

Stochastic modelling of scientific terms distribution in publications

Rimantas Rudzkis, Vaidas Balys, Michiel Hazewinkel

MKM 2006

The Fifth International Conference on MATHEMATICAL KNOWLEDGE MANAGEMENT

THE PROBLEM

The problem of automatic keywords assignment to scientific publications is considered.

Automatic tools designed for management of everyday-growing volumes of scientific knowledge are necessary.

Conventional ways of management became obsolete mostly due to the shift from paper-based publishing to e-publishing.

There is a sound request for such tools from both academic and publishing-related societies.

Keywords is a traditional widely used means for classifying and indexing of scientific texts as well as searching for relevant information in data bases.

THE IDEA

M. Hazewinkel proposed an idea to solve the problem by using so called identification clouds.

Identification cloud of a keyword is a list of scientific terms that are likely to be found in a text that is highly relevant to this keyword.

The keyword – cloud relation is explicit and represents a kind of metaknowledge of scientific field that is of great value itself.

Clouds could be used to solve a number of other scientific texts processing and knowledge extraction related problems.

Cloud items have weights assigned that are proportional to their discriminative power.

Clouds should be as compact as possible. On the other hand, the efficiency of algorithms based on clouds should be comparable to the results of alternative ones.

THE IDEA

A substantial amount of researches related to our problem have been conducted since early 1990s though in rather different fields.

The idea of contextual track of a keyword is quite a common concept. However, term-context relations are often hidden in the state-of-the-art plain text classification algorithms.

We prefer the supervised learning approach, which needs pre-labelled (pre-classified) corpus. The algorithms perform by analysing positive and negative examples of classes and building discriminative rules.

The unsupervised learning methods need no prelabelled data while they reveal latent topics (represented by distributions of words) by analysing distributional patterns of co-occurring data. However, there may be some problems with interpretation of 'unnamed' topics as well as mapping them to the predefined classes.

IDENTIFICATION CLOUDS. NOTATIONS AND DEFINITIONS

Let K be the set of all possible keywords, V be the set of scientific terms, and N be the set of natural numbers.

A vector $a = (a_1, \dots, a_d)$, $a_i \in V$ which is generated by chronologically numerating elements of text that belong to V corresponds to each scientific publication. We call these vectors the articles and by A denote the set of all articles.

We assume that every article $a \in A$ consists of $q = q(a)$ continuous homogenous nonintersecting parts and each of them is classified by a single element $w \in K$. So, intervals of indices $I_j(a) \subset N$ and keywords $w_j(a)$, $j = \overline{1, q}$ correspond to these parts and $\bigcup_{j=1}^q I_j(a) = \{1, 2, \dots, d(a)\}$.

Let an article $a \in A$ and a set of indices $I \subset N$ be chosen randomly so that the part of an article $a_I = \{(a_\tau, \tau), \tau \in I\}$ is homogenous and it is attributed to the class η .

Let Y be the set of all possible values of a_I .

NOTATIONS AND DEFINITIONS

Since (a, I, η) is the result of a random experiment, the following probability distributions are defined:

$$Q(w) = \mathbb{P}\{\eta = w\},$$

$$P(y) = \mathbb{P}\{a_I = y \mid |I| = d(y)\},$$

$$P(y|w) = \mathbb{P}\{a_I = y \mid |I| = d(y), \eta = w\},$$

where $w \in K$, $y \in Y$, $d(y) = \dim y$, $|I| = \text{card } I$.

Functional

$$\psi_w(y) = P(y|w)/P(y), \quad y \in Y$$

defines the concept of the identification cloud of a keyword $w \in K$.

If η and $|I|$ are independent, after observing a_I , the a posteriori probability of the random event $\{\eta = w\}$ satisfies the equality

$$\mathbb{P}\{\eta = w \mid a_I = y\} = Q(w) \cdot \psi_w(y).$$

STATISTICAL INFERENCE

Now we can define a Bayes classifier which minimizes mean classification losses

$$\hat{\eta} = \arg \min_{w \in K} \sum_{v \in K} P(a_I|v)Q(v)l(w, v),$$

where $l(w, v)$ is as usual the loss function, $l(\cdot, \cdot) \geq 0$ and $l(w, w) = 0$.

