# Topic Models in R

**Bettina Grün**
Wirtschaftsuniversität Wien

**Kurt Hornik**
Wirtschaftsuniversität Wien

**Abstract**

Topic models are a popular method for modeling the term frequency occurrences in documents. The fitted model allows to better estimate the similarity between documents as well as between a set of specified keywords using an additional layer of latent variables which are referred to as topics. The R package **topicmodels** provides basic infrastructure for fitting topic models based on data structures introduced in the text mining package **tm**. The package includes interfaces to two algorithms for fitting topic models provided by David M.˜Blei in C and one algorithm by Xuan-Hieu Phan in C++.

*Keywords*:˜text analysis, topic model, variational inference, R.

## 1. Introduction

Topic models are generative models which specify a probabilistic model for the term frequency occurrences for documents in a given corpus. This signifies that topic models are bag-of-word models, i.e., it is assumed that the information in which order the words occur in a document is negligible. This assumption is also referred to as *exchangeability* assumption for the words in a document (Blei, Ng, and Jordan 2003). In order to model dependencies between words, i.e., to allow related words to occur more likely together in a document, topics are introduced as latent variables. The content of each document is based on certain topics and these underlying topics induce a certain word distribution for the document. Each document therefore has its own topic distribution.

The Latent Dirichlet Allocation (LDA; Blei *et˜al.* 2003) model is the basic model where topics are assumed to be uncorrelated. The Correlated Topics Model (CTM; Blei and Lafferty 2007) is an extension of the LDA model where correlations between topics are allowed. An introduction to topic models is given in Steyvers and Griffiths (2007) and Blei and Lafferty (2009). C code for fitting LDA (`http://www.cs.princeton.edu/~blei/lda-c`) and CTM (`http://www.cs.princeton.edu/~blei/ctm-c`) is available under the GPL from the first author of the papers introducing these models. For estimation the variational EM algorithm is used. For each model a standalone program is provided which requires as input a data format where each document is succinctly represented as a sparse vector of word counts.

Other implementations of topic models—especially of the LDA model—are available. GibbsLDA++ (Phan, Nguyen, and Horiguchi 2008, `http://gibbslda.sourceforge.net/`) is available under the GPL . It is a C/C++ implementation of LDA using Gibbs Sampling for parameter estimation and inference. The Matlab Topic Modeling Toolbox 1.3.2 (Griffiths and Steyvers 2004, `http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm`) is free

for scientific use. A license must be obtained from the authors to use it for commercial purposes. It also uses Gibbs sampling for parameter estimation. The standalone program lda (Mochihashi 2004, http://chasen.org/~daiti-m/dist/lda/) provides standard variational Bayes estimation. The authors of the lda package claim that according to some experiments their package runs about 4 to 10 times faster than Blei's code. MALLET (McCallum 2002, http://mallet.cs.umass.edu) is released under the CPL and is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling using LDA, information extraction, and other machine learning applications to text.

This R package provides an interface to the code for fitting a LDA topic model and a CTM with variational EM written by Blei and to the code for fitting a LDA topic model with Gibbs sampling written by Phan.

## 2. Topic models specification and estimation

For both models—LDA and CTM—the number of topics $k$ has to be fixed a-priori. LDA and CTM assume the following generative process for a document $\mathbf{w}$ containing $N$ words of a corpus $\mathbf{D}$:

**Step 1:** The proportions $\boldsymbol{\theta}$ of the topic distribution for the document $\mathbf{w}$ are determined.

**LDA:** Draw $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha})$.

**CTM:** Draw $\boldsymbol{\eta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and determine $\boldsymbol{\theta}$ by

$$\boldsymbol{\theta} = \frac{\exp\{\boldsymbol{\eta}\}}{\sum_i \exp\{\eta_i\}}.$$

The dimensions of the parameters correspond to the number of topics $k$.

**Step 2:** For each of the $N$ words $w_n$:

(a) Choose a topic $z_n \sim \text{Multinomial}(\boldsymbol{\theta})$.

(b) Choose a word $w_n$ from a multinomial probability distribution conditioned on the topic $z_n$: $p(w_n|z_n, \boldsymbol{\beta})$.

