Desidero ringraziare il Professor Lijoi, relatore di questa tesi, per l'attenzione, la guida e il sostegno che hanno accompagnato lo sviluppo di questo lavoro.

Dedico questo lavoro a chi ha reso questi anni fonte di gioia e crescita.
A Mamma e Papà, che con il loro amore, la loro fiducia e il loro sostegno costante hanno reso possibile ogni passo di questo percorso. Se oggi posso guardare a questo traguardo con orgoglio, è soprattutto grazie a loro.
Ad Ale, che mi è sempre stato vicino, pronto a sostenermi e a volere soltanto il meglio per me. Non avrei potuto desiderare un fratello migliore.
A Sofia, il sorriso che illumina le mie giornate, che con il suo affetto ha reso questo cammino molto più bello. Grazie per il sostegno totale, per le risate e per ogni momento insieme. Questo risultato appartiene anche a te.
A Fuma e Fede, compagni di viaggio di lunga data, amici di sempre e pietre miliari della mia vita.
Infine, ai miei amici più stretti, Fra, Leo, Marco e Pie, che hanno reso questi anni indimenticabili.
A tutti voi, che con la vostra presenza, il vostro affetto e la vostra vicinanza avete reso questo percorso unico, va la mia gratitudine più profonda.

Table of Contents

1 INTRODUCTION

The communication of political bodies and central banks has undergone a profound transformation over time, and specifically, in recent decades, the language of monetary policy has changed. In the past, it was characterised by opacity and technical jargon, aimed at maintaining an aura of confidentiality around policy decisions. However, since the 2000s, the awareness that communication itself represents a fundamental economic policy tool for each country and, consequently, also for the European Union, has emerged. In the case of the European Central Bank (ECB), this procedure was especially significant: the European Union was still developing and was in a complex and constantly evolving institutional context, where public speeches, press conferences and every official declaration had progressively assumed a central and ever-increasing function in influencing market expectations and building the institution's credibility in the eyes of investors, citizens and the rest of the world. In the most significant and turbulent phases of the history of the European Union, from the global financial crisis of 2008, to the European sovereign debt crisis, up to the Covid-19 pandemic, the speeches and the choice of words used by the European Central Bank, have often preceded or accompanied the monetary policy decisions of those periods, thus ensuring that the words of policy makers became an integral part of the institution's toolkit. Consequently, language does not simply have the function of explaining decisions already taken but has transformed into a guidance mechanism that is capable of influencing the behaviour of economic actors even before concrete measures are adopted. In a complex institutional context characterised by recurring economic crises, these official speeches enhance the comprehension and prediction of market expectations. The growing attention paid to the language of the ECB has stimulated numerous studies, but the literature remains limited. Many have focused on limited periods, such as the global financial crisis, or have used qualitative approaches and sentiment indicators, while systematic, long-term analysis is lacking: an analysis that explores the entire archive of the institution's speeches, connecting emerging issues to monetary policy decisions through advanced quantitative tools. Despite the growing importance of this phenomenon, relatively little attention has been devoted to systematic analysis in the language of central banks. In recent years, with the advent of techniques such as Natural Language Processing, the possibility has opened up to address this area in a new way. Textual analysis tools, such as topic modelling, allow researchers to identify recurring topics within large corpora of texts and data in an automated manner, and to study their evolution over time. The Latent Dirichlet

Allocation model represents a pioneering tool in the diffusion of topic modelling and is one of the most widespread and versatile methods thanks to its probabilistic structure, which allows it to describe each document as a combination of latent topics, and each topic as a probability distribution over a set of words. Its application to the speeches of the ECB allows us to analyse what the recurrent topics are in the language of the institution and to what extent these topics can anticipate monetary policy decisions. The present work, therefore, lies at the intersection of theory and practice. On the one hand, it systematically delves into the mathematical and operational basis of LDA, discussing its mechanisms, hyperparameter choices and limitations. On the other hand, it applies the model to a concrete case study: the set of official ECB speeches between 1999 and 2025, to explore the relationship between communication and interest rate decisions. The focus will be on the theoretical foundations of the model, but the dual path allows the study not only to show the validity of the LDA as a tool for textual analysis in the economic field, but also to provide empirical evidence on a topic of political and economic relevance. There are three specific objectives of the empirical case study:

I. Identify the main recurring topics in the ECB's speeches using LDA in Python and assess their economic interpretability.
II. Describe the evolution of these issues over time, distinguishing different periods and contexts (pre-crisis, financial crisis, debt crisis, post-2015, pandemic, recent phase).
III. Evaluate the associations between the relative presence of certain topics with interest rate decisions, using an integrated methodology of Principal Component Analysis and multinomial regressions.

The remainder of this thesis is structured in a concise and rigorous manner: the second chapter presents the theoretical and methodological framework of the LDA, explaining its mathematical foundations, basic assumptions and possible applications. The third chapter presents the empirical analysis, describes the dataset and the preprocessing procedures. The fourth and last chapter delves into the results of the analysis: the descriptive analysis of the topics and their temporal evolution, the Principal Component Analysis and the multinomial regressions, aimed at testing the association between communication and rate decisions.

2 THE LATENT DIRICHLET ALLOCATION MODEL

2.1 Why we need Latent Dirichlet Allocation

In the analysis of modern text corpora, each document is often represented as a vector of term frequencies across the entire vocabulary. Since a document corpus vocabulary typically contains tens of thousands of unique terms, this representation is high-dimensional: the dimensionality of the feature space corresponds to the size of the vocabulary. In addition, each single document uses only a relatively small fraction of possible terms, meaning that most entries in the vector are zeros. This property is known as sparsity (Blei et al., 2003), and an example is a document of a few thousand words, which may contain just ten per cent of the words available in the corpus vocabulary. High-dimensional and sparse representations are challenging to work with. Similarity measures between documents become unreliable, statistical models are prone to overfitting, and the resulting vectors are not directly interpretable (Chandra et al., 2023; Gkioulekas & Zickler, 2011). Topic modelling offers a statistically principled way to reduce dimensionality, discover latent thematic structure that is not annotated a priori, and provide interpretable, human-readable summaries of large collections. It does so by representing each document as a mixture of a small number of latent "topics" and each topic as a probability distribution over words. In practice, this means that even a long document combines multiple topics, for example a single central bank speech might be composed of 40% "inflation" 35% "financial stability" and 25% "climate risk. The general modelling problem is to replace the sparse, extremely high-dimensional "bag-of-words" vector with a parsimonious, low-dimensional representation that captures those topics and their relative weights. An important practical issue is that the number of topics, denoted by K, is not known a priori. The consequences are that if we choose a value of K too small, distinct topics are forced to merge into overly broad categories, while for values of K too large, coherent subjects may be fragmented into artificial subtopics (Griffiths & Steyvers, 2004). Fortunately, in empirical applications, measures such as perplexity and topic coherence are often employed to guide model selection, complemented by considerations of interpretability.

LDA explicitly encodes the idea that documents exhibit multiple topics to varying degrees, while topics themselves are probability distributions over vocabulary items. In this chapter, we will aim to present LDA rigorously but accessibly, covering:

- its generative structure and probabilistic foundations

- the role and interpretation of the Dirichlet priors on document-topic and topic-word distributions
- the approximate inference machinery that makes the model practical
- its strengths and limitations.

2.2 Notation and Probabilistic Preliminaries

In the Latent Dirichlet Allocation (LDA) model, we consider a vocabulary of size V (the number of unique words). We assume a corpus of M documents and a fixed number of latent topics K. Each document d contains N_d word tokens. For instance, the sequence “apple”, “is” consists of two *word tokens* and two *word types*, whereas “apple”, “is”, “apple” consists of three *word tokens* but only two *word types*. We denote the n-th word in document d by w_{dn}. This is an observed word, typically represented as a discrete index $1, \dots, V$ or a one-hot vector of length V. Each word token is associated with a latent topic assignment $z_{dn} \in \{1, \dots, K\}$ indicating which topic generated that word. Each document d has a document-topic distribution:

$$\boldsymbol{\theta}_d = (\theta_{d1}, \dots, \theta_{dK}),$$

a probability vector over K topics. By construction, its components satisfy $\theta_{dk} \geq 0$ and $\sum_{k=1}^{K} \theta_{dk} = 1$. Hence $\boldsymbol{\theta}_d$ lies in the $(K-1)$ simplex, defined as:

$$\Delta^{K-1} = \{\theta \in \mathbb{R}^K : \theta_k \geq 0, \sum_{k=1}^{K} \theta_k = 1\}.$$

Although $\boldsymbol{\theta}_d$ is represented as a vector in $\mathbb{R}^K$, the simplex has intrinsic dimension $K-1$ due to the linear constraint. In the standard formulation of LDA, the topic-word distributions $\boldsymbol{\phi}_k$ are treated as fixed parameters to be estimated. In contrast, we focus on the fully Bayesian formulation, where each topic k has a topic-word distribution:

$$\boldsymbol{\phi}_k = (\phi_{k1}, \dots, \phi_{kV}),$$

a probability vector over the V words in the vocabulary. By construction, its components satisfy $\phi_{kv} \geq 0$ and $\sum_{v=1}^{V} \phi_{kv} = 1$. Hence $\boldsymbol{\phi}_k$ lies in the $(V-1)$ simplex, defined as:

$$\Delta^{V-1} = \{\phi \in \mathbb{R}^V : \phi_v \geq 0, \sum_{v=1}^{V} \phi_v = 1\}.$$

Although $\boldsymbol{\phi}_k$ is represented as a vector in $\mathbb{R}^V$, the simplex has intrinsic dimension $V-1$ due to the linear constraint. This means that both $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$ are structured probability distributions restricted to their respective simplices, rather than arbitrary vectors. Having established this distinction, we now clarify which elements of the model are observed and which are latent. Specifically, w_{dn}, M, V, K, N_d are given (observed data and corpus parameters), while z_{dn}, $\boldsymbol{\theta}_d$, and $\boldsymbol{\phi}_k$ are latent random variables or parameters to be inferred.

The LDA model relies on two main types of distributions: "*Categorical*" and "*Dirichlet*". The categorical (or multinomial for a single trial) distribution is used for drawing discrete indices like topics and words. For example, if $z_{dn} = k$, then the observed word w_{dn} is drawn from the categorical distribution $\boldsymbol{\phi}_k$ associated with topic k:

$$P(w_{dn} = v \mid z_{dn} = k) = \phi_{k,v}.$$

Likewise, given a document's topic proportions $\boldsymbol{\theta}_d$, the topic assignment is drawn as

$$P(z_{dn} = k \mid \boldsymbol{\theta}_d) = \theta_{d,k}.$$

Both $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$ are drawn from Dirichlet distributions. The Dirichlet distribution is a continuous probability distribution over vectors of non-negative real numbers whose components add up to one. This characteristic makes the distribution suitable for representing proportions or probability assignments. As a consequence, it is possible to think of the Dirichlet distribution as a "distribution of distributions". It is frequently used in Bayesian statistics as a prior for models involving categorical or multinomial variables.

