolis-Hastings step. The computational bottleneck is then the matrix inversion required to resample Q, giving the complexity cubic dependence on K.

By way of an empirical comparison of scaling, results are now shown for a larger experiment generated in the same way as the previous one.

- $\alpha = 0.5$, the parameter for the Dirichlet prior over document topic distributions
- $\eta = 0.1$, the parameter for the Dirichlet prior over topic word distributions
- $\rho = 0.7$, the TFM parameter
- $D = 100$, the numbers of documents
- $N_d \sim \text{Poisson}(200)$, the number of words in document d
- $K = 10$, the number of topics
- $M = 2500$, the size of the dictionary
- $T = 250$, the number of time periods

Figure 4.3(a) demonstrates that both MHWG and variational inference scale relatively well; inference is still relatively quick even for this more realistic corpus size. The convergence rate for MHWG is more significantly impacted by the change in problem size. It also appears that in larger problems MHWG suffers slightly more from issues of local



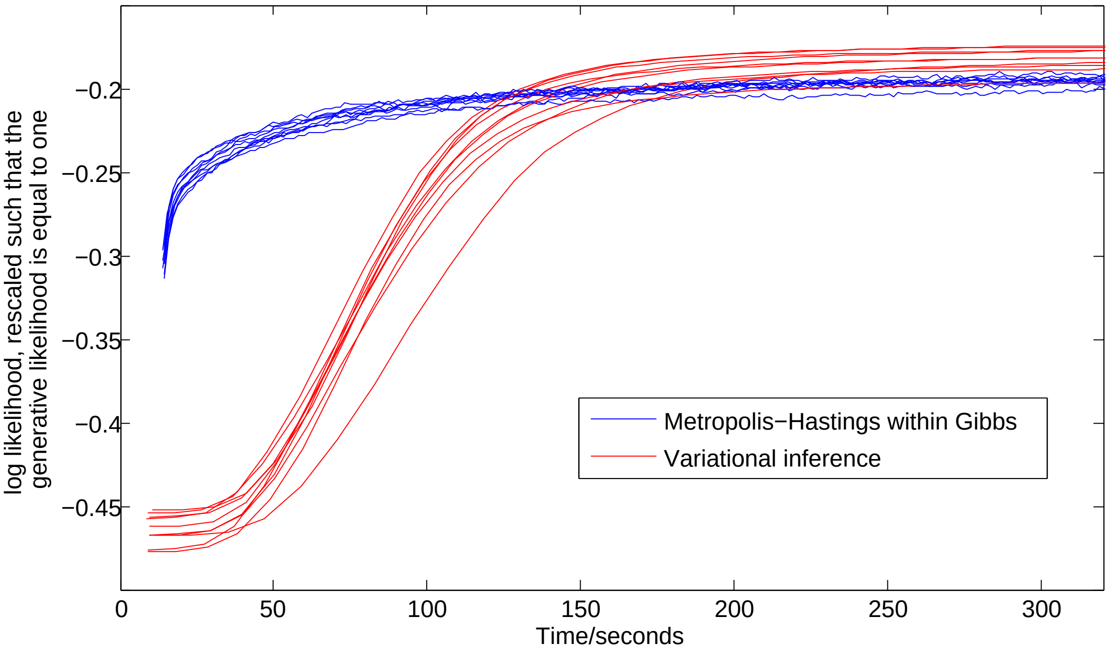
(a)

(b)

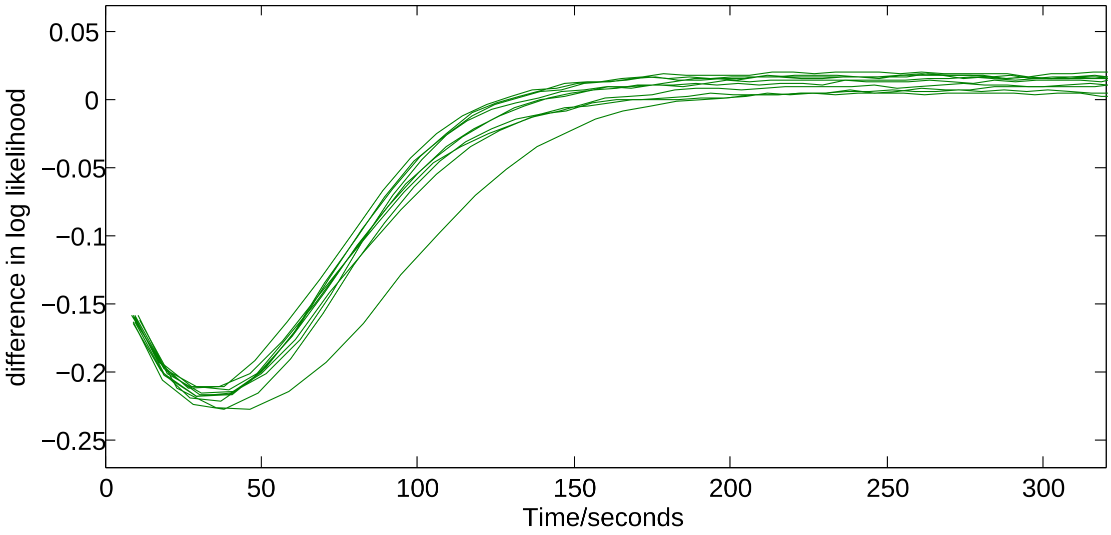
Figure 4.2: (a) Rescaled likelihood of generated training corpora given a point estimate of the parameters from the two inference methods. (b) The difference between the two methods for each sampled corpus. The plots begin after one iteration; the likelihood of the randomized start point of the algorithms is significantly lower.

optima, as shown by the lower rescaled converged log likelihood. Figure 4.3(b) shows MHWG underperforming in log likelihood for every one of these larger corpora.

The above experiment is now repeated, varying the transition density parameter h to check that it is not the value of this parameter which causes the difference in performance. The results in figure 4.4 show this not to be the case (indeed they show



(a)



(b)

Figure 4.3: (a) Rescaled likelihood of the larger generated training corpora given a point estimate of the parameters from the two inference methods. The three changes in problem size are in number of documents $D = 100$, size of dictionary $M = 2500$ and number of topics $K = 10$. (b) The difference between the two methods for each sampled corpus. The plots begin after one iteration; the likelihood of the randomized start point of the algorithms is significantly lower.

that for a wide variety of choices of distribution the MHWG approach is approximately equally effective). For reasons of both scalability and better optimization performance, variational inference is therefore used for all experiments on real corpora.

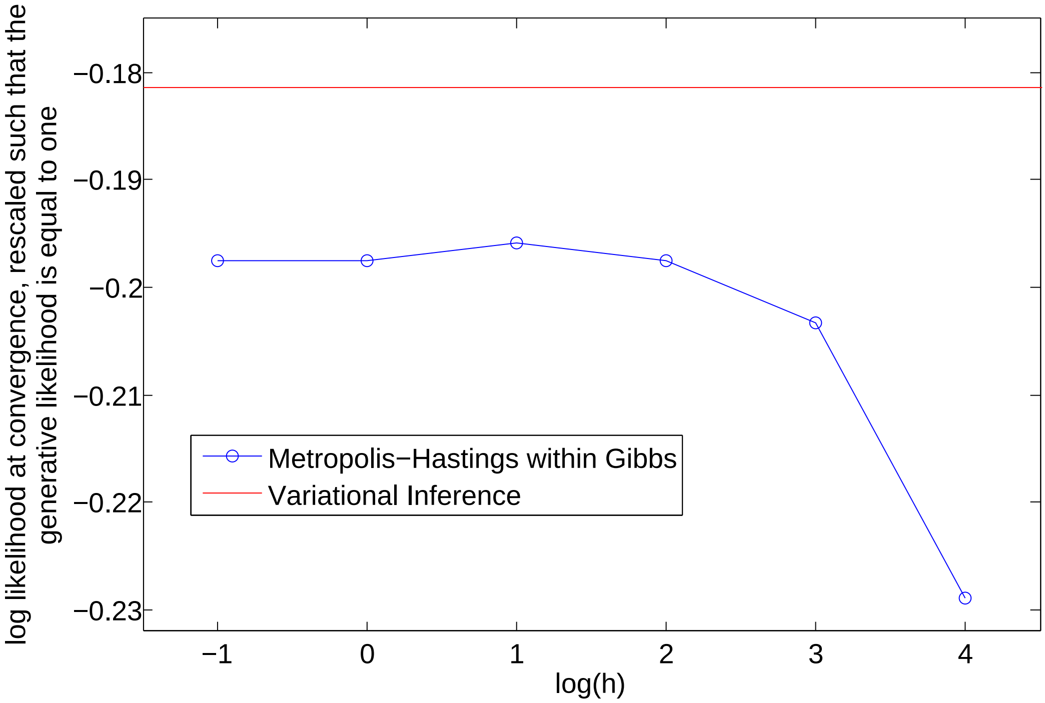
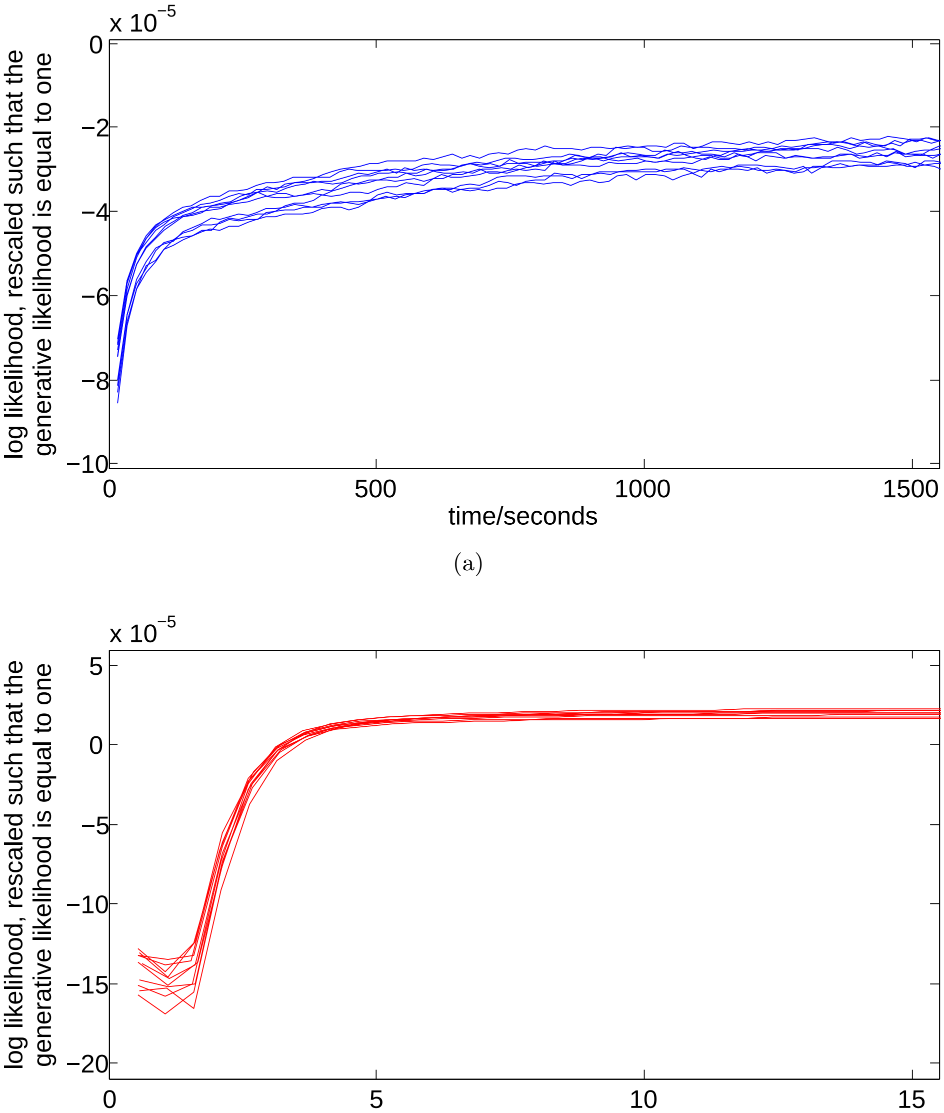


Figure 4.4: A comparison of the log likelihood of a corpus at convergence for varying values of the parameter h in the Metropolis-Hastings step. No value of h gives as strong a performance as that achieved using variational inference.

4.2.1 Comparing inference methods for sLDA

Part of the aim of this thesis is to compare TFM to sLDA. It is therefore equally important to determine the effectiveness of inference algorithms for sLDA. Variational inference is easier in sLDA than TFM because it is possible to apply analytic updates for every variable. Gibbs sampling in sLDA can be fully collapsed, but the resampling procedure for the topic tokens z proves to be costly. On page 63 the complexity of resampling z is found to be $O\left(\sum_d N_d K^3\right)$. To test the effectiveness of collapsed Gibbs sampling against variational inference ten corpora are generated using the comparable sLDA model (see section 3.4.2), with the following parameters.

- $\alpha = 0.5$, the parameter for the Dirichlet prior over document topic distributions
- $\eta = 0.1$, the parameter for the Dirichlet prior over topic word distributions
- $\rho = 0.7$, the TFM parameter
- $D = 100$, the numbers of documents
- $N_d \sim \text{Poisson}(200)$, the number of words in document d
- $K = 10$, the number of topics
- $M = 2500$, the size of the dictionary
- $T = 250$, the number of time periods



(a)

(b)

Figure 4.5: Figures showing the convergence of (a) collapsed Gibbs sampling and (b) variational inference for sLDA on a number of sampled corpora. These are plotted on separate axes because of the dramatic difference in convergence times. Note the initial decrease in likelihood in some of the series in (b). This is because the likelihood is given for a point estimate of parameters. Variational optimization tightens its bound on the posterior by decreasing the expression in equation 3.17 rather than the likelihood.

The convergence of the models is compared by finding the log likelihood of the corpus given a point estimate of the latent parameters at each iteration and recording the time taken to reach that iteration.

Figure 4.5 shows variational inference converging not only to better optima but also with a convergence time orders of magnitude shorter than collapsed Gibbs sampling. The difference between these methods is far greater than in the case of TFM, though it is important to note that the MHWG algorithm used for TFM has significant differences from the collapsed Gibbs method for sLDA. The difference is down to the resampling process for z. In TFM this is no harder than in Gibbs sampling for LDA, while in sLDA it requires matrix inversion for each word position, giving complexity of $O\left(\sum_d N_d K^3\right)$. TFM also requires matrix inversion (in its resampling process for Q, see appendix B.2) but only once per iteration, giving complexity of $O\left(K^3\right)$.

4.3 Choosing hyperparameters for topic factor modelling

In the case of real data, the hyperparameter settings are not known. There are a number of options available for setting them. The prior can be symmetric or asymmetric with respect to permutations of topics and/or words. The parameter for a symmetric Dirichlet prior is a single constant while asymmetric priors have instead a vector of parameters. In this case cross validation would become significantly more challenging because there are a great variety of possible priors to choose from. In some applications hand chosen priors could be a way to incorporate domain specific knowledge.

A Bayesian treatment of the priors could be used, by placing a distribution over the parameter vector itself. Wallach et al. [2009b] found an advantage in using asymmetric priors for LDA, in particular asymmetric priors on the document-topic distribution. They did this by placing a Dirichlet prior on the base measure for the generative Dirichlet itself. The advantage gained seems to be both relatively small and corpus dependent. The idea of a prior on the hyperparameter is redolent of Bayesian non-parametrics. Indeed non-parametric topic modelling has been attempted [Teh et al., 2006]. Given the additional challenges involved in inference in the non-parametric case and the small benefits over a parametric approach with effective parameter selection, this thesis uses only a parametric approach.

There are two factors motivating the choice of hyperparameters. Firstly, hyperparameter settings should give rise to high likelihood on held-out data, which is addressed in the rest of this section. Secondly, the inferred parameters should also represent an interpretable description of the corpus. Interpretability in β and θ relies on their distribution of probability mass being neither to concentrated nor too diffuse. For θ, the distribution for each document should be sufficiently concentrated that each document has significant contributions from only a small number of topics but not so concentrated that mixing of topics within a single document is unlikely. And similarly for β, the distribution for each topic should be concentrated on relatively few representative words but not so few that they cannot richly describe some recognizable topic. The set {customer, financial, insurance, investment} is more meaningful than {customer, financial}. The hyperparameter ρ determines the significance of the topic content of a document in its time series. Interpretability in this case requires only that ρ be larger than zero so that the topic content does contribute to the explanation of the times series.

To test the impact on inference of varying symmetric hyperparameter settings ten corpora are generated using hyper-parameters $\alpha_g = 1$, $\eta_g = 0.1$, and $\rho_g = 0.7$ and with the problem dimensions as follows.

- $D = 100$, the number of documents
- $K = 10$, the number of topics
- $N_d \sim \text{Poisson}(200)$, the number of words in document d

The latent variables θ^g, β^g, and Q^g are stored to find the generative log likelihood against which the inferred parameters can be benchmarked.

These benchmark figures are compared with the log likelihood of the inferred parameter set along with the log likelihood for held-out data (see section 3.5). For each corpus data are selected to hold out in two ways: text by sampling a single word per document to leave out (without replacement), and time series data by leaving out a randomly selected time interval. The model is trained for the whole corpus with each sample in turn excluded. The measure of success is the average of the log likelihood per held-out item as described in section 3.6. This sampling method is used rather than a systematic leave-one-out scheme for reasons of efficiency (training time for a single corpus is of the order of minutes). Sampling of the held-out portion is repeated 50 times to give reasonable coverage of the corpus. This measure on held-out data is presented to show results free from overfitting.

The results in figure 4.6 demonstrate that the inference algorithm is able to find solutions of comparable quality to the true generative parameters. Moreover, by changing the parameters it can be seen that the effect of inaccurate parameters is not that significant. Indeed, setting α, the parameter for the prior over document topic distributions, higher or lower than α_g by as much as a factor of ten doesn't prevent an effective latent parameterization from being found. The same can be said for changes in the TFM parameter ρ of as much as 0.1. Setting ρ or α too high causes overfitting to the time series data and a corresponding drop in out-of-sample performance in the likelihood of time series and text respectively. Note that the change in likelihood when ρ is changed can in part be attributed to the change in the variance of the idiosyncratic component $\sqrt{1-\rho^2}\ \epsilon_{d,t}$ rather than to finding a better inferred model of thematic structure per se. With lower

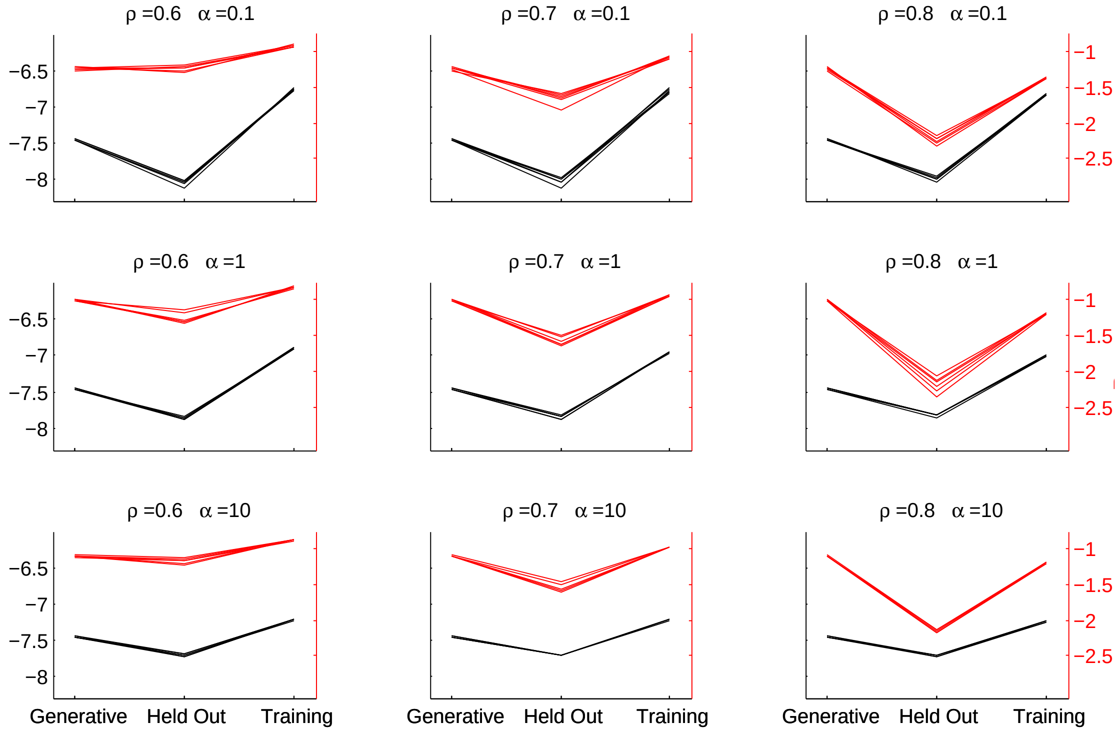


Figure 4.6: Log likelihood results for training and held-out data with inference conducted using varying hyperparameters. The likelihood of the true generative parameters of the ten different corpora are given alongside. The black series represent the average log likelihood of the text data and the red series the log likelihood of the time series intervals.

