SCIENTIFIC PUBLICATIONS OF THE STATE UNIVERSITY OF NOVI PAZAR
Ser. A: Appl. Math. Inform. and Mech. vol. 15, (1) 2023, 1-19

Independent topical structures identification: Theoretical approach and proof-of-concept study

Marcus Spies

Abstract: This paper introduces a new approach to the identification of topical structures from text data using independent component analysis (ICA). The approach resembles current probabilistic topic modelling approaches in some respects, however introduces an axiomatic definition of topic structures as independent feature functions over a given vocabulary. In addition, the identification of such structures from data is decoupled from estimates of topical compositions of the training documents. Considerations motivating these choices are discussed, and a proof-of-concept study on a small corpus is presented to demonstrate feasibility and interpretative features of the approach. As computational approach to ICA, a method based on distance covariance is used.

Keywords: applied statistics, text mining

1 Introduction and Overview

Probabilistic topic modelling refers to a collection of unsupervised or semi-supervised machine learning techniques for identifying themes or topics in corpora of text documents. Two prominent and popular approaches to probabilistic topic modelling are LDA [4, 10] and STM [24, 23]. One basic underlying assumption of these techniques is that topics are probability distributions over terms (in the linguistic sense of word forms, while words occurring in documents are called tokens). A further key assumption is that topics as inferred from a statistical model can be combined into mixture distributions of topics to describe the content of single documents. The main usage perspective contains search or categorization applications that allow to retrieve documents by descriptors derived from topics rather than just by keyword search.

In the present paper, an approach to topic models is proposed that modifies the first of the above assumptions. Topics will be derived from *latent feature functions on the domain of terms in a corpus vocabulary.* The identification of these latent

Manuscript received December 15, 2022.; accepted July 15, 2023.
Author holds chair of Knowledge Management at LMU University of Munich

features builds on the method of independent component analysis (ICA, [15]) which is frequently used in signal processing due to its capability regarding blind source separation of signals.

Thus, the proposed identification procedure for topics will be based on the additional assumption that *different topics should represent independent latent features*. Intuitively, independence of topics correspond to non-overlapping semantic content which seems important for meaningful interpretation of topic mixtures in documents.

Moreover, as shown in a proof-of-concept study, topics as latent features can express semantic contrasts or similarity of terms. This can assist evaluation and support use cases needing identification of semantic profiles in test documents.

Regarding topic supported document retrieval, a predictive approach is proposed that is capable of identifying key latent topic features in given documents, but also allows test documents to have largely flat topic profiles. This relaxes the mixture distribution assumption in purely probabilistic approaches and offers flexibility for documents whose content does not match common characteristics found in the corpus data being analyzed.

In order to motivate the approach put forward in the present paper, the next section 2 will briefly summarize some properties of current probabilistic approaches that can cause problems on the inference or application side, in particular for small data sets. Section 3 presents basic properties of independent component analysis and identification methodologies as far as relevant. A proof-of-concept study in section 4 will conclude this paper.

One key result will be that the independent components found are characterized by long-tail empirical distribution functions. This is a marked difference to the prevailing pattern of short-tail topic-term distributions resulting from most of the known probabilistic approaches. Advantages of long-tail distributions in topic modelling include facilitation of identifiying characteristic semantics of identified topics and enhanced transferability of components inferred to documents beyond the training corpus.

2 Introducing topic structure independence – Problem statement and overview

Many approaches to probabilistic topic modelling, including LDA [4, 10] and STM [24, 23], rest on parametric Bayesian inference for a user-defined number of topics for inferring multinomial topic term as well as document topic distributions that maximize the probability of corpus data. LDA is built on prior Dirichlet distributions while STM uses logistic normal priors for the document topic distributions. The latter approach is shared also with CTM [3], to which STM adds the capability to process document metadata for better prediction of topic prevalence and also for localized predictions of variability in topic wording (content covariates). These predictions are made possible through additional hyperparameters govering document

topic prevalence variability and occurrence correlations. Further hyperparameters on the level of each topic for each vocabulary item are added to support varying topic wording depending on a discrete variable. Both LDA and CTM choices of prior distributions have been generalized to non-parametric Bayesian inference allowing for unlimited topic resources in the HDP [34] and DILN [21] approaches, which, respectively, use a hierarchical Dirichlet or infinite logistic normal stochastic process for the inference of topical structures.

A common property of Dirichlet and multivariate logistic normal distribution based topic models is the estimation of document-topic probabilities (where the number of topics is fixed in advance in LDA, CTM or STM, and virtually unlimited in HDP or DILN) and term probability vectors in overall *generative* models of documents viewed as grouped data in a collection (referred to as corpus in linguistics).

Empirical work focussing on a collection of small corpora shows that under commonly used Dirichlet hyperparameter settings posterior document-topic multinomials tend to approach singleton focussed probability mass functions (pmfs) in the course of Bayesian learning [26]. Moreover, posterior document-topic models exhibit smoothing of the original document term corpus probabilities [27]. In the case of small corpora and/or short documents, the smoothing effect is often substantial and thus may limit usability of a topic model in terms of practical interpretability and specificity. Finally, a model validation study in [27] shows that a key property deriving from Dirichlet priors, namely proportion neutrality [9], is hardly justifiable given real corpus data, while locally neutral elements in topics could be safely identified.

Motivated by these findings, the present paper proposes an alternative approach to identifying topic structures in text documents. The key problem addressed here is the topic structure correlation problem. More specifically, several approaches including CTM, DILN or STM [3, 21, 24] have been addressing topic *correlations on the level of topic mixing proportions* which explain topic cooccurrences on the document level. However, the question whether the mixing components (topic term distributions) themselves exhibit communalities has not explicitly been addressed.