For maximum likelihood estimation of the LDA model the (marginal) log likelihood of the data is maximized with respect to the model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$:

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{\mathbf{w} \in \mathbf{D}} \log\left(p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})\right).$$

For the CTM model the model parameters are given by $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}$.

The quantities $p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ for the LDA model and $p(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$ for the CTM cannot be computed tractably. Hence, a *variational EM* procedure is used for estimation. The missing data in the EM algorithm are the hidden variables $\boldsymbol{\theta}$ and $\mathbf{z}$ for LDA and $\boldsymbol{\eta}$ and $\mathbf{z}$ for CTM. In the variational EM algorithm variational parameters are used to approximate the posterior of the hidden variables given the model parameters and an observed document. For the LDA model the variational parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ for a given document $\boldsymbol{w}$ are determined by

$$(\boldsymbol{\gamma}^*, \boldsymbol{\phi}^*) = \arg\min_{(\boldsymbol{\gamma}, \boldsymbol{\phi})} \mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})||p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})).$$

$\mathrm{D_{KL}}$ denotes the Kullback-Leibler (KL) divergence. For the variational distribution it holds that

$$q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\phi}) = q_1(\boldsymbol{\theta}|\boldsymbol{\gamma}) \prod_{n=1}^{N} q_2(z_n|\boldsymbol{\phi}_n).$$

$q_1()$ is a Dirichlet distribution with parameters $\boldsymbol{\gamma}$ and $q_2()$ is a multinomial distribution with multinomial parameters $\boldsymbol{\phi}_n$.

Analogously for the CTM the variational parameters are determined by

$$(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*, \boldsymbol{\phi}^*) = \arg\min_{(\boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\phi})} \mathrm{D_{KL}}(q(\boldsymbol{\eta}, \boldsymbol{z}|\boldsymbol{\lambda}, \boldsymbol{\nu}^2, \boldsymbol{\phi})||p(\boldsymbol{\eta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})).$$

Since the variational parameters are fit separately for each document the variational covariance matrix can be assumed to be diagonal, i.e., $\boldsymbol{\nu}^2$ consists only of the diagonal elements. The variational distribution is given by

$$q(\boldsymbol{\eta}, \boldsymbol{z}|\boldsymbol{\lambda}, \boldsymbol{\nu}^2, \boldsymbol{\phi}) = \prod_{i=1}^{k} q_1(\eta_i|\lambda_i, \nu_i^2) \prod_{n=1}^{N} q_2(z_n|\boldsymbol{\phi}_n).$$

For CTM $q_1()$ is a univariate Gaussian distribution with mean $\lambda_i$ and variance $\nu_i^2$ and $q_2()$ denotes again a multinomial distribution with multinomial parameters $\boldsymbol{\phi}_n$.

For the LDA model it can be shown with the following equality that the variational parameters result in a lower bound for the log likelihood:

$$\log p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = L(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) + \mathrm{D_{KL}}(q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})||p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})).$$

Maximizing the lower bound $L(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ is equivalent to minimizing the KL divergence between the variational posterior probability and the true posterior probability.

For estimation the following steps are repeated until convergence of the lower bound of the log likelihood:

**E-step:** For each document find the optimizing values of its variational parameters $\{\boldsymbol{\gamma}, \boldsymbol{\phi}\}$ for the LDA model and $\{\boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\phi}\}$ for the CTM.

**M-step:** Maximize the resulting lower bound on the log likelihood with respect to the model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for the LDA model and $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}$ for the CTM.

Under the assumption that the variational posterior probability is a good approximation of the true posterior probability it can be used for inference (e.g., for new documents).

For Gibbs sampling in the LDA model draws from the posterior distribution $p(\boldsymbol{z}|\boldsymbol{w})$ are obtained by sampling from

$$p(z_i = j|\boldsymbol{w}, \boldsymbol{z}_{-i}) \propto \frac{n_{-i,j}^{(w_i)} + \delta}{n_{-i,j}^{(.)} + V\delta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,.}^{(d_i)} + k\alpha}.$$

(cf.˜ Griffiths and Steyvers 2004; Phan *et˜al.* 2008).

