Statistical Science
2019, Vol. 34, No. 2, 280–300
https://doi.org/10.1214/18-STS686

The Importance of Being Clustered: Uncluttering the Trends of Statistics from 1970 to 2015

Laura Anderlucci, Angela Montanari and Cinzia Viroli

Abstract. In this paper, we retrace the recent history of statistics by analyzing all the papers published in five prestigious statistical journals since 1970, namely: *The Annals of Statistics*, *Biometrika*, *Journal of the American Statistical Association*, *Journal of the Royal Statistical Society, Series B* and *Statistical Science*. The aim is to construct a kind of "taxonomy" of the statistical papers by organizing and clustering them in main themes. In this sense being identified in a cluster means being important enough to be uncluttered in the vast and interconnected world of the statistical research. Since the main statistical research topics naturally born, evolve or die during time, we will also develop a dynamic clustering strategy, where a group in a time period is allowed to migrate or to merge into different groups in the following one. Results show that statistics is a very dynamic and evolving science, stimulated by the rise of new research questions and types of data.

Key words and phrases: Model-based clustering, cosine distance, textual data analysis.

1. INTRODUCTION

It is hard to date the birth of statistics as a modern science. Certainly, in the past 45 years, remarkable new ideas and contributions to a rich variety of topics were stimulated by the rise of new research questions and new types of data, and disseminated by a wider access to highly-performing electronic devices and scientific journals.

In this work, we retrace the recent history of the adult-stage of statistics by analyzing the contributions published in some of the most prestigious statistical journals from 1970 to 2015.

Laura Anderlucci is Senior Assistant Professor, Department of Statistical Sciences, University of Bologna, via Belle Arti 41, 40126 Bologna, Italy (e-mail: laura.anderlucci@unibo.it). Angela Montanari is Professor of Statistics, Department of Statistical Sciences, University of Bologna, via Belle Arti 41, 40126 Bologna, Italy (e-mail: angela.montanari@unibo.it). Cinzia Viroli is Professor of Statistics, Department of Statistical Sciences, University of Bologna, via Belle Arti 41, 40126 Bologna, Italy (e-mail: cinzia.viroli@unibo.it).

Since classification into distinct entities is a fundamental step to discover meaningful information and to create new knowledge, we aim at constructing a "taxonomy" of the considered statistics papers by organizing and clustering them according to main topics. As the topics of research are many, heterogeneous and they evolve over time, we also develop a dynamic clustering strategy: a group in a decade can migrate or merge into different groups in the following decade; the birth and the (potential) death of topics are allowed as well. Of course, it is very hard to disentangle all the possible subfields of the statistical research. Suppose a certain number of topics are identified in a period of time: despite their unavoidable degree of internal heterogeneity and their mutual linkage, paraphrasing the title of our work, being clustered is important because it means being uncluttered in the vast and interconnected world of statistics. In other words, a cluster identifies an aggregation of related papers around a relevant statistical topic. In so doing, we assume that a statistical paper is developed around a single research topic. Although we believe that in few cases it can be a restrictive assumption, this unique association is a fun-

damental condition to create a clear taxonomy of the most important research themes.

Information about the papers is collected as textual data, from their titles and abstracts, and it is stored in a high-dimensional document-term matrix. The final data consist of very short and sparse texts: the documents represented by the title and abstract have an average length of 55 words with a degree of sparsity of 99.7% (which means that the 99.7% of matrix cells are zeros).

These data features raise the need of a novel statistical model, extendable in a dynamic fashion to analyze how statistics has evolved over time.

Conventional statistical models for analyzing textual data developed in the context of information retrieval, natural language processing and machine learning are not adequate for this purpose. One of the most popular method, mixtures of unigrams models (Nigam et al., 2000), is based on the natural idea of modeling the word frequencies as multinomial distributions and consider a unique association between topics and documents. The approach works very well in a supervised context, where some prior information on the topics is known, but it is not appropriate in our unsupervised setting because of the very high level of sparsity that forces the proportions to be all equal and very close to zero. The undesired consequence is that all documents end up being clustered in the same group, regardless the number of mixture components. When data are very sparse, several authors have shown the superiority of mixtures of von Mises–Fisher distributions for text classification (see Zhong and Ghosh, 2005, Banerjee et al., 2005) provided that the textual data are *directional*, that is, the frequencies of the documents are normalized to 1 according to the L2 norm. Our proposal will move along this line, by relaxing the assumption of directional data.

More sophisticated versions that consider multiple-topic documents are the latent semantic indexing (Deerwester et al., 1990), the probabilistic latent semantic models (Hofmann, 1999), Latent Dirichlet Allocation Model (Blei, Ng and Jordan, 2003, Chang and Blei, 2009, Sun et al., 2009) and more recently elaborated proposals based on graphical models to incorporate information about the co-authorship network (see, for instance, Bouveyron, Latouche and Zreik, 2018 and references therein). In the recent work by Varin, Cattelan and Firth (2016), citations are used to cluster (via agglomerative hierarchical methods) and to rank (via quasi-Stigler model) statistical journals.

In this work, we will embrace the setting of a single association topic-document. As we discussed above, this is fundamental in order to build a clear taxonomy, but there are also further reasons. First of all, our documents are very short and, with only a few terms—having positive frequency—available from abstracts and titles, it is very hard and certainly imprecise to disentangle many subtopics. Second, in the perspective of studying the dynamic of the scientific main research areas from 1970 to 2015, we prefer to work under the interpretation of a unique prevalent topic per document. This will force the analysis towards general scientific areas, rather than specific subfields, and macro-topics can be better investigated and analyzed during the large period of time of 45 years.

In a similar perspective, recently Ji and Jin (2016) analyzed the network of coauthorship and citations of all statistical research papers published in four of the top journals in statistics from 2003 part of 2012 with the aim of identifying hot topics and key authors of the statistical community. Even if their analysis is conducted for a shorter period of time we will show that many findings are consistent with our analysis.

The paper is organized as follows. After describing the data, our model is presented in Section 3, together with a strategy to get a reasonable level of the precision and results from an empirical study; model estimation is also described. In Section 4, results on the classification of the considered statistical papers are presented for the whole period of time 1970–2015. Finally, in Section 5, we extend the proposed model in a dynamic fashion through a semi-supervised mixture model, so as to describe the evolution over time of the recent scientific research in statistics.

2. THE DATA: STATISTICS IN 1970–2015

2.1 Data Collection

The study is based on the articles published on five top statistical journals during the period 1970–2015: *The Annals of Statistics* (*AoS*), *Biometrika* (*Bka*), *Journal of the American Statistical Association* (*JASA*), *Journal of the Royal Statistical Society, Series B* (*JRSSB*) and *Statistical Science* (*StS*). These journals have been selected for both their historical importance and their highest citation metrics, in terms of the Article Influence Score (AIS) and of the five-year Impact Factor, among all the statistical journals ranked by the ISI Web of Knowledge database. These selected journals are very different: for instance, AoS publishes

TABLE 1
Number of statistical papers published in the five journals by period of time

Journals	1970–1979	1980–1989	1990–1999	2000–2009	2010–2015	**1970–2015**
AoS	610	864	920	809	508	**3711**
Bka	738	722	613	753	425	**3251**
JASA	1431	1218	1393	1123	744	**5909**
JRSSB	356	395	498	467	234	**1950**
StS	0	69	151	239	192	**651**
Total	**3135**	**3268**	**3575**	**3391**	**2103**	**15,472**

mostly methodological papers, JASA is historically divided into a Methodological part and an Applications and Case Studies part, but reviews may be also considered. StS presents *the full range of contemporary statistical thought* and it may treat the research themes, which are particularly relevant and of interest, at a more accessible level. These peculiarities offer a general and complete picture of the most relevant research themes of the statistical community.

More precisely, we considered information contained in titles and abstracts of all the articles published in these journals starting from 1970 or from the first available issue (dated 1973 for AoS and 1986 for StS). Data have been collected by downloading the freely available bibliography files from the journal websites in RIS format for *Statistical Science* and in BIB format for the other four journals. Then, the bibliography files have been imported in `R` by using the package `RefManageR` in order to produce a single textual file for each article containing only the title followed by the abstract. Author names and the other editorial information were not considered. Since our aim is to identify the most relevant topics in the statistical research, we excluded from the analysis the editorial frontmatter articles, the book reviews, the series of papers entitled "*A conversation with...*" published in *Statistical Science* during the whole period (1986–2015), the interviews with authors narrating career and life rather than statistics and the series *Studies in the History of Statistics and Probability* published in Biometrika in the first decades. We also excluded the discussion or comments to the articles, replies and rejoinders when they were not accompanied by an abstract. Overall, we collected information on 15,472 articles, which are summarized in Table 1 by journal and five periods of time: 1970–1979, 1980–1989, 1990–1999, 2000–2009 and 2010–2015.

2.2 Data Management

The 15,472 textual files were imported in `R` with the library `lsa` so as to produce a document-term matrix containing the term frequencies of each paper. Raw data were processed by stemming in order to reduce inflected or derived words to their unique word stem. Moreover, we removed numbers and we filtered the terms by a list of English stopwords, that includes the most common words in English, such as adverbs and articles. The whole procedure was automatically performed by using the options available in the `R` function `textmatrix` (library `lsa`). In addition to the default stopwords of the package, we added a list of generic words that are not generally common in English, but that are certainly widespread in the statistical language, such as "variable", "statistics", "analysis", "data" and "model". At the end of this filtering process we ended up with a final document-text matrix of 15,472 rows, corresponding to the papers, and of 15,036 columns, corresponding to the final reduced stemmed terms.

In order to measure the importance of a term in the whole collection of documents, we have weighted each frequency by the so-called Inverse Document Frequency (IDF), which is the logarithm of the total number of documents divided by the number of documents where each term appears. This commonly used normalization (Salton and McGill, 1986) allows us to give more importance to the terms that are contained in the documents but are in general rare.

3. CLUSTERING STATISTICS

The basic idea of this work is to cluster papers according to their weighted term frequencies, in order to identify the main relevant topics of statistics since 1970. Obviously, it is very hard to disentangle all the possible subfields of the statistical research. Statistical topics are many, they are naturally interconnected and they evolve over time. However, when a certain

number, say k, of clusters are identified, they certainly aggregate similar subtopics of the research. In other words, a cluster identifies an aggregation of related papers around a broad statistical theme and we assume clusters identify the main relevant topics. The internal degree of heterogeneity will depend on the total number of groups k.