In fact, our empirical work indicates that topic term distributions across topics, viewed by terms, tend to correlate [26]. In the case of LDA, this contributes to the observed tendency of generated topics towards focussing strongly on terms with high document frequency (occurring in many documents). From a user perspective, a consequence of these observations is that topics generated using the above approaches, from an interpretative point-of-view, may be difficult to distinguish or to assign to perceived document content in clear way. *Dependencies between the mixture components (topic term distributions)* may lead to a lack of clear semantics of identified probabilistic topics.

Therefore, an alternative topic term component identification procedure is proposed in in this paper that ensures independence of topic structures as components in a statistical topic model.

The approach established in the sequel is based on independent component analysis (ICA, [15]), which is widely used in various areas of data analytics and, in particular, signal processing. Starting, thus, from the postulate of stochastic independence of term distributions across topics, the goal is to find topics that cover *different semantic dimensions of the document content under study*. Stochastic independence on the topic level implies that mixtures of topics in a document should be interpretable in a meaningful way as the components of the mixtures are complementary.

However, some additional steps are needed to make an ICA model work in a way that can be compared to the outputs of a topic model in the sense described above. First, while ICA estimation procedures have a probabilistic justification, independent components as identified are basically scale functions on the domain of vocabulary words. The empirical cumulative distribution of each of these functions may have a different best fitting distribution type. So, the components identified are not forced to follow a common distributional model like in usual probabilistic topic models (where often multinomial distributions are used).

Moreover, an ICA approach to topic modelling needs an additional module for estimating or predicting the content of documents from training, testing or validation data. The reason for this is that the ICA based topic estimation procedure as proposed is decoupled from document topic assignments as provided by a joint estimation procedure in the approaches discussed above.

To address this issue, a linear predictive approach to estimating topical content of documents is proposed in section 4. It consists of using a sequence of regularized least-squares estimates of component weights in order to predict a document vector in the sense of multilinear regression. A sequential procedure using the lasso variable subset selection methodology [8] allows a model assessment without referring to the full set of assumptions on error term distributions and homoscedascity in linear regression.

Finally, a procedure is needed to assess meaning of topics as identified in a qualitative sense. For achieving this, an interpretation of feature functions in line with standard linguistic approaches building on [25] is proposed in section 4.3.2.

3 Independent component analysis

Independent component analysis is a data processing technique building on data matrix decomposition methods like singular value decomposition (SVD) or principal components (PCA), see [14], ch. 14.7. The key distinguishing feature of ICA is that the latent variables it identifies are stochastically independent instead of only being uncorrelated as after applying a PCA analysis [15]. This allows for applications in signal processing and other domains that require separation of unknown sources from a mixed signal. An everyday example is the separation of speech input from multiple speakers by the human auditory system.

More formally, given a $n \times p$ matrix X of n observations in p variables, the objective is to identify a random vector of independent latent variables $s_1, \ldots, s_p$ with n instances, arranged in a $p \times n$ matrix S, such that an orthogonal $p \times p$ mixing matrix A applied to these reproduces X,

$$X = S^T A \tag{1}$$

It should be noted that the requirement of stochastic independence refers to the p components in the model, not the n instances. In fact, in the common example of speech signal source separation, instances are sound signal properties with a temporal ordering exhibiting systematic dependencies reflecting natural language phonology etc.

In general, the dimensionality of the latent components matrix S is $n \times q$ with $q \leq p$. The case $q = p$ is useful in signal processing in order to recover the same number of independent sources as there are signals observed. However, in other domains, a smaller number of generating independent sources would be searched than the true or conjected number of observation sources. This also holds true in document / term processing, as the number of latent topical components in an unsupervised approach to topic identification is unknown, see section 4. ICA has no dimension reduction capability by itself. In usual applications, the data matrix X is supposed to come doubly centered and pre-whitened using any standard algorithm as discussed in [17], notably comprising PCA and thus also dimension reduction. As shown in [15], ch. 7, having a centered and whitened data matrix X implies that the ICA mixing matrix A is orthogonal.

The crucial point in estimating an ICA model given a centered and whitened data matrix X, then, is to choose an objective function describing statistical independence in order to solve for the independent components S and the mixing matrix A. Several different approaches have been pursued.

One key approach to identifying an ICA model, see [5] and [15], rests on the mutual information in random vectors composing S or X which is known to vanish iff the components are independent [6]. Moreover, for a given random vector s with non-singular covariance matrix, define *negentropy* as the difference of the differential entropy of a multivariate Gaussian variable with the same covariance matrix minus the differential entropy of s [15]. Negentropy is always non-negative by the maximum entropy distribution theorem [6], assuming throughout component random variables with finite first and second moments. Then, the mutual information for a candidate model containing S with orthogonal transformation matrix A can be expressed using the difference in negentropies of S vs its components plus a term depending on the model covariance matrix, see [5], eq. 2.14. More generally,[15], p. 223, show that maximizing the information theoretic departure from multivariate normals (called *non-gaussianity* in [15]) is equivalent to minimizing mutual information between components in S. Based on these principles, comparisons of observed random vectors Y and candidate model components S with orthogonal transfor-

mation matrix A allows to define a contrast function assessing the closeness of the candidate components to independence. [5] and [15] derive algorithmic approaches to solving for an ICA model using higher cumulants of the distributions involved. An efficient implementation using an online learning approach is derived in [16].