Formally, a Dirichlet distribution, denoted as $Dir(\boldsymbol{\gamma})$, is defined by a parameter vector $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_K)$ with strictly positive entries. It can be regarded as the natural multivariate extension of the Beta distribution, and its conjugacy with the multinomial distribution explains its widespread use in Bayesian modelling (Bishop, 2006). Conjugacy refers to the property for which the posterior distribution belongs to the same family as the prior distribution, thus simplifying Bayesian updating.

If $\mathbf{x} = (x_1, \dots, x_K)$ lies in the $K-1$ simplex ($x_i \geq 0$ and $\sum_i x_i = 1$), the Dirichlet density is:

$$p(\mathbf{x} \mid \boldsymbol{\gamma}) = \frac{\Gamma(\sum_{i=1}^{K} \gamma_i)}{\prod_{i=1}^{K} \Gamma(\gamma_i)} \prod_{i=1}^{K} x_i^{\gamma_i - 1}, \qquad \mathbf{x} \in \Delta^{K-1},$$

and the density is zero for any $\mathbf{x} \notin \Delta^{K-1}$. Here $\Gamma(\gamma_i)$ is the Gamma function and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)$ are known as the Dirichlet hyperparameters (or concentration parameters). In the symmetric case all components are set to the same value ($\gamma_i = \gamma$), while in the asymmetric case, they may differ. As already discussed, the Dirichlet distribution is the conjugate prior of the multinomial distribution, which is why LDA uses Dirichlet priors for the multinomial topic and word distributions.

The names of the parameters are arbitrary, but in this thesis about Latent Dirichlet Allocation, $\boldsymbol{\alpha}$ denotes the Dirichlet prior concentration parameter on each $\boldsymbol{\theta}_d$ (controls document-topic proportions), and $\boldsymbol{\beta}$ denotes the Dirichlet prior concentration parameter on each $\boldsymbol{\phi}_k$ (controls topic-word distributions), where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ play the same role as $\boldsymbol{\gamma}$ in the general definition above. As a consequence, assume

$$\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha}) \quad \text{for each document } d,$$

where $\boldsymbol{\alpha}$ is a K-dimensional hyperparameter vector ($\alpha_i = \alpha \ \forall\ k$ in the symmetric case), and

$$\boldsymbol{\phi}_k \sim Dir(\boldsymbol{\beta}) \quad \text{for each topic } k,$$

where $\boldsymbol{\beta}$ is a V-dimensional hyperparameter vector ($\beta_v = \beta \ \forall$ v in the symmetric case). Intuitively, $\boldsymbol{\alpha}$ controls how concentrated or spread out the topic mixture for a document is, and $\boldsymbol{\beta}$ controls how concentrated each topic's word distribution is.

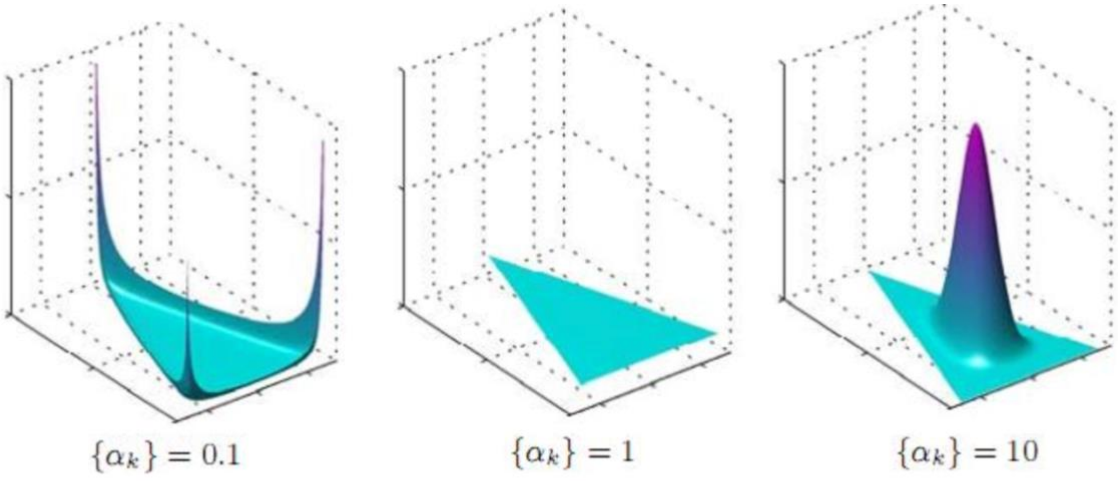


Figure 2.1 (Aldebakel et al.) Visualisation of the Dirichlet distribution over the 2-simplex for three different values of the symmetric concentration parameter α_k (with K=3). When α_k = 0.1, the distribution is concentrated near the simplex vertices, favouring sparse probability vectors. For

α_k= 1, it is uniform over the simplex, assigning equal probability to all configurations. When α_k= 10, the distribution is peaked near the centre, favouring nearly uniform vectors.

The use of Dirichlet priors together with categorical distributions fully specifies the generative mechanism of LDA. To complete the probabilistic formulation, one further assumption is required: the bag-of-words assumption, which simplifies the modelling of documents by disregarding word order. It corresponds in probabilistic terms to the exchangeability of word observations. Exchangeability means that the joint probability distribution is invariant to the ordering of words, and under de Finetti's theorem, any infinitely exchangeable sequence can be represented as conditionally i.i.d. draws given a latent variable. In Latent Dirichlet Allocation, this latent variable is the document-specific topic mixture $\boldsymbol{\theta}_d$, drawn from a Dirichlet prior with parameter $\boldsymbol{\alpha}$. Given $\boldsymbol{\theta}_d$**,** words are generated independently by sampling a topic $z_{dn} \sim \text{Cat}(\boldsymbol{\theta_d})$ and then a word $w_{dn} \sim \text{Cat}(\boldsymbol{\phi}_{z_{dn}})$, but we will focus more on the generative structure of the model in the following. section. This assumption simplifies modelling but also implies that LDA ignores word order and syntactic structure. LDA also assumes that documents themselves are exchangeable, which justifies the use of shared hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ across the entire corpus.

2.3 Formal definition of the Latent Dirichlet Allocation Model

Generative Model

LDA is a generative model for documents. It defines a hypothetical random process by which you could generate synthetic corpora. While the previous section introduced the necessary notation and probabilistic preliminaries, here we summarise the generative procedure of the fully Bayesian formulation of LDA:

1. **Topic distributions:** For each topic $k = 1, \dots, K$, draw a topic-word distribution

$$\boldsymbol{\phi}_k \sim \text{Dir}(\boldsymbol{\beta}).$$

2. **Document proportions:** For each document $d = 1, \dots, M$, draw a document-topic distribution

$$\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\alpha}).$$

3. **Words in documents:** For each word position $n = 1, \dots, N_d$ in document d:

   a. Draw a topic assignment

$$z_{dn} \sim \mathrm{Cat}(\boldsymbol{\theta}_d),$$

   b. Draw the word

$$w_{dn} \sim \mathrm{Cat}\left(\boldsymbol{\phi}_{z_{dn}}\right).$$

The generative specification described above induces a joint probability distribution, conditional on the latent parameters $\boldsymbol{\theta}_{1:M}$ and $\boldsymbol{\phi}_{1:K}$ over all topic assignments $\mathbf{z}$ and words $\mathbf{w}$ in the corpus. This distribution can be expressed as:

$$p(\boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{\theta}_{1:M}, \boldsymbol{\phi}_{1:K}) = \prod_{d=1}^{M} \prod_{n=1}^{N_d} p\,(z_{dn} \mid \boldsymbol{\theta}_d)\, p(w_{dn} \mid z_{dn}, \boldsymbol{\phi}_{1:K})$$

with $p(z_{dn} = k \mid \boldsymbol{\theta}_d) = \theta_{d,k}$ and $p(w_{dn} = v \mid z_{dn} = k, \boldsymbol{\phi}_k) = \phi_{k,v}$. Since the latent parameters $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$ are unknown, a Bayesian formulation places prior distributions on them, and, specifically, Dirichlet priors are imposed on the document-topic distributions $\boldsymbol{\theta}_d$ and on the topic-word distributions $\boldsymbol{\phi}_k$. Formally, the full joint probability is:

$$p(\,\boldsymbol{\theta}_{1:M}, \boldsymbol{\phi}_{1:K}, \boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}\,)$$
$$= \prod_{k=1}^{K} p\,(\boldsymbol{\phi}_k \boldsymbol{\beta}) \prod_{d=1}^{M} \left(p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) \prod_{n=1}^{N_d} p\,(z_{dn} \mid \boldsymbol{\theta}_d)\, p(w_{dn} \mid z_{dn}, \boldsymbol{\phi}_{1:K}) \right),$$

where $p(\boldsymbol{\phi}_k \mid \boldsymbol{\beta})$ denotes the Dirichlet density for topic k and $p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha})$ the Dirichlet density for document d. Substituting:

$$p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_{d,k}^{\alpha_k - 1}, \qquad p(\boldsymbol{\phi}_k \mid \boldsymbol{\beta}) = \frac{1}{B(\boldsymbol{\beta})} \prod_{v=1}^{V} \phi_{k,v}^{\beta_v - 1},$$

where $B(\boldsymbol{\alpha})$ and $B(\boldsymbol{\beta})$ are the beta-function normalising constant of the Dirichlet distributions:

$$B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma\,(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}, \qquad B(\boldsymbol{\beta}) = \frac{\prod_{v=1}^{V} \Gamma\,(\beta_v)}{\Gamma(\sum_{v=1}^{V} \beta_v)}.$$

The marginal probability distribution of the topic assignments in a document, obtained by integrating out its document-topic proportions $\boldsymbol{\theta}_d$, follows a Dirichlet-multinomial form depending only on the topic counts within the document. This distribution has a closed expression but once extended to the full corpus the marginal likelihood of the observed words becomes substantially more complex. Formally, the likelihood of the data is

$$\iint \sum_{\boldsymbol{z}} p(\,\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}\,)\, d\boldsymbol{\theta} d\boldsymbol{\phi}.$$

While the integrals over $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ can be evaluated analytically due to Dirichlet-multinomial conjugacy (Dickey, 1983), the remaining summation over the exponentially many topic assignments $\boldsymbol{z}$ renders the expression intractable. For a document of length N_d, there are K^{N_d} *p*ossible topic configurations, making exact computation computationally prohibitive. Therefore, exact inference in LDA is not feasible for realistic text collections and we must use approximate methods (Blei et al., 2003).