$\boldsymbol{z}_{-i}$ denotes the current topic memberships of all words without the $i^{\text{th}}$ word and $n_{-i,j}^{(w_i)}$ the number of words $w_i$ currently assigned to topic $j$ without the $i^{\text{th}}$ word. $V$ is the size of the vocabulary. . implies that summation over this index has been made. $\delta$ denotes the parameter of the prior distribution for the word distribution of the topics. Please note that in this model formulation $\alpha$ also is a parameter of a prior distribution, in this case for the topic distribution of the documents. The predictive distributions of the parameters $\theta$ and $\beta$ given $\boldsymbol{w}$ and $\boldsymbol{z}$ are given by

$$\hat{\theta}_j^{(w)} = \frac{n_j^{(w)} + \delta}{n_j^{(\cdot)} + V\delta},$$
$$\hat{\beta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_j^{(\cdot)} + k\alpha}.$$

The data preprocessing step involves selecting a suitable vocabulary, i.e., the columns of the document-term matrix. The most frequent terms or those terms with the highest term-frequency inverse document frequency (TFIDF) scores might for example be selected (Blei and Lafferty 2009). With respect to model selection an appropriate number of topics cross-validation of the predictive likelihood might be employed.

## 3. Example

The application of the package **topicmodels** is demonstrated on the abstract collection of the Journal of Statistical Software (up to 2010-01-13).

The JSS data is available as a list matrix in the package **corpus.JSS.papers** which can be installed and loaded by

```
> install.packages("corpus.JSS.papers",
+                  repos = "http://datacube.wu.ac.at/", type = "source")
> data("JSS_papers", package = "corpus.JSS.papers")
```

Alternatively package **OAIHarvester** can be used to harvest the meta information of the papers published at JSS from its web page.

```
> require("OAIHarvester")
> x <- oaih_list_records("http://www.jstatsoft.org/oai")
> JSS_papers <- oaih_transform(x[, "metadata"])
> JSS_papers <- JSS_papers[order(as.Date(unlist(JSS_papers[, "date"]))), ]
> JSS_papers <- JSS_papers[grep("Abstract:", JSS_papers[, "description"]), ]
> JSS_papers[, "description"] <- sub(".*\nAbstract:\n", "",
+                                    unlist(JSS_papers[, "description"]))
```

The data set contains 326 documents. Before analysis we transform it to a `"Corpus"` using package **tm** (Feinerer, Hornik, and Meyer 2008). HTML markup in the abstracts for greek letters, subscripting, etc., is removed using package **XML**.

```
> set.seed(1102)
> library("topicmodels")
> library("XML")
> remove_HTML_markup <-
+ function(s) {
+     doc <- htmlTreeParse(s, asText = TRUE, trim = FALSE)
+     iconv(xmlValue(xmlRoot(doc)), "", "UTF-8")
+ }
> corpus <- Corpus(VectorSource(sapply(JSS_papers[, "description"],
+                                      remove_HTML_markup)))
```

LDA and CTM are bag-of-words models, i.e., a term-document matrix indicating which terms appear how often in which document is sufficient as input. The data set is transformed to a matrix with the documents in the rows and terms in the columns. Each entry indicates how often each term occurred in the document. Function `DocumentTermMatrix()` from package **tm** can be used and the terms can be stemmed and the stopwords and numbers removed using the control argument.

```
> dtm <- DocumentTermMatrix(corpus,
+    control = list(stemming = TRUE, stopwords = TRUE, minWordLength = 3,
+      removeNumbers = TRUE))
```

Infrequent terms can be removed using `removeSparseTerms()`. We only include terms which are present in at least 1% of the documents.

```
> dtm <- removeSparseTerms(dtm, 0.99)
> dim(dtm)

[1] 326 870
```

In the following the data set is divided into a training and a testing data set. The models can be fitted using functions `LDA()` and `CTM()`. The number of topics is set to $k = 10$ in both cases. An initial value for $\alpha$ has to be provided for LDA. By default the parameter $\alpha$ is estimated. Alternatively it can also be fixed by setting as argument `control = list(alpha = "fixed")`. Please note that in the current implementation it holds that $\alpha_k \equiv \alpha$ for all topics.

```
> jss_LDA <- LDA(dtm[1:250,], control = list(alpha = 0.1), k = 10)

> jss_CTM <- CTM(dtm[1:250,], k = 10)
```