ρ the variance of this part is larger and the likelihood is more forgiving towards outliers. Another point of interest is that an inaccurate value of α doesn't significantly impact the likelihood of time series entries. And likewise the impact of the choice of ρ on the likelihood of held-out time series is minimal.

In contrast to the overestimated parameters, wherever the parameters are underestimated an improvement in out-of-sample performance is seen. Lower values of the hyperparameters seem to afford a degree of regularization to the inference process. As such, when dealing with real data one should tend towards underestimating ρ and α. Given the low sensitivity of the method to the choice of hyperparameters and the benefits of underestimating, a more sophisticated parameter selection method is not necessary. Similar results can be seen for the choice of η, the parameter for the prior over topic term distributions. However, for reasons of interpretability this decision can be simplified. Values of η significantly deviating from 0.1 do not give rise to models with realistic ratios of word frequencies. That is, if η is too high the topics are too concentrated on single words to be meaningful, and if too low they allocate the probability mass to more words than might reasonably be attributed to a single topic.

4.4 Summary of results from synthetic experiments

This chapter detailed the experiments conducted using corpora generated from the topic models themselves. These serve as evidence that the inference process works as intended, and give a point of comparison for inference on empirical data in chapter 5. The most important experimental result was the apparent superiority of solutions found using variational inference. This motivates the use of variational, rather than Gibbs based, methods in the experiments on equity data in chapter 5 and on foreign exchange data in chapter 7. The experiments also provide the basis for decisions of hyperparameter settings for real data. However, as described at the beginning of this chapter, the relevance of experiments on synthetic corpora is conditional on the the model being appropriately specified for real data. The next chapter provides evidence of the effectiveness of TFM for equity data.

Chapter 5

Experiments with FTSE 100 Data

This section describes experiments based on equity data. The text was drawn from publically available resources concerning companies which made up the FTSE 100 index on 1st October 2012 and the time series are constructed from the prices of those companies during the years 2010 and 2011. The results show some value in using topic factor modelling as opposed to independent modelling of text and time series. To the authors' knowledge this is the first time a joint topic model has been used to analyse financial text and time series data. The results from this chapter were presented at Business Analytics in Finance and Industry [Staines and Barber, 2014]. This chapter also contains a comparison of topic factor modelling to the equivalent formulation of supervised latent Dirichlet allocation. The improvement on sLDA motivates the use of TFM in its place, despite the complications in inference.

5.1 Real world data: the FTSE 100

The FTSE 100 index comprises 100 large cap stocks listed in the UK. The constituents are the companies with highest market capitalization having a full listing on the London Stock Exchange, as well as fulfilling some other eligibility criteria (for details see [FTSE Group]). The set of stocks for the experiments in this thesis are the constituents of the FTSE 100 index, as of 1st October 2012, which existed without significant change in corporate structure for every trading day in the years 2010-2011. That is, excluding International Consolidated Airlines, Glencore, and Polymetal International. Prices were taken (adjusted for stock dividends and splits) from Yahoo Finance [Yahoo.com].

The text data was taken from Bloomberg company profiles and the investor relations portion of corporate websites. This hand selection of data may introduce some bias to the process. Many corporations have an “at a glance” section from which text could simply be copied, but often this was too short or presented in a multimedia format. The data set used is thus not as homogeneous as would be preferable. The ideal text corpus for this section would be analyst reports on each company in some set, written by the same analyst so that all differences in themes presented are attributable to economic features rather than the style of presentation. Unfortunately such analysis is expensive and providers of such data are in general unwilling to share even old reports. Out of necessity then, the corpus used here was constructed by hand for this thesis. These data have been made available online [Staines, 2014].

5.2 Pre-processing steps

Reducing the size of the data can help to speed up inference and find more relevant structure. Written text contains content which has little relevance to the thematic content of the documents. The text data are sanitized by first removing all non-alphabet characters. A set of common stop words are also removed. These are function words such as “the”, “at” or “is” which don’t help to identify themes in a document. Stemming, the process of aggregating instances of words from the same root, can also help to reduce the size of the document-term matrix. Stemming techniques are not employed because of the tendency to over- or under-stem. A more drastic option is filtering with respect to TF-IDF [Baeza-Yates and Ribeiro-Neto, 1999]. The discriminative words are likely to have higher TF-IDF, that is they appear frequently in relevant documents (high term frequency) but are relatively rare in the corpus as a whole (high inverse document frequency). In the case of the FTSE 100 corpus the size of the vocabulary is not prohibitive, but in the case of larger corpora a TF-IDF filter could be used to cut the size of the problem while excluding only minimal discriminative information. After removing stop words the vocabulary size for the corpus used is 2654 words.

The historical price is taken for each day in 2010, adjusted for dividends and splits. The adjustments are made by adding back the dividend per share every time the stock goes ex-dividend, and by multiplying by the ratio after the date of any stock split. This removal of discontinuities in share price gives a return figure which represents the true economic rate of return from all sources (dividends and capital gains). The log return is then computed as given in equation 2.20. The log returns are centred and normalized because the modelling aim pertains to the correlation, rather than variance or mean of the returns. This standardization, with look-ahead adjustments to the returns, is permissible here because what is presented is a post-hoc analysis without time labels on the text. This is discussed further in section 3.2.

5.3 Assessing the inferred model

Where the true parameterization is not available (i.e. any real data sets) there is no obvious target likelihood. One can only hope to find a local optimum in the likelihood of the training corpus. It would also be reassuring to find average log likelihood on held-out data comparable to those found when inferring parameters with a synthetic corpus. The table below is an attempt to provide a reference point for assessing the results on real data. It gives the mean likelihood of held-out data, averaged over the held-out data and over 20 different start points. The corpus used is the same size as the FTSE 100 corpus and is generated from $K = 10$ topics. The generative parameters are the same as those used for inference: $\alpha = 1$, $\eta = 0.1$, and $\rho = 0.7$.

	per word	per time series interval
mean log likelihood ($K = 1$)	-7.752	-66.19
(standard deviation in the mean)	(0.008)	(0.07)
mean log likelihood ($K = 5$)	-7.860	-65.48
(standard deviation in the mean)	(0.006)	(0.06)
mean log likelihood ($K = 10$)	-7.786	-64.84
(standard deviation in the mean)	(0.004)	(0.08)
mean log likelihood ($K = 25$)	-7.742	-63.95
(standard deviation in the mean)	(0.002)	(0.05)
mean log likelihood ($K = 50$)	-7.731	-63.25
(standard deviation in the mean)	(0.003)	(0.06)

Table 5.1: A table showing log likelihood of held-out synthetic data for reference. The standard deviation in the mean is with respect to new (stochastic) initializations.

Restarting from different initial values might give more confidence that a good optimum has been found. However, in the case of TFM it appears that the dynamics of inference don't allow for gains to be made in this way. In table 5.1 the parameterizations found tend to have the same likelihood regardless of the stochastic start point as seen by the small standard deviations relative to the mean values.

Another point of reference is the likelihood from competing models. In this case there has historically been no attempt to jointly model the two portions of the corpus, but comparisons can be made to independent models of each data type. For this purpose LDA is most suitable for comparative, independent modelling of the text. The same hyperparameters are used for LDA as for TFM. As an independent model of the time series portion of the corpus, probabilistic PCA [Tipping and Bishop, 1999] is used. Finally, a Gaussian with mean zero and covariance given by the empirical covariance of the training data is used as a naive benchmark.

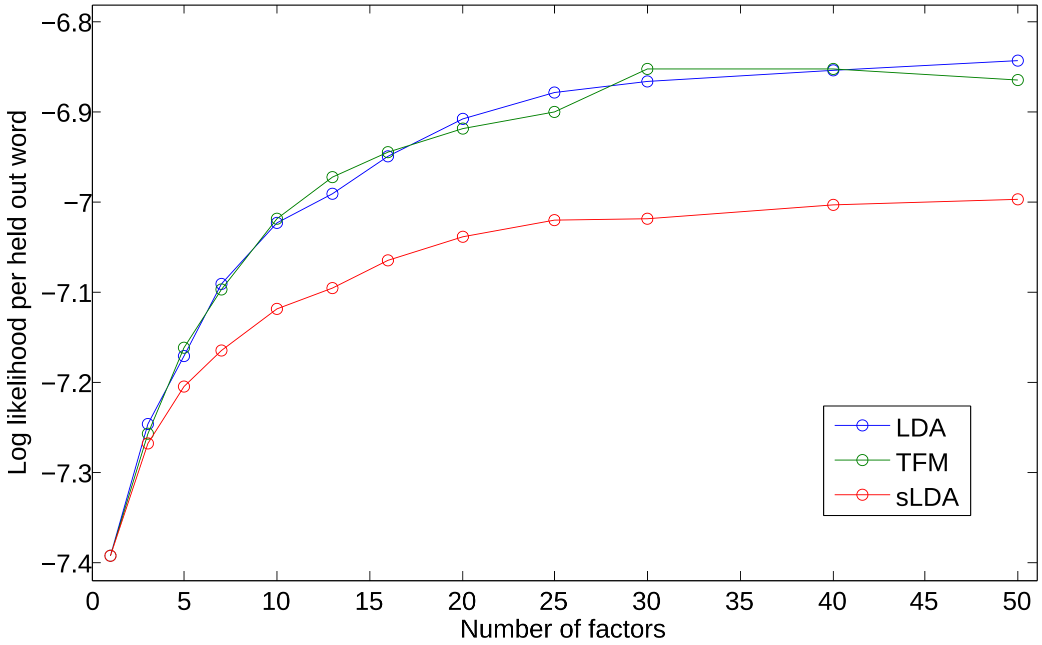
For comparison to a joint model sLDA is applied with the time series being generated by the response variable. The generative model is identical to TFM with the exception that r is dependent on the mean word label $\overline{z}_{d,k} = \frac{1}{N_d} \sum_{n=1}^{N_d} z_{d,n,k}$ rather than the topic weights. More detail are given in section 3.4.2. The log likelihood for sLDA is calculated in the same way as for TFM: using the log likelihood of a Gaussian with covariance given by the expectation, under the variational posterior, of the covariance of the training data. The same parameters are used for both sLDA and TFM:

- $\alpha = 1$, the parameter of the prior over document topic distributions
- $\eta = 1$, the parameter of the prior over topic term distributions
- $K = 10$, the number of topics
- $\rho = 0.7$, the text/time series balance parameter.

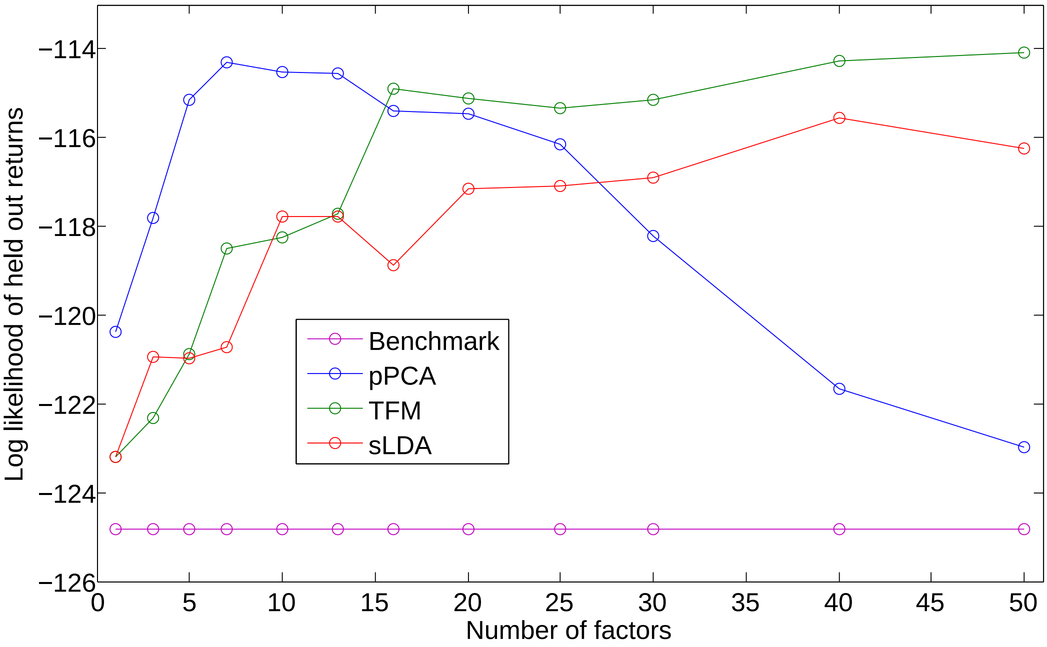
5.4 Results

To test the success of these competing models for a range of values of K each model is trained on the corpus first excluding one randomly sampled word per document (sampled without replacement) then excluding one whole time period. 100 words are sampled per document, and each of the 252 days in the corpus is held out in turn. Figure 5.1 shows the average log likelihood of held-out data for both data types.

Figure 5.1(a) shows that the likelihood of held-out text increases for all three topic models as the number of topics is increased, up to around 25 topics. Most importantly, sLDA suffers relative to LDA while TFM does not. That neither model is able to outperform LDA is somewhat surprising. Fitting to the time series data affects sLDA in a way to which TFM is more robust. This could be a result of the freedom to allocate weight to topics unused for modelling text. Under TFM, a topic can influence the distribution of returns without being heavily represented in the text tokens. Under sLDA the influence of a topic in the returns data can only be increased by allocating elements of z in the text to that topic. This is important because the text written about companies might not mention all of the relevant factors in correlation of equity prices.



(a)



(b)

Figure 5.1: A figure showing (a) the mean log likelihood of held-out text and (b) the mean log likelihood of the returns on a held-out day. The benefit of using the joint model is shown in the strong performance of both the TFM and sLDA on held-out returns when large numbers of factors are used. However, sLDA is seen to underperform in text likelihood. This supports the idea that TFM is a more appropriate model for the data.

For example it may not be apparent in the text whether a stock is aggressive or defensive. Other relevant factors might not be mentioned either because they are assumed by the authors to be obvious, or because they have not been considered by the authors of the corpus.

In figure 5.1(a), for LDA, sLDA and TFM, the log likelihood of held-out text is higher than for the synthetic data in figure 4.1(b) across all values of K. That these values are slightly higher reflects the fact that the structure in real text is more predictable than that categorical data sampled from LDA.

Figure 5.1(b) shows that for time series likelihood sLDA was approximately as successful as TFM. The best likelihood of held-out time series data is achieved by TFM with 50 factors. This was, however, not dramatically better than using pPCA with a small number of components. To some extent, this can be attributed to poor data quality in terms of homogeneity and extent of the corpus. With superior text data, TFM could be hoped to outperform pPCA. Notably, the joint models both improve when more factors are permitted. They appear to be uncovering more of the complex thematic structure in the data. By comparison, pPCA peaks in likelihood for small numbers of topics, with the likelihood of held-out data decreasing as the model complexity increases. It does not have the same protection from overfitting that topic modelling with Dirichlet priors offers. All three methods are dramatically better than the benchmark (a zero mean Gaussian with covariance given by the historical covariance), reflecting the importance of robust estimation for financial data.

The likelihood of held-out real data is significantly lower than was found for the synthetic data in table 5.1. This is because of the well-known long tails in financial data, illustrated for this corpus by figure 5.2. Extreme events have small log likelihood in a Gaussian model so drag down the average log likelihood for real data. Artificial data from Gaussian generative models are highly unlikely to include returns many standard deviations away from zero.

It is important to recognise that the behaviour of trained models is highly dependent on the corpus when time series data is used. Because the returns are heavy tailed fitting to extreme events can skew the output significantly. For instance, if the corpus included a day in which two companies happened to both announce excellent results their stocks would be more likely to be identified as closely related. The problem lies in the nature of financial time series: the signal-to-noise ratio is low and non-stationarity means that the number of data points is never sufficient to satisfactorily model a high dimensional space. The Gaussian model isn't sufficiently tuneable to fit arbitrary distributions over

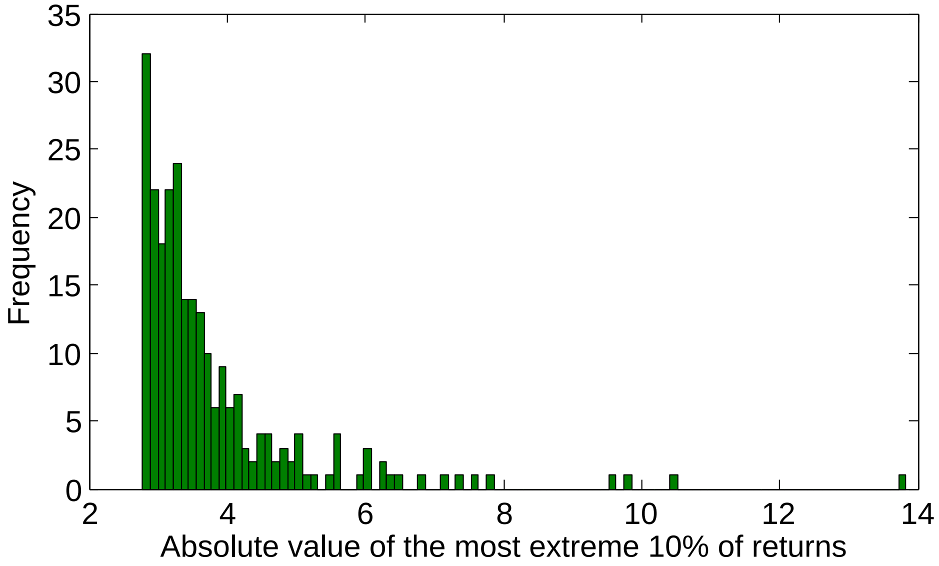


Figure 5.2: A figure showing the heavy tails in the returns data. The data are standardized. An extreme event of over 13 standard deviations in 24,444 entries demonstrates highly non-Gaussian behaviour.

the space. By comparison, a categorical distribution over a discrete variable, as used for the text variable, can be tuned to fit any possible distribution.