LDA Visualization

It is often helpful to visualise LDA with a graphical model in plate notation. The figure below shows LDA's structure (three hierarchy levels: corpus, document, word):

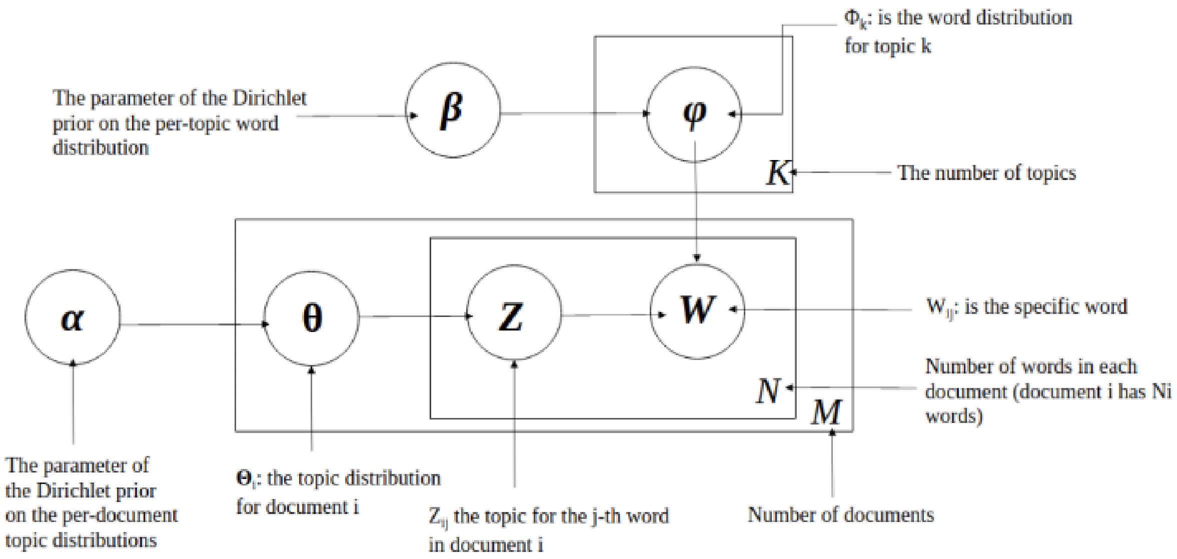

Φ_k: is the word distribution for topic k
The parameter of the Dirichlet prior on the per-topic word distribution
β
φ
K
The number of topics
α
θ
Z
W
W_{ij}: is the specific word
N
M
Number of words in each document (document i has Ni words)
The parameter of the Dirichlet prior on the per-document topic distributions
Θ_i: the topic distribution for document i
Z_{ij} the topic for the j-th word in document i
Number of documents

Figure 2.2 (Zankadi et al. 2022). The boxes ("plates") indicate replication: the outer plate (M) represents documents, the outer plate (K) represents topics, and the inner plate (N_d) represents *repeated word-generating steps within each document. Each document d has a Dirichlet prior* $\boldsymbol{\alpha}$ *generating* $\boldsymbol{\theta}_d$*, and each topic k has a Dirichlet prior* $\boldsymbol{\beta}$ *generating* $\boldsymbol{\phi}_k$*. For each word token* w_{dn}*, a topic* z_{dn} *is chosen from* $\boldsymbol{\theta}_d$ *(arrow* $\boldsymbol{\theta} \rightarrow \mathbf{z}$*), and the word* w_{dn} *is then chosen from* $\boldsymbol{\phi}_{z_{dn}}$ *(arrows* $\boldsymbol{\phi} \rightarrow \mathbf{w}$ *and* $\mathbf{z} \rightarrow \mathbf{w}$*).*

LDA can also be understood geometrically as a matrix factorisation or dimensionality-reduction model. Each topic $\boldsymbol{\phi}_k$ is a point (vector) in the V-dimensional word simplex. Visualise a $(V - 1)$-simplex whose corners represent degenerate distributions that put all probability on a single word. Each $\boldsymbol{\phi}_k$ lies somewhere inside this simplex (typically near some corners if $\boldsymbol{\beta}$ is small). For each document d, the empirical word distribution is modelled as a convex combination of the K topic vectors $\{\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_K\}$, with weights given by the document's topic proportions $\boldsymbol{\theta}_d$. Thus the set of all distributions that LDA can produce for documents (given fixed topics) is the convex hull of the $\boldsymbol{\phi}_k$ vectors, a "topic simplex" (a K-vertex polytope) embedded within the word simplex. For example, with $K = 3$ and $V = 3$ the word simplex is a triangle; the three topics are points inside that triangle, and their convex hull is the topic triangle. Each document's word distribution $p(\boldsymbol{w} \mid$ d) lies somewhere inside the topic triangle. If a document uses primarily one topic, its $\boldsymbol{\theta}_d$ is near a unit vector and its word distribution is near that topic's corner. If it mixes topics evenly, it lies more centrally. A symmetric Dirichlet prior leads to smooth distributions over the topic simplex (the concentric contour lines).

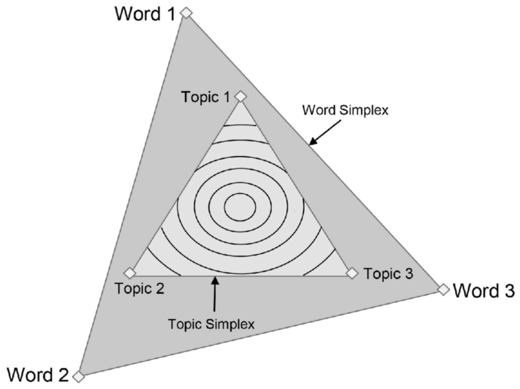


Figure 2.3 (Lopea, 2017) Geometric interpretation of LDA in the case K=3 topics and V=3 words.

Additional properties, strengths, and limitations of LDA

Having introduced the full generative structure, we discuss the main strengths and limitations of Latent Dirichlet Allocation:

Strengths:

I. Interpretability. Topics often correspond to coherent themes and are easily understandable via top-word lists. Unlike earlier methods such as Latent Semantic Analysis, which produce dense, difficult-to-interpret factors, LDA's probabilistic topics and sparse distributions make it easy to assign semantic labels.

II. Admixture structure. LDA does not limit documents to only one topic but captures multi-faceted content by modelling documents as a mixture of multiple topics.

III. Scalability. Inference via variational Bayes or collapsed Gibbs sampling scales linearly with the number of tokens and topics. Moreover, LDA's generative nature means it is easy to incorporate new documents after training (for example, inferring topics for new documents without retraining the whole model, using a fixed set of learned topics). Inference algorithms for LDA can be adapted with only minor changes to different data types or extended models, underscoring that LDA provides a tractable basis for large-scale unsupervised learning (Blei, 2012).

IV. Extensibility. LDA's modular design has led to numerous variants (e.g., correlated topic models, hierarchical LDA, dynamic topic models) that adapt the core structure to specific needs (Blei & Lafferty, 2007).

V. Probabilistic grounding. It provides a coherent generative probabilistic model for text, which allows the utilisation of Bayesian inference techniques and offers a clear semantics to model components.

Limitations:

I. Identifiability**.** In LDA, different combinations of topics can explain the same dataset equally well, so there is no guaranteed "true" solution. We may add some constraints, but in practice, identifiability issues are managed by assessing stability across runs.

II. Label switching. Topics are exchangeable under symmetric priors, and this symmetry implies that without imposing artificial identifiability constraints, one cannot attribute an identity to a particular topic across different runs of the algorithm. Label switching is especially apparent in MCMC sampling: the sampler's state space contains *K!* symmetric modes corresponding to each possible labelling of the topics. All these modes represent the same mixture model solution, just with permuted labels. While label switching does not affect the model's fit or predictive power, it complicates interpretation and post-processing.

III. Bag-of-words assumption. Perhaps the most often-cited limitation of LDA is its bag-of-words assumption. LDA treats each document as an unordered collection of words, meaning it ignores grammar, word order, syntax, and even semantics beyond word co-occurrence. Furthermore, ignoring word context means LDA sometimes groups words in a topic that co-occur but have distinct senses (polysemy issues).

IV. No topic correlation. Standard LDA also assumes that topics are a priori independent in documents, which can be restrictive. The Dirichlet document-topic prior treats each topic's proportion as drawn independently (aside from the normalisation constraint). This leads to the absence of a mechanism in basic LDA for one topic to be more likely given the presence of another. In reality, topics in corpora often exhibit correlations. For example, documents about sports might frequently be about health (think of sports medicine), whereas a document about sports is unlikely to also be about medieval art. LDA cannot capture such patterns. Treating the topics as independent aspects is often an oversimplification of reality, but one can overcome it by using more complex graphical models at the cost of more difficult inference.

V. Inference instability. The process of fitting LDA (be it via EM, variational Bayes, or Gibbs sampling) can suffer from instability due to local optima (Blei, 2016). The likelihood surface of LDA is highly multimodal, partly because of the label switching symmetry and partly due to the high-dimensional, non-convex nature of the parameter space, so different initialisations can yield different topics (Newman et al., 2010). To reduce this problem, people usually run the model several times and compare the results to see if the topics are consistent.

2.4 Inference

Having introduced the generative structure of Latent Dirichlet Allocation, we now turn to the problem of inference. The objective is to recover the latent structure from the observed corpus $\boldsymbol{w}$. This requires computing the posterior distribution:

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{z} \mid \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})}.$$

As we have already discussed, this posterior distribution is computationally intractable in the general case. Given the intractable nature of exact posterior inference, numerous approximate inference techniques have been developed for LDA. We will discuss two of the main techniques: Collapsed Variational Bayes and Collapsed Gibbs Sampling.

Collapsed Variational Bayes

The collapsed approach analytically marginalises the continuous variables $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$ from the joint distribution, focusing inference on the discrete topic assignments $\boldsymbol{z}$. Under the variational framework, the true posterior distribution $p(\boldsymbol{z} \mid \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is intractable. To address this problem, we can introduce a tractable family of approximating distributions, known as *variational posteriors*. These are parameterized distributions $q(\mathbf{z})$ designed to approximate the true posterior as closely as possible while remaining computationally manageable. In the collapsed setting, the variational posterior is factorised across documents and word positions as:

$$q(\boldsymbol{z}) = \prod_{d=1}^{M} \prod_{n=1}^{N_d} q\,(z_{dn}).$$

so that each topic assignment z_{dn} is governed by an independent variational distribution. The expression "*collapsed variational formulation*" refers precisely to this marginalisation of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, which reduces the number of variables to be approximated. It has an important advantage: the collapsed variational formulation yields a tighter evidence lower bound (ELBO) than alternative variational approximations, as the effects of the marginalised parameters are correctly evaluated in CVB while standard VB approximates them (Ishiguro et al., 2017). Nevertheless, the resulting expressions still contain terms that cannot be evaluated in closed form. To make inference tractable, Collapsed Variational Bayes employs approximations based on Taylor expansions of these intractable expectations. The simplest variant, CVB0, applies a zeroth-order approximation, while higher-order versions (commonly second-order) improve accuracy at the cost of additional computation. In what follows we consider two settings:

I. Symmetric priors: $\alpha_k = \alpha > 0$ and $\beta_v = \beta > 0$ for all k, v (thus $\sum_{v=1}^{V} \beta_v = V\beta$).