3.1 Mixtures of Cosine-Based Distance Densities

Mixture models allow to decompose a heterogeneous population into a finite number of subgroups with a homogeneous density function (Fraley and Raftery, 2002, McLachlan and Peel, 2000). In our case, modeling the component densities of term frequencies is a hard task due to the peculiar characteristics of the data. Each document is characterized by a Gdimensional vector of nonindependent term frequencies with a relevant degree of sparsity. The natural model for identifying the topics is the mixture of unigrams models (Nigam et al., 2000), which is essentially a mixture of multinomial distributions estimable by an EM algorithm. However, although in general it is an efficient estimation model, results on our (big) data are seriously affected by the amount of zeros and they are extremely sensitive to the initialization, which, very frequently, leads the algorithm to be trapped in local unsatisfactory points after very few iterations.

As an alternative, a zero-inflated distribution could be employed to model sparsity. We investigated mixtures of zero-inflated Poisson, Bernoulli and Negative Binomials, but such choices did not offer a satisfactory approximation to the observed distributions for two principal reasons. First, the theoretical zero-inflated models involve a very large number of parameters, since (at least) two different parameters have to be estimated for each term and each group, namely the zero-inflation and the location parameters; as a consequence, the fit is computationally unfeasible. Second, these univariate distributions would be fitted to each observed set of term frequencies, interpreted as a variable, therefore ignoring the semantic dependence among the terms.

Due to these difficulties we changed our perspective from density-based estimation to distance-based clustering models. Distance-based densities have been successfully used by several authors (see Mallows, 1957, Fligner and Verducci, 1986, Diaconis, 1988) in the context of ranking data and then adapted for classification in a mixture-based perspective by Murphy and Martin (2003). These models can be viewed as special cases of the so-called "probabilistic D-clustering" (Ben-Israel and Iyigun, 2008). In the context of ranking data, several distance measures have been used, for example, Euclidean, Kendall, Spearman and Cayley's distances. None of them provides a useful measure for sparse textual data, since they may be highly affected by the high proportion of zeros. A prominent measure of distance overcoming these difficulties is based on the cosine similarity, because it considers only the nonzero elements of the vectors, allowing to measure the dissimilarity between two documents in terms of their subject matter. Given two p-dimensional documents, say $\mathbf{x}$ and $\mathbf{y}$, the cosine distance of the two corresponding frequency vectors is

$$d(\mathbf{x},\mathbf{y}) = 1 - \frac{\sum_{h=1}^{p} x_h y_h}{\sqrt{\sum_{h=1}^{p} x_h^2}\sqrt{\sum_{h=1}^{p} y_h^2}}, \tag{1}$$

where x_h and y_h denote the frequency of word h in document $\mathbf{x}$ and $\mathbf{y}$, respectively. This measure is not affected by the amount of zeros and is a normalized synthesis of the p-variate terms of the documents. Since the elements of $\mathbf{x}$ and $\mathbf{y}$ are positive or null frequencies, it is easy to prove that the distance ranges between 0 and 1.

Given the cosine distance $d(\mathbf{y},\boldsymbol{\xi})$ of a generic document $\mathbf{y}$ from a reference centroid, say $\boldsymbol{\xi}$, we define a probability density function for the random variable $\mathbf{y}$ as

$$f(\mathbf{y};\boldsymbol{\xi},\lambda) = \psi(\lambda)e^{-\lambda d(\mathbf{y},\boldsymbol{\xi})}, \tag{2}$$

where λ is a positive precision (or concentration) parameter, with $\lambda > 0$, and $\psi(\lambda)$ is a normalization constant such that $f(\mathbf{y};\boldsymbol{\xi},\lambda)$ is a proper density function. When $\mathbf{y}$ and $\boldsymbol{\xi}$ are distributed on the surface of a unit hypersphere, so that they are directional, the density in (2) is the von Mises–Fisher distribution and its normalization constant analytically exists as a function of the modified Bessel function of the first kind and order $p/2-1$ (Mardia and Jupp, 2000). Mixtures of von Mises–Fisher distributions have been largely used by many authors in the information retrieval community for clustering direction data under the assumption that the direction of a text vector is more important than its magnitude (see, for more details, Banerjee et al., 2005, McLachlan and Peel, 2000, Zhong and Ghosh, 2005).

In our data problem, many words are removed as either stopwords or very widespread statistical terms. Moreover, as will be explained in Section 4, the analysis will be performed on a subset of selected variables, chosen according to their entropy in order to remove biases due to the high dimensionality and to the

presence of irrelevant words. In this perspective, data cannot be normalized into directional data and, therefore, analyzing the absolute values of the frequencies is preferable. In order to perform clustering, we consider a mixture of k cosine distance-based density functions:

$$f(\mathbf{y};\boldsymbol{\xi},\lambda)=\sum_{i=1}^{k}\pi_i\psi(\lambda)e^{-\lambda d(\mathbf{y},\boldsymbol{\xi}_i)} \tag{3}$$

with positive mixture weights π_i, summing to unity, $\sum_{i=1}^{k}\pi_i=1$, and component varying centroid vectors $\boldsymbol{\xi}_i$. Notice that in this case the normalization quantity $\psi(\lambda)$ cannot be derived in closed form, thus making the estimation of λ hard. In the next section, we will show a strategy to get a reasonable value for the precision parameter λ.

3.2 Role of the Precision Parameter

The precision parameter λ is taken common among the mixture components for theoretical and practical reasons. First, observe that the precision acts as a scaling of the distances. For high values of λ, even small differences between distances induce relevant differences in the density values. In this case, a small difference between the cosine distance of a document $\mathbf{y}$ from two centroids, say $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$, is over-weighted by λ and it produces a relevant difference in terms of likelihood, favoring the posterior clustering to the component with smallest distance. On the contrary, when $\lambda\to 0$ the importance of the distances vanishes and the densities converge to the uniform distribution. In other terms, as λ increases, the mixture is forced to produce more homogeneous and "purer" clusters (in order to have a good fit this implies to have more components). Fixing λ across components implies all clusters have the same degree of internal homogeneity.

A more theoretical insight about the crucial role of λ for clustering derives by imposing a consistency relation between distances and mixture posterior classification. More precisely, for a generic document y, let i' denote the component with the minimum distance, that is, $i':\min_{1\le i\le k}\{d(y,\xi_i)\}=d(y,\xi_i')=d_{i'}$. Then the following definitions and propositions establish a formal consistency relation between the value of λ and the posterior classification, which justifies the choice of a common dispersion parameter for all the mixture components.

DEFINITION 1. Given the mixture model (3), the *consistent clustering rate*, say $1-\alpha$, is the probability of allocating y to the component to which it has the minimum distance

$$1-\alpha=\Pr(z_{i'}=1|y,d_{i'}),$$

where z is the hidden allocation vector of length k with value one in correspondence of the component membership and zero otherwise.

In a similar manner, we may define the inconsistent clustering rate:

DEFINITION 2. Given the mixture model (3), the *inconsistent clustering rate* is defined as

$$\alpha=\Pr(z_{i'}\neq 1|y,d_{i'}).$$

The following results establish a formal relation between λ and the inconsistent clustering rate.

PROPOSITION 1. *Given $d_i=d(y,\xi_i)$ for $i=1,\dots,k$ and $i':\min_{1\le i\le k}\{d(y,\xi_i)\}=d_{i'}$, the inconsistent clustering rate for the model (3) is inversely and non linearly related to λ through the formula*

$$\alpha=\frac{\sum_{i=1}^{k}\pi_i e^{-\lambda(d_i-d_{i'})}-\pi_{i'}}{\sum_{i=1}^{k}\pi_i e^{-\lambda(d_i-d_{i'})}}. \tag{4}$$

To prove the proposition observe that by definition we have

$$\begin{aligned}1-\alpha&=\frac{\pi_{i'}f(y|z_{i'}=1)}{\sum_{i=1}^{k}\pi_i f(y|z_i=1)}\\&=\frac{\pi_{i'}\psi(\lambda)e^{-\lambda d_{i'}}}{\sum_{i=1}^{k}\pi_i\psi(\lambda)e^{-\lambda d_i}}\\&=\frac{1}{1+\sum_{i\neq i'}\frac{\pi_i}{\pi_{i'}}e^{-\lambda(d_i-d_{i'})}}.\end{aligned} \tag{5}$$

Now equation (4) derives by observing that $\sum_{i\neq i'}\frac{\pi_i}{\pi_{i'}}e^{-\lambda(d_i-d_{i'})}+1=\sum_{i=1}^{k}\frac{\pi_i}{\pi_{i'}}e^{-\lambda(d_i-d_{i'})}$. Generally, as λ increases α decreases and viceversa. More formally is the following proposition.

PROPOSITION 2. *Given the relationship* (4),

$$\lim_{\lambda\to\infty}\alpha(\lambda)=0,\qquad \lim_{\lambda\to 0}\alpha(\lambda)=1-\pi_{i'}. \tag{6}$$

The first limit derives by observing that $\sum_{i=1}^{k}\pi_i e^{-\lambda(d_i-d_{i'})}=\pi_{i'}+\sum_{i\neq i'}^{k}\pi_i e^{-\lambda(d_i-d_{i'})}\ge\pi_{i'}$ for all $\lambda>0$ because $(d_i-d_{i'})>0$ for $i\neq i'$. The second limit derives directly from the right-hand side part of equation (5).

Figure 1 (first panel) shows the relation between λ and the average consistent clustering rate in the dataset with all the 15,472 documents and $k=10$ components. The second panel of the figure shows the best value of λ to get an average consistent clustering rate of 0.90 (circles points) and 0.95 (triangle points), respectively on the same data as k ranges between 2 to 30.

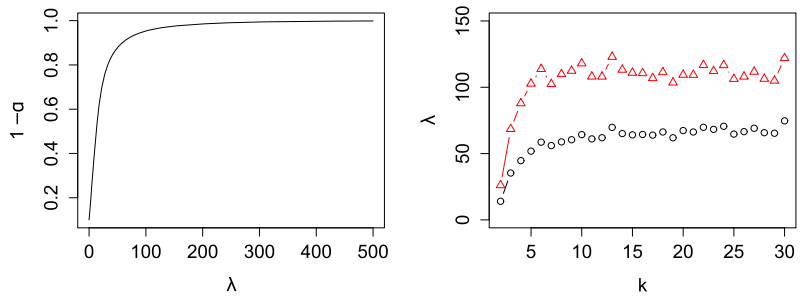

FIG. 1. *The effect of* λ: *in the first panel the relation between* λ *and* $1-\alpha$ *is shown*; *the second panel shows the value of* λ *corresponding to an average consistent clustering rate of* 0.90 (*circles points*) *and* 0.95 (*triangle points*) *as* k *varies*.