5.5 Qualitative analysis of topic content

Estimating parameters for a 20 factor TFM gives rise to a highly interpretable topic breakdown of the data set. In table 5.2 summaries are given of four topics found by the inference process, along with topics from LDA (run with identical parameters) which appear to correspond. It is important to note that there is nothing concrete connecting these topics to the corresponding topic from the other model. They are merely contributions to the two models which appear to have a semantic connection to a human reader. Where a theme is highly prevalent in the set of stocks examined, as is the financial sector in the FTSE 100, the prevalence of the vocabulary of that theme is sufficient to allow LDA to identify it effectively. This is clearly the case for the financial sector in the FTSE 100 corpus. Note also that the use of more cautious and humble language (not normally the realm of corporate publicity materials!) by financial companies in light of the events of recent years makes it easy for a topic model to identify this theme using only text.

Another strong presence in the FTSE 100 is mining. It is natural then that both models generate topics relating to mining. However, TFM does a better job of identifying this theme. The presence of intruder companies in the LDA column of 5.2 gives evidence of this. For instance GKN and AstraZeneca appear in the top 5 associated stocks but

		LDA	TFM
Topic 1 (Energy)	Most probable words	gas, oil, growth, deliver, market, production, strong, distribution	gas, market, power, engineering, product, global, service, oil
	Associated companies	Tullow Oil Bunzl Capita BG Group BP	Weir Group Compass Group Shell Smiths Group Meggitt
Topic 2 (Consumer)	Most probable words	product, service, brand, leading, Tinto, Rio, quality, building	product, brand, market, tobacco, food, care, home, portfolio
	Associated companies	CRH Admiral SAB Miller Tesco Aviva	Reckitt Benckiser Imperial Tobacco RSA Admiral Tate and Lyle
Topic 3 (Mining)	Most probable words	copper, mining, development, coal, world, resource, operation, gold	mining, coal, copper, Africa, ore, largest, operation, iron
	Associated companies	Randgold Eurasian Anglo American GKN AstraZeneca	Anglo American Randgold Kazakhmys Xstrata Tate and Lyle
Topic 4 (Financial)	Most probable words	service, UK, customer, banking, financial, management, insurance, investment	management, customer, financial, insurance, investment, UK, business, banking
	Associated companies	HSBC RBS Standard Life Barclays Old Mutual	Prudential Aviva Schroders Legal and General Standard Life

Table 5.2: A comparison of some topics inferred using TFM and LDA. The topics are matched and given labels by hand. For each topic, the eight most probable words (those w for which the inferred $\beta_{k,w}$ is greatest) and the five stocks (with greatest $\theta_{d,k}$) are included.

have nothing to do with mining. The text alone doesn't give sufficient identity to the topic. Under TFM a much higher weight is allocated to this topic by the mining stock distributions. The weight of Tate and Lyle allocated to this topic under TFM is not as incongruous as it might appear; Tate and Lyle is a commodities business (agricultural products and food ingredients, the refined sugar brand having been sold in 2010).

The financial and mining topics are in some sense the easiest, since the prevalence of companies and the similarities of language between companies makes them easy to identify. In more challenging cases, such as the consumer products/insurance topic in table 5.2, TFM seems to be producing more coherent groupings of words and companies (note for example the name of Rio Tinto, a mining company, appearing in the LDA list of probable words). There are of course a number of nonsense topics whose weight is low for all but a few of the companies in the data set. These are more prominent in the LDA output; two LDA topics have $\theta_{d,k} > 0.1$ for all but one company. In that case the topic simply approximates the content of that company's associated text, giving no insight into the thematic structure of the corpus.

5.6 Applying TFM to data summarization and covariance estimation

The parameters inferred from a corpus provide a representation of the thematic content of each company in the corpus. They provide a way to visualize, navigate or summarize the information that the corpus represents. In practice this might be useful as a primer for analysts in financial institutions who aren't familiar with the companies in the corpus (particularly when the number of companies is large). The summaries of themes could be used by economists to aid their explanatory models of past events.

One way to leverage the representation is to use it to compare companies. The thematic structure is in essence a description of the relationships between companies. This can be formalized by constructing a distance metric on the document topic content $\vec{\theta}_d$. For example,

$$d(\vec{\theta}_d, \vec{\theta}_{d'}) = \sqrt{\sum_k \left(\theta_{d,k} - \theta_{d',k}\right)^2} \tag{5.1}$$

gives a measure of the difference between company d and company d'. A matrix of these distances can be used for clustering. For example, it could be used to find sector clusters or to identify singleton companies which could be allocated extra weight in a portfolio in the hope that their greater degree of independence would deliver greater diversification

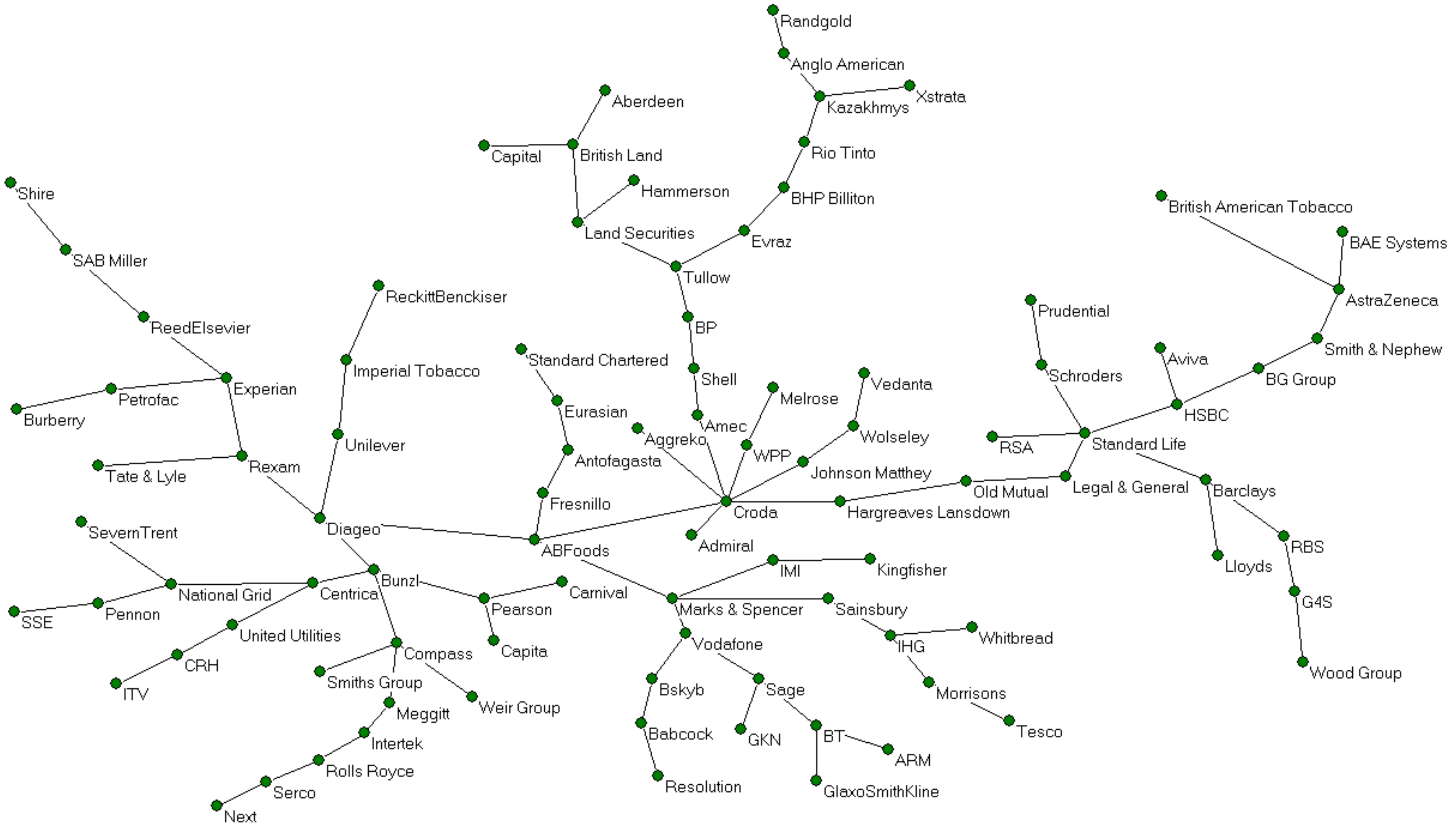


Figure 5.3: The minimum spanning tree of the constituents of the FTSE 100 using a distance matrix constructed from θ. Some obvious industry clusters include utilities, mining, and financial companies.

benefit. Figure 5.3 shows the spanning tree over all companies with minimum total distance between connected nodes. This provides an alternative to the spanning tree and associated hierarchical clustering methods using metric distances based only on historical correlation [Bonanno et al., 2001]. It is worth recognising at this point that these methods can poorly summarize correlation structure, ignoring significant links because of the tree constraint. This is the reason for the more recent work using less drastic filtering of the links, such as finding the optimal filter constraining the remaining graph to be embedded on a given surface [Tumminello et al., 2005]. For either embedded graphs or trees the relationships summarized can be enriched by using TFM to associate the nature of the captured relationships with interpretable topics. One might annotate a graph with the topic most strongly shared between two companies (which could be defined by, for example, $\max_k \min\{\theta_{d,k}, \theta_{d',k}\}$). This could reinforce the quantitative method by allowing better human interpretation.

Perhaps most significantly, a trained TFM can be used to construct a forecast for future covariance. TFM as presented generates standardized data with unit variance. It can therefore be used to find only the correlation. The covariance estimate must be constructed from this by multiplying the correlation prediction by the standard deviations σ_d of the assets.

$$\mathrm{cov}(d, d') = \mathrm{corr}(d, d') \times \sigma_d \sigma_{d'} \tag{5.2}$$

TFM was designed not to learn these standard deviations because they are implied by the prices available in options markets. Observation of the so called “volatility smile” most likely gives a better range for the future volatility than any that could be estimated from historical data. Option implied correlations are far harder to extract and have been shown not to give the same benefits given by option implied volatility [DeMiguel et al., 2013].

While the model above generates independent $r_{k,t}$ the inferred parameters show an average correlation of 0.262. This is not surprising, since equities show positive correlation with very few exceptions. A significant contribution in correlation comes from this factor correlation, so this should be included in the forward prediction rather than directly taking the generative covariance of the model. A covariance forecast of the form given in equation 3.33 is required. This is indeed the motivation for the form of the covariance prediction for held-out data.

This predictor gives a better estimate of the correlation in the following year than simply taking the empirical covariance to be the prediction. The KL-divergence between the forecast density and the empirical distribution for the following year is only 25.40 while for the historical covariance matrix it is 47.11. Far more evidence would be required to justify this predictor reliably, but robustification by removing components not associated with textual similarity (and as a result assumed to be noise) is attractive and deserves attention. It may be useful as a shrinkage target if not a forecast in its own right.

5.7 Summary of equity results

The key findings of this chapter were that TFM outperforms sLDA on both text and time series data, making it the best extant topic model for this corpus. It failed, however, to significantly outperform the best methods for independent modelling of the two portions of the corpus. While this can in part be blamed on data quality, it is disappointing in light of the performance gains found in modelling the joint corpus using matrix factorization methods in chapter 6. As well as the numerical results, qualitative assessment of the inferred output provided compelling justification for this type of joint modelling. This chapter marks the end of development of the topic modelling approach to mining text and time series data in favour of the second approach developed in this thesis: matrix factorization.

Chapter 6

Constrained Matrix Factorization: A Discriminative Approach to the Topic Factor Modelling Problem

For the most interesting problems in topic factor modelling a full generative model is not necessary. This chapter describes a novel, discriminate method which attempts to capture some of the same information which makes TFM successful. This method, which is referred to as matrix factorization in this thesis, focuses on modelling $p(R \mid w)$ while ignoring the generative process for text. It is in essence a supervised analogue of the matrix factorization methods discussed in section 2.2. This simplicity allows it to achieve superior performance while reducing the time costs of training dramatically.

6.1 Framing the topic factor modelling problem as constrained matrix factorization

LDA can be thought of as an attempt to find an approximate factorization of the document-term matrix into two lower rank matrices: the document-topic matrix θ and the topic-term matrix β. Topic factor modelling has to find this same factorization while also decomposing the time series returns into a document-topic matrix, that same θ from the text model, and a topic-return matrix Q. This joint factorization was chosen to give rise to a full generative model with its resultant flexibility. If an application doesn't require the power of a generative model, or doesn't concern itself with decomposing the thematic structure of text or generalizing to new text, then a simpler model might be preferable.

The aim is still to find a low-rank factorization of the return data. As before, some latent structure made up of θ and Q must be found from the observed variables w and R. The size of the data (D, M and N_d) is given by the corpus but the number of topics K is another fixed value which must be chosen. As a start point, take the approximation of a data matrix by a matrix product (see approximation 2.5). Since their interpretation is related to the latent variables in TFM, the matrix of factors is denoted Q and the weights θ.

$$\vec{r}_d \approx \sum_k \theta_{d,k} \vec{q}_k \tag{6.1}$$

The difference between this simple factorization and TFM originates chiefly from the dependence of the topic document matrix θ on the text for that document. This dependence could be built into a model in a far simpler way by making the document topic matrix a deterministic function of the text data. Recall the bag-of-words representation X with elements

$$x_{d,m} = \frac{1}{N_d} \sum_{n=1}^{N_d} I[w_{d,n} = m]. \tag{6.2}$$

Then θ should be some function of X. The document-term matrix X is renormalized so that the rows sum to one (to eliminate the impact of varying document size) and θ constructed using the softmax of the product of X and a parameter matrix β with dimension K by M. The parameter matrix is again named for its correspondence to variables in TFM.

$$\theta_{d,k} = \frac{\exp\left(\sum_m x_{d,m}\beta_{k,m}\right)}{\sum_j \exp\left(\sum_m x_{d,m}\beta_{j,m}\right)} \tag{6.3}$$

The empirical correspondence between the two variables is shown in section 6.4.

The approximation of R by θQ can be tightened from a randomly chosen start point by minimizing the total squared error.

$$\sum_{d=1}^{D} \sum_{t=1}^{T} \left(r_{d,t} - \sum_k \theta_{d,k} q_{k,t} \right)^2 \tag{6.4}$$

The gradients of this objective can be computed analytically and are given in appendix C. An efficient gradient following method, such as L-BFGS [Nocedal, 1980] can be used directly on the joint parameter space. Note that the start point needs to be randomized to give asymmetry, just as for LDA.

An alternative choice of summary function is simply the product of the two matrices.

$$\theta_{d,k}^{\mathrm{RRMR}}(W) = \sum_m x_{d,m}\beta_{k,m} \tag{6.5}$$

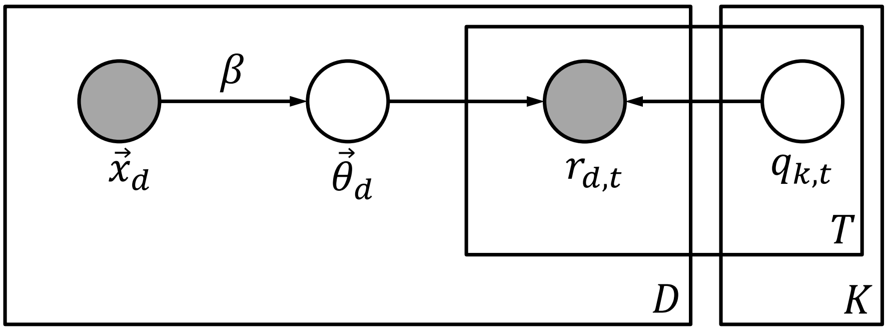


Figure 6.1: The graphical model for a generative process corresponding to the constrained matrix factorization described in this chapter. The shaded nodes are the observed reweighted vector of word frequencies in document d, $\vec{x}_d$, and the corresponding returns at time t, $r_{d,t}$. The returns are dependent on the factor returns $q_{k,t}$ and the weight vectors $\vec{\theta}_d$ constructed using the word frequencies and the parameter matrix β. The structure of the model is significantly simpler than the TFM but does not generalize to new text. For an explanation of graphical models see section 1.4.

This reduces the problem to reduced-rank multivariate regression (RRMR, see for example [Izenman, 1975]). In RMRR multivariate regression is performed with the restriction that the regression parameter matrix has rank smaller than that of either the regressor matrix or the dependant variable matrix. RRMR is closely related to the matrix factorization method described here, but permits negative weights. This can be harder to interpret. Essentially RRMR allows additional freedom but is less structured. In common with the methods in section 2.2, both of the matrix factorization method and RMRR break the data matrix into a factors and weights. The difference is simply that the weights are dependent on the text data.

Minimizing the squared error is equivalent to maximum likelihood for a Gaussian model with isotropic noise. This suggests a conditional probability distribution for R of

$$r_{d,t} \mid Q \sim \mathcal{N}\left(\sum_k \theta_{d,k} q_{k,t}, \sigma^2_{\text{CMF}} \mathbf{I}\right). \tag{6.6}$$

The maximum likelihood setting of σ^2_{CMF} for a given factorization is simply the variance of the residuals. To generalize to new text there must also be a generative process for R. Here, drawing $R_{t>T}$ from a multivariate Gaussian with mean zero and variance taken from the inferred values of $R_{1:T}$ is proposed. The probability of a new time series interval is then given by

$$\vec{r}_{t>T} \sim \mathcal{N}\left(\vec{0}, \theta(X,\beta)^{\mathsf{T}} \text{cov}\left(Q\right) \theta(X,\beta)\right) \tag{6.7}$$

where $\vec{0}$ is the vector of zeros. This likelihood is related to the one used to measure the success of TFM (see equation 3.33).

This matrix factorization approach can also be thought of as a feed-forward neural

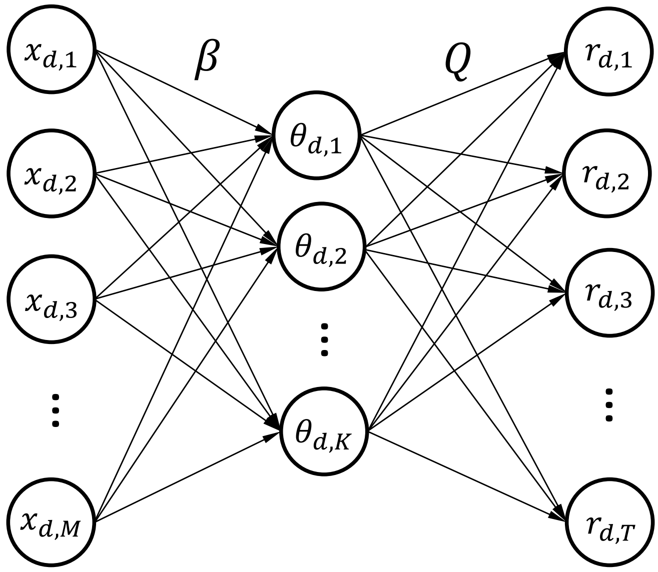


Figure 6.2: Matrix factorization as a neural network. The learning algorithm seeks to explain the relationship between text and time series data using least squares regression.

network (for more detail on neural networks see, for example, [Bishop, 1995]). Figure 6.2 shows a graphical representation of this. The hidden layer is constructed deterministically from inputs X using the function θ, parameterized by β. The output layer R is stochastic with distribution given by 6.6.