II. Asymmetric document-topic prior: $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$ with $\alpha_k > 0$ and symmetric $\beta_v = \beta > 0$ (thus $\sum_{v=1}^{V} \beta_v = V\beta$)

All counts are leave-one-out, i.e., they exclude the current token (*dn*) and are denoted with a superscript $\neg dn$. Moreover, define:

- $N_{d,k}^{\neg dn}$: expected count of tokens in document d assigned to topic k, excluding the current token.
- $N_{k,w_{dn}}^{\neg dn}$: expected count of tokens of word type w_{dn} assigned to topic *k*, excluding the current token.
- $N_{k}^{\neg dn} = \sum_{v=1}^{V} N_{k,v}^{\neg dn}$: total expected count of tokens assigned to topic *k*, excluding the current token.
- $V_{dk}^{\neg dn}$; $V_{wk}^{\neg dn}$; $V_{k}^{\neg dn}$: associated variances.
- V: vocabulary size.

CVB0

Symmetric prior case:

$$\gamma_{dnk} \propto \frac{N_{k,w_{dn}}^{\neg dn} + \beta}{N_{k}^{\neg dn} + \mathrm{V}\beta} \cdot (N_{d,k}^{\neg dn} + \alpha).$$

Asymmetric document-topic prior case:

$$\gamma_{dnk} \propto \frac{N_{k,w_{dn}}^{\neg dn} + \beta}{N_{k}^{\neg dn} + \mathrm{V}\beta} \cdot (N_{d,k}^{\neg dn} + \alpha_k).$$

CVB with second-order Taylor expansions:

Symmetric prior case:

$$\gamma_{dnk} \propto \frac{N_{k,w_{dn}}^{\neg dn} + \beta}{N_{k}^{\neg dn} + \mathrm{V}\beta} \cdot \left(N_{dk}^{\neg dn} + \alpha\right) \cdot exp\left\{-\frac{V_{dk}^{\neg dn}}{2(N_{dk}^{\neg dn} + \alpha)^2} - \frac{V_{wk}^{\neg dn}}{2(N_{k,w_{dn}}^{\neg dn} + \beta)^2} + \frac{V_{k}^{\neg dn}}{2(N_{k}^{\neg dn} + V\beta)^2}\right\}.$$

Asymmetric document-topic prior case:

$$\gamma_{dnk} \propto \frac{N_{k,w_{dn}}^{\neg dn} + \beta}{N_{k}^{\neg dn} + \mathrm{V}\beta} \cdot \left(N_{dk}^{\neg dn} + \alpha_k\right) \cdot exp\left\{-\frac{V_{dk}^{\neg dn}}{2(N_{dk}^{\neg dn} + \alpha_k)^2} - \frac{V_{wk}^{\neg dn}}{2(N_{k,w_{dn}}^{\neg dn} + \beta)^2} + \frac{V_{k}^{\neg dn}}{2(N_{k}^{\neg dn} + V\beta)^2}\right\}.$$

Intuitively the term $\frac{N_{k,w_{dn}}^{\neg dn}+\beta}{N_{k}^{\neg dn}+\mathrm{V}\beta}$ represents the probability of the word w_{dn} under topic k (topic-word compatibility). The term $\left(N_{dk}^{\neg dn}+\alpha\right)$ or $\left(N_{dk}^{\neg dn}+\alpha_k\right)$ represents how strongly document d tends to use topic k (document-topic preference).

In practice, Collapsed Variational Bayes is implemented as an iterative algorithm: the update rules for γ_{dnk} are applied repeatedly across all tokens in the corpus and at each iteration, the variational parameters for each token assignment z_{dn} are updated using the current estimates of the counts $N_{dk}^{\neg dn}$, $N_{k,w_{dn}}^{\neg dn}$, $N_{k}^{\neg dn}$. After one full pass over the corpus, the counts are refreshed, and the process is repeated until convergence, typically assessed by monitoring the change in the evidence lower bound (ELBO) or by checking stability of the variational distributions.

Collapsed Gibbs Sampling

Another common alternative way to perform inference instead of Variational Bayes is to use Markov Chain Monte Carlo methods: Collapsed Gibbs Sampling (CGS) (Griffiths and Steyvers, 2004) takes advantage of the Dirichlet-Multinomial conjugacy to analytically integrate out the continuous parameters $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$, focusing the sampling process entirely on the topic assignment variables. In Collapsed Gibbs Sampling, the conditional probability of assigning topic k to the token at position n in document d given the observed word w_{dn}, is:

Symmetric prior case:

$$P(z_{dn}=k \mid \boldsymbol{z}^{\neg dn},\boldsymbol{w},\alpha,\beta) \propto \frac{N_{k,w_{dn}}^{-dn}+\beta}{N_{k}^{-dn}+\mathrm{V}\beta}\cdot\left(N_{d,k}^{\neg dn}+\alpha\right).$$

Asymmetric on document-topic prior case:

$$P(z_{dn}=k \mid \boldsymbol{z}^{\neg dn},\boldsymbol{w},\alpha,\beta) \propto \frac{N_{k,w_{dn}}^{-dn}+\beta}{N_{k}^{-dn}+\mathrm{V}\beta}\cdot\left(N_{d,k}^{\neg dn}+\alpha_k\right).$$

The collapsed formulation delivers multiple benefits over non-collapsed alternatives. First of all, it eliminates the strong correlations between $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, and $\boldsymbol{z}$. Secondly, it focuses sampling on discrete assignments rather than continuous distributions. Moreover, it maintains ergodicity and detailed balance properties of MCMC. Ergodicity means that if you observe a random process long enough, the average behaviour over time will be the same as the average over all possible outcomes. Even though un-collapsed Gibbs Sampling over all latent variables is possible, empirical evidence consistently demonstrates superior mixing properties for the collapsed variant in practice.

It is interesting to note that the conditional probability used to update topic assignments in CGS is mathematically identical to the update rule employed in CVB0. They differ in how this probability is used, because CGS draws a single topic assignment for each token at each iteration, introducing stochasticity, and CVB0 updates a full discrete probability distribution over topics for each token deterministically. CVB0 can be viewed as the deterministic "infinite-sample" limit of CGS: if CGS were to draw infinitely many samples for each token from its conditional probability, the resulting empirical distribution would match the CVB0 update exactly. By using the full probability distribution rather than a single sample, CVB0 eliminates sampling noise and can converge faster, while CGS retains the benefit of asymptotic exactness (Asuncion et al., 2009).

How to choose the hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

We have already discussed how the Dirichlet distribution works, and the effect of the parameters: the parameter of the document-topic distribution controls the expected number of topics significantly represented within a document, while the parameter of the topic-word distribution governs the distribution of words within each topic. Smaller values encourage sparsity, and larger values encourage more uniformity.

It is possible to estimate the parameters with a priori knowledge, nonetheless, they are not fixed constants and should be treated as random variables with prior distributions, often Gamma distributions to ensure positivity and to take advantage of conjugacy properties (Wallach et al., 2009). This allows the model to infer their values directly from the data rather than relying on ad-hoc settings. When implementing Collapsed Gibbs Sampling, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be updated alongside the topic assignments within each iteration. At the beginning, initial values are chosen, most commonly the symmetric case $\boldsymbol{\alpha} = \left(\frac{50}{K}, \dots, \frac{50}{K}\right)$ and $\boldsymbol{\beta} = (0.01, \dots, 0.01)$, but in practice they should be based on the observed sparsity and coherence of the topics. In addition, each word token is randomly assigned to a topic (Griffiths & Steyvers, 2004). Even though the initial counts are based on random allocations, they serve as the starting point for an iterative refinement process. As in CGS, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be updated, and since their conditional posterior distributions are not available in closed form, methods such as Metropolis-Hastings or slice sampling are typically employed for these updates (Wallach et al., 2009). With successive iterations, the inference procedure updates topic assignments and hyperparameters so that the sampled or estimated values progressively approximate the underlying posterior distribution. In the simplest case, the document-topic prior $\boldsymbol{\alpha}$ is symmetric, but in more flexible settings one allows $\boldsymbol{\alpha}$ to be asymmetric, with different values $(\alpha_1, \dots, \alpha_K)$: this means we believe some specific topics are a priori more likely to appear than others. During learning, one may find an asymmetric $\boldsymbol{\alpha}$ where some α_k are higher (topics that occur in many documents), and some are lower (rare topics). Wallach et al. (2009) showed that using an asymmetric α learned from data often improves held-out likelihood, since it can allocate probability mass appropriately (e.g., a "background" topic of stop words might get a high α_k because it appears in every document, whereas very specific topics get low α_k). Asymmetric $\boldsymbol{\beta}$ could be used if one has prior knowledge that certain words are generally more likely in any topic. However, typically $\boldsymbol{\beta}$ is kept symmetric and learned as a single scalar. Like $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ can be tuned by cross-validation or learned by maximum likelihood. The learned value often ends up in the range of a few tenths for good topic coherence. The hyperparameters are one reason LDA outperforms PLSA (Probabilistic Latent Semantic Analysis): PLSA effectively has an implicit $\boldsymbol{\alpha} \to 0$ for each document (no sharing of topic mixture prior) and no $\boldsymbol{\beta}$ prior, which can lead to overfitting on training data (Hofmann, 1999).

How to choose the number of topics

Selecting the appropriate number of topics K can be done following different criteria. It is important to understand how this decision significantly impacts the interpretability and effectiveness of the resulting model: too small a value of K may lead to overly broad topics, and too high values to fragmented or redundant topics. Newer and more complex expansions of LDA make it possible to let the model directly infer the number of topics based on the data, but in this work, we will focus on K chosen a priori. However, this decision can still be guided by different methods, such as minimising the "perplexity" or maximising the "topic coherence score".

Perplexity (held-out likelihood): A common quantitative measure is held-out perplexity, which is based on the model's predictive likelihood for unseen data. Perplexity is defined similarly to how it is in language modelling. In the context of Latent Dirichlet Allocation, perplexity measures how well the model predicts a held-out set of documents. Given a held-out test set of documents D_{test}, perplexity is:

$$\text{perplexity}(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d\in D_{\text{test}}} \log p(\boldsymbol{w}_d)}{\sum_{d\in D_{\text{test}}} N_d}\right\}.$$

Intuitively, perplexity is the geometric mean of the inverse likelihood per word. It can be thought as a way of asking how surprised the model is when it sees new text, and looking at the formula, it is possible to understand that lower values of perplexity mean the model is assigning higher probability to the held-out documents, indicating better generalisation. This idea is not recent: the approach has been widely adopted in early studies of topic modelling, notably by Griffiths and Steyvers (2004), who relied heavily on perplexity to decide how many topics to use when training LDA on collections such as scientific abstracts. It is also important for the practical applications, that perplexity remains computationally tractable when used with approximate inference techniques such as variational Bayes or Gibbs sampling. When comparing models (with different K or different algorithms), the dataset is divided into a training and a test set. The model computes perplexity on the test set, and the values tends to decrease as K increases (since a larger model can fit the training data more flexibly), but after a point, perplexity on held-out data will start to rise if the model is overfitting, due to too many topics: the model may encode noise or very corpus-specific patterns that do not generalize. In practice, one might train LDA with various K values and pick the K that minimises held-out perplexity. It is worth noting that perplexity, being based on likelihood, sometimes does not correlate with human judgments of topic quality (Chang et al., 2009), the words in a topic may not be interpretable as a group, and therefore not useful to the classification of documents, but still have low values of perplexity. Still, it remains a convenient objective metric for model selection in LDA, since it directly relates to predictive power.