A different approach to ICA is discussed in [14], ch. 14.7, based on [13], using sampling product densities and generalized additive models. For the corresponding implementation in R, see [12]. [1] propose an ICA approach based on kernel methods and canonical correlation. In this paper, another approach is focused which is based on distance covariance [30].

3.1 Basic elements of distance covariance

The concept of energy distance became known to the data science community mainly due to an innovative approach to hierarchical clustering in [29] that defines a new clustering criterion combining within and between cluster distances. The combination is defined such that it has a precise and specific minimum for pointwise identical samples. This property addresses many shortcomings of other hierarchical clustering approaches. Since then, the more encompassing framework of enery statistics has been developed and has evolved substantially, see [30]. Within the framework, numerous innovative approaches to independence testing, goodness-of-fit statistics and related fields of data science have been contributed [30].

A key concept of energy statistics was introduced in [32] – distance covariance. Distance covariance is a dependence evaluation function generalizing the statistical concepts of covariance and correlation of random vectors. The important benefit of the generalization is distance covariance provides an equivalent characterization of stochastic independence as opposed to a one-way implication only as it holds between independence and zero covariance or correlation. For a comprehensive text on the theory, see [30].

More specifically, let X and Y being real-valued random vectors in $\mathbb{R}^p$ and $\mathbb{R}^q$ with finite absolute expectations, and let $(X^{(n)}, Y^{(n)})$ be an n-element *sample from the joint distribution* of (X, Y). Note that X and Y may be of different dimensionalities.

Székely, Rizzo and Bakirov [32] define a dependence measure with properties generalizing statistical covariance based on the characteristic functions $f_X()$ and $f_Y()$, where $f_X(\boldsymbol{s}) = E(\exp(i\ \langle \boldsymbol{s}, X\rangle)$, and the characteristic function over the joint space $\mathbb{R}^{p+q}$. Then, the distance covariance of X, Y is

$$\mathcal{V}^2(X,Y) = \int_{\mathcal{R}^{p+q}} |f_{X,Y}(s,t) - f_X(s)f_Y(t)|w(s,t)\mathrm{d}t\mathrm{d}s$$

([32], eq. 2.2), where $|f|^2$ denotes $f\overline{f}$, and $w(s,t)$ is a weight function chosen such that integrability, scale equivariance and rotation invariance conditions are met. In fact, while the relationship between independence and product of characteristic

functions is well known, the choice of the appropriate weighting function to guarantee these invariances is based on fundamental theoretical developments in energy statistics, see [32] and [30], appendix A, for a summary of the historical background. More specifically, the fundamental lemma established in [29] allows to deduce that the adequate weight function is composed of dimension specific constants divided by $\|s\|^{p+1}\|t\|^{q+1}$ (where $\|\cdot\|$ denotes Euclidean norm), for details, see [30], p. 193.

In addition, the limitation to RVs with finite absolute expectations can be dropped using another energy statistics concept, the α-coefficient bounded in the open interval $(0, 2)$. The generalization implies modifications to several components in the above formula, see [32], p. 2784. For univariate variables, $\alpha = 2$ reduces distance covariance or correlation to the usual covariance or correlation. As a side remark, it is important to note that distance covariance is *not* the covariance of distances in the literal sense, see [30], p.202.

From this definition, several options for calculating distance covariance sample statistics have been deduced (see [30], ch. 12). A surprisingly straightforward option is to use the entries of the doubly centered Euclidean distance matrices from $X^{(n)}$ ($Y^{(n)}$). Building on [32], theorem 1, [30] show that sample distance covariance may be expressed as sum of cross products of these entries, for details, see [30], ch. 12, and [32], p.2776. An important generalization of this estimate is the definition of an unbiased sample statistic from appropriately centered distance matrices (using the available degrees of freedom in the dominators instead of the sample size) together with an unbiased estimator of distance covariance, see [30]. This approach is also used in the ICA estimation algorithm based on distance covariance in [20], to which we return in the next section.

Moreover, evaluating distance covariances X and Y with itself, distance correlation with suitable properties can be defined on the basis of distance covariance in complete analogy to conventional statistics.

Essential uses of the theory are established in [33], where a statistical independence test based on distance covariance is defined. [31] extend the scope of the theory to random vectors with respect to stochastic processes and give several examples of resulting data analyses. More recently, stimulated by a theory on distance covariance in metric spaces [18], the relationship of energy statistics with kernel methods became subject of several influential developments as summarized in [30], ch. 14. A generalization of the theory to multiple samples possible, in this context, [2] gives a further method for defining distance covariance over multiple variable subsets using the Möbius transform.

3.2 Independent Component Analysis via Distance Covariance

Based on these definitions, distance covariance has been used as criterion for finding an optimal solution in independent component analysis by [20].

Their approach consists of constructing a hypothesis test with the null hypothesis stating the equality of the joint characteristic function over the entire random vector

of the latent variables $S_k, k \in \{1, \ldots p\}$ with the product of the single-dimension characteristic functions $\prod_{k=1}^{p} \varphi_k(S_k)$. Finding a mixing matrix A and independent latent variables $_{1,\ldots,p}$ in the sense of ICA from empirical data can thus be cast as a minimization problem on the difference between a joint and a product of marginal characteristic functions. As we saw above, this is equivalent to minimizing distance covariance.