Regularization

Since the size of the parameters in this model is large compared to typical data sizes it is important to guard against overfitting to prevent the solution being dominated by noisy text. This can easily be achieved by adding a regularizing term to the objective. A regularized objective is proposed:

$$\sum_{d,t}\left(r_{d,t} - \sum_k \theta_{d,k} q_{k,t}\right)^2 + \zeta_1 \sum_{k,m} \beta_{k,m}^2 + \zeta_2 \sum_{k,t} q_{k,t}^2 \tag{6.8}$$

where ζ are regularizing parameters to be chosen by cross validating to optimize mean held-out likelihood. The best choice for these parameters will change with the corpus and the chosen number of topics. Because of the two independent parameters, cross validation is very costly. This is mitigated to some extent by the speed of optimization.

One shouldn't expect every word to be discriminative for every topic. For instance, the word "management" might not give either positive or negative information about the weight allocated to an energy related topic. For this reason a sparsity encouraging form of regularization on β might be appropriate. Then the topic-word weights would only be non-zero on some pertinent subset of the dictionary. The first regularization term is in that case replaced with $\zeta_1 \sum_{k,m} |\beta_{k,m}|$. This is referred to in this thesis as sparse matrix factorization. The regularization parameter is again to be determined by cross validation, which must be performed a second time in its entirety since the second parameter ζ_2 may have different optimal values to the non-sparse case.

6.2 Results using the FTSE 100 corpus

As a first test of the effectiveness of this new approach to the problem the experiment from section 5.4 is repeated, testing TFM, sLDA and pPCA against both standard and sparse matrix factorization using the likelihood of held out returns. It is clear from figure 6.3 that matrix factorization represents an improvement on the model used in TFM. The non-sparse version indeed shows material outperformance versus pPCA for a number of factors ranging between 3 and 50. While TFM demonstrated that text data could inform a time series prediction, an appropriate model is also necessary to capture this benefit. Matrix factorization seems to offer that. It is interesting to note that, while a topic modelling approach has highest likelihood for held-out data with larger numbers of topics, matrix factorization peaks at 10 factors, approximately the same as the 7 factors which proved the optimum for pPCA . The non-sparse version of matrix factorization has strong performance for higher factor numbers. The sparse version, by comparison, gives a likelihood that falls off greatly at higher factor numbers and is beaten by the non-sparse likelihood for any number of factors greater than 3.

6.3 Generalizing to new companies

One of the key strengths of linking text and time series by matrix factorization is that the model can easily generalize to new documents. This allows calculation of $\vec{\theta}_{d'}$ for a new company d' whose stock price time series is not included in the corpus. It is thus possible to predict likely correlation of companies for which time series data are not available. This could be hugely useful for determining the risk properties of investments in private equity, or opportunities arising from IPOs or spinoffs.

To demonstrate this application the parameters of the model are inferred using 2010 time series and text, excluding one company from the corpus and using $K = 20$ topics and setting the regression parameter by cross-validation. The 2011 correlation is then predicted using expression 6.7. An alternative method using human input might chose a comparable company, taking the historical correlation of that company with the rest of the corpus as a prediction of the future correlation for the new company. In this case comparable companies are identified by finding another FTSE 100 constituent labelled with the same sub-sector in their Bloomberg profile. Both correlation predictions are

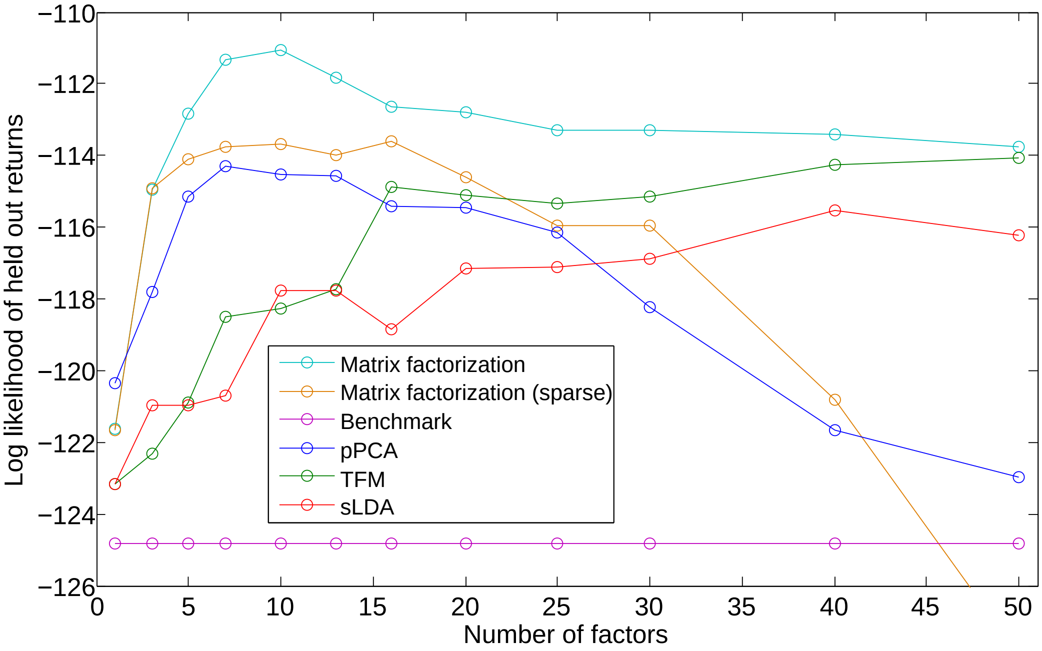
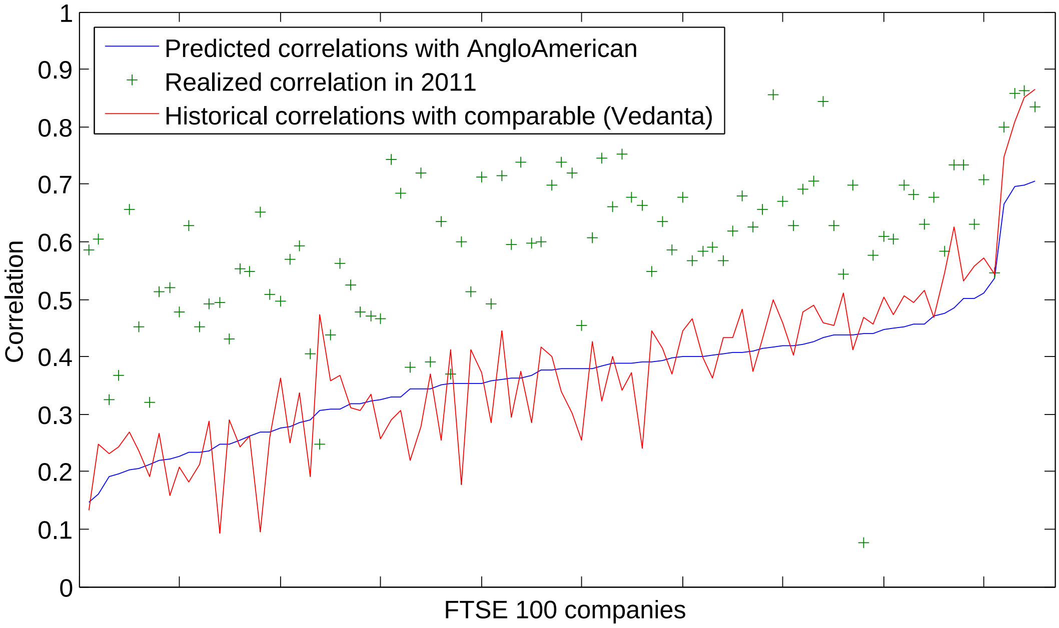


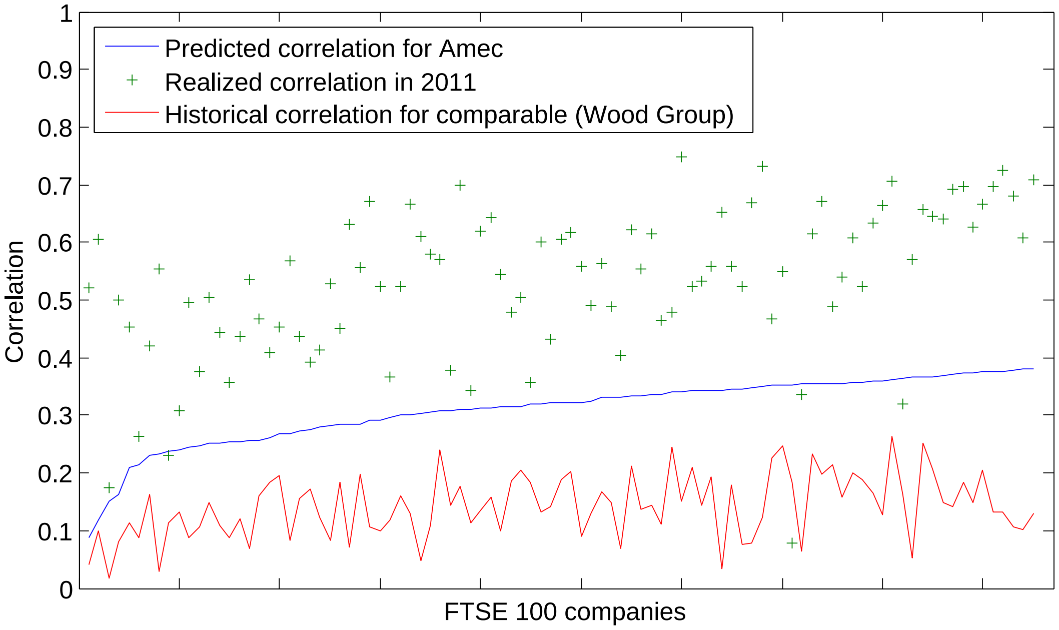
Figure 6.3: The log likelihood of matrix factorization with two kinds of regularization. Probabilistic PCA is included for comparison, as well as a benchmark based on the empirical covariance of the training data.

plotted against the values that were realized in 2011 in figure 6.4 for two examples. It should first be noted that correlation within the FTSE 100 rose between 2010 and 2011 so any prediction based on recent historical data will be an underestimate. The matrix factorization predictor for Anglo American in factor 6.4(a) is similar to the comparable company based method because of the similarity in their text. Predictions from both methods demonstrate an ability to identify, in a relative sense, the companies which will have higher and lower correlation with a held-out company. Figure 6.4(a) shows the potential value of a more sophisticated method than drawing on the historical correlation for a single comparable company. Wood Group, used as a comparable for Amec, had low correlation with all of the companies in the corpus in 2010. The matrix factorization predictor is able to avoid this large underestimation by using information from the broader corpus. An analyst producing an estimate of the likely future correlation would of course be able to construct a better predictor than the single company comparable in this case (perhaps noting the historically low correlations for Wood Group), but matrix factorization is already robust to this type of issue without human input.

In figure 6.5 the results are shown for a larger group of companies. The squared error in correlation prediction is relatively large because of the increase in correlation across all equities which is mentioned above. Matrix factorization gives a predictor with error similar to the historical comparable method in all cases, and significantly better in the case of Amec. Using a comparable company is contingent on analyst input to



(a)



(b)

Figure 6.4: The predicted correlations and subsequent realized correlation of (a) Anglo American, and (b) Amec with the rest of the FTSE 100. The companies are sorted in order of predicted correlation from matrix factorization. The red lines are the historical correlations of a comparable company (Wood Group and BHP Billiton respectively) with each other company. The historical correlation of these comparables could also be used as a predictor of the correlation. The green points show the empirical correlation in the following year. (b) highlights the risk of relying on historical data; Wood Group had unusually low correlation with the rest of the FTSE 100 in 2010.

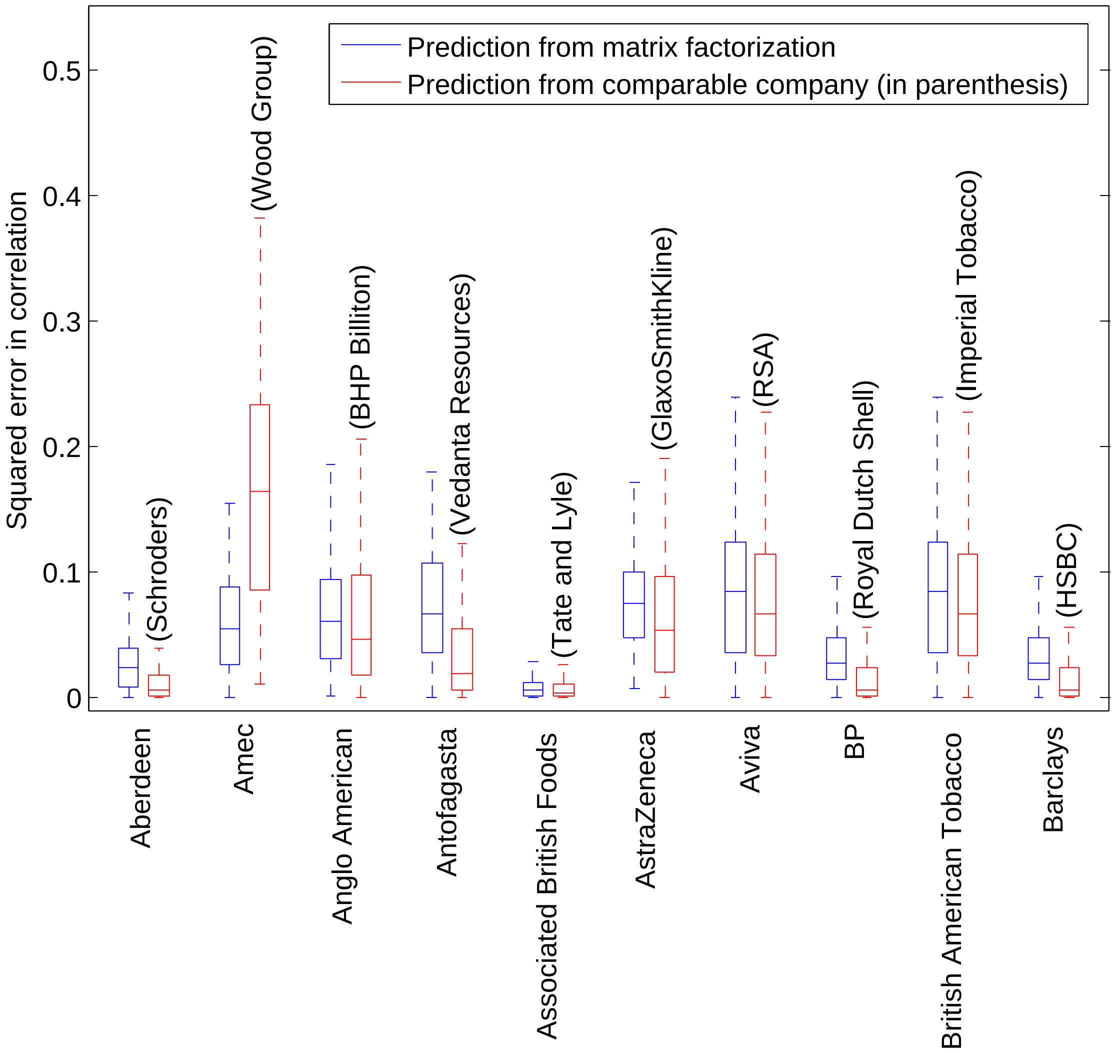


Figure 6.5: Box plots of the squared error in correlation predictions for the 95 companies in the corpus (excluding the new company and the comparable) based on matrix factorization (in blue) and the historical correlation for a comparable company (in red). The boxes show the quartiles of the error, and the whiskers the maximum and minimum error. The comparable companies used are noted in parenthesis next to their respective box plots.

identify appropriate comparables (on the data available on Bloomberg in this case) and on the existence of a suitable comparable. In many cases a new company may not have a direct comparable (note that only half of the first 20 companies in the FTSE 100 have direct comparables also in the index). In that case it might be necessary for an analyst to construct a time series from financial reports or some weighted combination of share prices. While matrix factorization can only reflect relationships present in the text data for new companies, this appears to be a highly effective and quick method for estimating correlation in the absence of historical price data.

6.4 Comparison to topic factor modelling

This matrix factorization approach is far simpler than the topic modelling used in the rest of this thesis, as shown by the graphical model in figure 6.1. The parameters, though, are closely related to the parameters in TFM and can be used for many of the same topic modelling type applications. The document time series are approximated by a linear combination of K time series which correspond to the topic time series from TFM. The parameters $\beta_{m,k}$ show the strength of relationship between each word and topic. They should thus be related to their namesake in TFM, which is made up of distributions over the dictionary for each topic. The aggregation of the document content $\vec{\theta}_d$ defines the contribution from each of the K time series, functions exactly like the $\vec{\theta}_d$ in TFM.

Just as with TFM, the output of matrix factorization can be summarized by identifying the most relevant words or documents to a factor. The most relevant words are those with maximum weight in $\vec{\beta}_k$ and the most relevant documents those with maximum weight in $\vec{\theta}_k$. For the FTSE 100 corpus these factors seem to be just as interpretable as topics arising from topic modelling. This in itself is telling, since this simple model has no sense of semantic relationships other than the impact they have on the correlation of document time series.

Interpretation	Top words
Factor 1: Utilities	networks, services, waste, north, electricity, water
Factor 2: Energy	delivering, oil, generation, power, turbines, gas
Factor 3: Finance/Insurance	life, customers, financial, management, investment, insurance
Factor 4: Technology	software, designs, chips, applications, digital, semiconductor
Factor 5: Tobacco	cigarettes, market, leaf, farmers, companies, tobacco
Factor 6: (Gold) Mining	Côte d'Ivoire, Senegal, gold, Randgold, deposit, ounce

Table 6.1: The highest weighted words in each factor, for matrix factorization applied to the FTSE 100 corpus using 20 factors. The factors are hand labelled with their apparent identities.