It is obvious that $\psi_w(\cdot)$ can be substituted for $P(a_I|\cdot)$.

In case the loss function is trivial, i.e.,

$$l(w, v) = \begin{cases} c, & w \neq v \\ 0, & w = v, \end{cases}$$

where $c > 0$, it follows that

$$\hat{\eta} = \arg \max_{w \in K} \psi_w(a_I)Q(w).$$

DISCRIMINANT ANALYSIS. NON-PARAMETRIC ESTIMATION

Suppose that we know keywords of n homogenous texts and also have observations of some parts of these texts.

The sample X is composed of n pairs, $X = (y(1), \eta(1)), \dots, (y(n), \eta(n))$, where $\eta(i) \in K$, $y(i) \in Y$.

The simplest non-parametric estimators of functionals Q and ψ :

$$\widehat{Q}(w) = \frac{1}{n} \sum_{j=1}^n 1_{\{\eta(j)=w\}},$$

$$\widehat{\psi}_w(y) = \frac{1}{\widehat{Q}(w) \cdot k} \sum_{j \in J(y)} 1_{\{\eta(j)=w\}}.$$

Here $J(y) \subset \{1, \dots, n\}$ is a set of k indices of those observations that are closest to y . As usual in k -nearest neighbours algorithms, $k = k(n)$ depends on the size of the sample, $k \rightarrow \infty, k/n \rightarrow 0$, as $n \rightarrow \infty$.

ASSUMPTIONS FOR PARAMETRIC ESTIMATION

Let the index $\tau \in I$ be a random variable. For all $w \in K, v \in V$

$$P(v) = \mathbb{P}\{a_\tau = v\}, \quad P(v|w) = \mathbb{P}\{a_\tau = v|\eta = w\}.$$

Assumption 1 (conditional stationarity and independence). Let for all $y \in Y$ and $w \in K$ the following equality hold

$$P(y|w) = \prod_{i=1}^d P(y_i|w)$$

Assumption 2 (conditional stationarity and Markovian property). Let for all $y \in Y$ and $w \in K$ the following equality hold

$$P(y|w) = P(y_1|w) \prod_{i=1}^{d-1} [P(y_i, y_{i+1}|w)/P(y_i|w)],$$

where $P(v, u|w) = \mathbb{P}\{a_\tau = v, a_{\tau+1} = u|\eta = w\}$

STATISTICAL INFERENCE

For every $v \in V$, $w \in K$ let us denote $\psi_w(v) = P(v|w)/P(v)$.

In case of assumption 1 we can easily derive Bayesian classifier:

$$\hat{\eta} = \arg \max_{w \in K} \left[Q(w) \prod_{\tau \in I} \psi_w(a_\tau) \right].$$

In case of assumption 2 we get a bit more complex expression:

$$\hat{\eta} = \arg \max_{w \in K} \left[Q(w) \psi_w(a_r) \prod_{i=r}^{m-1} [\psi_w(a_i, a_{i+1}) / \psi_w(a_i)] \right].$$

Here $\psi_w(v, u) = P(v, u|w)/P(v, u)$, $v, u \in V$.

We see that in case of assumption 2 it is enough to define functional ψ only for single terms and the pairs of them.

PARAMETRIC ESTIMATION OF ψ IN CASE OF ASSUMPTION 1

Suppose we have the same sample X . Let $d(j)$ denote a dimension of an observed vector $y(j)$.