The algorithm in [20] then rests on a theorem showing that distance covariance in this application may be minimized sequentially by considering distance covariance of a single component w.r.t. all components following S_k in the sequence $k+1, \ldots, p$ thus dividing the minimization into $p-1$ subproblems (lemma 2.1 in [20] shows that the sum of these $p-1$ optima are an upper bound to the true distance covariance). Note that this approach makes use of the definition of distance covariance between objects of different dimensionalities. This approach is aligned with the method for adapting estimates of the mixing matrix for the whitened data case (i.e., an orthogonal matrix) which uses a similar sequential decomposition of a p-dimensional rotation. Initial values for the minimization given a random rotation matrix and observed whitened data Y are obtained by taking YW^{-1}, i.e. by demixing the data using the inverse of the mixing matrix.

In addition, [20] propose an estimator based on a smooth kernel approximation of the component-wise S_k CDFs interpolating the ECDF from the sampled values. This estimator is referred to as PITdCovICA in [20] and used in the simulations reported below based on the corresponding `R` package `steadyICA` [22].

4 A proof-of-concept study

In this section, a proof-of-concept study is reported in order to clarify the ICA-based approach to document corpus analysis and its implications on topic modelling.

4.1 Data and ICA solution

A small corpus of German election platform documents from 2013-15 has been used. This corpus is comparatively well curated and fully preprocessed including morphological analysis. This results, in particular, in a vocabulary of highly reduced size compared to using no or shallow preprocessing only. Details on the corpus, LDA analyses and related evaluations may be found in [28].

The ICA based analysis starts from setting a preferred number of independent components. Following the results as cited, 12 components were searched. This reflects a reasonable guess on the number of politics domains addressed in the electoral campaigns represented.

In terms of data, ICA, if based on document vector or bag-of-words approach [19], starts from a term-document matrix (TDM). In the present study, this matrix was built from raw frequencies. Thus, each raw data row represents a term frequency vector across documents. Each raw data column represents a document

vector. Common information retrieval preprocessing techniques are available for TDMs related to normalization etc. of term or inclusion of inverted document frequencies [19]. Such preprocessing steps are often considered recommendable due to the extreme sparsity of a usual TDM caused by the absence of a large majority of vocabulary terms from single documents. Examining the effects of such preprocessing steps on ICA analyses remains a task for future research.

A standard prerequisite of ICA is double centering of data. In the present application, column centering means subtracting the column means from each term frequency value in the TDM. The effect of centering is to greatly reduce the sparsity of this matrix. Mapped values of zero frequency terms due to centering differ between documents, but, for all such terms, are identical within a document. Moreover, *if* a term has zero raw frequency in a given document *then* it has negative centering value for this document. As a result from these two facts, centering based statistics allow a more detailed characterization of documents than raw term occurrences.

A further transformation of the input matrix to ICA is whitening, and this is required in several optimization approaches to ICA, including the R package `steadyICA` used for the present analysis [22]. Whitening (see [17] for essentials and algorithms) consists of a decorrelating transform similar to principal components with the additional requirement of standardizing variances. Note that, in the present setting, whitening affects the columns of the TDM and thus gives a new representation of the documents as decorrelated variables with unit variance each.

Given this setup, the objective of the ICA analysis is to model the term occurrence dependencies across whitened documents by a (small) set of independent components given by random variates ranging over the set of all corpus terms (the vocabulary). In the ICA literature, a data column often represents a discrete time series, e.g. in blind source separation of digital speech signals. In the present application, however, a data column is not time-indexed, therefore a symmetric ICA estimate will be required that does not take into account any ordering assumption across data columns.

The ICA estimation approach chosen was, in line with earlier paragraphs in the present paper, based on minimizing distance covariance. This minimization is performed on preprocessed input data using the training documents as variables and the vocabulary terms as observation elements. In the case of the present proof-of-concept study, a $n \times p$ term document matrix was obtained based on $n = 5296$ terms from 38 input variables corresponding to the available training documents. The data whitening process was used to reduce data dimensionality to $p = 12$ leading to a 5296×12 preprocessed input matrix X to the ICA procedure.

Minimizing distance covariance, in this setup of ICA, then, amounts to estimating independent components based on pairs of complementary subsets of variables (document vectors) as fully described in [20] and, using a different approach based on Moebius transforms, in [2]. For the present study, the implementation of the approach from [20] in the R package `steadyICA` [22] is used. The ICA result is a

system matrix S^T of the same dimensionality as the input matrix and a $p \times p$ mixing matrix.

4.2 Exploring a reference ICA solution

An ICA solution computed over the centered and whitened document-term matrix X identifies a given number of independent components and mixing coefficients. In the present simulations, the ICA model trained from the whitened input document data using the `steadyICA` [22] optimiziation converges and the model fit to the whitened training data using the model assumption $X \approx S^T.A$ is close to perfect (root MSE 7.6131*10^-13). The 12 ICA components found show a correlation matrix that comes highly close to the identity matrix (norm of the matrix difference equals 7.1094*10^-14) as would be expected under the independence assumption and the unit variance constraint.

Each ICA component in the present analysis can be interpreted as a real-valued feature function on the discrete domain of the vocabulary terms from the training document corpus. The features obtained from ICA (after applying a dimension reduction together with whitening as detailed above) can be inspected more closely by examining their empirical distribution functions. Best fits approximating these functions with mixture distributions have been calculated with the `FindDistributionParameters` and related functions in [35] and further checked and validated with the *fitdistrplus* R-package [7]. The resulting approximated continuous distributions are dominated by

- 4 Cauchy distributions with close-to-zero location parameters and scale parameters between .17 and .20. By definition, expected values and variances of these distributions are indeterminate.
- 3 Student t distributions with degree-of-freedom parameters between 1.0 and 2.0 and varying location parameters close to zero. This implies that these fitted distributions all have a finite expected value (tightly close to zero), but, as common for this class of distributions, indeterminate variance.
- 5 mixture distributions mostly involving dominating proportions of a Cauchy distribution (between .76 and .96). The remaining components are Gamma, LogNormal, InverseGaussian and again Cauchy distributed. None of the fitted mixtures has a finite mean or variance.