Topic coherence: To address the disconnect between perplexity and human interpretability, researchers use topic coherence measures, which evaluate how semantically meaningful the top words of each topic are. One of the most established coherence metrics is the UMass coherence, introduced by Mimno et al. (2011). It evaluates whether high-probability words in a topic tend to co-occur in the same documents, based on corpus-level word co-occurrence statistics. For a topic k consisting of the top M words $\{w_1, w_2, \dots, w_M\}$, the UMass coherence is defined as:

$$\text{Coherence}(k) = \sum_{m=2}^{M}\sum_{l=1}^{m-1}\log\left(\frac{D(w_m, w_l) + \epsilon}{D(w_l)}\right).$$

Where $D(w_m, w_l)$ is the number of documents in which both words w_m and w_l appear, $D(w_l)$ is the document frequency of w_l and ϵ is a small smoothing factor (typically $\epsilon = 1$) to avoid taking the log of zero. For example, if a topic's top words are "bank, money, loan, credit, ...", these tend to appear in similar documents, so the topic is coherent, whereas a topic with top words "money, cell, algorithm, bank" might have low coherence because those groups of words do not all co-occur frequently (and still have low value of perplexity). A coherence score can be aggregated across topics, often by taking the mean (or by ensuring most topics have acceptable coherence). This enables the comparison of multiple models trained with different values of K. Implementing LDA on a large corpus of data takes time, so it is always better to already have an idea of the range of the number of topics possible, and the comparison then permits researchers to select the exact value of K that maximises average topic coherence. When comparing different K, one might observe that as K grows very large, perplexity often continues to decrease while average topic coherence plateaus or even declines, due to over-fragmentation. In practice, topic coherence is used alongside perplexity as they use different criteria: perplexity reflects statistical generalisation, coherence captures the human-oriented semantic validity of the topics. Usually, in settings where interpretability is more important, coherence is often regarded as the more informative metric.

3 EMPIRICAL METHODOLOGY

3.1 Introduction to the case study

In this chapter, we move from theory to practice, seeking to understand what can be learned by analysing the speeches of the European Central Bank (ECB) using the LDA model. The ECB's choice is not random: over time, the bank's communication has become increasingly central to the dynamics of financial markets. Initially, public statements primarily served to explain decisions that had already been made, but over the years, especially during the global financial crisis and the subsequent sovereign debt crisis, speeches and press conferences have been utilised as genuine monetary policy tools, capable of guiding market expectations. The point investigated here is whether there is a link between the recurring topics in speeches and the interest rate decisions taken in the days following meetings. In other words, if a certain topic (such as "inflation", "growth", "financial markets") appears more frequently in speeches preceding a meeting, is it possible that this may anticipate a certain type of rate decision?

A multi-stage procedure was followed to investigate this aspect. First, the corpus of speeches was collected and subjected to linguistic cleaning, eliminating texts that were too short or had encoding problems. Next, the LDA model was estimated in order to identify a set of main topics and associate each speech with a probability distribution across the different topics. The results were then aggregated at the meeting level, considering the speeches delivered in the weeks preceding each meeting. This made it possible to link the relative presence of the topics with the decisions actually taken by the ECB.

The goal is not only to demonstrate that LDA can be successfully applied to a complex corpus such as that of the ECB, but also to see whether this approach offers interesting insights into the relationship between communication and monetary policy decisions. It is therefore a case study that combines method and substance: on the one hand, it shows the operational steps for using LDA, and on the other, it attempts to draw some useful evidence on the role of European policymakers' speeches. The analysis is divided into two parts: a descriptive phase, aimed at illustrating the evolution of the topics over time and their distribution across different periods of European monetary policy, and an inferential phase, based on the regressions, aimed at assessing the ability of the topics to anticipate monetary policy decisions.

3.2 Data and Preprocessing

The empirical material used comes from the public archive of official speeches of the European Central Bank (ECB) (it is possible to find it in the link to the repository GitHub in Appendix 9), which collects speeches by members of the Governing Council and other representatives of the institution since 1997. To analyse the data, i.e. for the entire process of collection, cleaning, preprocessing, implementation of LDA, and regressions, I used Python, using standard libraries for processing textual and numerical data.

The original dataset includes a total of 2,928 speeches. After an initial cleaning phase that eliminated records with no textual content or consisting solely of empty strings, the number of documents was reduced to 2,798. Subsequently, automatic language detection was performed using the *"langdetect"* package. Speeches too short to be classified were marked as *"short"*. Limiting the analysis to English only, the final corpus comprises 2,612 speeches, spanning a time period from February 7, 1997, to August 23, 2025.

Descriptive statistics show an average text length of approximately 3,064 words, with a median of 2,745 words. The 10th percentile corresponds to just over 1,200 words, while the 90th percentile is above 5,100 words. These are therefore generally long and homogeneous speeches, with differences attributable to format (press conferences, short speeches, annual reports).

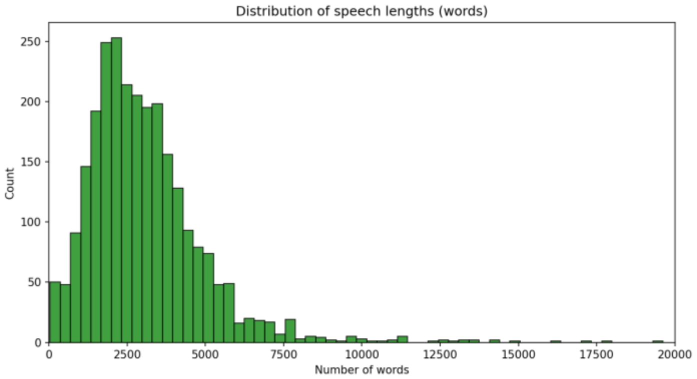


Figure 3.1 Distribution of the number of words for each speech. The highest density is concentrated between 2,000 and 4,000 words, with a median of approximately 2,745 words. There are some outliers of exceptionally high length (over 10,000 words), but these do not significantly alter the overall distribution.

Given the objective of thematic analysis with LDA, the texts were subjected to a processing pipeline that included:

- Normalisation and removal of symbols and numbers.
- Tokenisation and lemmatisation with the Python package "*spaCy*", limited to nouns, adjectives, verbs, and adverbs.
- Removal of English stop words and low-informative ECB-specific terms (e.g., ECB, monetary policy, conference).
- Construction of recurring bigrams (e.g., forward_guidance, asset_purchase).

During the analysis, all the documents marked as "*short*" (documents with fewer than 40 tokens after cleaning) were excluded to ensure robustness in model estimation, and at the end of this phase, the filtered vocabulary contained several thousand distinct lemmas.

In parallel with the text corpus, two quantitative datasets were collected and harmonised:

- EA-MPD (Euro Area Monetary Policy Database), from which the official dates of monetary policy meetings were obtained (312 meetings, 1999-2025).
- ECBDFR (ECB Deposit Facility Rate), used to codify rate decisions (hike, cut, hold) on the day of the meeting and in the following 7-day window.

3.3 Choosing the priors and number of topics

As we have discussed in section 2.4 of this thesis, there are several strategies for choosing these hyperparameters: the most common default is the symmetric α = 50/K, while for β, the most commonly used standard value is 0.01. In this study, values of α = 50/K were not explored, as they would have generated excessively high values of α ($\alpha > 1$) for our number of topics (intended to be between 15 and 35). This approach would have resulted in overly uniform document-topic distributions, reducing the model's ability to capture specific thematic concentrations within the ECB's communications. It was therefore considered more consistent with the context analysed to maintain symmetric and lower values of α, reflecting the expectation that each speech addresses a limited but significant set of topics. To choose the best α, a sensitivity exercise was conducted, testing different values (0.1, 0.3, 0.5 and the default K/50 used in the "*tomotopy*" package) in combination with multiple values of K (20, 25, 30). The analysis showed that α = 0.1 guarantees better results in terms of semantic consistency of the topics and more stable levels of perplexity, thus suggesting a more parsimonious structure, with documents tending to focus on a few main topics.

Number of topics	**20**	**25**	**30**
Alpha			
0.1	0.45929091	0.46035966	**0.47513287**
0.3	0.45257599	0.46213337	0.47114421
0.5	0.4598342	0.46143233	0.46934775
K/50	0.45877032	0.46143233	0.46651124

Based on these considerations, the final model estimation was conducted by setting the symmetric value of α = 0.1 and β = 0.01, replicating the analysis on different values of K to identify the optimal configuration.

The final choice of the number of topics fell on $K = 30$. The analysis conducted with prior $\alpha = 0.1$ shows that coherence increases steadily as K increases, reaching its highest value at 30 topics (0.475), higher than the values observed for K =15, 20, 25, and 35 (*Appendix 2*). At the same time, perplexity remains low for almost all values of K but increases with K=35 (*Appendix 3*). All values can be seen in the table in *Appendix 4.* In order to verify the robustness of this choice, the same analysis was repeated for the alternative values of the document-topic prior ($\alpha = 0.3, \alpha = 0.5, \alpha = K/50$). The results (Appendix 5) show that for each of the higher values of α, coherence systematically reached its maximum at $K = 35$, while perplexity had its minimum at smaller values of K (15 or 20). The difference in coherence between $K = 30$ and $K = 35$ was modest (between 0.88% and 4.65% depending on the prior), but perplexity at $K = 30$ was consistently lower or very close to that observed at $K = 35$. This pattern suggests that, although larger models may marginally improve interpretability, they do so at the cost of statistical fit. Taken together, the results indicate that $K = 30$ represents a stable and robust compromise across different priors, ensuring sufficient topic granularity without incurring the perplexity deterioration observed for $K = 35$.

Looking at the qualitative assessment of the topics that emerged also confirms the validity of this choice: with 30 topics, the model is able to distinguish the main areas of the ECB's speeches (inflation, monetary policy, forward guidance, sovereign debt crisis, etc.) in an interpretable and consistent manner, without producing excessive fragmentation. Therefore, the value of K=30 represents a balanced compromise between quantitative robustness (following the results of coherence and perplexity) and qualitative readability of the topics and is adopted as the reference configuration for subsequent analyses in this case study. The table in *Appendix 1* shows the complete set of topics and their respective top-words.