Let us start from empirical estimates of $P(\cdot)$ and $P(\cdot|\cdot)$:

$$\tilde{P}(v) = \sum_{j=1}^n \sum_{k=1}^{d(j)} 1_{\{y_k(j)=v\}} / \sum_{j=1}^n d(j),$$

$$\tilde{P}(v|w) = \sum_{j=1}^n \sum_{k=1}^{d(j)} 1_{\{y_k(j)=v, \eta(j)=w\}} / \sum_{j=1}^n d(j) 1_{\{\eta(j)=w\}},$$

We obtain empirical cloud $\tilde{\psi}_w(v)$ by substituting the estimates for the corresponding probabilities.

A smoothing of functional $\tilde{\psi}_w$ is then performed as there may be too few observations for some $v \in V$.

PARAMETRIC ESTIMATION OF ψ IN CASE OF ASSUMPTION 1

The functional ψ_w determines the arrangement of V : $V(w) = (v_1, \dots, v_h)$, $h = |V|$, so that $\psi_w(v_1) \geq \psi_w(v_2) \geq \dots \geq \psi_w(v_h)$.

We skip the part of cloud $s < k < h - l$ for which $\tilde{\psi}_w(v_k)$ insignificantly differs from 1, as it is of very low discriminative importance.

Then we apply the chosen parametric model, e.g.,

$$\gamma(k, \theta) = \theta_0 + \theta_1 k^{-\theta_2}, \quad \theta = (\theta_0, \theta_1, \theta_2)$$
$$\log \psi_w(v_k) = \begin{cases} \gamma(k, \theta), & 1 \leq k \leq s, \\ \gamma(k, \theta^*), & h - l \leq k \leq h. \end{cases}$$

Estimates for θ (θ^* analogously) are defined by equality

$$\hat{\theta} = \arg \min_{\theta} \sum_{k=1}^s \left| \log \tilde{\psi}_w(v_k) - \gamma(k, \theta) \right|^{\beta}, \quad \beta \in \{1, 2\}$$

EXPERIMENTAL EVALUATION

The first experimental study was performed on rather poor database consisting of abstracts from the field of computer science.