These findings demonstrate that the empirical distributions of the component functions found by ICA comply well with non-normality requirement as derived in [5, 15]. Moreover, the fitted approximate probability densities are in the *long tail class* (i.e., decreasing algebraically rather than exponentially for x approaching $+/-\infty$) and exhibit indeterminate variances (and, in part, indeterminate expectations).

The quality of the approximations are satisfactory if inspected via probability plots, and log-likelihoods and AIC /BIC criteria improve substantially compared to fits of a Gaussian distribution only. However, these fitted distributions should not be taken as confirmed models of the components found by ICA. The main reason for this is that the empirical distributions as constructed from the ICA components respecting the unit sample variance constraint are more strongly populated in the modal areas than expected under the approximations found.

An illustrative plot of these findings for one ICA component is provided in Fig. 1.

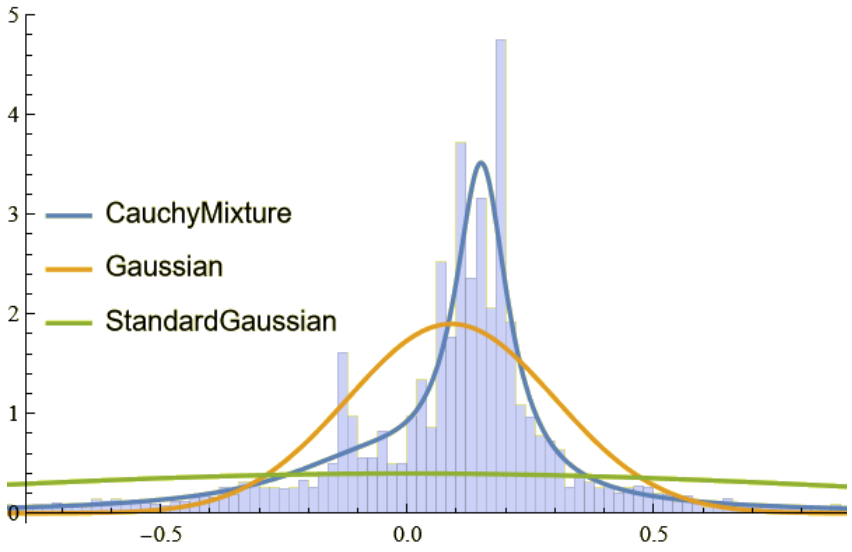


Figure 1. Estimated continuous density fit to the empirical distribution of an ICA component in the solution discussed in the text. The fitted density is a mixture of Cauchy distributions. For highlighting the skewness and long-tail properties, the plot of a normal distribution pdf with matching mode and approximating standard deviation is added. As a baseline for comparison, a plot of the standard Gaussian is also added.

To sum up, in contrast to topic-term distributions as obtained within a probabilistic topic modelling approach the independent components found by ICA show a great diversity of non-Gaussian distributions with mostly long tail properties for approximated population distributions.

4.3 Evaluation of candidate topical structures

Based on a converged ICA solution, two essential steps are needed in order to enable an interpretation in the sense of a topic model like LDA or STM (see earlier sections and references [4, 24]).

- A method and a procedure are needed to assess the relevance or importance of the latent components (interpreted as latent topics) for new documents using the same vocabulary as the training documents. This corresponds to finding document topic distributions in standard probabilistic topic modelling.

- A method and a procedure are needed to describe the latent components by relevance or irrelevance of observations (here: terms from the vocabulary). This is an interpretive step that may need a heuristic approach. In probabilistic topic modelling, the step is performed by estimating topic term distributions.

4.3.1 Finding ICA model based document topic descriptors

Given a converged ICA model, the coefficients of a particular linear combination of ICA components in response to a given input vector (corresponding to a document in the present setting) can generally be obtained from computing a linear model with the system matrix S^T of ICA in the role of a design matrix and a document vector of length n acting as response vector. The linear model, then, describes the orthogonal projection of the input vector (a document term vector in whitening space, in the case of a training document) on the space spanned by the ICA components. This amounts to finding optimal (in the least squares sense) mixing coefficients of the ICA components in matrix S for predicting an arbitrary data (document) vector of appropriate dimensionality.

However, such an approach is of limited value in the present study where training document vectors contain frequencies (before preprocessing as described). Thus, in view of testing and validation beyond model training, a more appropriate approach to predicting latent component (topical structure) content of a document is to use a generalized linear model (GLM) of the Poisson distibution family.

Using Poisson GLMs for assessing topical content of document term frequencies as predicted from ICA generated independent components is therefore the method proposed in the present study. Basically, selecting relevant components that allow to describe topical content in an end-user friendly way becomes then a model selection problem.

A highly renowned approach for model selection in GLMs is based on the theory of the bias-variance tradeoff, see[14] ch. 7, and formulated as a sequence of parameter regularization model solution in [8]. The approach is available in the `R` package **glmnet** where parameter vectors can be regularized on the L_1 norm (lasso approach), the L_2 norm (ridge approach) or a convex combination of these two options (elastic net approach), for details, see [8] and the **glmnet** documentation. For all regularization options, the software produces a sequence of linear or suitably generalized linear regressions with penalized constraints according to the chosen regularization objective.