From these results it is clear that, in order to have the same goodness of fit, a mixture model with higher λ would generally require more components. Moreover, the inverse relation between λ and the the inconsistent clustering rate in (4) can be used to approximate the precision parameter as explained in the next section.

3.3 Model Estimation

A computational problem of the mixture model (3) is related to the estimation of the normalization constant $\psi(\lambda)$, that cannot be derived neither analytically nor numerically due to the complexity of the cosine distance and to the high dimension of the multiple integral to be evaluated. This translates into the problem of estimating the precision parameter λ. According to Proposition 1, a way to get a reasonable value for the precision parameter is to fix a desirable level for the consistent clustering average rate among all the observations, such as $1-\alpha = 0.95$. For a given value of α, λ can be obtained by solving equation (4) over the sum of all the observations. The other parameters of the mixture model (3) can be easily estimated via a conventional EM algorithm by maximizing the complete-data log-likelihood. We denote by $y_j = (y_{j1}, \ldots, y_{jp})$ the vector of weighted IDF frequencies of the jth ($j = 1, \ldots, n$) document and by $z_j = (z_{j1}, \ldots, z_{jk})$ the latent allocation variable, that takes the value 1 in correspondence of the cluster membership and zero otherwise. The complete log-likelihood is

$$\ell_C(\boldsymbol{\xi}, \boldsymbol{\pi}; \mathbf{y}, \mathbf{z}) = \sum_{j=1}^{n}\sum_{i=1}^{k} z_{ji}\big(\log \pi_i + \log \psi(\lambda) - \lambda d(y_j, \xi_i)\big). \tag{7}$$

The two steps of the EM algorithm are:

E-Step: Compute the posterior probabilities as

$$\hat{z}_{ji} = \frac{\pi_i e^{-\lambda d(y_j,\xi_i)}}{\sum_{i=1}^{k} \pi_i e^{-\lambda d(y_j,\xi_i)}}.$$

M-Step:

(a) Compute λ by solving

$$\sum_{j=1}^{n}\sum_{i=1}^{k} \frac{\pi_i}{\pi_{i'}} e^{-\lambda(d_{ij}-d_{i'j})} = \frac{n}{1-\alpha},$$

where $d_{ij} = d(y_j, \xi_i)$.

(b) Compute via numerical optimization methods the values of the centroids as

$$\xi_i = \operatorname*{argmin}_{\xi} \sum_{j=1}^{n} \hat{z}_{ji}\, d(y_j, \xi).$$

(c) Compute the weights as

$$\pi_i = \sum_{j=1}^{n} \hat{z}_{ji}/n.$$

The algorithm converges quickly, but, it is sensitive to its starting values. To avoid to get stuck in local maxima, we initialized it with the solution of the spherical k-means based on the cosine distance (Dhillon and Modha, 2001, Maitra and Ramler, 2010).

3.4 Empirical Study

The performance of the proposed method is evaluated and compared with other possible approaches in an empirical simulation study.

We consider three balanced clusters (for total sample size of 1500 units/documents) and a set of 1500 variables/words, 150 of which are relevant for clustering purposes. Simulated data have been generated with a high level of sparsity in order to reflect the peculiarity of the real data. The average proportion of null frequencies with respect to the total number of entries in a single dataset is about 99%.

Two scenarios are analyzed. In the first scenario, the whole set of 1500 words is considered. In the second scenario, we reduce the dimensionality by performing

TABLE 2

Simulation study: *in scenario* 1, *original data with no feature selction are clustered*; *in scenario* 2, *reduced data* (*feature selection via entropy values*) *are clustered. Average Adjusted Rand index* (*with standard errors in brackets*) *for different methods. Numbers in squared brackets indicate the valid cases out of* 100 *runs*

Scenario	Cosine	Cosine—SVD	Euclidean	Unigrams
1 - $p = 1500$	$0.974\ (0.001)_{[100]}$	$0.974\ (0.001)_{[100]}$	$0.000\ (0.000)_{[100]}$	$0.090\ (0.025)_{[90]}$
2 - $p = 150$	$0.980\ (0.001)_{[100]}$	$0.980\ (0.001)_{[100]}$	$0.002\ (0.000)_{[100]}$	$0.269\ (0.041)_{[55]}$

feature selection. In particular, we are interested in selecting the words that help identifying homogeneous groups, that is, that are differentially present across the documents. A classic and popular measure, able to capture the *noise* in a single distribution, is *Shannon's entropy* (Shannon, 1948); the entropy H for a term h is calculated as

$$H_h = -\sum_{j=1}^{n} \frac{f_{jh} \log f_{jh}}{\log n}, \quad 0 \le H_h \le 1, h = 1, \ldots, p,$$

where f_{jh} is the relative frequency of word h in document j, n is the number of documents. Values of H_h close to zero refer to words that are very rare in the considered set of documents; these terms are not informative for clustering and they can also be potentially insidious because of their excess of zeros. Therefore, a natural way to perform variable selection is to set a lower bound for entropy values or to retain the features with higher entropy.

In the second scenario, the first 150 words (equal to the true number of relevant terms) decreasingly ordered according to their entropy values are only retained.

For each setting, one hundred repetitions are run and on each generated dataset, several clustering methods are compared, namely:

- Mixtures of cosine distance-based densities on the raw data;
- Mixtures of cosine distance-based densities on the data obtained from the singular value decomposition of the original ones (Latent Semantic Analysis);
- Mixtures of Euclidean distance-based densities (i.e., mixture of Gaussian distributions, diagonal, equal volume and shape, via `Mclust` R function);
- Mixture of Unigrams.

We have also fitted mixtures of Gaussian distributions, allowing for the maximum flexibility in terms of density shapes (namely, a so-called `VVV` model: ellipsoidal, varying volume, shape, and orientation), but the number of parameters is probably too high to be successfully estimated and the clustering algorithm could not return any solution.

Table 2 contains the average Adjusted Rand Index (ARI) values for every clustering method according to the two scenarios. Results show that our proposed method does not perform differently in the two settings in terms of clustering accuracy: reducing the number of variables by discarding those that are in principle not relevant for clustering purposes has not worsened the results, it has actually slightly improved the capability of recovering the "true" cluster memberships. In fact, despite the fact that the entropy only accounts for individual—rather than joint—word distribution, simulation results show that in a very sparse dataset with many variables this aspect is negligible, because low-entropy variables are those that are almost zero everywhere.

Mixtures of Euclidean distance-based densities often fail to resort to the original clustering structure, yielding to ARIs close to zero. Mixture of Unigrams either do not reach the convergence or allocate all the documents to the same group, thus performing poorly in terms of accuracy. This is probably due to the high sparsity (i.e., on average 99%) of the data. Mixtures of cosine distance-based densities prove to perform very well, on both raw and SVD-transformed data; in other words, the good clustering performance is preserved, regardless whether accounting for relations between terms or not.

4. OVERVIEW OF MAJOR CLUSTERED STATISTICAL TOPICS FROM 1970 TO 2015

The richness of the statistical contributions of the past 45 years is so broad that no clustering can exhaustively describe the variety of topics and ideas; even if it could, the results would be unintelligible anyway.

The clustering procedure described in Section 3 can help disentangling the principal trends that characterized the statistical research of the last half century. The

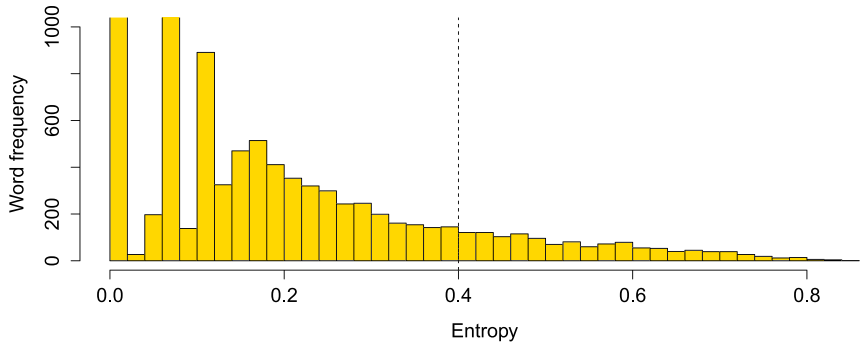

FIG. 2. *Histogram of the word entropies calculated in the complete dataset. The dashed line denotes a cutoff of* 0.4.

idea is to obtain a picture of the top 25 topics that have led the research since 1970. As previously stated, it will not be an exhaustive list: it would rather represent the most important 25 topics (or, at least, important enough to become separate clusters) that have been discussed in the literature so far.

The five chosen journals are standard reference in the statistical literature, and their articles, given the high number of citations, can be considered representative of the main research trends in Statistics. However, each journal has a very distinctive character and the discussed themes could be different. For this reason, we decided to also separately zoom in on the periodicals: distinct analyses allow to identify and distinguish transversal topics, highlighted by every journal, from more specific subjects, that only appeared in a subset.

4.1 Analysis

The global document-term matrix is very large (15,472 × 15,036) and sparse (about 99.76% are zero, as many rare or misspelled words appear in a paper only once in 45 years); these characteristics make it hardly tractable as it is: a variable selection is needed so as to narrow the dimensionality.

As described in Section 3.4, entropy is a measure able to capture the noise in a single distribution; since the number of variables that are relevant for clustering purposes is not known, a natural way to perform variable selection is to set a lower bound for entropy values. In particular in our data, among the total of 15,036 words, 5140 terms (around the 33%) appear just one time in 45 years in the totality of 15,472 papers and they have practically zero entropy. Other 2060 terms appear only two times and their average entropy is 0.0405. Therefore almost half of the words (7200 terms that are the 48% of the total) are so rare that they are basically irrelevant for clustering. By inspecting the entropy distribution of all the words contained in the complete dataset (see Figure 2), we considered only the terms with $H_h \geq 0.40$, it being the approximated middle point between the minimum and maximum observed entropy. The so reduced dataset still presents about the 97.6% of the entries null.

Table 3 contains the number of words for each journal, with and without imposing a constraint on the minimum entropy. Variable selection so performed allowed to deal with much smaller but still meaningful data sets.