3.4 Econometric and dimensionality-reduction framework

To investigate the extent to which the topics emerging from the LDA analysis are able to anticipate the ECB's monetary policy decisions, a multinomial logit regression model was estimated, in which the dependent variable takes three categories: hold, hike, or cut (hold as the base category). Firstly, to avoid problems of multicollinearity and complicated interpretability, not all 30 topics were used, but only a subset of independent variables was initially selected. Subsequently, to reduce the dimensionality of the model and overcome the problem of multicollinearity among the 30 topics identified by LDA, Principal Component Analysis (PCA) was applied, because this strategy permits a large set of correlated variables to be transformed into a small number of orthogonal components capable of capturing the main sources of variation in the dataset.

4 RESULTS AND DISCUSSION

4.1 Descriptive analysis of topics

To interpret the results of the analysis carried out using the LDA model, it is very useful to observe how the identified topics evolve and are distributed over time: this analysis can verify the consistency between various topics and various macroeconomic events and can help explain how the European Central Bank's communications have accompanied the different phases of monetary policy over the last 25 years.

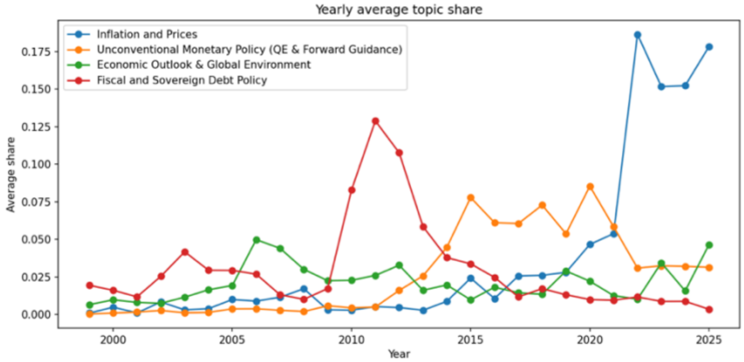


Figure 4.1 Yearly average share of four relevant topics from 1999 to 2025.

Figure 4.1 shows the average annual trend in four key areas: "*Inflation and Prices"*, "*Unconventional Monetary Policy (QE & Forward Guidance)", "Economic Outlook & Global Environment*", and "*Fiscal and Sovereign Debt Policy*". By "average share", we mean the average annual share of each topic, calculated as the arithmetic mean of the thematic distributions of all the speeches delivered in that year. This choice allows the progress of the themes over time to be represented in a synthetic and legible way, avoiding the excessive fragmentation that would result from reporting each speech individually. The issue of inflation has been steadily growing since the beginning of the period and has become increasingly important in recent years, growing especially with the return of inflationary pressures, particularly during the COVID period and the energy crisis. Attention to sovereign debt is mainly focused on the two years between 2010 and 2012, coinciding with the government bond crisis. Similarly, the topic of unconventional monetary policy, which emerged after 2011, became an integral part of communication, and its importance continued to grow, peaking in 2015 and 2020. Finally, a more general topic such as "*Economic*

Outlook and Global Environment" remained relatively constant in importance throughout the analysis period.

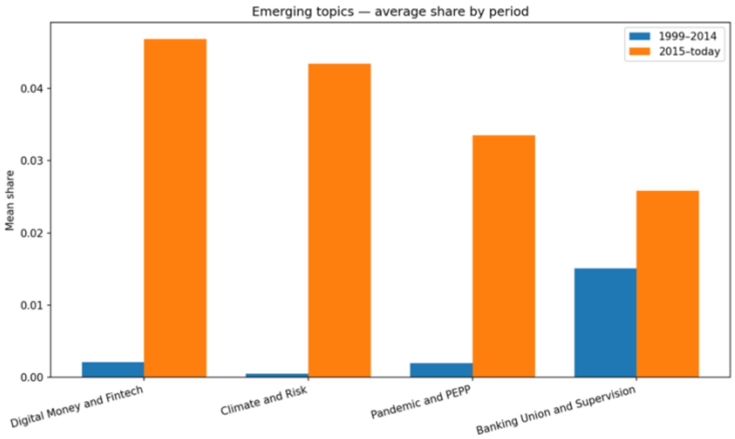


Figure 4.2 Average share by period 1999-2014 or 2015-present of four emerging topics.

Figure 4.2 shows how the emergence of new topics over the last 15 years is evident thanks to the LDA analysis. Topics such as digital currency, climate and risk, the pandemic, and banking union and supervision are relatively recent, and all show a marked increase in the frequency of speeches compared to the previous period, 1999-2014. In particular, the European Central Bank has begun to systematically discuss digitisation and climate in line with structural changes in the economy and the world. Since 2020, the pandemic has introduced a whole new language, linked to exceptional instruments such as the PEPP, while the Banking Union has become increasingly present following the strengthening of the institution's supervisory role. The data confirms that the European Central Bank's communications evolve in line with historical events and the priorities that arise from them.

4.2 Multinomial Logit Regression

The "topics" extracted via Latent Dirichlet Allocation (LDA) must not be understood as qualitative or dummy variables, but rather as continuous variables. In particular, each speech of the ECB is represented by a probability distribution on the K topics identified by the model, i.e. by a vector of values between 0 and 1 which add to 1. These values therefore correspond to the relative share of each topic within the discourse.

To link the speeches to monetary policy decisions, the values of the topics have been aggregated at the meeting level: for each Governing Council meeting, the average of the topic shares of speeches delivered during the preceding decision has been calculated. Thus, each meeting is defined by a set of independent variables reflecting the average distribution of the discussed topics. For example, if in the 45 days (period used in the calculations) before a meeting, on average 40% of the content of the speeches relates to inflation, 25% to unconventional monetary policy, 20% to liquidity operations and 15% to the ECB strategy, these values directly constitute the independent variables to be associated with that meeting in the model.

The choice of topics to use as independent variables in the first regression follows two complementary criteria:

I. Economic relevance: the topics most closely related to monetary policy have been made a priority (for example, "Inflation and Prices", "Monetary Policy and Liquidity Operations", "Non-Conventional Monetary Policy").
II. Robustness statistics: to lighten multicollinearity and high interpretative complexity, a subset of topics was initially taken, and subsequently a principal component analysis (PCA) was applied, in order to summarise the information coming from all 30 topics in the form of a limited number of orthogonal components.

Explanation of the model

The estimated model is a multinomial logit in which the dependent variable Y_m is the monetary policy decision made in meeting m, with three possible categories:

- 0 = Hold
- 1 = Cut
- 2 = Hike

Let $\mathbf{x}_m = (x_{m1}, x_{m2}, \dots, x_{mJ})$ be the vector of independent variables for meeting m, where each x_{mj} corresponds to the average share of the j topic in the speeches preceding the meeting. The model takes the form:

$$P(Y_m = k \mid \boldsymbol{x}_m) = \frac{exp(\beta_{k0} + \boldsymbol{\beta}_k{'}\boldsymbol{x}_m)}{1 + \sum_{l=1}^{2} exp\,(\beta_{l0} + \boldsymbol{\beta}_l{'}\boldsymbol{x}_m)}, \quad k = 1{,}2$$

and

$$P(\,Y_m = 0 \mid \boldsymbol{x}_m\,) = \frac{1}{1 + \sum_{l=1}^{2} exp\,(\beta_{l0} + \boldsymbol{\beta}_l{'}\boldsymbol{x}_m)}.$$

In that specification, the coefficients $\boldsymbol{\beta}_k$ measure how changes in topic shares affect the log-odds of observing outcome k (*cut* or *hike*) relative to the base category (*hold*). β_{k0} is the intercept of the equation, i.e. the baseline log-odds of a cut or hike versus keeping rates the same, when all independent variables are set to zero.

The selected independent variables of the multinomial logit regression are "*Inflation and Prices*", "*ECB Strategy and Price Stability", "Monetary Policy Operations & Liquidity",* and *"Unconventional Monetary Policy (QE & Forward Guidance)".* A dummy variable was also introduced to distinguish the period after 2015. The regression results (Appendix 6) show statistically significant associations for some variables. It is important to note, first of all, that a topic such as "*Inflation and Prices"* is strongly associated with non-neutral decisions, i.e., that the centrality of this issue in discussions increases the probability of both a rate hike ($coef. 19.81, p = 0.001$) and a rate cut ($coef. 13.16, p = 0.010$) relative to hold. The topic of Monetary Policy Operations and Liquidity is also positively and significantly associated with the probability of a rate cut ($coef. 14.53, p = 0.016$), while it is not significant for hikes, which confirms the importance of market operations during monetary easing. On the other hand, the theme of "*Unconventional Monetary Policy (QE & Forward Guidance)"* is negatively and significantly associated with the probability of a rate cut ($coef. = -36.71$, $p = 0.013$), consistent with the function of these extraordinary measures in maintaining accommodative conditions that may reduce the need for additional cuts relative to holding. No statistically significant evidence emerges for "*ECB Strategy and Price Stability*", nor for the post-2015 dummy variable. In conclusion, the model (McFadden's Pseudo-$R^2 \approx 0.129$) suggests that the rhetoric of the European Central Bank in its statements and speeches not only accompanies, but in some cases may help to anticipate interest rate decisions, with particular emphasis on the role of inflation as a central driver of monetary policy, as might be expected a priori.

The analysis continued with the implementation of PCA: Figure 4.3 shows the cumulative explained variance of PCA, where the first 11 components explain about 70% of the total variance, a threshold commonly adopted in applied econometrics as a criterion for dimensionality reduction. In this way, it was possible to reduce the complexity of the model without excessive loss of information.

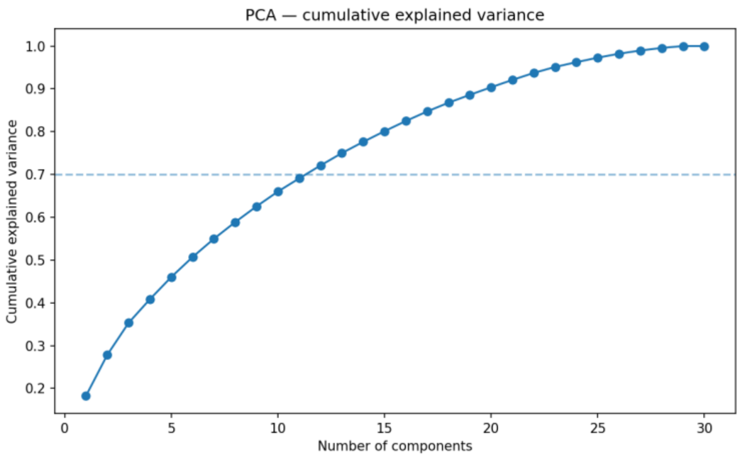


Figure 4.3 Cumulative explained variance

Appendix 7 reports the loadings of the 30 topics extracted with LDA on the first 11 principal components. The interpretation of the principal components derives from the analysis of the loadings, that is, the coefficients that connect each topic to each component. For a component PC_j:

$$PC_j = \sum_{k=1}^{K} w_{jk} x_k,$$

Where x_k is the share of topic k, w_{jk} is the loading of topic k and K is the total number of topics. In practice, loadings indicate how much a topic helps to define a certain component. If a component presents multiple topics with high loading and of the same sign, it can be interpreted as a latent factor that summarises those themes, highlighting that they tend to move together. If, however, some topics have positive loading and others negative, the component is interpreted as: high values of the component indicate a greater weight of the topics associated with positive loading, while low values reflect a prevalence of the topics with negative loading.