An advantage of the lasso regularization is the capability to zero out predictor variables in runs having more restrictive values of the penalty constraint. In the

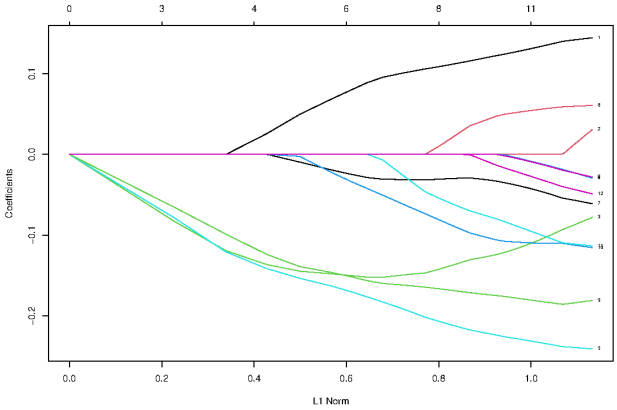
case of the present analysis aiming at identifying only the most relevant contributing ICA system components to the prediction of the content of unseen data, this is a valuable property which therefore has guided the simulations. Figure 2(a) shows the unfolding of contributions from latent ICA components (numbered in the legend on the right-hand side) to the model prediction as the L_1 norm is relaxed in a simulation for one document of our corpus. To repeat, the generalized linear model used to predict term frequencies from the ICA components is from the Poisson family. Moreover, the crossvalidation plot for one single test document in Fig. 2(b) reveals meaningful information on the relevant topic components and a plausible number of them to consider for interpretation. A reasonable minimum of Poisson deviance is already attained by including seven parameters in the model. The boxplot-like symbols for the dispersion of model deviances also show a comparably low estimated variance in the region of log regularization penalties delineated by dotted lines. Effectively, thus, a solution combining low deviance and moderate variance can be selected. It should be noted that negative parameter values indicate a stronger contribution of terms assigned negative values in the corresponding ICA component viewed as feature function. More detail on the meaning of component values in relationship to semantic features is explained in subsection 4.3.2.

Comparing the Poisson GLM outputs for several documents it was found that that weights for some components have noticeable covariance across documents, and more strongly so if they originate from the same political party. Moreover, some documents show an essentially flat component weight profile, and, in fact, these are very generic in nature and do not dwell on any details of the decision needs at the time of the election being covered. It should be noted that flat document profiles would not be recognizeable in conventional probabilistic model modelling since the weights would need to conform to probabilistic additivity. More generally, in the conventional approaches, uninformative documents are as important for the overall optimization of likelihood and thus may contribute to weaken topic information relevant to end users.

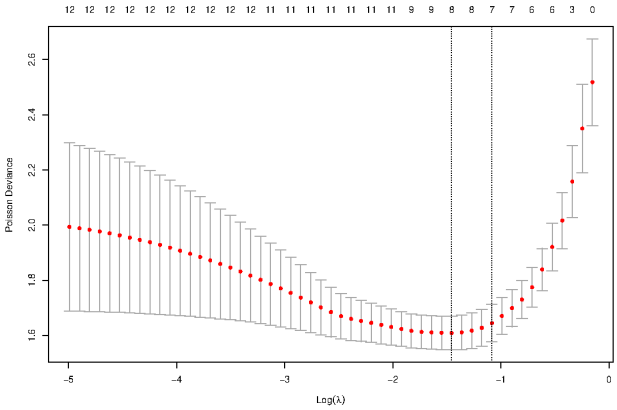
4.3.2 Identifying Semantic profiles from ICA model components

To obtain information about a potential semantic focus of an ICA component, the following approach is suggested – In order to identify discriminating terms, a useful metric is the variance of the ICA component feature functions evaluated for a given term. Moreover, in order to identify semantically influential terms, a useful metric is the norm of the value vector for a given term across the ICA component feature functions. Evaluating these tentative analysis metrics, intersecting elements in the 95 percent quantiles of both metrics and cleaning up the resulting term list, a list comprising 228 terms having ICA vectors with very high cross component variance and very high norm is found.

Inspecting the results of this index search, a semantic focus of each ICA component can be found by inspecting terms and contexts at the high end of the compo-



(a) Contibution paths of ICA component coefficients in a penalized Poisson GLM using the Lasso (L1) regularization.



(b) Crossvalidation of the GLM sequence in Fig. 2(a) shows a range of optimal parameter selection and shrinking coefficients.

Figure 2. Evaluation steps of a Poisson generalized linear model describing influences of ICA components on term frequencies in a single document using the **glmnet** approach [8].

nent's feature function. Evaluation of terms at the low end of the feature function can be understood as a collection of potentially opposing or even conflicting issues with political themes addressed in the terms at the high end of the respective ICA feature function. This view of a balanced feature function as pointing to semantic differences stands in a long tradition of linguistic research, see [25] and also reflects cognitive psychological feature based models of similarity, see [11].