It appears that the economic interpretation of the factors is clearer than the topics from TFM. This could be because they are driven by prices rather than semantic content in the documents, making their meaning more coherent from the perspective of price impact. Table 6.1 shows a selection of columns from matrix factorization run with 20

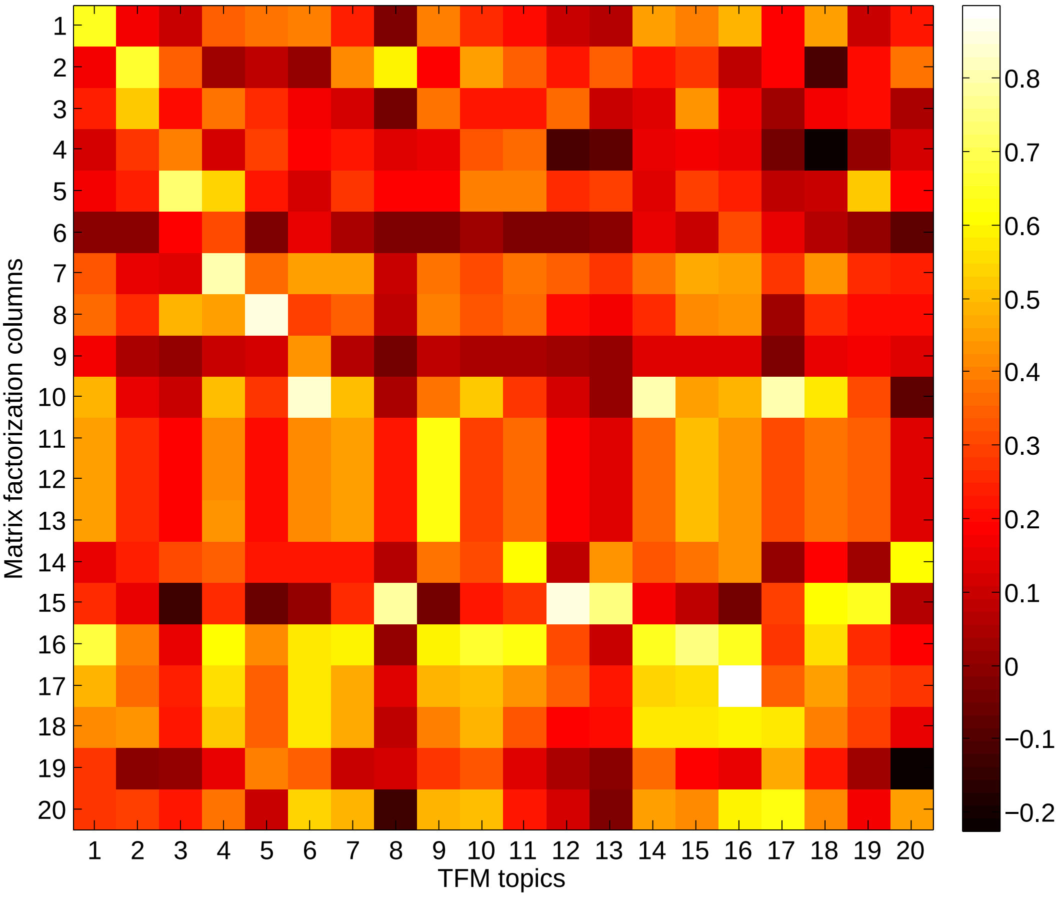


Figure 6.6: Heat map of the correlations between the 20 topic time series for TFM and matrix factorization on the FTSE 100 corpus. The matrix factorization columns are sorted by the TFM topic number of the highest correlation (i.e. such that the diagonal elements would be one if the two methods gave identical output). Note the high correlations near the diagonal and the repeated rows where matrix factorization gives rise to near identical factors.

factors. Notice that their meaning is more specific than the topics in TFM. For example, there is a topic wholly dedicated to tobacco companies. Unsurprisingly, this factor has highest weight $\theta_{d,k}$ for British American Tobacco and Imperial Tobacco. In TFM this theme is represented as a part of a broader consumer products topic (see table 5.2).

Another feature in matrix factorization is repeated thematic content. Factor 6 in table 6.1 actually appears almost identically four times. This is an artefact of quadratic regularization. If a topic has strong explanatory power (as is the case with the mining theme in the FTSE 100) then repeating it can contribute the same explanatory power while suffering a smaller regularization penalty. Using TFM or matrix factorization with weaker regularization gives rise to nonsense topics which do not contribute strongly to any time series. It is not obvious whether nonsensical or repeated features are preferable.

In section 6.1 it is mentioned that the time series are built up using a linear combination of factor time series and so are related to the topic time series from TFM. This

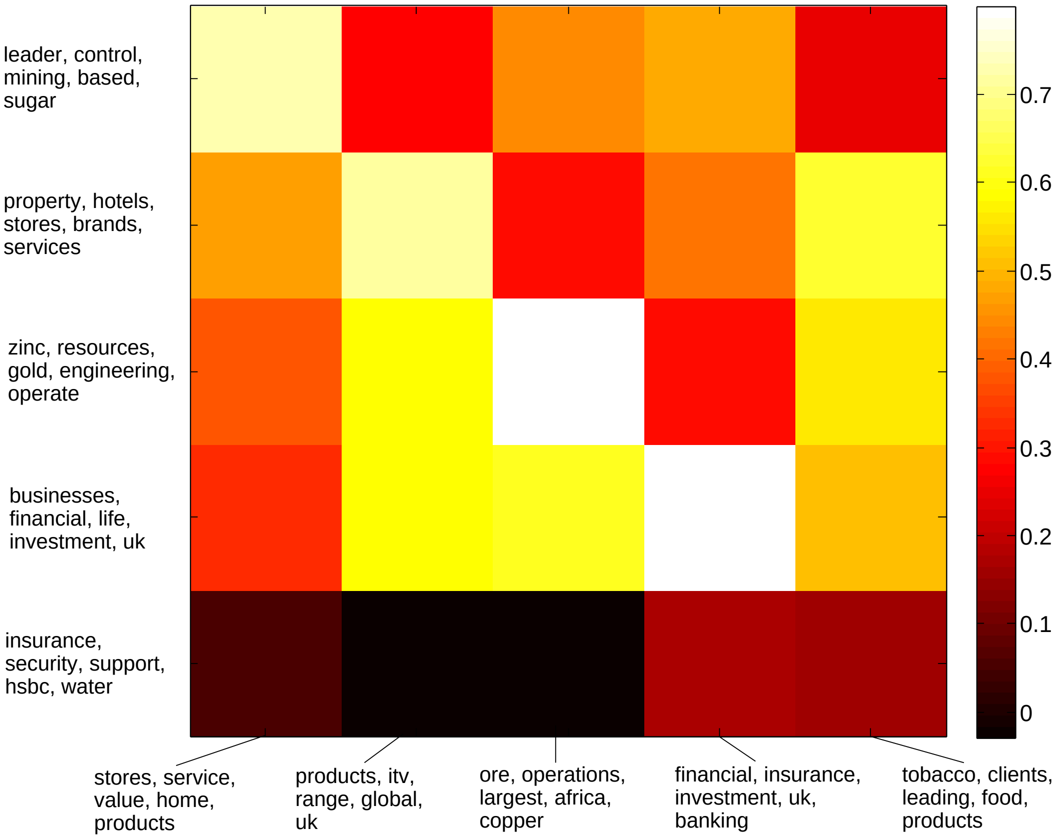


Figure 6.7: Heat map of topic correlations for 5 topics inferred by TFM and matrix factorization including labels of the top 5 words from each.

relationship is tested by running both methods on the FTSE 100 corpus and checking the correlations between topic time series. A heat map of the correlations for $K = 20$ is shown in figure 6.6. This shows strong links (correlation as high as 90%) between some topics. One can see in this comparison the repeated topic versus nonsense topic behaviour that was identified qualitatively in the top words for topics. In particular, rows 11-13 represent content from matrix factorization that is almost identical. Some of the factors do not correlate strongly with any of the TFM topics because they describe very specific subsets of the corpus (even a single stock) while the TFM topics tend to be more general.

As the number of topics is reduced this behaviour becomes easier to see. Figure 6.7 shows strong correspondence between the thematic structures found, including some correspondence between the top words. Note the diagonal of very high correlation, showing almost one to one correspondence of topics. Some correspondence in the text (for example between the first topic and first column) occurs deeper in θ than the top 5 words. This is natural given that the thematic structure is necessarily vague in an example with K so small. The themes contained in the first row and column relate to retail and manufacturing industries. Those in the second row and column to service

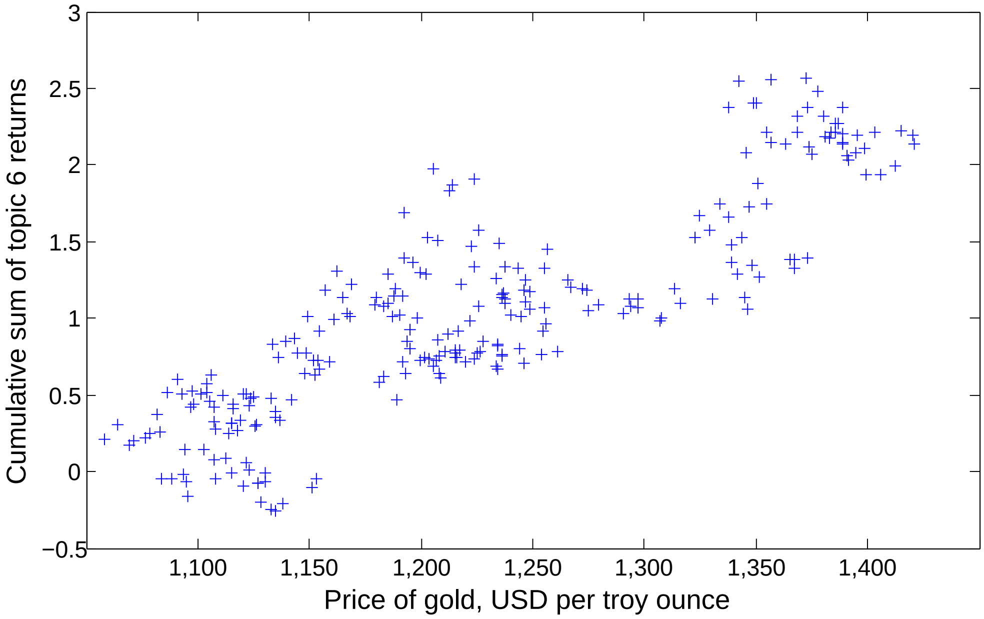


Figure 6.8: Plotting the cumulative returns alongside other financial data from outside the corpus can be suggestive of interpretations for the topics.

industries such as hospitality and entertainment. It is only the final column and row that fail to show semantic similarity. This disconnect is borne out by their low correlation.

The interpretation of topic time series can be bolstered by comparison with financial data not included in the original corpus. In figure 6.8 the cumulative sum of the topic returns for topic 6 from table 6.1 is plotted against the price of gold for the same time period. This is strongly supportive of the identity assigned to it using the words with greatest weight. It also shows that the themes identified by factorizing the returns matrix can have real economic relevance.

6.5 Summary of matrix factorization results

This chapter described and tested a simpler method for uncovering the thematic structure of the relationship between text and time series, using reduced-rank matrix factorization. This method is closely related to the historically popular matrix factorization methods set out in chapter 2. While simpler than the topic modelling approach to which the first portion of this thesis is dedicated, matrix factorization gave better results on held-out data and was interpretable to a similar degree. As such, in the next chapter the focus is on using matrix factorization for the applications in portfolio construction.

Chapter 7

Experiments with Foreign Exchange Data and a Case Study in Applications to Portfolio Management

This chapter contains experiments conducted using time series constructed from foreign exchange rates. The text data is made up of global macro summaries written by economists at Citi. The first results concern only the 2013 data, and show the value of matrix factorization in this context. They also seem to indicate that the foreign exchange data present a less obvious thematic structure than do equity data. The results from applying TFM to foreign exchange data at the topic modelling workshop were presented at Neural Information Processing Systems 2013 [Staines and Barber, 2013]. The second set of experiments demonstrates the application of matrix factorization to portfolio construction. There is some evidence that gains in risk adjusted return might be achieved by incorporating text into covariance prediction for portfolio construction.

7.1 Foreign exchange data

The foreign exchange rate XYZABC_t refers to the price of the base currency XYZ in units of the quoted currency ABC at time t. In the foreign exchange market each transaction crosses one of these currency pairs. Complete time series data would thus comprise $D(D-1)$ time series where D is number of currencies. It is possible to reduce the size of a data set in this space by considering only the prices of the currencies against a single base currency, giving $D-1$ time series. One can then find any cross currency pair, i.e. any pair not including the base currency, by assuming non-existence of triangular arbitrage. That is, assuming that one may never profit by instantaneous transaction free conversion of currency. This proves to be relatively robust [Fenn et al., 2009] and trading in many currencies is predicated on this assumption in any case, with pairs being cross traded through a more common currency (typically the US dollar).

In this chapter corpora are constructed by combining foreign exchange rate time series with text describing the outlook for the global economy. This is a less obvious relationship than the link between company descriptions and share price time series used in chapter 5, but the premise of exploring the joint corpus is the same. The themes described in analysis of a company should be related to the themes driving the value of that country's currency.

The time series are constructed from the exchange rates for a set of 14 currencies (AUD, BRL, CAD, CHF, EUR, GBP, HKD, INR, JPY, MXN, NZD, SEK, SGD, and ZAR) against the US dollar. These were chosen by taking the set of 20 currencies with highest trading turnover in 2013 and removing those for which it was difficult to construct complete data (KRW, NOK, TRK, CNY and RUB). Daily time series for exchange rates in these currencies are made available in Federal Reserve release H10 [Federal Reserve System]. These are indicative buying rates at noon each day. The spread and any differences between indicative and tradable prices are ignored for simplicity. These prices are used to find the log return in each exchange rate, and the time series is split into yearly segments for the years 2003-2013.

The text comprises sections from the "Global Economic Outlook and Strategy" report prepared by Citi Research. Only the sections of the report on the economies above are taken. As in the equity example, the data is treated by removing non alphabet characters and stop words. Only one report is taken per year: the last one published. These are used to construct a corpus per year with vocabulary sizes ranging between 585 and 1357 words.

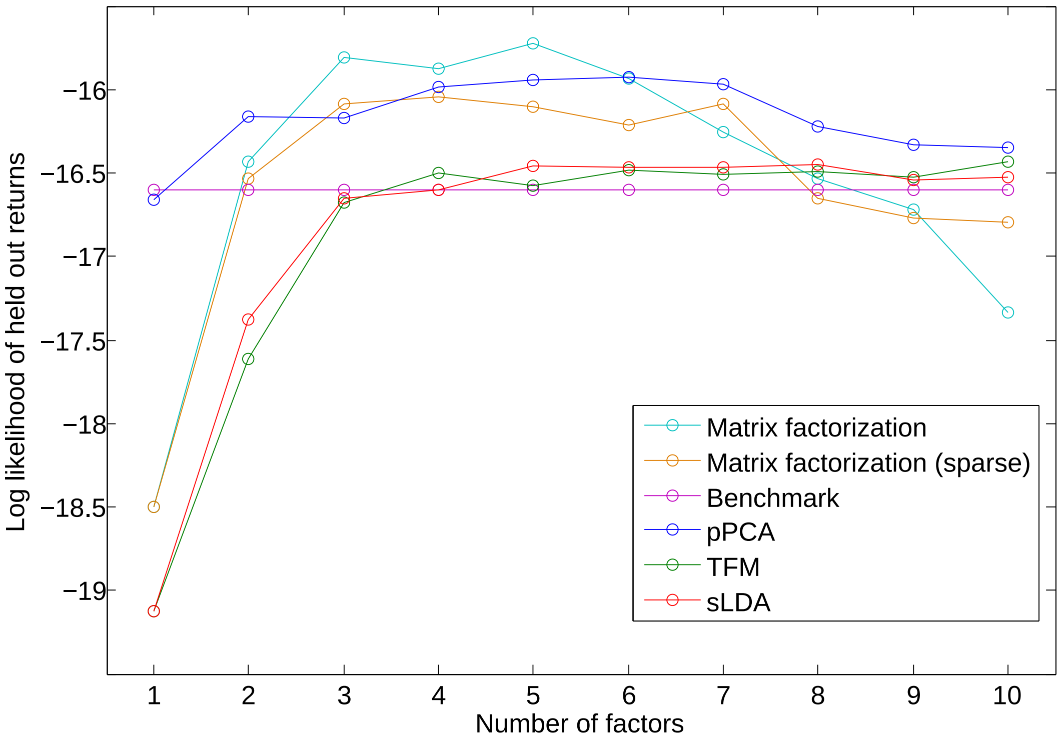


Figure 7.1: A comparison of the performance of matrix factorization, topic modelling and benchmark methods on the foreign exchange corpus. This can be compared to the results in figure 5.1 for the equity corpus, where the ratio of number of days to number of assets was far lower.

7.2 Results using a corpus of foreign exchange data

The foreign exchange corpus has far fewer documents per day, and a more complex correlation structure with some negative correlations. It is therefore important to verify that the same gains observed with equity data can be replicated. This is tested using a corpus comprising the text and time series from 2013. The latent parameters for both matrix factorization methods, TFM, sLDA and pPCA are found with the returns for each day held out in turn. The average likelihood of the held out day are compared to a benchmark of a Gaussian with covariance given by the empirical covariance of the corpus excluding the held-out day. A fuller discussion of the evaluation method is given in section 3.6. Figure 7.1 shows these results for a variety of values of K. This replicates the results shown in figure 5.1 for the new corpus.

Figure 7.1 shows results for methods incorporating text are not as successful as in the case of the equity data. Indeed, topic modelling barely outperforms the benchmark based purely on the covariance of the training data. This could be explained by the smaller ratio of the dimension of the data to the number of training points. With relatively more training data in this case, the empirical distribution could be a better reflection of the structure, and the gains to be made from robust methods smaller. It is also possible that the weaker link between the text and time series in this case limits the value of incorporating text. Nonetheless matrix factorization appears to show some benefit, outperforming pPCA in the range of 3-5 topics.

	Top words
Topic 1	prices, lower, risks, weak, view, expect, ongoing
Topic 2	growth, negative, exports, reflecting, demand, wage, external
Topic 3	GDP, forecast, annual, down, public, revision, result
Topic 4	rate, expected, recent, monetary, inflation, data, high
Topic 5	growth, view, continue, demand, negative, year, house

Table 7.1: The words with highest weight in each column of the topic-term matrix for the foreign exchange 2013 corpus. The topics are far less interpretable than for the FTSE 100 corpus.

Table 7.1 shows the top words for matrix factorization with 5 topics. The thematic structure is harder to interpret than in the case of equity data. This could be because the real structure is more subtle and less semantically coherent. It could also be another feature of the slight disconnect between text and time series in this case. It is possible to read some meaning into most of the word groupings however. One interesting feature of the highest weight words is their negativity, particularly topic 1. This is a feature of the wider text (words with the root “weak” appear 25% more frequently than those with the root “strong”) reflecting Citi’s cautious view at the beginning of 2013.

Topic 1 seems to be focused on financial distress, and is heavily weighted on the euro, British pound and South African rand. The words in topic 2 seem to relate to exports. Topic 3 is heavily weighted by only the Brazilian real and Mexican peso. The top word is GDP, and the rest seem consistent with a focus on GDP, a major driver of exchange rates for emerging market currencies. Topic 4 is heavily weighted by the euro, British pound, and Australian and New Zealand dollars. The top words suggest a topic relating to rates. The final topic is not heavily weighted by any currency, and its top words don’t strongly suggest any meaningful theme.

Overall, the semantic meaning of the topics is slightly disappointing compared to the clear themes when using the equity corpus. However, the factorization found clearly does represent meaningful economic relationships, and perhaps with expert examination could be better interpreted. A correlation forecast can be constructed and examined in just the same way as before. The maximum spanning tree of the predicted correlation in figure 7.2 contains a number of links corresponding to close geographical (and therefore

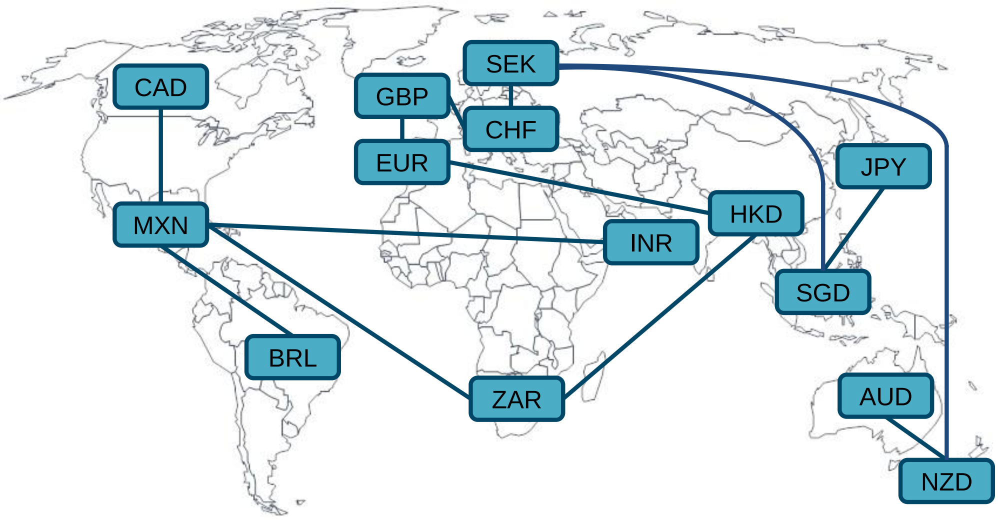


Figure 7.2: This figure shows the maximal spanning tree of predicted currency covariance for 2013. Note the resemblance to geographic proximity and the apparent clustering of emerging market currencies.

in some senses economic) proximity.