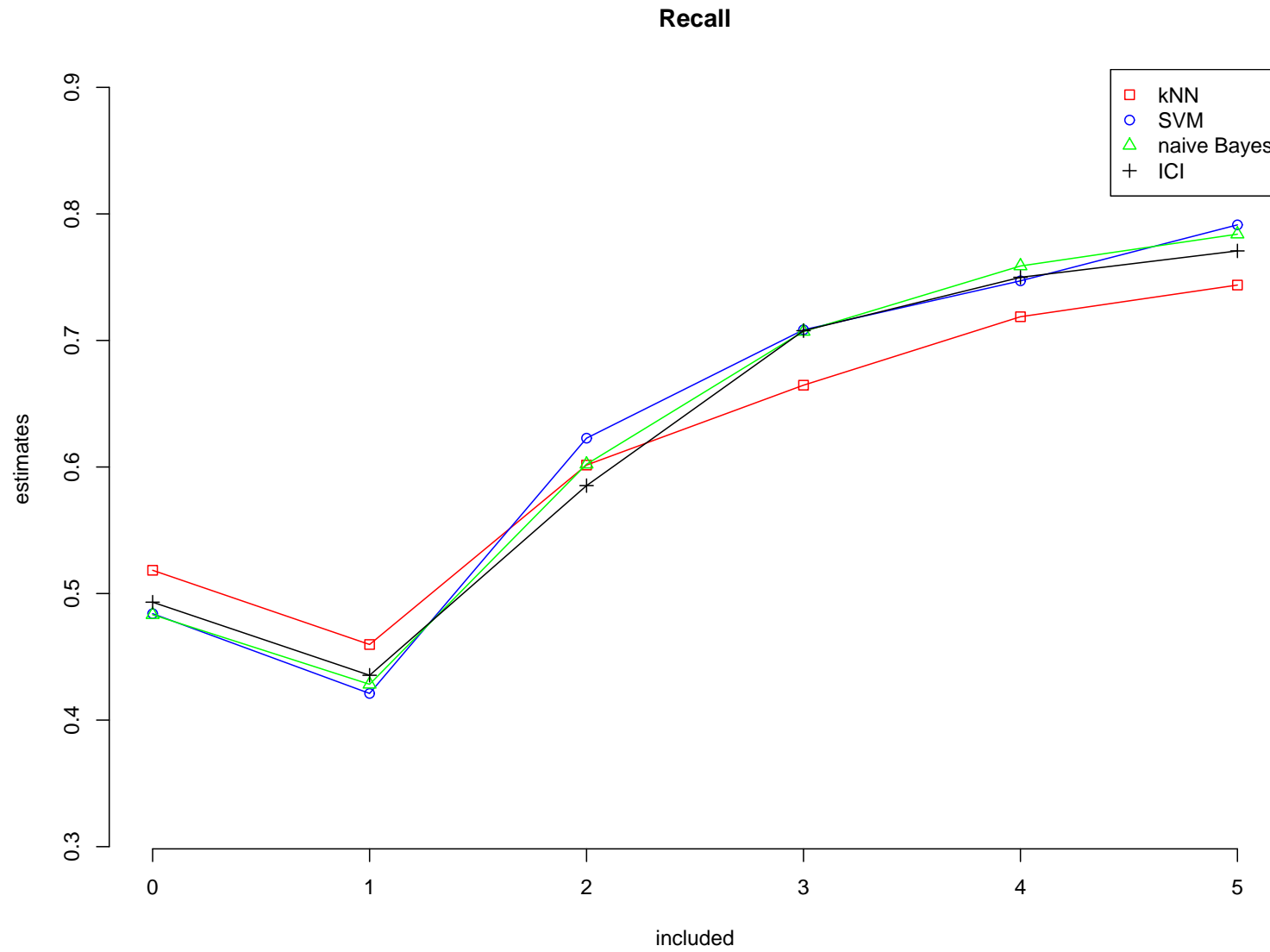
The new database consisting of approximately 14000 full text articles from the field of probability theory and mathematical statistics was kindly provided by IMS and VTEX Ltd. Comprehensive experimental studies are not finished yet, so we present some preliminar results.

- 2904 abstracts
- 2211 scientific terms
- 42 keywords (avg. 1.4 per abstract)
- no splitting into homogeneous parts, assumption of independence
- fixed number of highest-ranked keywords are chosen
- 10-fold cross-validation