A closer look into the semantic profiles found reveals that key themes or issue around the electoral campaign addressed by various parties are clearly identified,

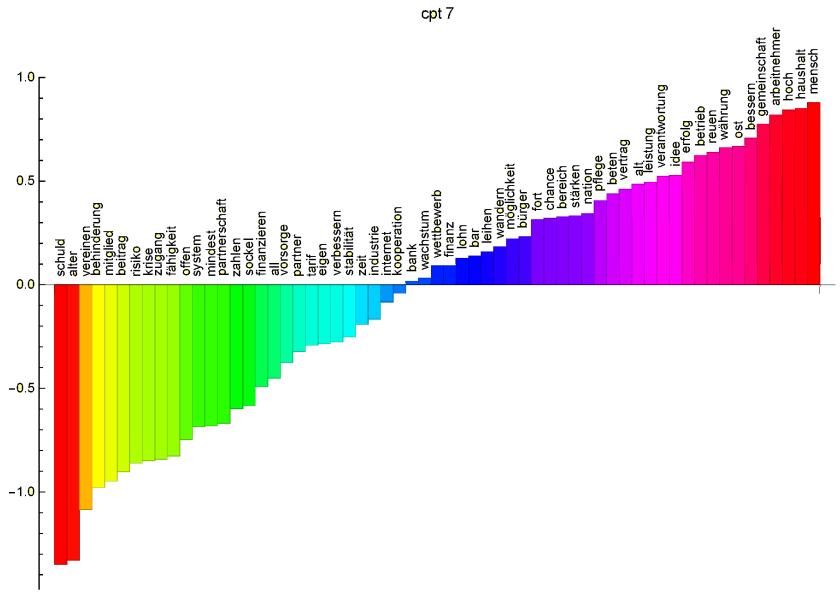


Figure 3. Bar chart of highest and lowest valued terms in a component of the ICA solution as discussed in the text.

e.g. consumer and environment protection, the increasing European dimension of political decisions relating to economy, labour and education, and perspectives of urban and rural development, to mention but a few examples. A detailed discussion would be out of the scope of the present publication channel and is deferred to a future paper.

5 Conclusion

A new approach to statistical topic structures modelling based on independent component analysis (ICA) has been introduced. In the present approach, topic structures are defined as independent random variables. Motivating considerations for choosing this approach have been discussed.

The interpretation of independent random variables as topic structures is motivated by their estimation from training document vectors using ICA. Each independent component, then, works as a feature function over the terms of the vocabulary of the training document corpus. The independence of the features obtained corresponds to non-overlapping semantic perspectives on this vocabulary. Each perspective forms a set of semantic profiles that can be viewed as meaningful distinctions between similar sets of terms in the sense of linguistic and psychological approaches to similarity and semantic contrast judgements.

A proof-of-concept study has been presented that demonstrates the technical feasibility and some interpretative properties of the proposed approach. A salient

theoretical advantage of independent semantic profiles emerging from this study is a clearer definition of the distinction between different topics and their semantics. Moreover, occurrence correlations of topical structures across documents become more easily interpretable if these structures are non-overlapping (which cannot be guaranteed in other approaches as discussed in the paper). Another advantage is that any documents used for training or testing / validation may show indifferent topical profiles – as it is likely in most realistic training text copora. Indifferent topic profiles, however, are not in the scope of existing probabilistic modelling approaches due to the Kolmogorov axioms which need to be valid for any document topic distributions.

As several approaches to estimating independent components are available together with state-of-the-art implementations, a motivation was given for choosing the approach based on minimizing distance covariance. Further studies taking into account also different input formats of the document corpus data are needed to evaluate possible performance or output advantages of any ICA estimation procedures in more detail. Energy based tests of goodness of fit [30] may also play an important role in modelling the independent component densities as estimated by ICA, since most of them, by theory and by the results reported here, fall short of reasonable approximations covered by conventional goodness-of-fit significance testing.

On the level of plausibility checking of the outcomes of the study, some promising results have been found. There seems to be a realistic interdisciplinary perspective of using the approach presented in end-user focussed tools for terminology extraction or document description.

References

[1] Francis Bach and Michael Jordan. "Kernel Independent Component Analysis". In: *Journal of Machine Learning Research* 3 (2002), pp. 1–48.

[2] Martin Bilodeau and Aurélien Guetsop Nangue. "Tests of Mutual or Serial Independence of Random Vectors with Applications". In: *Journal of Machine Learning Research* 18.74 (2017), pp. 1–40. URL: http://jmlr.org/papers/v18/16-184.html.

[3] David M. Blei and John D. Lafferty. "A Correlated Topic model of science". In: *The Annals of Applied Statistics* 1.1 (2007), pp. 17–35. DOI: 10.1214/07-AOAS114.

[4] David M. Blei, Andrew Ng, and Michael Jordan. "Latent Dirichlet allocation". In: *JMLR* 3 (Jan. 2003), pp. 993–1022.

[5] Pierre Comon. "Independent Component Analysis, a new concept". In: *Signal Processing*. Vol. 36. Elsevier, 1994, pp. 287–314. DOI: 10.1016/0165-1684(94)90029-9. URL: https://hal.archives-ouvertes.fr/hal-00417283.

[6] Thomas Cover and Joy Thomas. *Elements of Information Theory*. Hoboken, NJ: Wiley-Interscience, 2006.

[7] Marie Laure Delignette-Muller and Christophe Dutang. "fitdistrplus: An R Package for Fitting Distributions". In: *Journal of Statistical Software* 64(4) (2015), pp. 1–34.

[8] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software; Vol 1, Issue 1 (2010)* (Feb. 2010). URL: http://dx.doi.org/10.18637/jss.v033.i01.

[9] Bela A. Frigyik, Amol Kapila, and Maya R. Gupta. *Introduction to the Dirichlet Distribution and Related Processes*. Tech. rep. University of Washington, Dpt. Electrical Engineering, 2010.