7.3 Practical application to portfolio optimization

In this section the output from matrix factorization is used to construct a covariance forecast, combining the correlation output with the historical standard deviation using equation 5.2. This is subsequently used to construct mean-variance optimal portfolios of investments. These are the portfolios with highest level of return for a each level of risk (as measured by variance). Assuming Gaussian returns, such portfolios can be constructed from estimates of mean return and covariance, and represent the full set of portfolios that a rational, risk averse investor would hold.

This type of assessment of inferred structure, while less rigorous than the likelihood based evaluation of previous chapters, serves to demonstrate how one might apply the methods of this thesis in a practical context. The availability of relatively long history allows full assessment for foreign exchange data. Similar historical corpora are harder to obtain for equities. The results from matrix factorization are compared to a number of other approaches. Alternative covariance forecasts are used from the historical sample covariance and a robust estimate by shrinkage [Ledoit and Wolf, 2001]. Finally, since it has been shown that gains from mean-variance optimization are typically offset by estimation error losses [DeMiguel et al., 2009], these are also compared to the naive $1/N$ portfolio.

The strategy to be tested is a funded carry trade portfolio as seen from the perspective of a US investor. A carry trade in foreign exchange markets is a strategy which involves borrowing a low interest currency and exchanging it for a higher interest currency to lend at a higher rate before unwinding the trade, hopefully having sustained losses due to currency movements smaller than the gains given by the positive carry. This can be made into a funded carry trade by backing it with a cash amount (which may even be in the low interest currency so that the short leg simply involves selling this backing currency rather than borrowing). The carry trade thus enhances the cash portfolio, resulting in a greater wealth than could have been achieved by lending in the first currency. The profitability of such trades violates uncovered interest rate parity, so was identified by economists as an anomaly. More recent work proposes that the excess returns are compensation for negative skew in the return profile or poor performance in times of financial stress [Burnside et al., 2011].

The portfolio tested is made up of dollar funded carry trades. It is constructed by either lending or borrowing in each of the currencies available. A long trade in a generic foreign currency XYZ starting on day t and lasting for τ days proceeds as follows.

1. The international currency is purchased at a rate USDXYZ_t

2. Interest is earned on the international currency balance while it is held, accruing to a final value of $\text{USDXYZ}_t(1+I_{XYZ})^{\frac{\tau}{365}}$ where I_{XYZ} is the prevailing annual rate of interest in the international currency.

3. Dollars are repurchased at the new exchange rate, giving a final dollar balance of $\frac{\text{USDXYZ}_t}{\text{USDXYZ}_{t+\tau}}(1+I_{\text{XYZ}})^{\frac{\tau}{365}}$

The short trade follows a similar procedure, but the international currency is borrowed and converted into dollars so that interest is both accrued on the dollar position and paid on the international currency borrowing. The absolute daily profit or loss per dollar notional on a long trade on day t is thus given by

$$\text{Profit} = \frac{\text{USDXYZ}_t}{\text{USDXYZ}_{t+1}} \times (1+I_{\text{XYZ}})^{\frac{1}{365}} - 1. \tag{7.1}$$

And the equivalent figure for a short trade (assuming the funded dollar position is also held) by

$$\text{Profit} = 2(1+I_{\text{USD}})^{\frac{1}{365}} - \frac{\text{USDXYZ}_t}{\text{USDXYZ}_{t+1}} \times (1+I_{\text{XYZ}})^{\frac{1}{365}} - 1. \tag{7.2}$$

Higher international interest rates increase the profit on long positions, and reduce the profit on short positions. A rising dollar against the foreign currency increases the profit on short trades and decreases the profit on long trade. Unless exchange rates move significantly, both trades earn from the differential in exchange rates: the “carry”. Therefore long trades are placed in currencies with higher interest rates and short trades in currencies with lower interest rates than the dollar. If a portfolio is rebalanced only once a year the relevant interest rates to examine are the spot 12 month rates in the two currencies at the start of the year. BBA LIBOR [British Banking Association] is used for the spot rates in all currencies for which it is available. Where no LIBOR rate is available, an applicable consensus reported rate is used in its place [Banco Central do Brasil; Hong Kong Association of Banks; National Stock Exchange of India Limited; Banco de México; Sveriges Riksbank; Association of Banks in Singapore; South African Reserve Bank]. The rates are taken from the first available day of each year.

To construct the portfolios then, at the beginning of each year either a long or short trade can be placed in each of the international currencies. On the last day of the year all trades are unwound. All lending and borrowing is assumed to be free from credit risk and transactions are assumed costless. Furthermore, it is necessary to assume that all quoted historical rates are tradable, and that lending and borrowing were available at the same rate. For these reasons the returns quoted will always be overstated. This is not a significant problem since the aim is to demonstrate the impact of different covariance forecasts on a portfolio, rather than to demonstrate the profitability of currency portfolios.

Mean variance optimal portfolios are constructed for each year with a target portfolio return of 5%. That is, assuming known mean annual return for asset d of μ_d and covariance matrix Σ, the combination of portfolio weights per dollar a_d in asset d which has minimal variance with an expected portfolio return of 5%. This can be found by the method of Lagrange multipliers where the objective is the variance of the portfolio value v.

$$\mathrm{Var}(v) = \sum_{d,d'} a_d \Sigma_{dd'} a_{d'} \tag{7.3}$$

The constraints are the funding constraint

$$\sum_d a_d = 1 \tag{7.4}$$

and the target return

$$\mathbb{E}[v] = 1 + \sum_d a_d \mu_d = 1.05. \tag{7.5}$$

For mean returns the available 12 month rate is taken in each currency. In other words

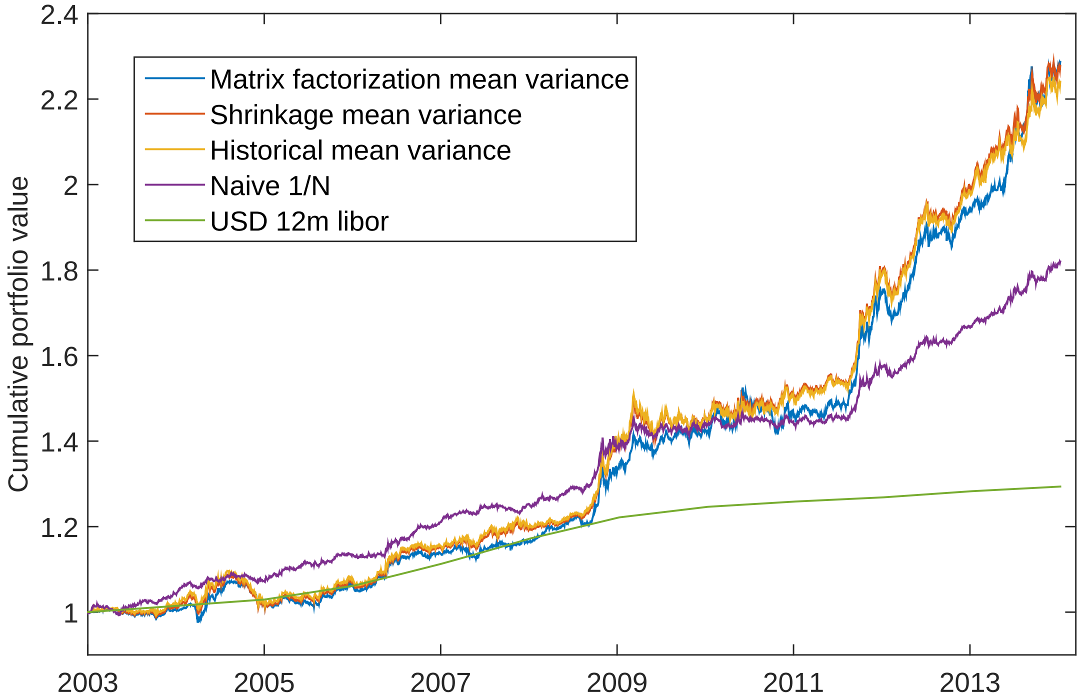


Figure 7.3: The cumulative value of mean variance optimal portfolios constructed using covariance estimates from matrix factorization, shrinkage, and historical methods. The cumulative return on a deposit earning LIBOR and a naively diversified portfolio are included for reference. Note the similarity of the returns on all of the mean variance optimal portfolios.

the distribution over the exchange rate at the end of the year is assumed to have a mean equal to the current exchange rate. The daily covariance is forecast in one of three ways: the correlation from matrix factorization shown in expression 6.7 multiplied by historical variances to give a covariance matrix; the sample covariance for the past 12 months; and the historical covariance matrix shrunk towards a one-factor model as recommended by Ledoit and Wolf [2001]. The matrix factorization model is applied using 5 topics and with regularization parameters set by cross validation as before. Each daily covariance matrix is then annualized, assuming independent daily returns.

Results

The cumulative value of these portfolios are given in figure 7.3, where the foreign deposits are assumed to accrue interest linearly over the 12 month deposit period. The same values are plotted for a portfolio with equal proportions allocated to each currency. Finally, a portfolio invested directly in dollar instruments yielding LIBOR is also included. This has no currency risk and provides a point of reference for the rate of return on the carry trade portfolios.

Portfolio construction	Annualized return	Annualized s.d.	Sharpe ratio	Daily VaR 95%	Max drawdown %
Matrix factorization	7.64 %	5.14 %	1.03	0.55 %	10.8 %
Shrinkage	7.60 %	5.04 %	1.04	0.54 %	8.64 %
Historical	7.45 %	5.12 %	1.00	0.56 %	9.89 %
1/N	5.48 %	3.05 %	1.02	0.34%	4.25 %
USD 12m LIBOR	2.34 %	-	-	-	-

Table 7.2: Table of statistics describing the results of constructing a carry trade portfolio with a mean variance optimal portfolios constructed using covariance estimates from matrix factorization, shrinkage, and historical methods. The results for a naive 1/N portfolio and the annualized return on a LIBOR deposit are also included. The data cover the years 2003-2013. The value-at-risk (VaR) is the amount lost on the worst day in the 95th percentile of historical daily returns. The max drawdown is the minimum, over the whole back test, of the portfolio value as a fraction of the previous maximum portfolio value (i.e. the biggest peak to trough loss).

The similarity of the value of mean variance portfolios through time makes clear that the differences between the portfolio weights are small. Shrinkage in particular proves only slightly different from using the historical covariance. All three appear to outperform the 1/N portfolio in terms of return but underperform it in terms of all risk measures. They rely on return estimation which is extremely challenging (and beyond the scope of this thesis. Table 7.2 shows some key features of the results. Shrinkage gives rise to the best risk/return results according to Sharpe ratio, but matrix factorization is only slightly behind. Intuitively matrix factorization and shrinkage attempt to make the covariance forecast more robust in the same way: by identifying an explanation of the covariance with rank less than the covariance matrix. The added benefit of matrix factorization is that by using text data this reduced rank model has semantically meaningful labels. The explanation is not merely a numerical construction but can also be interpreted in human terms.

The risk numbers for the methods differ somewhat. This means that the differences in return could be attributed to increased realized volatility rather than to superior performance. In practice one might wish to try to adjust for this. Some better impression of performance can be taken from risk adjusted return statistics, such as the quoted

Sharpe ratio. It is worth noting that all mean variance portfolios exceeded the target returns used to derive the portfolio weights. This is a result of the simplistic estimate of mean returns. A more elaborate method of expectations could easily be applied. Another way to improve the matrix factorization method would be to construct the covariance forecast using implied volatilities from derivative markets, rather than historical variance.

The daily variances of the mean variance portfolios are all around 0.3%, making the value-at-risk at the 95% level super-Gaussian. This is evidence of the negative skew used to explain the apparent excess returns from carry trade portfolios. It is worth reiterating that the returns are overstated in this experiment, making this simple carry trade look more attractive than it might be in practice. The method of portfolio construction has also benefited from the devaluation of the dollar during the ten years studied . More typically the Sharpe ratio of carry trade portfolios is found to be around 0.9 [Burnside et al., 2011; Handley, 2008].

7.4 Summary of foreign exchange results

This chapter showed the results of applying both the topic modelling and matrix factorization approaches to mining text and time series data taken from the foreign exchange market. The link between the text and time series was not as obvious, so one would expect this to be in some sense a more challenging application than the equity case. The majority of the conclusions found using the equity corpus in chapter 5 also held true for the foreign exchange corpus. Most significantly, matrix factorization again proved superior to TFM in terms of held-out data. The latter part of the chapter provided an example of applying thematic structure discovery to a financial problem: portfolio construction. This helps to show the motivation for development of these methods. A portfolio built using the covariance estimate from matrix factorization showed the same improvement in Sharpe ratio as using a shrinkage estimator for the covariance, a popular method in the financial industry.

Chapter 8

Conclusions and Further Work

This chapter contains some elaboration on opportunities to improve on the models for finding structure in joint corpora of text and time series data developed throughout this thesis. It also describes extensions to more diverse data sources and applications. The final section summarizes the contribution made by this thesis, and highlights some strengths of the work.

8.1 Opportunities for further work

This thesis describes topic factor modelling largely from the perspective of financial analytics. The combination of text and time series data might easily have applications in other areas. The first obvious opportunity for further work is then to build and explore joint corpora in these new domains. Below are given a selection of joint corpora which might yield interesting results.

- **Traffic to a website and the website contents** - One might expect that the traffic to a website through time would be determined by the popularity of its thematic content. Relationships between websites might thus be effectively explored using topic factor models. This might be of value in information retrieval.

- **Book sales and textual content** - Like the web traffic, the sales of a book are likely to covary highly with books on similar topics. Analysis of literary corpora could thus be enriched by adding the information of their changing popularity through time.

- **Time series of medical observations and self-reported symptoms** - Strategies for coping with chronic illness might be improved by better understanding the differences between patients. Topic factor modelling could provide a way to use all available data to determine the relationships between patient experiences.

- **Infection rates and reports on public health** - Trends might be easier to identify by enriching numerical statistics with expert opinion.

As well as applications, there is great scope for methodological improvements in the simultaneous analysis of text and time series data. Topic models are highly modular in nature and it is easy to imagine improvements to the generative process of extensions to more structured problems. In particular, Bayesian nonparametrics has added greatly to the possibilities of topic modelling [Blei et al., 2004; Teh et al., 2006; Roy et al., 2007]. It allows increased flexibility in models and can find averages over models to help protect from misspecification. A nonparametric approach to topic factor modelling might permit direct inference of hierarchical structure and would resolve the problem of choosing a number of topics. Even without nonparametrics, more structured models are both achievable and potentially useful. Below are described two particular areas of interest.

8.1.1 Alternative correlation structures

One weakness of topic factor modelling is that negative correlation in the document time series can only be introduced by negative correlation between topic time series. This weakness is also shared by the constrained matrix factorization method. Since negative correlations between assets are possible, even with this restriction, it is not immediately clear why it is a weakness. The issue is with interpretation. Effective representation of the thematic structure of a corpus may require two assets to have a relationship of opposite sign to the same topic. Imagine a topic k_1 representing the price of oil. A producer of oil d_1 has a share price with positive correlation with this topic ($\theta_{d_1,k_1} > 0$); as oil prices rise their reserves increase in value. A freight company d_2 who cannot pass on costs to their customers has a share price with negative correlation with the same topic ($\theta_{d_2,k_1} < 0$). Positively constraining θ means that the same relationship can only be represented by two topics, an “oil” topic and a “negative oil” topic.

Since equities are overwhelmingly positively correlated, this situation has relatively low relevance for the FTSE 100 data set. For smaller capitalization stocks, or for corpora outside the equity space, however, a modified model might be needed. For TFM, this could be achieved, for example, by allowing negative θ with Gaussian priors and drawing

$z_{d,n}$ from a categorical distribution with

$$p(z_{d,n} = j) \propto \exp(\theta_{d,j}). \tag{8.1}$$

Adjusting the matrix factorization approach would require simply a change in the function used to construct θ. For instance, using reduced-rank multivariate regression might prove valuable. Even when the data set doesn't suggest negative correlation between equities there may be benefits to allowing topics which have a negative impact on correlation.

Another possible improvement might be a model with non-diagonal correlation matrix for R. It would also be possible to introduce correlation between topics as in the correlated topic model [Blei and Lafferty, 2006a] to uncover relationships between the co-occurrence of risk factors. An improvement to the text model could be to implement a background distribution over words, or a set of factors, which don't contribute to the time series. This would allow isolation of themes from the text data which don't have economic impact (vocabulary attributable to differing authors or sources of text data for instance).

8.1.2 More realistic marginal distributions

As discussed in section 3.3, a Gaussian generative model misstates the risk associated with changes in asset price movements. Both the topic modelling and matrix factorization approaches developed in this thesis are predicated on Gaussian generative models of time series. In either case this can be resolved by replacing the Gaussian model with something more realistic for financial data. This would help to make the models more robust to extreme events. For matrix factorization the change in generative model corresponds to optimizing some other loss function than the squared error, dependant on the non-Gaussian distribution chosen. In the case of TFM the impact on the complexity of inference would likely be prohibitive. This highlights another benefit of simpler models. They leave more flexibility, in terms of time and memory costs, to extend the model. This is shown in practice by the sparse matrix factorization model in this thesis, which is easy implementable.

8.1.3 Temporal topic factor modelling

In all the work in this thesis, all text has been taken to be equally relevant at each time point. Of course in finance and many other applications text is most relevant to a certain

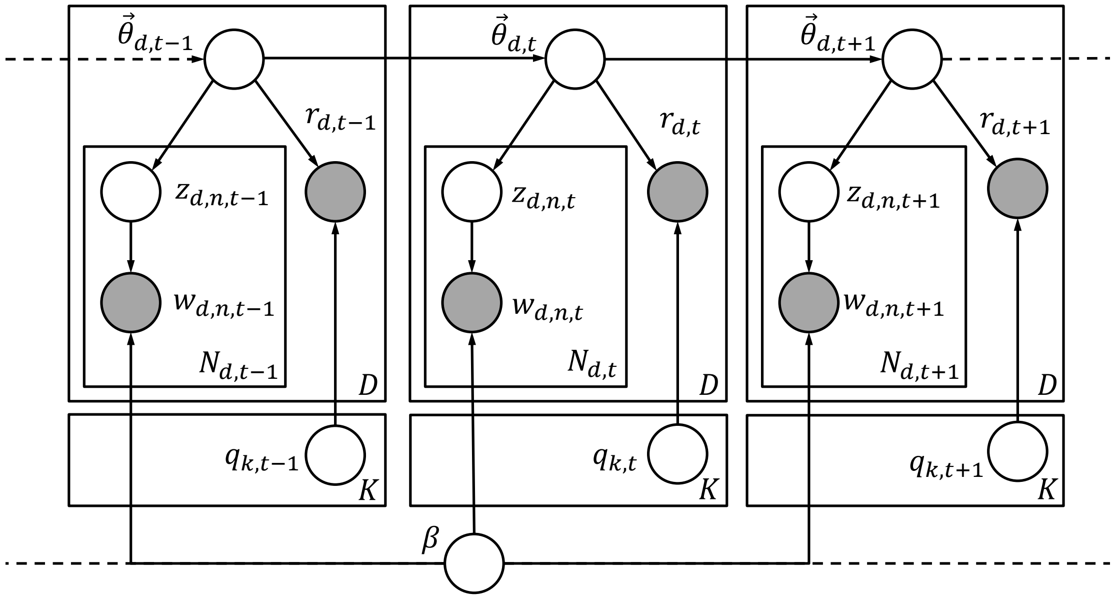


Figure 8.1: A proposed Bayesian network for a temporal topic factor model. The observed data arriving at each time interval are the shaded nodes. The text content of the topics β are constant through time, but the weights of the documents $\vec{\theta}_{d,t}$ are allowed to drift.

point in time. Indeed, text may be created in a timely manner so that it is not available until mid-way through data collection. To use a single period text model (including all models presented in this thesis) would represent look ahead bias in the time series modelling if any of the text was not available from the start. Further, particular text may be most relevant for a short time. Imagine text drawn from news articles. The thematic structure driving prices and news stories in the financial news media should be expected to change through time. To capture these temporal dynamics a temporal topic model should be used.