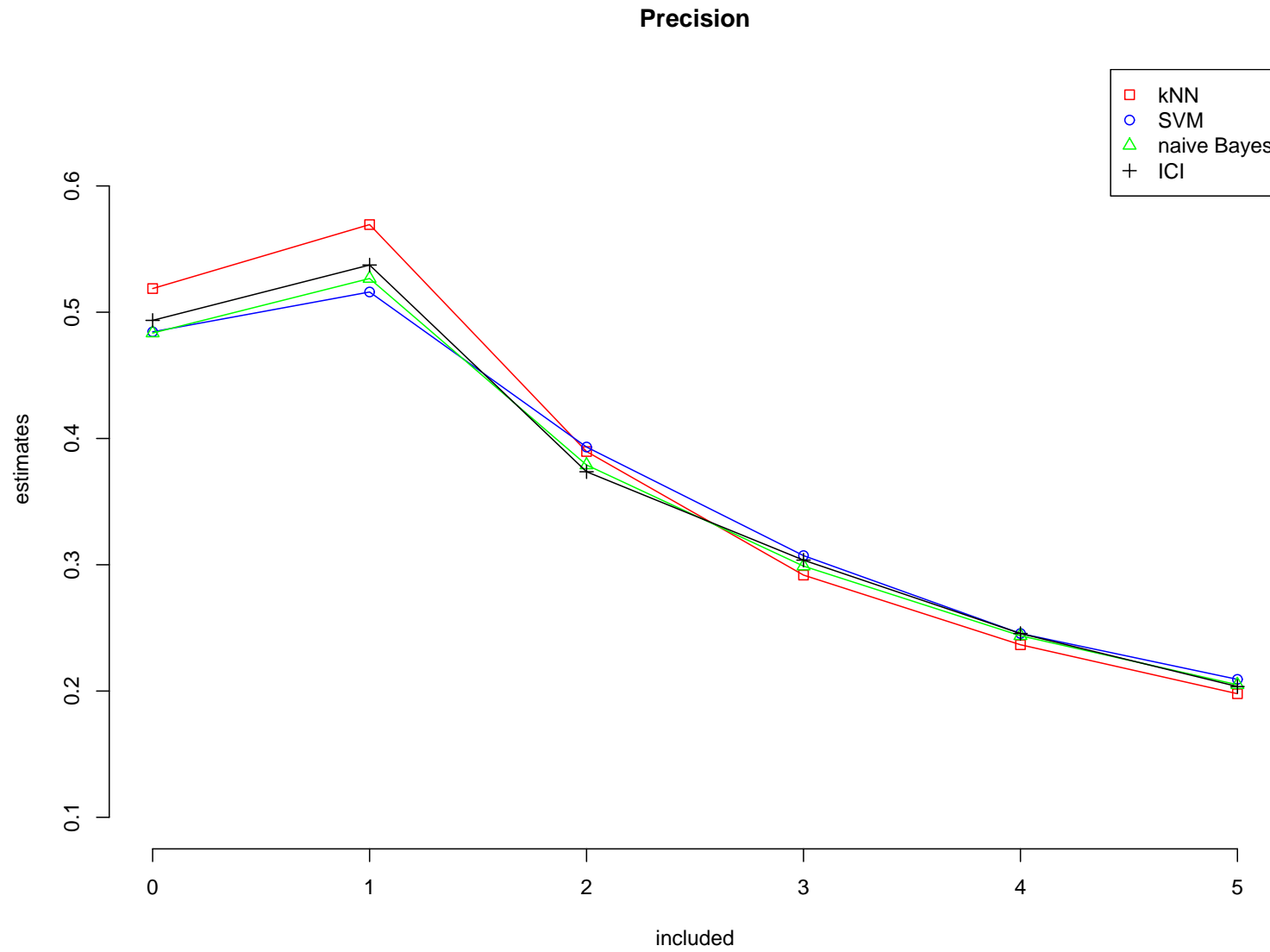
KEYWORDS

admissibility, asymptotic distribution, asymptotic efficiency, asymptotic expansion, asymptotic normality, bootstrap, brownian motion, censored data, central limit theorem, confidence interval, consistency, contact process, convergence rate, density estimation, edgeworth expansion, efficiency, empirical process, exponential families, gaussian processes, invariance principle, large deviation, law of iterated logarithm, local time, markov chain, markov process, martingale, maximum likelihood, maximum likelihood estimation, nonparametric regression, optimal stopping, order statistic, percolation, poisson process, random walk, rate of convergence, regression, regular variation, robustness, stationary process, stochastic integral, stopping time, time series, weak convergence