For each of the six datasets, we used the cosine-distance k-means algorithm (with $k = 25$, and five different runs) as initial values for the EM algorithm. The desirable average consistency rate in (a) step of the estimation algorithm in Section 3.3 was fixed at $1 - \alpha = 95\%$.

Since the idea was to identify the top 25 topics, the number of groups was set, rather than chosen by some information criteria. For this reason, a few abstracts may have ended up being assigned to a group that

TABLE 3
Number of words for each journal with (first row) and without (second) a threshold on the minimum entropy value H

No. of words	AoS	Bka	JASA	JRSSB	StS	All journals
$H \geq 0.40$	703	632	1021	604	483	**1272**
Unconstrained H	5874	5460	10,769	4561	4349	**15,036**

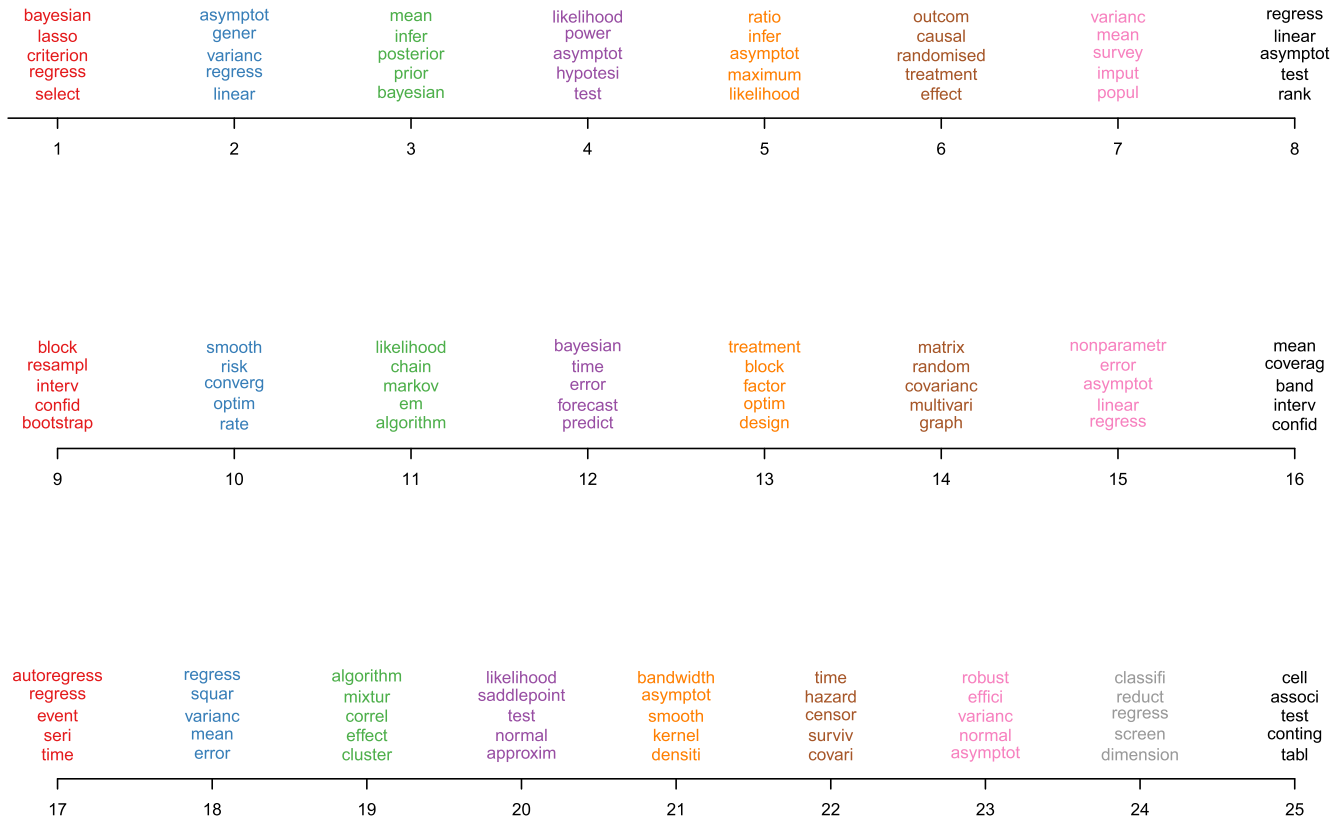

FIG. 3. *Most frequent words (from bottom to top) for the estimated 25 groups of the five considered data.*

is meaningfully not close, but that still represents the nearest one. To describe the homogeneity of the clusters we defined and computed a cohesion index C_i as

$$C_i = \sqrt{1 - \bar{d}_i^{\,2}}, \quad 0 \le C_i \le 1,$$

where $\bar{d}_i$ is the average cosine distance between all papers within cluster i; the closer C_i is to 1, the more homogeneous cluster i is.

We also tried to cluster the data by fitting a mixture of unigrams, but the model allocated all the documents to a single component, failing to find a grouping structure, due to the still very high sparsity of the data (99.76%). Differently, the algorithm fitting a mixture of Gaussian distributions could not reach the convergence.

4.2 Results

To fully characterize the estimated groups and to identify the corresponding topics, we considered the five most frequent words (according to the IDF-corrected frequencies) contained in a group and the most representative paper having the minimum cosine distance from the corresponding centroid. Figure 3 and Table 4 show the results for the analysis conducted on the papers published in the five considered journals from 1970 to 2015.

The balloon plot of Figure 4 gives a picture on the top 25 leading topics published in the five considered journals. The size of the balloons is proportional to the group dimension and the colors are shaded by cohesion index, that is, the lighter the color, the less homogeneous a cluster is. The cluster position has been computed by multidimensional scaling.

It is not surprising that the topic *hypothesis testing* is the biggest group, while other topics, such as *rank data analysis* and *contingency methods* contain less contributions. Some popular research themes are missing. Think at robust estimation, nonparametric methods or classification methods. They have been clearly absorbed by the other estimated groups, that in this sense represent the more important broad themes in the last 45 years. For instance, the major contributions on nonparametric methods are contained in the *density estimation* and *regression models* groups. A more detailed view of the research topics of major interest in the statistical literature can be obtained by a complete analysis on the journals considered separately.

TABLE 4
The most representative paper in each group for the five considered journals

Cluster	Paper
1	J. Huang, P. Breheny and S. Ma, *A Selective Review of Group Selection in High-Dimensional Models*, 2012, StS
2	T. Mathew, *Linear Estimation with an Incorrect Dispersion Matrix in Linear Models with a Common Linear Part*, 1983, JASA
3	F. J. Samaniego and A. Neath, *How to be a Better Bayesian*, 1996, JASA
4	K. V. Mardia, *A Multisample Uniform Scores Test on a Circle and Its Parametric Competitor*, 1972, JRSSB
5	J. D. Kalbfleisch and D. A. Sprott, *Application of Likelihood Methods to Models Involving Large Numbers of Parameters*, 1970, JRSSB
6	A. Richardson, M. G. Hudgens, P. B. Gilbert and J. P. Fine, *Nonparametric Bounds and Sensitivity Analysis of Treatment Effects*, 2014, StS
7	B. C. Arnold and R. A. Groeneveld, *Maximal Deviation between Sample and Population Means in Finite Populations*, 1981, JASA
8	L. Li and W. R. Schucany, *Some Properties of a Test for Concordance of Two Groups of Rankings*, 1975, Bka
9	A. C. Davison, D. V. Hinkley and G. A. Young, *Recent Developments in Bootstrap Methodology*, 2003, StS
10	E. Mammen and A. B. Tsybakov, *Smooth Discrimination Analysis*, 1999, AoS
11	F. Liang, C. Liu and R. J. Carroll, *Stochastic Approximation in Monte Carlo Computation*, 2007, JASA
12	K. Skouras and A. P. Dawid, *On Efficient Point Prediction Systems*, 1998, JRSSB
13	J. Kunert and R. J. Martin, *On the determination of optimal designs for an interference model*, 2000, AoS
14	K. Khare and B. Rajaratnam, *Wishart distributions for decomposable covariance graph models*, 2011, AoS
15	X. Chen and R. D. Cook, *Some insights into continuum regression and its asymptotic properties*, 2010, Bka
16	P. H. Peskun, *A New Confidence Interval Method Based on the Normal Approximation for the Difference of Two Binomial Probabilities*, 1993, JASA
17	J. H. Stock, *Estimating Continuous-Time Processes Subject to Time Deformation*, 1988, JASA
18	D. H. Richardson and D. M. Wu, *Least Squares and Grouping Method Estimators in the Errors in Variables Model*, 1970, JASA
19	R. Peck, L. Fisher and J. Van Ness, *Approximate Confidence Intervals for the Number of Clusters*, 1989, JASA
20	S. J. Skates, *On Secant Approximations to Cumulative Distribution Functions*, 1993, Bka
21	A. J. Izenman, *Review Papers: Recent Developments in Nonparametric Density Estimation*, 1991, JASA
22	D. Oakes, *The Asymptotic Information in Censored Survival Data*, 1977, Bka
23	C. D. Kershaw, *Asymptotic Properties of $\bar{w}$, an Estimator of the ED50 Suggested for Use in Up-and-Down Experiments in Bio-Assay*, 1985, AoS
24	K. Chen, K. Chen, H. G. Müller and J. L. Wang, *Stringing High-Dimensional Data for Functional Analysis*, 2011, JASA
25	H. H. Ku, R. N. Varner and S. Kullback, *On the Analysis of Multidimensional Contingency Tables*, 1971, JASA

Table 5 contains the top 25 leading topics published in the five journals since 1970, sorted by homogeneity. The first column shows the cluster id; the number of abstracts and the cohesion index for each group are reported in third and fourth column, respectively. Tables 6, 7, 8, 9 and 10 include similar information for AoS, Bka, JRSSB, JASA and StS, respectively.

Classic themes (like *hypothesis testing*, *confidence intervals*, *regression models*, *Bayesian analysis*, *design of experiments*, *time series*) formed separate clusters in all the journal-specific analysis. *Hypothesis testing* constitutes the biggest (with more than 2500 papers) and most homogeneous group of abstracts, its relevance is remarkable. Regression and graphical models are fairly big groups, carrying about a thousand papers each. Other topics in Table 5 can be found in many of the considered journals; these are broad concepts and of general interest, like—among others—*estimation algorithms*, *prediction analysis*, *density estimation*, *treatment effect*, *graphical models*, *variable selection*, *bootstrap methods* and *classification*.