Temporal elements in topic modelling are widely used to track trends in corpora with changing thematic structure [Blei and Lafferty, 2006b; Wang and McCallum, 2006]. A similar process would naturally be possible in the case of the topic factor model thanks to the modular nature of topic models. A graphical model for just such a temporal topic factor model is proposed in figure 8.1. The benefits of this in terms of the applications in finance are clear: timely knowledge of changes in the structure of relationships between assets would be of great value in assessing risk in real time. The challenge in this case would be in increased complexity of inference. Both Blei and Lafferty [2006b] and Wang and McCallum [2006] use approximate variational inference using the Kalman Filter, which might be applicable in this case too.

8.2 Summary of contribution and conclusions

This thesis contains descriptions of two methods for data mining using a combined data set of text and time series, as well as results on two real, financial corpora. The idea of shared thematic structure between text and time series is itself novel, as is the application to financial data. Recent literature has begun to explore shared structure between text and time series, but not yet shared correlations from thematic similarities. The methodology of TFM represents a contribution to the topic modelling literature, highlighting the challenges of finding shared structure between discrete and continuous variables and demonstrating the importance of correct model specification. The matrix factorization method also differs from other models already in existence. The applications to equity and foreign exchange data both show significant improvement in held-out likelihood of returns relative to independent models. This justifies further investigation into the use of text data to inform time series models in finance, supporting the conclusions of recent work which has succeeded in finding shared structure in time series and text [Shah and Smith, 2010; Kim et al., 2013]. The ability of such techniques to reflect economic reality outside of the corpus used, as demonstrated by figure 6.8, supports the idea that the structure discovered in this way may be of practical interest and not simply as an exercise in data mining (interdisciplinary researchers in this area should take note that the term “data mining” is frequently used as a pejorative in the financial community).

The inspiration for this thesis comes from the topic modelling community. Topic modelling, and in particular work on joint corpora of text and images, provided the conceptual revelation that it would be possible to find shared structure in text and time series. That a far simpler discriminative method ultimately proves better able to achieve many of the aims of the initial experiments in topic modelling is a lesson in parsimony. That is not to deny the value of more complex models with a more natural intuition. As well as quantitative measures of success, the understanding which researchers and practitioners have of models is a part of their value, and it is in that respect that topic modelling proves most useful in this work.

References

Association of Banks in Singapore, SIBOR, http://www.abs.org.sg/, accessed: 2nd Jan 2014.

R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval.* Addison-Wesley, 1999.

Banco Central do Brasil, BRAZIBOR, https://www.bcb.gov.br/, accessed: 2nd Jan 2014.

Banco de México, 28 day TIIE, http://www.banxico.org.mx/, accessed: 2nd Jan 2014.

D. Barber, *Bayesian Reasoning and Machine Learning.* Cambridge University Press, 2012.

C. Bishop, *Neural Networks for Pattern Recognition.* Clarendon Press, 1995.

D. Blei and M. Jordan, Modeling annotated data, in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 127–134, 2003.

D. Blei and J. Lafferty, Correlated topic models, in *Advances in Neural Information Processing Systems*, vol. 18, pp. 147–154, 2006.

D. Blei and J. Lafferty, Topic models, in *Text Mining: Classification, Clustering, and Applications*, A. Srivastava and M. Sahami, Eds., pp. 71–94. Chapman & Hall, 2009.

D. Blei and J. Lafferty, Dynamic topic models, in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120, 2006.

D. Blei and J. McAuliffe, Supervised topic models, in *Advances in Neural Information Processing Systems*, vol. 20, pp. 121–128, 2008.

D. Blei, A. Ng, and M. Jordan, Latent Dirichlet allocation, *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

D. Blei, D. Griffiths, M. Jordan, and J. Tenenbaum, Hierarchical Topic Models and the Nested Chinese Restaurant Process, in *Advances in Neural Information Processing Systems*, vol. 16, 2004.

G. Bonanno, F. Lillo, and R. Mantegna, High-frequency cross-correlation in a set of stocks, *Quantitative Finance*, vol. 1, no. 1, pp. 96–104, 2001.

M. Borga, Learning Multidimensional Signal Processing, Ph.D. dissertation, Linköping University, 1998.

British Banking Association, LIBOR, https://www.bba.org.uk/, accessed: 2nd Jan 2014.

W. Buntine and A. Jakulin, Discrete component analysis, in *Subspace, Latent Structure and Feature Selection Techniques*, ser. Lecture notes in computer science, vol. 3940, pp. 1–33. Springer-Verlag, 2006.

W. Buntine and A. Jakulin, Applying discrete PCA in data analysis, in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 59–66, 2004.

C. Burnside, M. Eichenbaum, and S. Rebelo, Carry trade and momentum in currency markets, *Annual Review of Financial Economics*, vol. 3, pp. 511–535, 2011.

J. Chang and D. Blei, Relational topic models for document networks, in *International Conference on Artificial Intelligence and Statistics*, vol. 5, pp. 81–88, 2009.

J. Chang, J. Boyd-Graber, C. Wang, S. Gerrish, and D. Blei, Reading tea leaves: How humans interpret topic models, in *Advances in Neural Information Processing Systems*, vol. 22, pp. 288–296, 2009.

S. Chib, Marginal likelihood from the Gibbs output, *Journal of the American Statistical Association*, vol. 90, pp. 1313–1321, 1995.

A. Cichocki, R. Zdunek, A. Phan, and S. Amari, *Non-negative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation.* John Wiley & Sons, 2009.

P. Comon, Independent component analysis, a new concept? *Signal Processing*, vol. 36, pp. 287–314, 1994.

P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications.* Academic press, 2010.

T. Conlon, H. Ruskin, and M. Crane, Cross-correlation dynamics in financial time series, *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 5, pp. 705–714, 2009.

P. Das, R. Srihari, and Y. Fu, Simultaneous joint and conditional modelling of documents tagged from two perspectives, in *Conference on Information and Knowledge Management*, pp. 1353–1362, 2011.

S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, Indexing by latent semantic analysis, *Journal of the American Society of Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

V. DeMiguel, L. Garlappi, and R. Uppal, Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *Review of Financial Studies*, vol. 22, no. 5, pp. 1915–1953, 2009.

V. DeMiguel, Y. Plyakha, R. Uppal, and G. Vilkov, Improving portfolio selection using option-implied volatility and skewness, *Journal of Financial and Quantitative Analysis*, vol. 48, no. 6, pp. 1813–1845, 2013.

A. Dempster, N. Laird, and D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B*, vol. 39, no. 1, pp. 1–38, 1977.

G. Doyle and C. Elkan, Financial topic models, in *NIPS Workshop, Applications for Topic Models*, 2009.

G. Doyle and C. Elkan, Accounting for burstiness in topic models, in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 281–288, 2009.

K. Drakakis, S. Rickard, R. de Fréin, and A. Cichocki, Analysis of financial data using non-negative matrix factorization, in *International Mathematical Forum*, vol. 3, no. 38, pp. 1853–1870, 2008.

J. Driessen, B. Melenberg, and T. Nijman, Common factors in international bond returns, *Journal of International Money and Finance*, vol. 22, no. 5, pp. 629–656, 2003.

E. Erosheva, S. Fienberg, and J. Lafferty, Mixed-membership models of scientific publications, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 1, pp. 5220–5227, 2004.

E. Fama and K. French, Multifactor explanations of asset pricing anomalies, *The Journal of Finance*, vol. 51, no. 1, pp. 55–84, 1996.

Federal Reserve System, Foreign Exchange Rates - H.10, http://www.federalreserve.gov/releases/h10/hist/, accessed: 19th March 2014.

D. Fenn, S. Howison, M. McDonald, S. Williams, and N. Johnson, The mirage of triangular arbitrage in the spot foreign exchange market, *International Journal of Theoretical and Applied Finance*, vol. 12, no. 08, pp. 1105–1123, 2009.

D. Fenn, M. Porter, S. Williams, M. McDonald, N. Johnson, and N. Jones, Temporal evolution of financial-market correlations, *Physical Review E*, vol. 84, no. 2, 026109. 2011.

FTSE Group, UK index rules, http://www.ftse.co.uk/Indices/UK_Indices/Downloads/FTSE_UK_Index_Series_Index_Rules.pdf, accessed: 2nd Jan 2014.

T. Geva and J. Zahavi, Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news, *Decision Support Systems*, vol. 57, pp. 212–223, 2014.

G. Gidofalvi and G. Elkan, Using news articles to predict stock price movements, University of California San Diego, Tech. Rep., 2003.

C. Granger, Investigating Causal Relations by Econometric Models and Cross-spectral Methods, *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

T. Grifiths and M. Steyvers, Finding scientific topics, *Proceedings of the National Academy of Sciences*, vol. 101, pp. 5228–5235, 2004.

I. Guyon and A. Elisseeff, An introduction to variable and feature selection, *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

N. Handley, Using the carry trade in a diversified portfolio, *JPMorgan Insights*, 2008.

R. Hisano, D. Sornette, T. Mizuno, T. Ohnishi, and T. Watanabe, High quality topic extraction from business news explains abnormal financial market volatility, *PloS one*, vol. 8, no. 6, p. e64846, 2013.

M. Hoffmann, D. Blei, and F. Bach, On-line learning for latent Dirichlet allocation, in *Advances in Neural Information Processing Systems*, vol. 23, pp. 856–864, 2010.

T. Hofmann, Probabilistic latent semantic indexing, in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57, 1999.

Hong Kong Association of Banks, HIBOR, http://www.hkab.org.hk/, accessed: 2nd Jan 2014.

T.-K. Hui, Portfolio diversification: a factor analysis approach, *Applied Financial Economics*, vol. 15, no. 12, pp. 821–834, 2005.

A. Hyvärinen and E. Oja, Independent component analysis: algorithms and applications, *Neural networks*, vol. 13, no. 4, pp. 411–430, 2000.

A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* John Wiley & Sons, inc., 2001.

A. Izenman, Reduced-rank regression for the multivariate linear model, *Journal of Multivariate Analysis*, vol. 5, no. 2, pp. 248–264, 1975.

M. Jockers, *Macroanalysis: Digital Methods and Literary History.* UIUC Press, 2013.

I. Jolliffe, *Principal component analysis.* Springer-Verlag, 2002.

S. Joty, G. Carenini, G. Murray, and R. Ng, Finding topics in emails : Is LDA enough? in *NIPS Workshop on Applications for Topic Models: Text and Beyond*, 2009.

H. Kim, M. Castellanos, M. Hsu, C. Zhai, T. Rietz, and D. Diermeier, Mining causal topics in text data: Iterative topic modeling with time series feedback, in *Proceedings of the ACM international conference on information and knowledge management*, vol. 22, pp. 885–890, 2013.

S. Kogan, D. Levin, B. R. Routledge, J. S. Sagi, and N. A. Smith, Predicting risk from financial reports with regression, in *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 272–280, 2009.

V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan, Mining of concurrent text and time-series, in *KDD Workshop on Text Mining*, 2000.

O. Ledoit and M. Wolf, Improved estimation of the covariance matrix of stock returns with an application to portfolio selection, *Journal of Empirical Finance*, vol. 10, pp. 603–621, 2001.

W. Li and A. McCallum, Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations, in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 577–584, 2006.

P. McCullagh and J. A. Nelder, *Generalized Linear Models.* Chapman and Hall, 1989.

D. Mimno and A. McCallum, Topic models conditioned on arbitrary features with Dirichlet-multinomial regression, in *Proceedings of the 24th conference on Uncertainty in artificial intelligence*, pp. 411–418, 2008.

T. Minka and J. Lafferty, Expectation-propagation for the generative aspect model, in *Proceedings of the 18th conference on Uncertainty in artificial intelligence*, pp. 352–359, 2002.

M.-A. Mittermayer, Forecasting intraday stock price trends with text mining techniques, in *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, vol. 3, pp. 64–73, 2004.

National Stock Exchange of India Limited, MIBOR, www.nseindia.com/, accessed: 2nd Jan 2014.

J. Nocedal, Updating quasi-Newton matrices with limited storage, *Mathematics of Computation*, vol. 35, pp. 773–782, 1980.

M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, The author-topic model for authors and documents, in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 487–494, 2004.

D. Roy, C. Kemp, V. Mansinghka, and J. Tenenbaum, Learning annotated hierarchies from relational data, in *Advances in Neural Information Processing Systems*, vol. 19, 2007.

H. Seung and D. Lee, Learning the parts of objects by non-negative matrix factorization, *Nature*, vol. 401, pp. 788–791, 1999.

N. Shah and N. A. Smith, Predicting risk from financial reports with supervised topic models, *Thesis, Carnegie Mellon University*, 2010.

J. Shen and B. Zheng, Cross-correlation in financial dynamics, *Europhysics Letters*, vol. 86, no. 4, 48005. 2009.

A. Singh and G. Gordon, A unified view of matrix factorization models, in *Machine Learning and Knowledge Discovery in Databases*, pp. 358–373, 2008.

South African Reserve Bank, JIBAR, https://www.resbank.co.za/, accessed: 2nd Jan 2014.

J. Staines, Personal webpages, 2014. [Online]. Available: http://www0.cs.ucl.ac.uk/staff/J.Staines/tfm.html

J. Staines and D. Barber, Topic factor models: uncovering thematic structure in equity market data, in *Business Analytics in Finance and Industry, Santiago Chile*, 2014.

J. Staines and D. Barber, Topic factor modelling: uncovering thematic structure in financial data, in *Neural Information Processing Systems 2013 workshops. Topic Models: Computation, Application, and Evaluation*, 2013.

M. Steyvers and T. Griffiths, *Probabilistic Topic Models.* Lawrence Erlbaum, 2006.

Sveriges Riksbank, STIBOR, http://www.riksbank.se/, accessed: 2nd Jan 2014.

Y. Teh, M. Jordan, M. Beal, and D. Blei, Hierarchical Dirichlet processes, *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

Y. Teh, D. Newman, and M. Welling, A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation, in *Advances in Neural Information Processing Systems*, vol. 19, pp. 1353–1360, 2007.

R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

M. Tipping and C. Bishop, Probabilistic principle component analysis, *Journal of the Royal Statistical Society, Series B*, vol. 61, no. 3, pp. 611–622, 1999.

M. Tumminello, T. Aste, T. D. Matteo, and R. N. Mantegna, A tool for filtering information in complex systems, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 30, pp. 10 421–10 426, 2005.

H. Wallach, Topic modeling: beyond bag-of-words, in *Proceedings of the 23rd international conference on Machine learning*, pp. 977–984, 2006.

H. Wallach, Structured topic models for language, Ph.D. dissertation, University of Cambridge, 2008.

H. Wallach, D. Mimno, and A. McCallum, Rethinking lda: Why priors matter, in *Advances in Neural Information Processing Systems*, vol. 22, pp. 1973–1981, 2009.

H. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, Evaluation methods for topic models, in *Proceedings of the 26th International Conference on Machine Learning*, pp. 139–147, 2009.

C. Wang, J. Wang, X. Xie, and W.-Y. Ma, Mining geographic knowledge using location aware topic model, in *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pp. 65–70, 2007.

C. Wang, D. Blei, and D. Heckerman, Continuous time dynamic topic models, in *Proceedings of the 23rd conference on Uncertainty in artificial intelligence*, pp. 579–586, 2008.

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

M. Welling, Y. Teh, and H. Kappen, Hybrid Variational/Gibbs collapsed inference in topic models, in *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, vol. 24, 2008.

Yahoo.com, http://uk.finance.yahoo.com/, accessed: 25th September 2012.

S. Yan, D. Xu, B. Zhang, and H.-J. Zhang, Graph embedding: a general framework for dimensionality reduction, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 830–837, 2005.

D. Yogatama, C. Wang, B. Routledge, N. Smith, and E. Xing, Dynamic models of streaming text, *Transactions of the Association for Computational Linguistics*, no. 2, pp. 181–192, 2014.

Appendix A

Variational Updates for Topic Factor Modelling

> This appendix details the updates to each variational parameter when inferring the latent variables in topic factor modelling. The model itself is described in section 3.2. A full variational inference algorithm involves initializing the variational parameters (with some asymmetry so that the topics can develop differences), then repeating these updates until convergence.

The variational distributions of each $\vec{\theta}_d$ and $\vec{\beta}_k$ are taken to be Dirichlet with parameter vectors $\vec{\phi}_d$ and $\vec{\gamma}_k$ respectively, z to be categorical $q(z_{d,n} = k|\lambda) = \lambda_{d,n,k}$, and $q_{k,t}$ to be Gaussian distributed with mean $\mu_{k,t}$ and variance $\sigma^2_{k,t}$. The complete variational distribution is thus given by:

$$\begin{aligned} q(\theta, \beta, z, Q) = &\prod_d \left(q(\vec{\theta}_d|\vec{\phi}_d) \prod_n q(z_{d,n}|\lambda) \right) \\ &\times \prod_k \left(q(\vec{\beta}_k|\vec{\gamma}_k) \prod_t q(q_{k,t}|\mu_{k,t}, \sigma_{k,t}) \right) . \end{aligned} \tag{A.1}$$

The variational objective is given by

$$\mathcal{L} = \mathbb{E}_q\big[\log\big(q(\theta, \beta, z, Q)\big)\big] - \mathbb{E}_q\big[\log\big(p(\theta, \beta, z, Q, w, R)\big)\big]. \tag{A.2}$$

Its gradients with respect to each variational parameter can be calculated. For λ, γ and $\{\mu, \sigma\}$, the minima with respect to each parameter can be found in closed form. In the case of ϕ however, no such simple update exists, and a gradient following method is needed.

A.1 Note: The digamma function

The expectation of the log of a component of a Dirichlet distributed vector with parameter vector $\vec{\alpha}$ can be expressed using the digamma function Ψ.

$$\psi_i(\vec{\alpha}) \equiv \mathbb{E}_q\big[\log(\theta_i)|\vec{\alpha}\big] = \Psi(\alpha_i) - \Psi\left(\sum_j \alpha_j\right) \tag{A.3}$$

The digamma is also equal to the first derivative of the log of a gamma function, yielding the alternative expression:

$$\vec{\psi}(\vec{\alpha}) = -\nabla_{\vec{\alpha}} \log\left(\frac{\Gamma\left(\sum_j \alpha_j\right)}{\prod_j \Gamma(\alpha_j)}\right). \tag{A.4}$$

This function $\vec{\psi}$ appears frequently in the variational inference expressions because of its relationship with the Dirichlet distribution. Its first derivatives are also required. These can be expressed in terms of trigamma functions Ψ'

$$[\psi'(\vec{\alpha})]_{ij} = \Psi'(\alpha_i)\delta_{ij} - \Psi'\left(\sum_k \alpha_k\right) \tag{A.5}$$

where δ_{ij} is Kronecker's delta.