[10] Bettina Grün and Kurt Hornik. "topicmodels: An R package for fitting topic models". In: 40 (2012), pp. 1–30. URL: http://www.jstatsoft.org/v40/i13/.

[11] Ulrike Hahn. "Similarity". In: *WIREs Cogn Sci* 5.3 (May 2014), pp. 271–280. ISSN: 1939-5078. DOI: 10.1002/wcs.1282. URL: https://doi.org/10.1002/wcs.1282.

[12] T. Hastie and R. Tibshirani. *Product Density Estimation for ICA using Tilted Gaussian DensityEstimates*. Tech. rep. https://cran.r-project.org/package=ProDenICA, 2022.

[13] Trevor Hastie and Robert Tibshirani. "Independent Components Analysis through Product Density Estimation". In: *Proc. Neural Information Processing Society*. 2002.

[14] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. 2nd ed. Springer Series in Statistics. New York, NY, USA: Springer New York, 2017.

[15] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. Wiley, 2001. ISBN: 0-471-40540-X. DOI: 10.1002/0471221317.

[16] Aapo Hyvärinen and Erkki Oja. "A Fast Fixed-Point Algorithm for Independent Component Analysis". In: *IEEE Trans. Neural Networks* 10(3) (1999), pp. 626–634.

[17] Agnan Kessy, Alex Lewin, and Korbinian Strimmer. "Optimal Whitening and Decorrelation". In: *The American Statistician* 72.4 (2018), pp. 309–314. ISSN: 1537-2731. DOI: 10.1080/00031305.2016.1277159.

[18] Russell Lyons. "Distance Covariance in Metric Spaces". In: *The Annals of Probability* 41.5 (2013), pp. 3284–3305. ISSN: 00911798. URL: http://www.jstor.org/stable/42919805.

[19] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2009. URL: http://www.informationretrieval.org/.

[20] David S. Matteson and Ruey S. Tsay. "Independent Component Analysis via Distance Covariance". In: *Journal of the American Statistical Association* 112.518 (Apr. 3, 2017), pp. 623–637. DOI: 10.1080/01621459.2016.1150851.

[21] John Paisley, Chong Wang, and David M. Blei. "The Discrete Infinite Logistic Normal Distribution". In: *Bayesian Analysis* 7.4 (2012), pp. 997–1034.

[22] Benjamin B. Risk, Nicholas A. James, and David S. Matteson. *steadyICA: ICA and Tests of Independence via Multivariate Distance Covariance*. Tech. rep. CRAN.R-project.org, 2015. URL: https://cran.r-project.org/package=stm.

[23] Margaret Roberts et al. *Package 'stm' – Estimation of the Structural Topic Model*. Tech. rep. https://cran.r-project.org/package=stm: CRAN.R-project.org, 2020.

[24] Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. "A Model of Text for Experimentation in the Social Sciences". In: *Journal of the American Statistical Association* 111.515 (July 2, 2016), pp. 988–1003. DOI: 10.1080/01621459.2016.1141684. URL: http://dx.doi.org/10.1080/01621459.2016.1141684.

[25] Ferdinand de Saussure. *Cours de Linguistique Générale*. Ed. by Tullio de Mauro. Critical Edition by Tullio Mauro. Payot, Paris, 1979.

[26] Marcus Spies. "Probabilistic topic models for small corpora – An empirical study". In: *TOTh 2017 Terminologie et Ontologie : Théories et Applications*. Ed. by Christophe Roche. Éditions de l'Université Savoie Mont Blanc. Éditions de l'université de Savoie, 2018, pp. 137–160.

[27] Marcus Spies. "Smooth or rough, neutral or biased – can topics uncover terminologies?" In: *TOTh 2018 Terminologie et Ontologie : Théories et Applications*. Ed. by Chr. Roche. Université de Savoie Mont-Blanc. Université de Savoie Mont-Blanc, Presse Universitaire, 2020, pp. 81–109.

[28] Marcus Spies. "Topic Modelling with Morphologically Analyzed Vocabularies". In: *Scientific Publications Of The State University Of Novi Pazar Ser. A: Appl. Math. Inform. And Mech.* 9.1 (2017), pp. 1–18.

[29] Gabor J. Szekely and Maria L. Rizzo. "Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method". In: *Journal of Classification* 22.2 (2005), pp. 151–183. ISSN: 1432-1343. URL: https://doi.org/10.1007/s00357-005-0012-9.

[30] Gabor Székely and Maria Rizzo. *The Energy of Data and Distance Correlation*. Monographs on Statistics and Applied Probability 171. Boca Raton, FL: CRC Press, 2023.

[31] Gábor J. Székely and Maria L. Rizzo. "BROWNIAN DISTANCE COVARIANCE". In: *The Annals of Applied Statistics* 3.4 (2009), pp. 1236–1265. ISSN: 19326157. URL: http://www.jstor.org/stable/27801540.

[32] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. "Measuring and Testing Dependence by Correlation of Distances". In: *The Annals of Statistics* 35.6 (2007), pp. 2769–2794. ISSN: 00905364. URL: http://www.jstor.org/stable/25464608.

[33] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. "Measuring and Testing Dependence by Correlation of Distances". In: *The Annals of Statistics* 35.6 (2007), pp. 2769–2794. ISSN: 00905364. URL: http://www.jstor.org/stable/25464608.

[34] Yee Whye Teh et al. "Hierarchical Dirichlet Processes". In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1566–1581. ISSN: 01621459. URL: http://www.jstor.org/stable/27639773.

[35] Wolfram Research. *Mathematica, Version 12.3*. Wolfram Research Inc., Champaign IL, 2020.