The term distribution for each topic as well as the predictive distribution of topics for a document can be obtained with `posterior()`. `posterior()` also has an argument `newdata` in order to be able to determine the predictive distribution of topics for new documents.

```
> post <- posterior(jss_LDA, newdata = dtm[-c(1:250),])
> round(post$topics[1:5,], digits = 2)
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,] 0.12 0.15 0.03 0.26 0.44 0.00 0.00    0 0.00  0.00
[2,] 0.00 0.00 0.16 0.00 0.00 0.27 0.00    0 0.40  0.16
[3,] 0.00 0.18 0.13 0.00 0.00 0.00 0.00    0 0.53  0.16
[4,] 0.06 0.00 0.28 0.00 0.30 0.00 0.05    0 0.31  0.00
[5,] 0.00 0.00 0.00 0.00 0.80 0.00 0.00    0 0.19  0.00
```

The five most frequent words for each topic can obtained by:

```
> get_terms(jss_LDA, 5)
```

```
     Topic 1  Topic 2  Topic 3     Topic 4     Topic 5    Topic 6     Topic 7
[1,] "test"   "model"  "model"     "statist"   "model"    "statist"   "program"
[2,] "method" "estim"  "algorithm" "distribut" "estim"    "data"      "comput"
[3,] "normal" "data"   "generat"   "program"   "data"     "softwar"   "statist"
[4,] "simul"  "method" "data"      "procedur"  "packag"   "packag"    "paper"
[5,] "propos" "packag" "packag"    "sampl"     "function" "implement" "calcul"
     Topic 8    Topic 9  Topic 10
[1,] "packag"   "packag" "model"
[2,] "data"     "model"  "packag"
[3,] "user"     "data"   "tool"
[4,] "function" "provid" "estim"
[5,] "graphic"  "method" "includ"
```

## 4. Summary and further research

The package **topicmodels** provides functionality for fitting topic models in R. It builds on and complements functionality for text mining already provided by package **tm**. It also amends package **lsa** (Wild 2009) available on CRAN which allows to perform latent semantic analysis. Topic models are an alternative method to analyze document-term matrices and have the advantage that they are based on a probabilistic generative model. The available tools for text mining in R are hence extended to contain these state-of-the-art techniques for analyzing document-term matrices and now conveniently allow to fit and compare different methods in order to develop a useful model. On CRAN also package **lda** (Chang 2009) is available which provides collapsed Gibbs sampling methods for topic models.

## References

Blei DM, Lafferty JD (2007). "A Correlated Topic Model of Science." *The Annals of Applied Statistics*, **1**(1), 17–35.

Blei DM, Lafferty JD (2009). *Topic Models*. Taylor and Francis.

Blei DM, Ng AY, Jordan MI (2003). "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, **3**, 993–1022.

Chang J (2009). *lda: Collapsed Gibbs sampling methods for topic models*. R package version 1.1, URL http://CRAN.R-project.org/package=lda.

Feinerer I, Hornik K, Meyer D (2008). "Text Mining Infrastructure in R." *Journal of Statistical Software*, **25**(5). URL http://www.jstatsoft.org/v25/i05/.

Griffiths TL, Steyvers M (2004). "Finding Scientific Topics." *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 5228–5235.

McCallum AK (2002). *MALLET: A Machine Learning for Language Toolkit*. URL http://mallet.cs.umass.edu.

Mochihashi D (2004). "A Note on a Variational Bayes Derivation of Full Bayesian Latent Dirichlet Allocation." URL http://chasen.org/~daiti-m/paper/lda-fullvb.pdf.

Phan XH, Nguyen LM, Horiguchi S (2008). "Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections." In "Proceedings of the 17th International World Wide Web Conference (WWW 2008)," pp. 91–100. Beijing, China.

Steyvers M, Griffiths T (2007). *Probabilistic Topic Models*. Lawrence Erlbaum Associates.

Wild F (2009). *lsa: Latent Semantic Analysis*. R package version 0.61.

## Affiliation:

Bettina Grün, Kurt Hornik
Institute for Statistics and Mathematics
WU Wirtschaftsuniversität Wien
Augasse 2–6
A-1090 Wien, Austria
E-mail: Bettina.Gruen@wu.ac.at, Kurt.Hornik@R-project.org
URL: http://statmath.wu.ac.at/~gruen/,
http://statmath.wu.ac.at/~hornik/