These synthetic components are then used as independent variables in a new model to include all of the European Central Bank's topics in our analysis, and the results (Appendix 8) indicate that certain components are significantly associated with European monetary policy decisions. The results show some significant patterns.

- PC6 (Economic research and household consumption vs. regulation and payments) has a negative and statistically significant coefficient for the cut vs. hold comparison: this implies that a greater emphasis on analytical and consumption issues reduces the likelihood of a rate cut, signalling that these topics are more associated with periods of stability.

- PC2 (inflation and climate risk vs. financial and fiscal stability) is positive and statistically significant for hike vs. hold: more talk of inflation and climate increases the likelihood of an increase in interest rates, consistent with the idea that the ECB reacts to inflationary pressures or risk factors. It is important to note that LDA does not distinguish between discussion of "low" or "high" inflation. Consequently, not all talks about inflation increase the probability of a rate cut relative to holding the same rate, and the broader context should be taken into account.
- PC3 (European integration and residual discourse vs. stability and global outlook) is negative and statistically significant: emphasis on stability and the international context tends to reduce the likelihood of a hike.
- PC4 (formal/protocol discourse vs. substantive policies) is positive and statistically significant: greater emphasis on formal and protocol language is associated with a higher probability of a hike, reflecting a more cautious decision-making environment.
- Finally, PC10 (European integration vs. consumption and payments) shows a marginally significant negative coefficient: references to micro issues reduce the propensity to hike.

Overall, the results confirm that the topics emerging from ECB speeches are not neutral but rather have predictive content regarding monetary policy choices. The model explains a significant portion of the variability (McFadden's Pseudo $R^2 \approx 0.14$), albeit with moderate explanatory power, highlighting the complexity of the decision-making mechanisms and the difficulty of capturing them solely through the topics of the speeches.

4.3 Discussion

The multinomial regressions demonstrate that certain thematic topics aggregated using Principal Component Analysis are significantly correlated with interest rate decisions. This is in line with previous studies that analysed ECB communication using sentiment indicators, but it extends their scope by applying a systematic topic modelling approach over the entire history of speeches. Topics such as inflation and price-related issues generally have a positive association with the probability of rate changes. The data show that they are strongly linked to both interest rate directions: cuts, suggesting that greater emphasis on these topics coincides with periods when the ECB adopts more expansionary measures, perhaps to offset deflationary pressures, and hikes, probably in periods of high inflation. On the other hand, components related to financial stability and sovereign debt seem to reduce the likelihood of rate hikes. This reflects greater concerns about the risks to stability during tightening phases. These results should be interpreted with caution: the predictive power of the model remains limited (McFadden's Pseudo $R^2 \approx 0.14$), and inference does not imply causality. Moreover, the analysis does not incorporate crucial macroeconomic variables, relying solely on textual data; therefore, the model should be considered complementary. However, the evidence of the analysis confirms that speeches serve as an important channel for understanding the Governing Council's priorities, partly anticipating the direction of decisions. Looking ahead, communication appears to be not only a tool for transparency but also a substantial element of monetary policy strategy. A possible future development could be to compare the language of the ECB with that of other central banks, such as the Federal Reserve or the Bank of England, or to extend the analysis to new areas, for example, climate policies, financial stability reports or official communications on digital currency. In this way, a more comprehensive view of the role that communication plays in the formation of market expectations across different institutional contexts would be obtained. Ultimately, the case study confirms the value of textual analysis as a tool for interpreting and anticipating central bank moves.

5 CONCLUSION

This thesis carried out the dual objective of essentially clarifying the theoretical framework of the Latent Dirichlet Allocation model in its fully Bayesian form, and of empirically showing how this tool can be easily applied to a large corpus of data. The case study also showed how the model allows for the analysis of the communication of the European Central Bank. On the theoretical level, we discussed the model, its limitations and benefits, the role of priors and the main approaches of inference, highlighting the trade-offs between accuracy, scalability, and transparency. On an empirical level, the analysis of the ECB texts confirms that the topics extracted follow and sometimes anticipate decisions on interest rate changes, from which significant but moderate associations emerge. This work has therefore tried to contribute both on a methodological level, to develop a clear and reusable procedure for complex institutional corpora, and on an empirical level. Future developments should include dynamic topic models, systematic comparisons with other central banks and high-frequency event studies.

APPENDIX

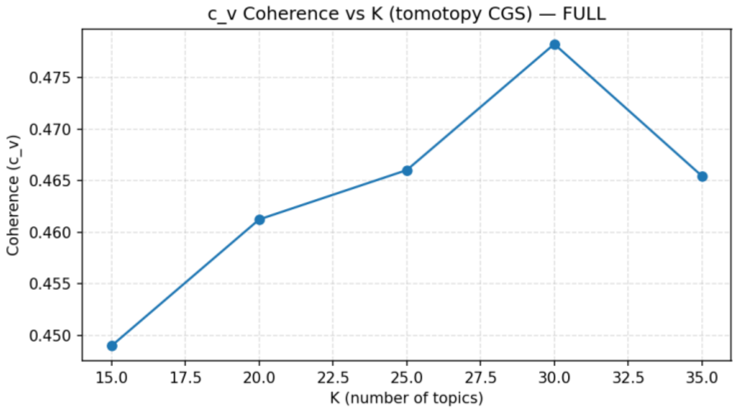
Appendix 1: Table of topics and the respective first 4 top-words

K	Words	Topic Name
1	risk, climate, climate_change, investment, transition, climate_relate, green, economy, impact, energy, finance, financial	Climate and Risk
2	credit, loan, non, capital, firm, lending, sector, finance, asset, cost, banking, large	Credit and Lending
3	banknote, banknote_coin, coin, cash_changeover, national, public, cash, currency, country, changeover, use, citizen	Cash and Banknotes
4	country, economic, member_state, exchange_rate, integration, convergence, union, process, currency, inflation, price, economy	European Integration and Convergence
5	market, financial, integration, system, cross_border, development, single, country, capital, service, economic, national	Fiscal Policy
6	market, union, banking_union, national, financial, capital, mechanism, resolution, single, country, supervision, level	Banking Union and Supervision
7	currency, international, exchange_rate, market, use, role, dollar, country, economic, reserve, financial, share	Currency and Exchange Rates
8	digital, payment, money, use, cash, financial, technology, system, transaction, risk, service, crypto_asset	Digital Money and Fintech
9	payment, sepa, service, market, retail_payment, use, national, payment_system, infrastructure, scheme, card, customer	Payment Systems and Infrastructure
10	economic, political, union, national, institution, single, common, people, integration, market, trust, citizen	European Integration and Institutions
11	inflation, price, shock, wage, inflation_expectation, rise, change, firm, energy, demand, cost, effect	Inflation and Prices
12	financial, market, risk, asset, stability, price, system, credit, institution, liquidity, investor, global	Financial Stability and Risk
13	such, other, however, particular, level, important, issue, first, provide, result, development, well	Communication and Rhetoric

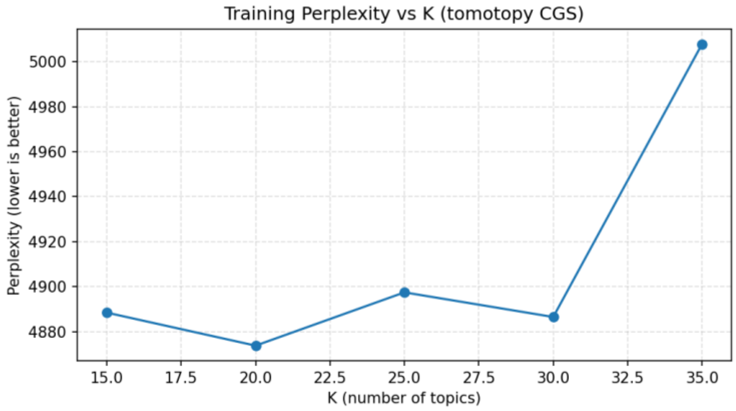
14	price_stability, strategy, inflation, economic, stability, price, decision, development, medium_term, interest_rate, govern_council, objective	ECB Strategy and Price Stability
15	fiscal, country, economic, government, reform, stability, debt, crisis, rule, financial, national, union	Fiscal and Sovereign Debt Policy
16	model, economic, economy, inflation, effect, price, money, analysis, research, work_paper, evidence, shock	Economic Research & Modelling
17	global, economy, international, country, trade, financial, globalisation, world, emerge, domestic, economic, market	Economic Outlook & Global Environment
18	interest_rate, low, household, rate, income, real, country, saving, asset, effect, investment, interest	Household Consumption and Savings
19	crisis, economy, financial, measure, support, take, challenge, first, recovery, condition, face, such	Crisis Management & Recovery
20	growth, economy, economic, productivity, market, reform, rate, sector, structural_reform, employment, country, firm	Structural Reforms & Competitiveness
21	financial, risk, supervisory, institution, supervisor, regulation, framework, supervision, authority, banking, level, standard	Prudential Supervision and Regulation
22	work, many, first, here, programme, well, very, thank, world, people, know, challenge	Formal and Protocol Discourse
23	market, liquidity, money, operation, rate, credit, measure, financial, collateral, interest_rate, provide, funding	Monetary Policy Operations & Liquidity
24	growth, economic, inflation, rate, development, outlook, expect, support, level, recovery, economy, measure	Economic Outlook
25	rate, interest_rate, market, asset_purchase, inflation, purchase, bond, effect, forward_guidance, instrument, programme, negative	Unconventional Monetary Policy (QE & Forward Guidance)
26	statistic, datum, financial, information, statistical, analysis, account, national, economic, use, quality, development	Statistics and Data
27	take, now, very, question, other, way, just, country, change, know, then, come	Generic Discourse / Residual

28	national, escb, decision, treaty, independence, system, govern_council, stability, institution, objective, task, member_state	ECB Governance and Independence
29	chart, rate, pandemic, model, base, late_observation, shock, source, note, impact, measure, estimate	Pandemic and PEPP
30	financial, stability, system, risk, crisis, macro_prudential, macroprudential, systemic_risk, sector, institution, framework, esrb	Financial Stability and Macroprudential Policy

Appendix 2: Graph of the coherence values for different number of topics



Appendix 3: Graph of the perplexity values for different number of topics



Appendix 4: Table of perplexity and coherence values for different number of topics and alpha=0.1

K	Perplexity	Coherence
15	4888.465559751200	0.44901980286116100
20	4873.784204112440	0.4612509022740650
25	4897.469317666590	0.46602922640382900
30	4886.520951195050	0.4782402814882010
35	5007.840297354600	0.4654686680059010