Contingency table, *dimension reduction*, *asymptotic properties of estimators* are examples of more specific arguments: their contribution is globally remarkable (hundreds of papers with a good degree of cohesion) despite their resonance echoed in only one of the five journals.

Although they did not result in the global clustering, other statistical topics (e.g., *variance estimation*, *robustness*, *spatial statistics*, *missing data*, *networks*) were relevant for JASA and for a few other periodicals. *The Annals of Statistics* has some impor-

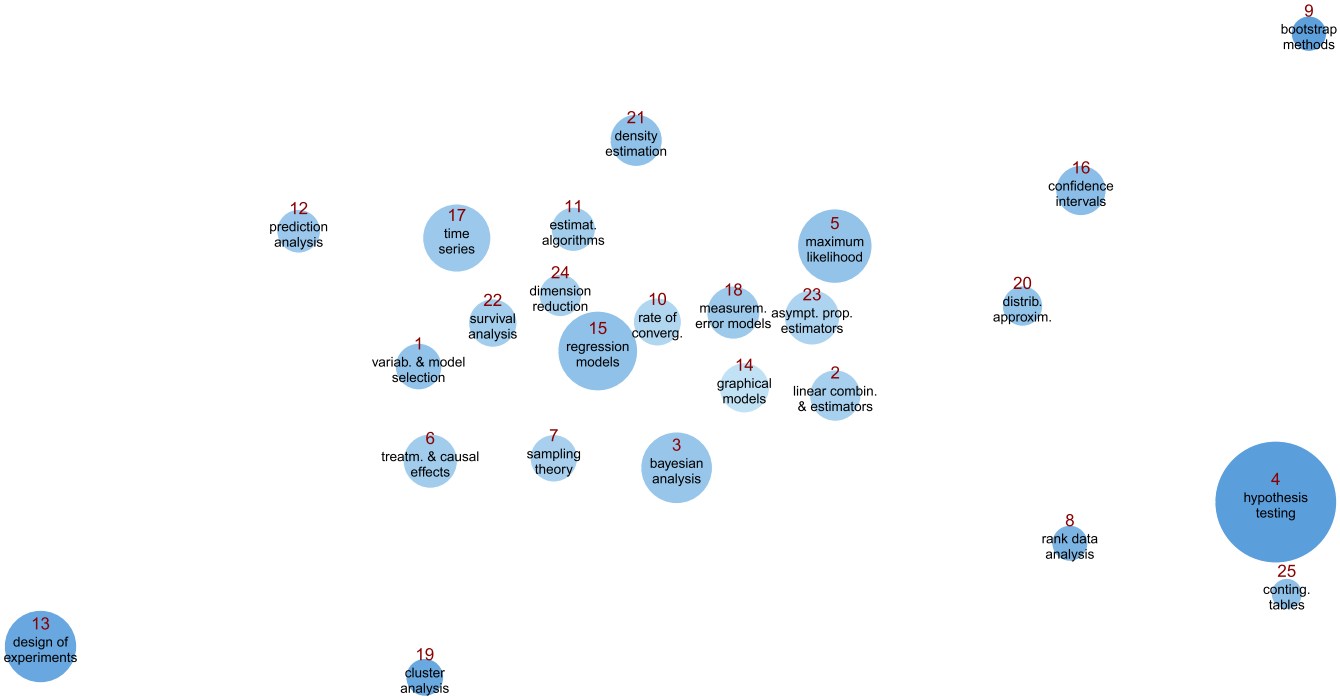

FIG. 4. *Balloon plot of the top* 25 *leading topics published in the five considered journals.*

TABLE 5

Top 25 *leading topics published in the five statistical journals* (The Annals of Statistics, Biometrika, Journal of the American Statistical Association, Journal of the Royal Statistical Society, Series B *and* Statistical Science), *from* 1970 *to* 2015

Cluster	Topic	No. of papers	Cohesion index
4	Hypothesis testing	2516	0.82
9	Bootstrap methods	214	0.82
13	Design of experiments	912	0.77
19	Cluster analysis and model-based clustering	247	0.77
8	Rank data analysis	225	0.72
16	Confidence intervals	439	0.67
15	Regression models	1105	0.65
21	Density estimation	477	0.65
25	Contingency tables	173	0.65
5	Maximum likelihood	970	0.64
1	Variable and model selection	397	0.63
20	Distribution approximation methods	313	0.62
3	Bayesian analysis	917	0.61
12	Prediction analysis	354	0.60
17	Time series	858	0.59
18	Measurement error models	553	0.57
22	Censored data and survival analysis	425	0.57
11	Estimation algorithm: EM, MCMC, ...	378	0.56
2	Linear combination and estimators	487	0.55
6	Treatment and causal effects	594	0.55
24	Dimension reduction	346	0.55
23	Asymptotic properties of estimators	568	0.53
7	Sampling theory	455	0.52
10	Rate of convergence determination	515	0.47
14	Graphical models	1034	0.31

TABLE 6
Top 25 *leading topics published in* The Annals of Statistics, *from* 1970 *to* 2015

Cluster	Topic	No. of papers	Cohesion index
1	Design of experiments: optimality	251	0.85
15	Quantile estimators	60	0.82
21	Graphical models	64	0.82
17	Bootstrap methodology	78	0.81
24	Hypothesis testing	525	0.81
5	Models and methods for rank data	72	0.75
9	Prediction functions and predictive models	59	0.73
19	Priors in Bayesian analysis	100	0.73
6	Density estimation	191	0.72
13	Bayesian analysis	127	0.72
20	Regression models	276	0.70
25	Confidence intervals	105	0.68
12	Maximum likelihood	240	0.67
3	Estimation algorithms: convergence and properties	72	0.63
16	Time series	132	0.62
18	Rate of convergence determination	161	0.61
22	Asymptotic properties of estimators	279	0.61
23	Probabilistic lower and upper bounds	102	0.60
2	Covariance matrix estimation	105	0.59
10	Optimal decision rule and optimality criteria	134	0.59
7	Linear models and combinations	131	0.57
4	Distribution approximations (saddlepoint, Laplace, ...)	59	0.54
8	Exponential family properties	78	0.53
11	Admissible minimax and other mean estimators	148	0.50
14	Random variable probabilistic results	162	0.44

tant clusters on probabilistic and inferential issues, Biometrika on population selection and inference, the JRSSB on generalized linear models and probability; *Statistical Science* is the only journal where genomic and genome-wide association studies groups appeared, together with clinical trials and, due to its peculiar character, statistics historical papers and Fisher's work reviews.

4.3 A Closer Look to the Top 25 Topics

Figure 4 graphically provides a simplified representation of the 25 most relevant topics in the statistical research of the last 45 years. The obtained groups have different size and present different degree of homogeneity; therefore, a more extensive description of the identified topics is provided in the following.

Group 1 collects articles and reviews of methods for model and variable selection, including lasso, penalty introduction and group methods. Many documents refer to the stability selection topic, on how to improve the performance of a variable selection algorithm as well as to control the stability of results after variable selection.

The second cluster deals with linear models, but also with both approximately and contaminated linear models. It includes works on linear combinations of variables and pattern of linear regressions among set of variables. Within the same group, also nonparametric regression problems are discussed and several articles are about the (generalized partially) linear single-Index models.

Cluster 3 includes more than 900 articles on the topic of Bayesian analysis: from the empirical Bayes to Bayesian procedures, from Bayesian analysis for regression models (e.g., Student t regression model) to Bayesian methods for model selection. Under the same broad theme, also some Bayesian aspects of nonparametric problems are casted, for example, nonparameteric regression and signal estimation.

Hypothesis testing is the biggest group (group 4) with more than 2500 articles and contains, indeed, a variety of more specific themes. For example, there are several works on tests in the context of high-dimensional problems, on testing for probability density functions, on both bootstrap and randomization tests. An important fraction of documents deals with

TABLE 7
Top 25 *leading topics published in* Biometrika, *from* 1970 *to* 2015

Cluster	Topic	No. of papers	Cohesion index
18	Hypothesis testing	614	0.84
23	Design of experiments	261	0.84
4	Cluster analysis	41	0.82
2	Priors in Bayesian analysis	77	0.77
15	ARMA models	75	0.74
24	Maximum likelihood	293	0.73
12	Prediction analysis	55	0.71
19	Correlation measures and estimation	72	0.71
14	Bayesian analysis	108	0.70
25	Distribution approximation methods	100	0.67
5	Smoothing splines	51	0.66
8	Time series	196	0.64
10	Regression models	221	0.64
21	Density estimation	65	0.64
1	Treatment effects	91	0.62
3	Confidence intervals	88	0.62
22	Censored data and survival analysis	120	0.62
11	Variance estimation	103	0.61
6	Estimation algorithms: EM, generalizations and MCMC	56	0.60
16	Missing data	69	0.58
17	Comparison and selection of populations	94	0.58
7	Mean square error of estimators	118	0.52
20	Conditional inference	91	0.50
9	Covariance matrix estimation	128	0.47
13	Classification methods	64	0.46

the differences between some parametric and nonparametric testing procedures. In addition, this cluster includes, among others, papers on tests for multivariate problems, on the test of Perlman vs. Likelihood Ratio Test, and on multisample uniform scores tests.

Cluster 5 on maximum likelihood is fairly big, as it includes almost a thousand papers. The topic is obviously wide and includes, among others, works on marginal, conditional and empirical likelihood, on the estimation of pseudo maximum likelihood, on applications of likelihood methods to models with a large number of parameters or in small samples, but also on the (asymptotic) properties of some ML estimators under both standard and nonstandard conditions.

The sixth group aggregates articles on treatment and causal effects, and the investigated research aspects are various: from sensitivity analysis and nonparametric bounds of treatment effects to model selection for the estimation of treatment effects, from the study of differential effects and generic biases in observational studies to the formulation of criteria for surrogate end points.

Cluster 7 on sampling theory contains contributions on different aspects of the research theme. For example, it includes works on some linear interpolation estimators of the total population, on the sequential importance sampling for fitting population dynamic models, on estimators for the finite population distribution function but also on new procedures for selecting good populations.

Cluster 8 on rank data analysis is fairly small and compact and collects articles on several features, including hypothesis testing, distributional shape (in heterogeneous and correlated groups), inferential problems, properties of estimators and real data applications.