A.2 Updating γ

The sum of the terms of the objective which change with respect to γ_k is given by

$$\mathcal{L}_{\gamma_k} = \mathbb{E}_q\Big[\log\big(q(\vec{\beta}_k|\vec{\gamma}_k)\big)\Big] - \mathbb{E}_q\Big[\log\big(p(\vec{\beta}_k)\big) + \sum_{d,n}\log\big(p(w_{d,n}|z_{d,n},\beta_{k,w_{d,n}})\big)\Big]. \tag{A.6}$$

The gradient of this can be expressed with the function defined in equation A.3.

$$\frac{\partial \mathcal{L}_{\gamma_k}}{\partial \gamma_{k,m}} = \psi'(\vec{\gamma}_k)\Big(\vec{\gamma}_k - \eta\vec{1} - \sum_{d,n}\lambda_{d,n,k}\vec{e}_{w_{d,n}}\Big) \tag{A.7}$$

where $\vec{1}$ is a vector of ones and $\vec{e}_m$ is the vector whose elements are $[\vec{e}_m]_i = \delta_{i,m}$. This is zero where

$$\vec{\gamma}_k = \eta\vec{1} + \sum_{d,n}\lambda_{d,n,k}\vec{e}_{w_{d,n}}. \tag{A.8}$$

This equation gives the update for each $\vec{\gamma}_k$.

A.3 Updating λ

The sum of the terms of the objective which change with respect to $\lambda_{d,n,k}$ is given by

$$\mathcal{L}_{\lambda_{d,n}} = \mathbb{E}_q\Big[\log\big(q(z_{d,n}|\lambda)\big)\Big] - \mathbb{E}_q\Big[\log\big(p(z_{d,n}|\vec{\theta}_d)\big) + \log\big(p(w_{d,n}|z_{d,n})\big)\Big]. \tag{A.9}$$

Then since the optimization must be constrained to the simplex satisfying

$$\sum_k \lambda_{d,n,k} = 1 \tag{A.10}$$

a Lagrangian $L_{\lambda_{d,n}}$ can be constructed

$$\begin{aligned} L_{\lambda_{d,n}} = {} & \mathbb{E}_q\Big[\log\big(q(z_{d,n}|\lambda)\big)\Big] - \mathbb{E}_q\Big[\log\big(p(z_{d,n}|\vec{\theta}_d)\big) + \log\big(p(w_{d,n}|z_{d,n})\big)\Big] \\ & + \Lambda\left(\sum_k \lambda_{d,n,k} - 1\right). \end{aligned} \tag{A.11}$$

This has gradients given by

$$\frac{\partial L_{\lambda_{d,n}}}{\partial \lambda_{d,n,k}} = \log(\lambda_{d,n,k}) - \mathbb{E}_q\big[\log(\theta_{d,k}) - \log(\beta_{k,w_{d,n}})\big] + \Lambda\vec{1}. \tag{A.12}$$

The stationary points in these give the updates

$$\lambda_{d,n,k} \propto \exp\Big(\mathbb{E}_q\big[\log(\theta_{d,k}) + \log(\beta_{k,w_{n,d}})\big]\Big). \tag{A.13}$$

The constant of proportionality is determined by the normalization of the variational distribution. Using the function defined in equation A.3, this can be expressed

$$\lambda_{d,n,k} \propto \exp\big(\psi_k(\vec{\phi}_d) + \psi_{w_{d,n}}(\vec{\gamma}_k)\big). \tag{A.14}$$

A.4 Updating μ and σ

The sum of the terms of the objective which change with respect to $\vec{\mu}_t = [\mu_{1,t}, \mu_{2,t}, \ldots, \mu_{D,t}]$ or $\vec{\sigma}_t = [\sigma_{1,t}, \sigma_{2,t}, \ldots, \sigma_{D,t}]$ is given by

$$\begin{aligned} \mathcal{L}_{\{\vec{\mu}_t, \vec{\sigma}_t\}} = {} & \mathbb{E}_q\left[\sum_k \log\big(q(q_{k,t}|\vec{\mu}_t, \vec{\sigma}_t)\big)\right] \\ & - \mathbb{E}_q\left[\sum_k \log\big(p(q_{k,t})\big) + \sum_d \log\big(p(r_{d,t}|\vec{\theta}_d, Q)\big)\right]. \end{aligned} \tag{A.15}$$

Then the gradients of this with respect to $\vec{\mu}_t$ and $\sigma_{k,t}$ are as follows.

$$\begin{aligned}\nabla_{\vec{\mu}_t}\mathcal{L}_{\{\vec{\mu}_t,\vec{\sigma}_t\}} &= \frac{\rho}{\sqrt{v(\alpha)}(1-\rho^2)}\left(\sum_d \mathbb{E}_q\left[\vec{\theta}_d\right]^{\mathsf{T}} r_{d,t}\right) \\ &\quad - \left(\sum_d \frac{\rho^2}{v(\alpha)(1-\rho^2)}\mathbb{E}_q\left[\vec{\theta}_d\vec{\theta}_d^{\mathsf{T}}\right] + I\right)\vec{\mu}_t \qquad \text{(A.16)} \\ \frac{\partial\mathcal{L}_{\{\vec{\mu}_t,\vec{\sigma}_t\}}}{\partial\sigma_{k,t}} &= \frac{1}{\sigma_{k,t}} + \sigma_{k,t} + \frac{\rho^2}{v(\alpha)(1-\rho^2)}\sum_d \mathbb{E}_q\left[\theta_{d,k}^2\right]\sigma_{k,t} \qquad \text{(A.17)}\end{aligned}$$

The expectations with respect to the variational Dirichlet are given by

$$\mathbb{E}_q\left[\vec{\theta}_d\right] = \frac{\vec{\phi}_d}{\Phi_d} \qquad \mathbb{E}_q\left[\vec{\theta}_d\vec{\theta}_d^{\mathsf{T}}\right] = \frac{\vec{\phi}_d\vec{\phi}_d^{\mathsf{T}} + \operatorname{diag}\left(\vec{\phi}_d\right)}{(\Phi_d+1)\Phi_d}. \qquad \text{(A.18)}$$

where $\operatorname{diag}\left(\vec{\phi}_d\right)$ is a diagonal matrix with elements equal to the elements of the vector $\vec{\phi}_d$ and $\Phi_d = \sum_k \phi_{d,k}$. The update for $\vec{\mu}_t$ thus takes a form similar to solving ridge regression.

$$\vec{\mu}_t = \left(\sum_d \frac{\rho}{\sqrt{v(\alpha)}}\mathbb{E}_q\left[\vec{\theta}_d\vec{\theta}_d^{\mathsf{T}}\right] + \frac{1-\rho^2}{\rho}\sqrt{v(\alpha)}I\right)^{-1}\left(\sum_d \mathbb{E}_q\left[\vec{\theta}_d\right]^{\mathsf{T}} r_{d,t}\right) \qquad \text{(A.19)}$$

The updates in standard deviation are given by

$$\sigma_{k,t} = \left(1 + \frac{\rho^2}{v(\alpha)(1-\rho^2)}\sum_d \mathbb{E}_q\left[{\theta_{d,k}}^2\right]\right)^{-\frac{1}{2}}. \qquad \text{(A.20)}$$

A.5 Updating ϕ

The sum of the terms of the objective which change with respect to $\vec{\phi}_d$ is given by

$$\begin{aligned}\mathcal{L}_{\phi_d} &= \mathbb{E}_q\left[\log\left(q(\vec{\theta}_d|\vec{\phi}_d)\right)\right] \\ &\quad - \mathbb{E}_q\left[\log\left(p\left(\vec{\theta}_d\right) + \log\left(p(r_{d,t}|\vec{\theta}_d, Q)\right) + \prod_n p\left(z_{d,n}|\vec{\theta}_d\right)\right)\right]. \qquad \text{(A.21)}\end{aligned}$$

The gradient with respect to $\vec{\phi}_d$ gives rise to an equation, unlike the other parameters, without a closed form solution.

$$\begin{aligned}\nabla_{\phi_d}\mathcal{L}_{\phi_d} = &\left(\vec{\phi}_d - \alpha\vec{1} - \sum_n \lambda_{d,n,k}\vec{e}_k\right)\psi'(\vec{\phi}_d) - \frac{\rho}{\sqrt{v(\alpha)}(1-\rho^2)}\sum_t r_{d,t}\vec{\nu}_t \\ &+ \frac{\rho^2}{2v(\alpha)(1-\rho^2)}\sum_t \frac{\left(2M_t\vec{\phi}_d + \Phi_d\vec{m}_t\right)}{\left(\Phi_d+1\right)\Phi_d} \\ &- \frac{\left(2\Phi_d+1\right)\left(\vec{\phi}_d{}^{\mathsf{T}} M_t\vec{\phi}_d + \vec{m}_t^{\mathsf{T}}\vec{\phi}_d\right)}{\left(\Phi_d+1\right)^2{\Phi_d}^2}\vec{1}\end{aligned} \tag{A.22}$$

where $M_t = \vec{\mu}_t\vec{\mu}_t^{\mathsf{T}} + \mathrm{diag}\left(\vec{\sigma}_t^2\right)$, $\vec{m}_t$ is the vector with elements $m_{k,t} = \mu_{k,t}^2 + \sigma_{k,t}^2$ and $\vec{\nu}_t$ is the vector with elements

$$\nu_{k,t} = \frac{\left(\Phi_d - \phi_{d,k}\right)\mu_{k,t} - \sum_{j\neq k}\mu_{j,t}}{\Phi_d^2}. \tag{A.23}$$

Appendix B

Gibbs Sampling for Topic Factor Modelling

This appendix describes the resampling process for inference in topic factor modelling using a Gibbs sampling based method. The resampling process for z marginalizes β, making this a partially collapsed method. Q, but not θ can be resampled directly from the complete conditional. For θ, a Metropolis-Hastings step is proposed. The need to constrain θ to the probability simplex and the high dimensionality of θ make this particularly challenging. As a result inference is harder than collapsed Gibbs sampling for supervised latent Dirichlet allocation. The whole process is properly called a partially collapsed Metropolis-Hastings within Gibbs method, and is refered to as MHWG in this thesis.

B.1 Resampling z

As with latent Dirichlet allocation, the model can be collapsed with respect to β. It is not possible to achieve the same simplification with respect to θ so the sample must be taken from the conditional $p(z|\theta)$, and θ resampled separately. The conditional probability of z is given by

$$\begin{aligned} p\left(z_{d,n} = k \mid z_{\setminus d,n}, w, \theta\right) &\propto p\left(z_{d,n} = k, z_{\setminus d,n}, w \mid \theta\right) \\ &\propto \int_{\beta} p(z_{d,n} = k, z_{\setminus d,n}|\theta) p(w|z_{d,n} = k, z_{\setminus d,n}, \beta) p(\beta) \\ &\propto \theta_{d,k} \frac{\left(\eta + \#\{k, w_{d,n}\}\right)}{M\eta + \#\{k\}} \end{aligned} \tag{B.1}$$

in which the functions # refer to counts of the occurrence of certain combinations of states in all word positions in the corpus excluding the one corresponding to the token being resampled.

B.2 Resampling Q

New samples of Q must also be taken from a distribution conditioned on θ. The relevant distribution is Gaussian with parameters found by completing the square. The vector $\vec{r}_t$ refers to the vector of time series elements at time interval t for all documents, and the vector $\vec{q}_t$ to the factor time series elements at that time.

$$\begin{aligned} p\left(\vec{q}_t \mid \vec{r}_t, \theta\right) &\propto p(\vec{q}_t)p(\vec{r}_t \mid \vec{q}_t, \theta) \\ &\propto \exp\left(\frac{-\left(\vec{q}_t - \frac{\rho}{\sqrt{v}}\vec{r}_t{}^{\mathsf{T}}\theta\right)^{\mathsf{T}}\left(I + \frac{\rho^2}{v}\theta\theta^{\mathsf{T}}\right)\left(\vec{q}_t - \frac{\rho}{\sqrt{v}}\vec{r}_t{}^{\mathsf{T}}\theta\right)}{2(1-\rho^2)}\right) \end{aligned} \tag{B.2}$$

Thus samples are taken from

$$\vec{q}_t \sim \mathcal{N}\left(\frac{\rho}{\sqrt{v}(1-\rho^2)}S(\vec{r}_t{}^{\mathsf{T}}\theta), S\right) \tag{B.3}$$

where $S = \left(I + \frac{\rho}{\sqrt{v}(1-\rho^2)}\theta\theta^{\mathsf{T}}\right)^{-1}$.

B.3 Resampling θ

The complete conditional over $\vec{\theta}_d$ is of the form

$$p(\vec{\theta}_d|z, r, Q) \propto p(\vec{\theta}_d)\prod_n p(z_{d,n}|\vec{\theta}_d)\prod_t p(r_{d,t}|\vec{\theta}_d, \vec{q}_t) \tag{B.4}$$

and is constrained to the K-dimensional probability simplex. This is difficult to sample from directly. However, since the above expression can easily be evaluated, a Metropolis-Hastings step can be performed in place of directly resampling from the conditional.

The transition density proposed is based on the (tractable) conditional that would be used for Gibbs sampling in LDA. To this is added a term to make the distribution more peaked around the previous value of $\vec{\theta}_d$. The parameter $h \geq 0$ determines the distribution over step size around the parameter space and should be tuned to allow quickest convergence of the Gibbs sampler.

$$H\left(\vec{\theta}'_d \mid \vec{\theta}_d\right) = \frac{\Gamma(K\alpha + N_d + h)}{\prod\limits_k \Gamma(\alpha + \#\{d,k\} + h\theta_{d,k})} \prod_k \left(\theta'_{d,k}\right)^{\alpha - 1 + \#\{d,k\} + h\theta_{d,k}} \tag{B.5}$$

A sample from this is then assessed against the previous value of $\vec{\theta}_d$ using the Metropolis-Hastings criterion.

$$\begin{aligned} u_{\text{crit}} &= \frac{p\left(\vec{\theta}'_d\right) \prod\limits_n p\left(z_{d,n} | \vec{\theta}'_d\right) \prod\limits_t p\left(r_{d,t} | \vec{\theta}'_d, \vec{q}_t\right)}{p\left(\vec{\theta}_d\right) \prod\limits_n p\left(z_{d,n} | \vec{\theta}_d\right) \prod\limits_t p\left(r_{d,t} | \vec{\theta}_d, \vec{q}_t\right)} \times \frac{H\left(\vec{\theta}_d | \vec{\theta}'_d\right)}{H\left(\vec{\theta}'_d | \vec{\theta}_d\right)} \\ &= \prod_k \frac{\Gamma\left(\alpha + \#\{d,k\} + h\theta_{d,k}\right) \theta_{d,k}^{h\theta'_{d,k}}}{\Gamma\left(\alpha + \#\{d,k\} + q\theta'_{d,k}\right) {\theta'}_{d,k}^{h\theta_{d,k}}} \\ &\quad \times \prod_t \exp\left(\frac{\left(r_{d,t} - \vec{\theta}_d^{\mathsf{T}} \vec{q}_t\right)^2 - \left(r_{d,t} - \vec{\theta}'^{\mathsf{T}}_d \vec{q}_t\right)^2}{2(1 - \rho^2)} \right) \end{aligned} \tag{B.6}$$

Then, sampling $u \in [0,1]$ from a uniform distribution, the resampled parameter $\vec{\theta}''_d$ is given by

$$\vec{\theta}''_d = \begin{cases} \vec{\theta}'_d & \text{if } u < u_{\text{crit}} \\ \vec{\theta}_d, & \text{otherwise.} \end{cases} \tag{B.7}$$

This process must be repeated independently for each document.

Appendix C

Gradient Descent for Matrix Factorization

This appendix shows the gradient calculations used to optimize the objective in the matrix factorization approach to mining text and time series data from chapter 6. These can be used as inputs to an efficient gradient following method. Limited memory BFGS is applied for the experiments in this thesis because of its ability to use some second derivative information while being tractable for high dimensional data.

The unregularized objective of matrix factorization is the total squared error.

$$\mathcal{L}^{MF} = \sum_{d=1}^{D} \sum_{t=1}^{T} \left(r_{d,t} - \sum_{k} \theta_{d,k} q_{k,t} \right)^2 \tag{C.1}$$

The gradient with respect to document-topic weights is given by

$$\begin{aligned} \frac{\partial \mathcal{L}^{MF}}{\partial \beta_{l,p}} &= -2 \sum_{d=1}^{D} \sum_{t=1}^{T} \left(r_{d,t} - \sum_{k} \theta_{d,k} q_{k,t} \right) \\ &\quad \times \left(x_{d,p} q_{l,t} \theta_{d,l} - \frac{\sum_{k} x_{d,l} q_{k,t} \theta_{d,k}}{\sum_{k} \exp\left(\sum_{m} x_{d,m} \beta_{k,m} \right)} \right). \end{aligned} \tag{C.2}$$

The gradient with respect to factor returns is given by

$$\frac{\partial \mathcal{L}^{MF}}{\partial q_{j,s}} = -2 \sum_{d=1}^{D} \left(r_{d,s} - \sum_{k} \theta_{d,k} q_{k,s} \right) \theta_{d,j}. \tag{C.3}$$

When a regularizer is added its gradient can be simply added to these. The second derivatives are not used since the Hessian matrix is too large to store but low rank approximations can be exploited by an appropriate optimizer. LBFGS can be applied, truncating to a sufficiently small number of previous search directions. This allows optimization over the joint parameter space, whose dimension of $K(M+D)$ would otherwise be prohibitive.

Acknowledgements

The completion of this thesis and any progress I have made as a graduate student owe a great deal to the opportunities I have been given. I must begin by thanking David Barber for his time and support; supervising a doctoral student who writes more fluent code than English is no small task. The examination this thesis was conducted with much rigour and sympathy by Edward Tsang and Mark Herbster. Without them this would be a less complete document. At UCL I have also benefited from advice from Ed Challis, Chris Bracegirdle and Tom Furmston. David Blei must be mentioned for inspiring this work, and for his availability both by email and at conferences. Thanks to Bryan Routledge, for the most helpful, constructive review of a paper I have ever received; to Richard Weber for his encouragement; and to Stuart O'Neil for forewarning me of some of my errors before my viva. For the provision of data and productive conversation, I am grateful to David Bieber and Citigroup. I owe to Rafael Molinero and Guillaume Bagnarosa my earliest understanding of the quantitative problems facing hedge funds, and to Ewan Kirk some of my appreciation of the value of the time granted to PhD students. I thank Philip Treleaven for his recognition of the demand for a "school for quants" and for the connections he has made for me. Thanks are due also to Yonita Carter for her help throughout my time at UCL. Ultimately, the life I have enjoyed as a graduate student is possible only with the funding of the EPSRC; to them, and to the taxpayers who ultimately foot the bill, I am most grateful.

A number of other people have helped me, directly or indirectly, through the process of living and working as a graduate student. Particular thanks to James Anderson for productive discussion on any number of topics, and for prompting my very enjoyable first foray into academic research. I'm indebted to Jamie Cutteridge and Szimon Daraszewicz for making the writing process less painful by their proofing help. For the emotional, grammatical, and gastronomic support of Charlotte Turnbull, I shall never be able to express adequate gratitude. And finally I move to the people who nudged me along the first steps of my intellectual journey: my grandfather John Goodman, from whom I first learned the value of empiricism; and my parents Alison and Martin Staines, who gave me so many of the opportunities that have led me to this point and who somehow foresaw great achievements in my childhood drawings.