Appendix 5: Tables of coherence and perplexity values for different number of topics and values of alpha

Coherence

Alpha	0.3	0.5	K/50
K			
15	0.4609722809267430	0.4392856671917440	0.4609722809267430
20	0.4558804991838080	0.4644101261303670	0.4577550787930190
25	0.4667481387943690	0.4534566227955990	0.4534566227955990
30	0.4640128851780180	0.4688983512391490	0.4724652021619750
35	0.4855484552265500	0.4770180666055050	0.4766458320241920

Perplexity

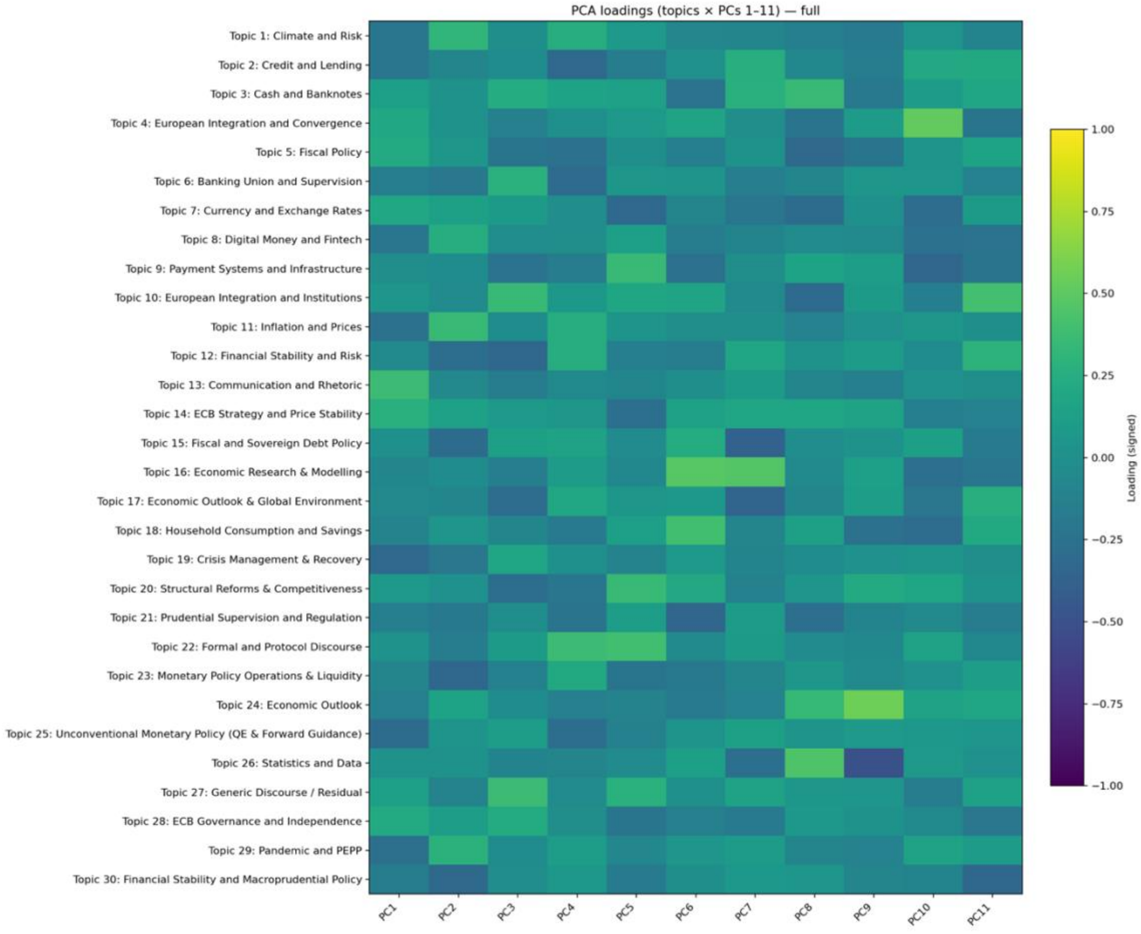
Alpha	0.3	0.5	K/50
K			
15	4816.048447608330	4887.207489611460	4816.048447608330
20	5008.402900375610	4801.453820813710	5082.840796649560
25	5025.464506630650	5055.215543981760	5055.215543981760
30	4888.788670838380	4983.096400605150	5069.689074724830
35	5039.049567294930	5011.757435950010	5061.3484421624600

Appendix 6: Regression results of four relevant topics. Baseline ‘hold’.

Observations = 304	Pseudo R² = 0.129		
Variable	**Coefficient (SE)**	**P>|z|**	**95% CI**
Cut vs Hold			
Intercept	-2.555 (0.467)***	0.000	[-3.470, -1.640]
Inflation and Prices	13.164 (5.119)**	0.01	[3.130, 23.198]
ECB Strategy and Price Stability	0.731 (3.220)	0.82	[-5.579, 7.041]
Monetary Policy Operations & Liquidity	14.533 (6.044)**	0.016	[2.687, 26.380]
Unconventional Monetary Policy (QE & Forward Guidance)	-36.718 (14.838)**	0.013	[-65.801, -7.635]
post2015	1.009 (0.870)	0.246	[-0.696, 2.714]
Hike vs Hold			
Intercept	-2.163 (0.439)***	0.000	[-3.024, -1.302]
Inflation and Prices	19.811 (5.765)***	0.001	[8.513, 31.110]
ECB Strategy and Price Stability	1.636 (2.865)	0.568	[-3.978, 7.251]
Monetary Policy Operations & Liquidity	-21.453 (13.542)	0.113	[-47.994, 5.088]
Unconventional Monetary Policy (QE & Forward Guidance)	-15.526 (14.200)	0.274	[-43.358, 12.306]
post2015	-0.725 (1.032)	0.482	[-2.748, 1.297]

*** p<0.01, ** p<0.05, * p<0.1

Appendix 7: PCA Loadings



Appendix 8: Regression results for the PCA analysis.

Observations = 304	Pseudo R² = 0.1358		
Variable	**Coefficient (SE)**	**P>|z|**	**95% CI**
Cut vs Hold			
Intercept	-2.049 (0.345)***	0.000	[-2.725, -1.373]
PC1	-0.196 (0.168)	0.243	[-0.526, 0.133]
PC2	0.188 (0.144)	0.192	[-0.095, 0.471]
PC3	0.022 (0.139)	0.876	[-0.250, 0.293]
PC4	0.256 (0.152)*	0.093	[-0.043, 0.554]
PC5	-0.057 (0.157)	0.716	[-0.365, 0.251]
PC6	-0.367 (0.180)**	0.042	[-0.720, -0.013]

PC7	-0.028 (0.183)	0.877	[-0.387, 0.330]
PC8	0.122 (0.176)	0.488	[-0.222, 0.466]
PC9	-0.073 (0.175)	0.678	[-0.416, 0.271]
PC10	-0.130 (0.224)	0.563	[-0.569, 0.309]
PC11	0.208 (0.207)	0.313	[-0.197, 0.613]
post2015	-0.506 (1.011)	0.617	[-2.488, 1.476]
Hike vs Hold			
Intercept	-2.327 (0.533)***	0.000	[-3.372, -1.281]
PC1	-0.213 (0.256)	0.404	[-0.715, 0.288]
PC2	0.778 (0.223)***	0.000	[0.341, 1.215]
PC3	-0.428 (0.191)**	0.025	[-0.801, -0.054]
PC4	0.515 (0.232)**	0.026	[0.060, 0.969]
PC5	0.066 (0.168)	0.695	[-0.264, 0.396]
PC6	-0.034 (0.179)	0.848	[-0.385, 0.317]
PC7	-0.079 (0.197)	0.688	[-0.466, 0.308]
PC8	0.139 (0.216)	0.522	[-0.285, 0.562]
PC9	0.111 (0.204)	0.586	[-0.288, 0.511]
PC10	-0.365 (0.203)*	0.072	[-0.763, 0.033]
PC11	-0.228 (0.245)	0.351	[-0.707, 0.252]
post2015	-2.208 (1.791)	0.218	[-5.719, 1.304]

*** p<0.01, ** p<0.05, * p<0.1

Appendix 9: GitHub repository of the source code used for the analysis https://github.com/marcofrova/lda_thesis_code

REFERENCES:

1. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.
2. Chandra, N. K., Canale, A., & Dunson, D. B. (2023). Escaping the curse of dimensionality in Bayesian model-based clustering. Journal of Machine Learning Research, 24(73), 1–33.
3. Bishop, C. M. (2006). Pattern recognition and machine learning. New York: Springer.
4. Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101(Suppl. 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101
5. Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 100–108). Association for Computational Linguistics.
6. Blei, D. M. (2016). Mixed-membership models (and an introduction to variational inference). Columbia University, Department of Computer Science.
7. Teh, Y. W., Newman, D., & Welling, M. (2007). A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In Advances in Neural Information Processing Systems, 19 (pp. 1353–1360).
8. Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2009). On smoothing and inference for topic models. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (pp. 27–34). AUAI Press. https://doi.org/10.48550/arXiv.1205.2662
9. Heinrich, G. (2005). Parameter estimation for text analysis [Technical report]. Retrieved from http://www.arbylon.net/publications/text-est.pdf
10. Dickey, J. M. (1983). Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. Journal of the American Statistical Association, 78(383), 628–637. http://dx.doi.org/10.1080/01621459.1983.10478022
11. Wallach, H. M., Mimno, D. M., & McCallum, A. (2009). Rethinking LDA: Why priors matter. In Advances in Neural Information Processing Systems, 21 (pp. 1973–1981).
12. Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 262–272).

13. Yurochkin, M., & Nguyen, X. (2016). Geometric Dirichlet Means algorithm for topic inference. arXiv preprint arXiv:1610.09034. https://arxiv.org/abs/1610.09034
14. Aldebakel, D., Cao, Y., Limon, A., & Zhang, R. (n.d.). Exploration of variational inference and Monte Carlo Markov Chain models for Latent Dirichlet Allocation of Wikipedia corpus [Project report, DSC180B A06, UC San Diego]. UCSD.
15. Zankadi, H., Idrissi, A., Najima, D., & Hilal, I. (2022). Identifying learners' topical interests from social media content to enrich their course preferences in MOOCs using topic modelling and NLP techniques. Education and Information Technologies, 28(5), 5567–5584. https://doi.org/10.1007/s10639-022-11373-1
16. Lopea, J. (2017). Speeding up calibration of Latent Dirichlet Allocation model to improve topic analysis in software engineering. https://doi.org/10.32920/ryerson.14665455.v1
17. Aletras, N., & Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In Proceedings of the 2013 Workshop on Continuous Vector Space Models and their Compositionality (pp. 13–22).
18. Gkioulekas, I., & Zickler, T. (2011). Dimensionality reduction using the sparse linear model. In Advances in Neural Information Processing Systems, 24 (pp. 271–279).