Bootstrap methods constitute group 9. It contains about 200 papers that provide several contributions to the topic: the use of bootstrapping for highly accurate parametric inference, theoretical properties of nonparametric bootstrap with unequal probabilities, bootstrap failures and remedies for superefficient estimators, bootstrap improvements of unstable classifiers, resampling for dependent data, more efficient compu-

TABLE 8
Top 25 *leading topics published in the* JASA, *from* 1970 *to* 2015

Cluster	Topic	No. of papers	Cohesion index
15	Hypothesis testing	1060	0.84
4	Forecast methods in applied contexts	83	0.82
8	Cluster analysis and model-based clustering	128	0.80
1	Rank data analysis	93	0.74
17	Missing data imputation	91	0.71
13	Design of experiments	206	0.68
20	Spatial data analysis	104	0.68
23	Confidence intervals	208	0.68
22	Contingency tables	104	0.67
5	Regression models	486	0.66
25	Variable and model selection	165	0.66
2	Nonparametric density estimation	134	0.65
6	Censored data and survival analysis	117	0.65
24	Measurement errors	270	0.63
18	Variance estimation	185	0.62
3	Bayesian analysis	306	0.61
19	Time series	408	0.61
7	Sampling theory	190	0.60
9	Smoothing methods	118	0.59
14	Distribution approximations	119	0.59
16	Treatment and causal effects	303	0.58
21	Maximum likelihood	258	0.57
11	Network and graphical models	167	0.44
10	Robust estimation	333	0.37
12	Risk rate estimation	273	0.36

tational methods, bootstrap model selection, to cite a few.

Cluster 10 focuses on the determination of (optimal) convergence rates of some estimators, that includes, for example, problems of density estimation, nonparametric regression, signal recovery, finite mixture models, or estimators of shape and scale parameters in distributions with regularly varying tails or extreme value index.

A whole cluster of estimation algorithms is identified in group 11. The documents mostly relate to MCMC and EM algorithms and their variants, like stochastic approximation Monte Carlo (SAMC) algorithm, stochastic EM, TM algorithm for the maximization of a conditional likelihood function, MM (majorize-minimize) algorithms, but also data augmentation algorithms and their sandwich variants.

Cluster 12 collects various documents about prediction: from the properties of point prediction systems to the best predictive estimator (BPE) in the context of small area, from frequentist prediction intervals and predictive distributions to prediction functions for categorical panel data, among others.

The research and the development of optimal experimental designs is the leading topic of cluster 13. Documents in this group deal, for example, with the determination of optimality for some (complete or incomplete) block designs, for interference models, for rational models and weighted polynomial regression but also with orthogonal and nearly orthogonal designs for computer experiments.

Graphical models constitute a rather numerous cluster, with about 1000 documents. The contributions are various, from both Bayesian and frequentist perspectives; some works are on the graph estimation with joint additive model, many others on the (flexible) covariance estimation in graphical Gaussian models and on Matérn class of cross-covariance functions for multivariate random fields, while only a few consider the graph estimation with matrix-variate instances.

Cluster 15 collects more than a thousand articles on regression models. As expected, the contributions are various, and deal with the fitting, the diagnostic, the depth, the significance and the asymptotic properties of a regression model. In addition, different kinds of regression model are discussed, including—but not

TABLE 9
Top 25 *leading topics published in the* JRSSB, *from* 1970 *to* 2015

Cluster	Topic	No. of papers	Cohesion index
6	Bootstrap methods	34	0.84
14	Design of experiments	156	0.84
15	Hypothesis testing	229	0.84
2	Extreme value analysis	31	0.78
18	Spatial statistics	49	0.77
19	Confidence intervals	53	0.73
24	Smoothing methods	46	0.73
21	Maximum likelihood	163	0.71
7	Priors in Bayesian analysis	65	0.68
11	Variable selection	30	0.68
9	Regression models	131	0.66
3	Bayesian analysis	88	0.65
23	Model selection methods and criteria	37	0.65
17	Treatment and causal effects	v63	0.64
22	Time series	152	0.64
25	Density estimation: semiparametric and nonparametric	63	0.64
8	EM and MCMC algorithm	84	0.63
16	Distribution approximations (saddlepoint, ...)	52	0.63
10	Measurement error problems	66	0.60
1	Generalized linear models	91	0.59
4	Dimension reduction and variable selection	50	0.57
5	Sampling theory	44	0.57
12	False discovery rate	39	0.54
13	Association and conditional independence	52	0.46
20	Multivariate normality assessment and extensions	82	0.45

limitedly to—continuum regression, least squares regression, isotonic regression, nonparametric regression, piecewise regression, models with incidental parameters or measurement errors.

Confidence intervals as a research topic are examined in group 16. Major contributions can be further grouped according to the objective of the interval estimation (variance, quantiles, proportions, mean, ...), but also to the parametric vs. nonparametric perspectives.

Time series is the subject of cluster 17 and includes, among others, papers on state-space models, on bilinear time series models, on the techniques for normalization and self-normalization of a time series, on regression models and spectral analysis for categorical time series, on the estimation of parameters in multivariate time series models and nonstationarity.

Cluster 18 includes statistical contributions in many contexts that have to deal with measurement errors, like linear autoregressive models, polynomial regressions, regressions with covariate measurement error, but also estimation problems for parametric and semiparametric measurement error models.

Cluster 19 aggregates documents on unsupervised classification, that goes from the proposal of new clustering approaches to the significance of clustering, from the choice of the number of clusters to the consistency of the clustering methods.

Distribution approximation methods constitute a separate group, collecting numerous contributions on saddlepoint approximations, tilted-exponential approximations, fully exponential Laplace approximations, to name a few, for several quantities, like normalizing constants, density and cumulative distribution functions, expectations and variances, marginal tail probabilities.

Units belonging to cluster 21 are papers about density estimation. In particular, nonparametric density estimation, via histogram, kernel, orthogonal series, and parametric density estimation are studied under several perspectives and in different contexts including, but not limitedly to, asymptotic properties, computational aspects, merging information, particular features of the support.

Cluster 22 collects almost 500 research papers about features and models for censored survival data. Sig-

TABLE 10
Top 25 *leading topics published in* Statistical Science, *from* 1970 *to* 2015

Cluster	Topic	No. of papers	Cohesion index
6	Bootstrap methodology	18	0.96
2	Network modelling	18	0.85
23	Confidence intervals	7	0.84
7	R. A. Fisher	20	0.83
22	Causal inference	19	0.82
14	Hypothesis testing	55	0.81
20	Clinical trials	18	0.81
5	Graphical models	18	0.78
1	Bayesian analysis	65	0.76
15	Priors in Bayesian analysis	20	0.76
3	Markov Chain Monte Carlo methods	21	0.75
13	Design of experiments	24	0.70
11	Estimation algorithms	34	0.69
19	Random and treatment effects	27	0.68
25	Regression models	41	0.67
24	Risk assessment problems	23	0.63
8	Missing data problems	16	0.61
10	Birnbaum Argument and likelihood	25	0.61
17	Prediction problems and methods	26	0.61
4	Analysis of genetic data and genomics	14	0.58
21	Analysis of temporal data: time series and survival analysis	22	0.53
9	Spatial statistics	26	0.50
12	Genome-wide association studies	22	0.46
16	Computing environments for data analysis	34	0.45
18	Reviews and statistics historical papers	38	0.34

nificance tests and statements of probabilistic inference for covariate parameters and the hazard function in specific contexts, multivariate survival functions in longitudinal studies, semiparametric survival models, Cox regression models with time-dependent covariates, model averaging for the Cox regression model, all constitute only a few examples of the variety of the contributions on the research topic.

Documents in group 23 are characterized by a more probabilistic flavor, as they are about the asymptotic properties of estimators. For example, many contributions study the asymptotic behavior of least squares estimators, others the asymptotic ancillarity for stochastic processes, some others the asymptotic relative efficiencies of certain location parameter estimates, to name a few.

Cluster 24 focuses on dimension reduction in different contexts: from pattern recognition to functional analysis, from multivariate responses to kernel regression, to give a few examples.

Finally, documents in cluster 25 involve contingency tables, with contributions on the estimate of interaction effects, on the analysis of latent structures, on the estimation of expected frequencies, on the analysis via log linear models, on the geometry, on the pattern of association and, of course, on hypothesis testing.

5. DYNAMIC CLUSTERING

The main statistical research topics naturally born, evolve or die during time. New interesting topics could emerge at some time, they may capture research interest and related developments, and at some point they could diverge into something else or they could disappear. Moreover, the life process of a topic may have a variable length.

Blei and Lafferty (2006) proposed an approach to perform dynamic topic modeling by fitting a Gaussian time series on the natural parameters of the multinomial distributions that represent the topics. The approach aims to analyze the topic evolution and it seems promising in terms of topic prediction; however it does not describe how new topics appear or disappear over time, mainly because it relies on the restrictive assumption of a fixed number of topics over time.

In order to both address this issue and describe the formation and the evolution of the topics, a dynamic clustering strategy based on semi-supervised mixtures is here developed (Ambroise and Govaert, 2000, Côme et al., 2009, Zhu et al., 2009, Vandewalle et al., 2013). We assume a forward evolution perspective, that means documents at the present temporal frame are projected on a future classification frame. Imagine several documents are classified around k_1 main topics in a certain temporal interval and a second set of documents is classified around k_2 topic clusters in a subsequent temporal interval. We are interested in studying the evolution of the first set of documents towards the future k_2 groups. In order to capture the dynamic forward process, the first-interval documents can be projected into the second-interval classification and the dynamic is considered "relevant" if the relative fraction of movement is above a specified cutoff.