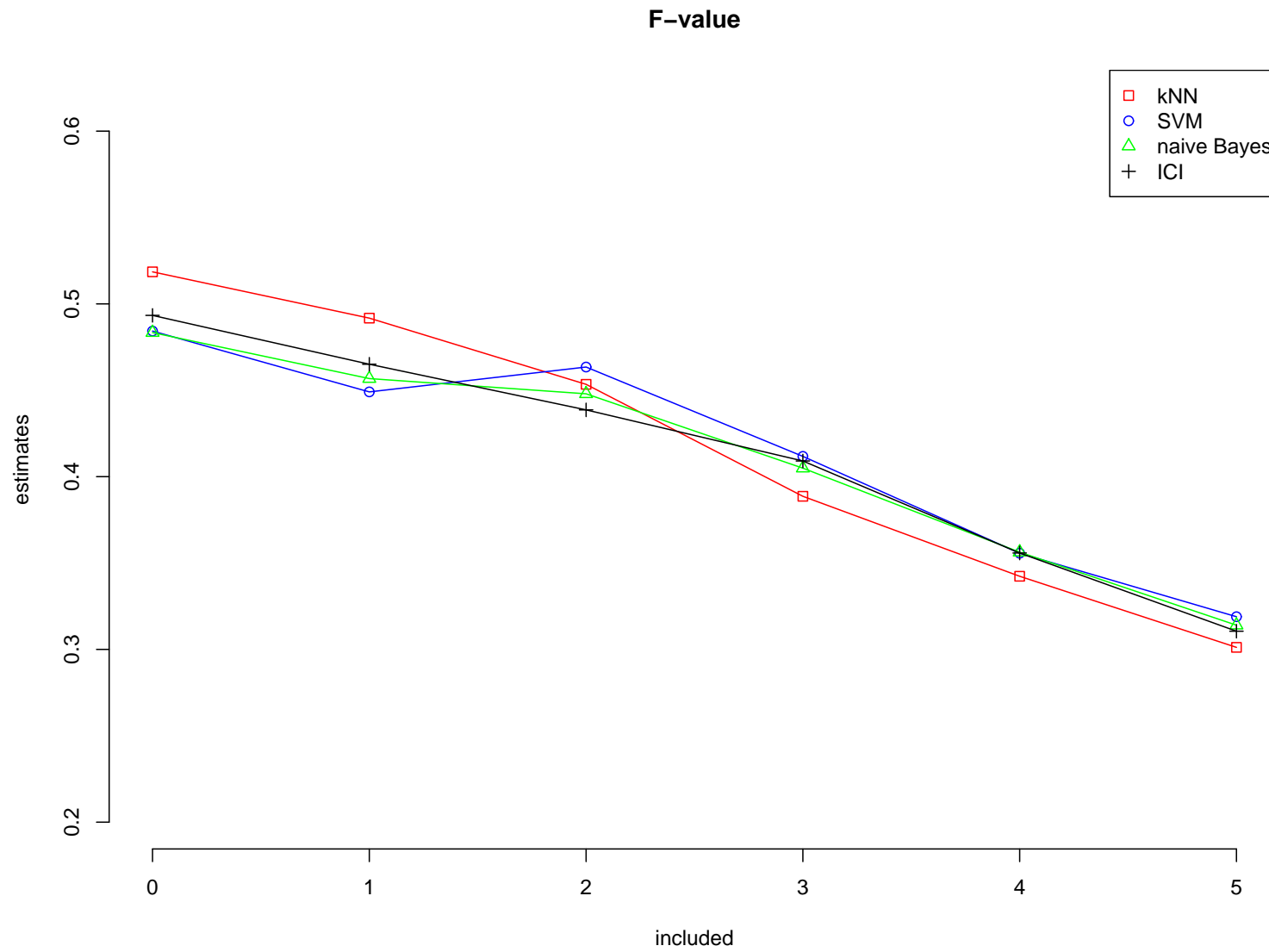
RESULTS



RESULTS



RESULTS



EXAMPLES

confidence interval — coverage (0.82), confidence interval (1.59), coverage probability (0.36), confidence (2.35), interval (2.43), bandwidth (0.36), accuracy (0.28), length (0.36), bootstrap (0.72), quantile (0.36), level (0.28)

time series — time series (0.96), spectral density (0.36), series (1.16), prediction (0.31), autoregressive (0.41), spectral (0.49), covariance (0.31), stochastic process (0.28), least squares (0.28)

density estimation — density estimation (0.61), kernel density (0.34), kernel estimate (0.25), density estimator (0.34), integrated (0.25), wavelet (0.25), mises (0.27), bandwidth (0.59), kernel (1.33), density function (0.27), minimax (0.52), selection (0.25), distance (0.43), support (0.25)

THANK YOU