More formally, suppose different sets of documents $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(T)}$, with dimensions $(n_1 \times p_1), \ldots, (n_T \times p_T)$ respectively, are collected in different time spans, say $1, \ldots, t, \ldots, T$. Assume $\mathbf{y}^{(t)}$ at time t is clustered into k_t groups and $\mathbf{y}^{(t+1)}$ into k_{t+1} groups. Since $p_t \neq p_{t+1}$ we first transform $\mathbf{y}^{(t)}$ into a "matched" version $\tilde{\mathbf{y}}^{(t)}$ such that it has dimension $n_t \times p_{t+1}$ by discarding the "missing in future" columns and artificially adding the eventual new term columns with frequency zero. Then a semi-supervised version of the mixture model (3) is fitted on $\mathbf{y} = (\tilde{\mathbf{y}}^{(t)}, \mathbf{y}^{(t+1)})$ with k_{t+1} components and known labels, $\mathbf{z}^{(t+1)}$, for $\mathbf{y}^{(t+1)}$. The log-likelihood to be maximized with respect to $\boldsymbol{\pi}$ only is

$$
\begin{aligned}
&\ell\big(\boldsymbol{\pi}; \boldsymbol{\xi}, \tilde{\mathbf{y}}^{(t)}, \mathbf{y}^{(t+1)}, \mathbf{z}^{(t+1)}\big) \\
&\quad = \sum_{j=1}^{n_t} \log \sum_{i=1}^{k_{t+1}} \pi_i \big(e^{-\lambda d(\tilde{y}_j^{(t)}, \xi_i)}\big) \\
&\quad\quad + \sum_{j=1}^{n_{t+1}} \sum_{i=1}^{k_{t+1}} z_{ji}^{(t+1)} \big(\log \pi_i - \lambda d\big(y_j^{(t+1)}, \xi_i\big)\big) \\
&\quad\quad + (n_t + n_{t+1}) \log \psi(\lambda).
\end{aligned} \tag{8}
$$

The dynamic parameter estimation can be obtained by an adapted version of the EM-algorithm previously described, where the allocation variable is $\mathbf{z}^{(t)}$ is latent and estimated in the E-step, while $\mathbf{z}^{(t+1)}$ is known. The precision parameter is taken fixed for all t. The M-step for $\boldsymbol{\pi}$ does not change, while the previously estimated centroid, $\boldsymbol{\xi}$, are taken as known. This naturally produces a classification of the first set of units $\tilde{\mathbf{y}}^{(t)}$ into k_{t+1} groups. The projected new classification, say $\tilde{c}_{t+1}$ can be compared to the classification originally obtained at t, c_t. The confusion matrix in Table 11 contains the frequencies of the documents collected at t that are allocated to a certain cluster at time t and $t+1$.

TABLE 11
Number of statistical papers $\mathbf{y}^{(t)}$ classified in k_t and k_{t+1} groups

	$\tilde{c}_{t+1}$			
c_t	1	2	$\cdots$	k_{t+1}
1	n_{11}	n_{12}	$\cdots$	$n_{1k_{t+1}}$
2	n_{21}	n_{22}	$\cdots$	$n_{2k_{t+1}}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
k_t	n_{k_t1}	n_{k_t2}	$\cdots$	$n_{k_tk_{t+1}}$

By denoting with $n_{uv}, u = 1, \ldots, k_t; v = 1, \ldots, k_{t+1}$ the number of papers that are allocated into the group u at time t and in group v at time $t+1$, the relative frequencies $f_{uv} = \frac{n_{uv}}{\sum_{v=1}^{k_{t+1}} n_{uv}}$ measure the migration of documents from cluster u to cluster v. When $f_{uv} > s$, where s is a cutoff value between 0 and 1, we can reasonably consider v as a pursuance of u and establish a dynamic connection between the two clusters.

5.1 Analysis and Results

To study the evolution of the leading statistical topics, we need to first cluster the abstracts separately for temporal intervals; the datasets are quite similar in terms of both the number of documents (except for the last period that has fewer abstracts since it extends for six years only) and the number of words. In order to reduce the dimensionality, a variable selection is performed, similarly to that described in Section 4. Words are sorted according to their entropy and in order to assure a comparable dimension problem over time, the first 700 words are retained: this allows us to consider terms with, on average, an entropy of at least 0.40. Table 12 reports the number of terms for each interval with (first row) and without (second row) a 0.40 entropy threshold.

The number of topics that have led each decade can be different, according to the themes that characterized the intervals; in order to fully and properly describe the temporal evolution, several models with different number of clusters are estimated on each dataset. In particular, we run the EM algorithm of Section 3.3, allowing for a number of clusters k varying from 2 to 20.

In principle, the number of clusters can be selected according to classical information criteria such as the Bayesian (BIC) and the Akaike (AIC) information criteria. However, since the normalization constant is un-

TABLE 12
Number of words for each time span, with and without a threshold on entropy value H

No. of words	1970–1979	1980–1989	1990–1999	2000–2009	2010–2015
Unconstrained H	5148	6525	7329	7907	6339
$H \geq 0.40$	536	642	798	860	774

known the conventional criteria cannot be used to compare models characterized by different λs. In order to overcome this limit and, therefore, to take advantage of the information criteria for model selection, we estimate a single precision parameter by averaging multiple λs. In particular, we run several k-means algorithm with an increasing number of groups, say from 2 to 20; for each classification $\hat{\lambda}$ is estimated according to equation (4), by considering $\alpha = 0.05$. The precision parameter to be used in the dynamic clustering for each dataset is, therefore, obtained by computing the mean of the 19 values. The classification and the cluster prototypes from the k-means clustering are used as initial values for the EM algorithm.

Once the documents of each interval are classified according to the most appropriate number of groups (according to the AIC they are 16,16,18, 20 and 15 in the five time intervals respectively), the dynamic clustering described in Section 5 can be performed. Group characterization is then possible by studying the most frequent words and the abstracts closer to prototypes.

Figure 5 represents the dynamic clustering. For each time span, displayed in columns, the corresponding clusters are plotted; if a topic persists in the following time span, an arrow will link that group with its subsequent. Topics may originate and disappear in the same decade, or the may evolve into something different. For a topic to survive it needs to have at least the 40% of its abstracts projected into a single cluster. Dashed arrows link groups whose relative fraction of movement is between 0.40 and 0.70, whereas solid arrows link groups whose percentage of projected documents is larger than 70%.

Some topics are evergreen: *hypothesis testing*, *time series*, *design of experiments*, *maximum likelihood*, *regression models* are themes that, despite having changed quite a lot in the past, have never lost their centrality in the statistical literature since 1970. *Bayesian analysis* covered the whole period as well; in the decade 1980–1989, a growing interest allowed to obtain a separate cluster on the *Bayesian prior*, that afterward converged to the general Bayesian one. Figure 6 shows the dynamic change of such major cluster size by plotting the average annual number of papers published for each decade.

Differently, *confidence intervals* was a leading topic for three decades, with a special attention to *bootstrap confidence interval* between 1980 and 1989. After that, *bootstrap methods* loomed out as a separate cluster.

Articles on *random effects* became particularly numerous since the decade 1990–1999, incorporating later also mixed and treatment effects. The estimation problem has been very relevant as well: at the very beginning there were three groups about estimation, including a group dedicated only to the *mean and variance estimation*. A separate flow is reserved to the *nonparametric literature*, that emerged in both density estimation and splines contexts.

The dynamic clustering proved to be able to identify the moment when topics like *bootstrap* and *estimation algorithms* (like EM or MCMC) became more popular: although the first theoretical contributions on these topics arose between 1970s–1980s, the scientific production sprang later, from the 1990s, helped by an increasing availability of computing machines.

Similarly, the development in many other areas of scientific research of the last 20 years led to a huge availability of *big data*; this translates to the urge of statistical tools able to deal with such data. In fact, classical multivariate methods found a renewed impulse: from the decade 2000–2009 topics like *cluster analysis*, *spatial analysis*, *dimension reduction*, *model and variable selection* and *classification methods* are expressed as separate trends, highlighting their importance in the recent statistical literature.

A related work is that of Kolar and Taddy (2016) in the discussion of Ji and Jin (2016), where topic analysis is yearly performed on the papers published in four top journals in the 2004–2012 interval. The number of groups is found to be equal to 15 and it is the same for every considered year. A proper comparison with the results presented here is not possible, as the considered intervals do not overlap perfectly and a different number of groups is considered for the 2000–2009 interval.

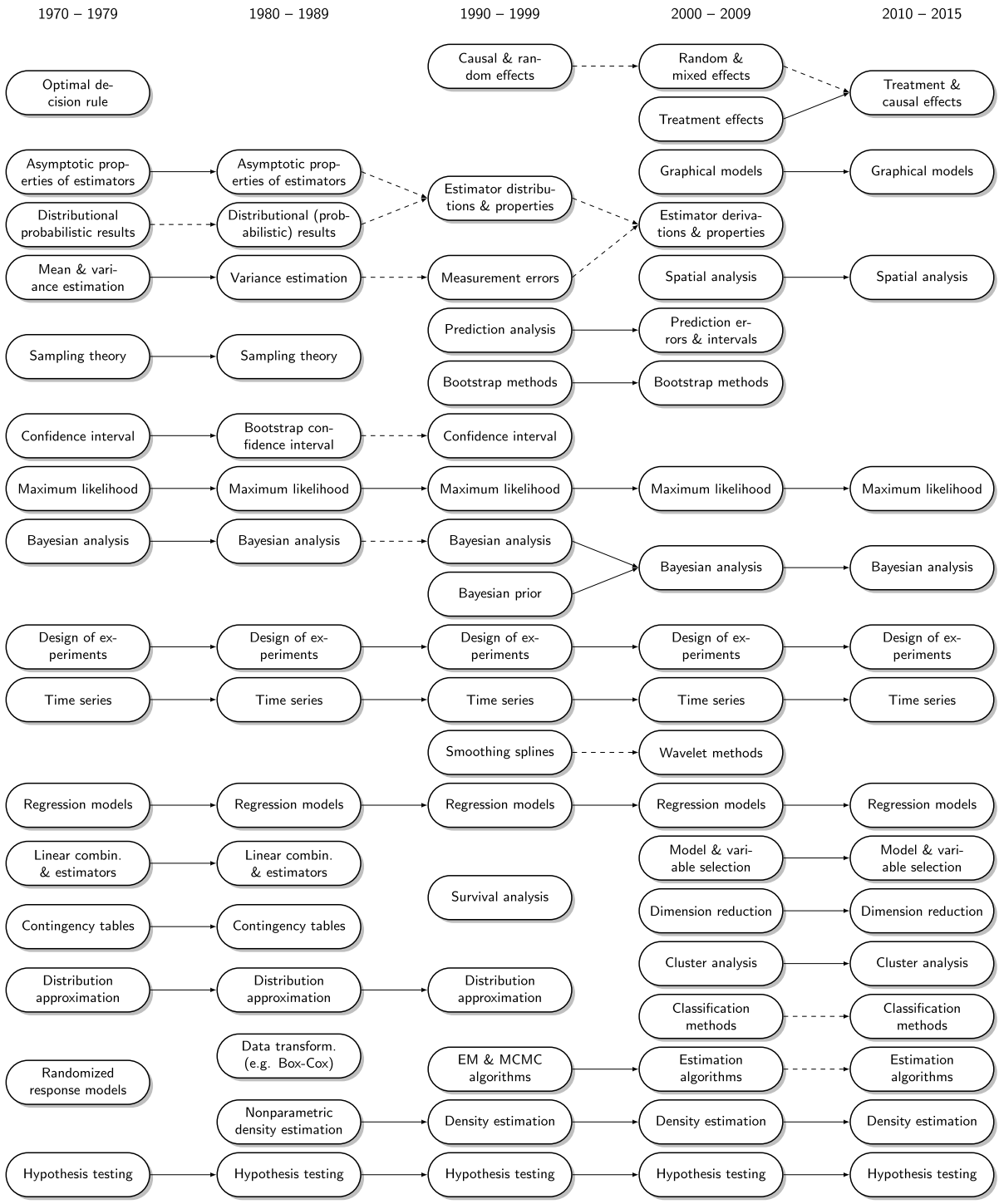

FIG. 5. *Dynamic clustering of abstracts from* 1970 *to* 2015. *Dashed arrows link groups whose relative fraction of movement is between* 0.40 *and* 0.70, *solid arrows between* 0.70 *and* 1.

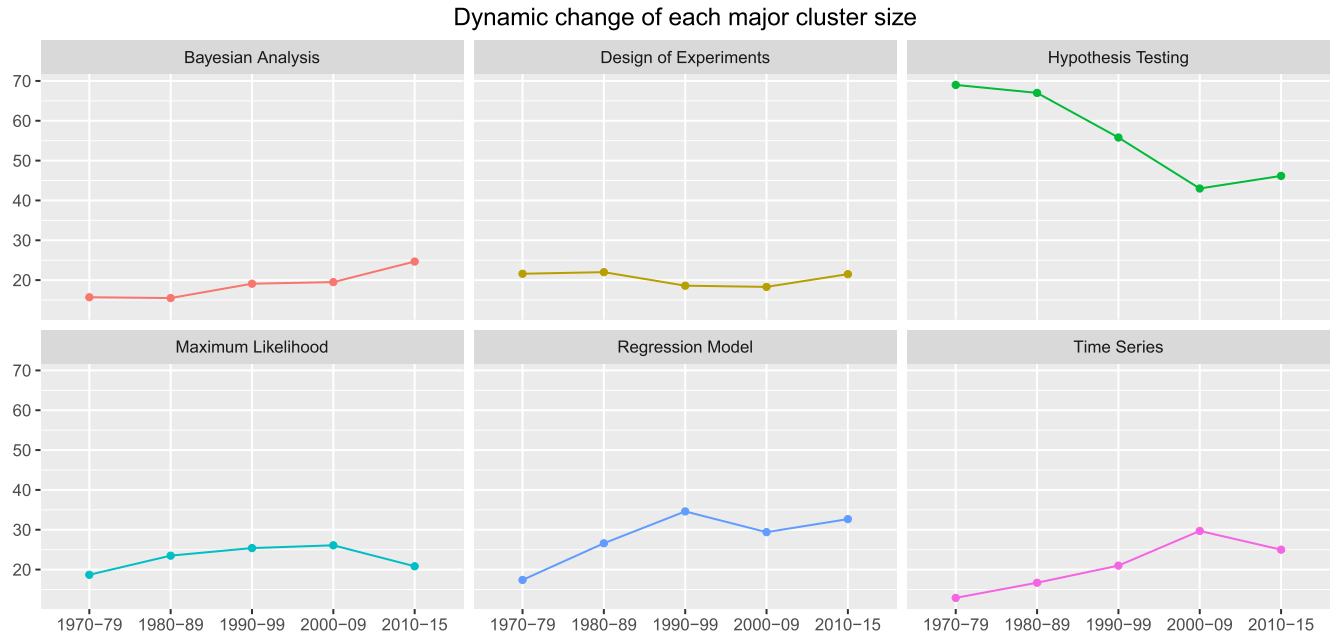

FIG. 6. *Dynamic change of each major cluster size. The points represent the average annual number of articles published every decade on each major topic* (Bayesian Analysis, Design of Experiments, Hypothesis Testing, Maximum Likelihood, Regression Model *and* Time Series).

However, similar results are found. In particular, both analysis give insights about an increasing interest for dimension reduction and variable selection, as well as a renewed interest for classification themes.

6. CONCLUSIONS

We presented a mixture of cosine distance-based densities that allowed us to identify and to describe the 25 most important topics published in five prestigious journals from 1970. The data were collected by considering the title and the abstract of each statistical article published in *The Annals of Statistics, Biometrika, Journal of American Statistical Association, Journal of the Royal Statistical Society, Series B* and *Statistical Science*.

The detected 25 clusters have different size and cohesion, according to the degree of heterogeneity and generality of each topic. In so doing, we obtained a sort of taxonomy of the main statistical research themes discussed in the last 45 years. In addition, we zoomed in on the considered journals: clustering the papers separately for the different journals allowed to distinguish the transversal topics from more specific themes, that resulted in a smaller set of periodicals.

Each classification is a simplification of the reality. Each summary brings new knowledge but, at the same time, implies losses. Our taxonomy is not an exhaustive list of the interconnected topics of the recent statistical research: some of the obtained groups are in fact very general (e.g., maximum likelihood, graphical models, regression models, ...) and include a variety of subthemes that are relevant and interesting, but not important enough to be uncluttered to form separate clusters (e.g., EM and MCMC algorithms joined the larger group of *estimation algorithms*).

Since the main statistical research topics naturally born, evolve or die during time, we also developed a dynamic clustering strategy that allowed to follow the projection of a statistical theme in the following decades. Data were organized in time intervals (1970–1979, 1980–1989, 1990–1999, 2000–2009, 2010–2015). Each period was characterized by a different number of groups, chosen via AIC from a sequence ranging from 2 to 20. Our approach did not aim to spot the true introduction of a statistical topic, rather to detect the moment when a certain theme became "popular" and "trendy", how it evolved, and also its "decadence". Our dynamic clustering strategy is based on semi-supervised mixtures and on mutual comparisons between the actual static classification and the predicted dynamic one in a forward perspective. The idea is pretty simple but really effective in detecting the topic evolution together with the description of how new topics appear or disappear over time.

REFERENCES

AMBROISE, C. and GOVAERT, G. (2000). EM Algorithm for Partially Known Labels. In *Data analysis*, *classification*, *and related methods*, 161–166. Springer, Berlin.

BANERJEE, A., DHILLON, I. S., GHOSH, J. and SRA, S. (2005). Clustering on the unit hypersphere using von Mises–Fisher distributions. *J. Mach. Learn. Res.* **6** 1345–1382. MR2249858

BEN-ISRAEL, A. and IYIGUN, C. (2008). Probabilistic D-clustering. *J. Classification* **25** 5–26. MR2429670

BLEI, D. M. and LAFFERTY, J. D. (2006). Dynamic topic models. In *ICML '06 Proceedings of the 23rd international conference on Machine learning* 113–120. ACM, New York.

BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

BOUVEYRON, C., LATOUCHE, P. and ZREIK, R. (2018). The stochastic topic block model for the clustering of vertices in networks with textual edges. *Stat. Comput.* **28** 11–31. MR3741634

CHANG, J. and BLEI, D. M. (2009). Relational topic models for document networks. In *International Conference on Artificial Intelligence and Statistics* 81–88. Avaialble at http://proceedings.mlr.press/v5/chang09a/chang09a.pdf.

CÔME, E., OUKHELLOU, L., DENŒUX, T. and AKNIN, P. (2009). Learning from partially supervised data using mixture models and belief functions. *Pattern Recognition* **42** 334–348.

DEERWESTER, S., DUMAIS, S., FURNAS, G., LANDAUER, T. and HARSHMAN, R. (1990). Indexing by latent semantic analysis. *J. Amer. Soc. Inform. Sci.* **41** 391–407.

DHILLON, I. S. and MODHA, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Mach. Learn.* **42** 143–175.

DIACONIS, P. (1988). *Group Representations in Probability and Statistics*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **11**. IMS, Hayward, CA. MR0964069

FLIGNER, M. A. and VERDUCCI, J. S. (1986). Distance based ranking models. *J. Roy. Statist. Soc. Ser. B* **48** 359–369. MR0876847

FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. MR1951635

HOFMANN, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 50–57. ACM, New York.

JI, P. and JIN, J. (2016). Coauthorship and citation networks for statisticians. *Ann. Appl. Stat.* **10** 1779–1812. MR3592033

KOLAR, M. and TADDY, M. (2016). Discussion of "Coauthorship and citation networks for statisticians" [MR3592033]. *Ann. Appl. Stat.* **10** 1835–1841. MR3592037

MAITRA, R. and RAMLER, I. P. (2010). A k-mean-directions algorithm for fast clustering of data on the sphere. *J. Comput. Graph. Statist.* **19** 377–396. MR2758308

MALLOWS, C. L. (1957). Non-null ranking models. I. *Biometrika* **44** 114–130. MR0087267

MARDIA, K. V. and JUPP, P. E. (2000). *Directional Statistics*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR1828667

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. *Wiley Series in Probability and Statistics*: *Applied Probability and Statistics*. Wiley Interscience, New York. MR1789474

MURPHY, T. B. and MARTIN, D. (2003). Mixtures of distance-based models for ranking data. *Comput. Statist. Data Anal.* **41** 645–655. MR1973732

NIGAM, K., MCCALLUM, A., THRUN, S. and MITCHELL, T. (2000). Text classification from labeled and unlabeled documents using em. *Mach. Learn.* **39** 103–134.

SALTON, G. and MCGILL, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

SHANNON, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27** 379–423, 623–656. MR0026286

SUN, Y., HAN, J., GAO, J. and YU, Y. (2009). Itopicmodel: Information network-integrated topic modeling. In *Ninth IEEE International Conference on Data Mining* 493–502.

VANDEWALLE, V., BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2013). A predictive deviance criterion for selecting a generative model in semi-supervised classification. *Comput. Statist. Data Anal.* **64** 220–236. MR3061900

VARIN, C., CATTELAN, M. and FIRTH, D. (2016). Statistical modelling of citation exchange between statistics journals. *J. Roy. Statist. Soc. Ser. A* **179** 1–63. MR3461568

ZHONG, S. and GHOSH, J. (2005). Generative model-based document clustering: A comparative study. *Knowledge and Information Systems* **8** 374–384.

ZHU, X., GOLDBERG, A. B., BRACHMAN, R. and DIETTERICH, T. (2009). *Introduction to Semi-Supervised Learning*. Morgan and Claypool, Williston, VT.