PR : 3253 ARTICLE IN PRESS

Pattern Recognition ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Soft memberships for spectral clustering, with application to permeable language distinction

Richard Nock[a,*], Pascal Vaillant[b], Claudia Henry[a], Frank Nielsen[c]

[a]Ceregmia-UFR Droit et Sciences Économiques, Universite des Antilles-Guyane, Campus de Schoelcher, BP 7209, 97275 Schoelcher, Martinique, France
[b]Département Lettres et Sciences Humaines, Celia/CNRS-Institut d'Enseignement Supérieur de Guyane, BP 792, 97337 Cayenne, Guyane, France
[c]LIX—Ecole Polytechnique, 91128 Palaiseau Cedex, France

ARTICLE INFO

Article history:
Received 19 September 2007
Received in revised form 10 April 2008
Accepted 24 June 2008

Keywords:
Spectral clustering
Soft membership
Stochastic processes
Text classification

ABSTRACT

Recently, a large amount of work has been devoted to the study of spectral clustering—a powerful unsupervised classification method. This paper brings contributions to both its foundations, and its applications to text classification. Departing from the mainstream, concerned with hard membership, we study the extension of spectral clustering to soft membership (probabilistic, EM style) assignments. One of its key features is to avoid the complexity gap of hard membership. We apply this theory to a challenging problem, text clustering for languages having permeable borders, via a novel construction of Markov chains from corpora. Experiments with a readily available code clearly display the potential of the method, which brings a visually appealing soft distinction of languages that may define altogether a whole corpus.

1. Introduction

This paper is concerned with unsupervised learning, the task that consists of assigning a set of objects to a set of $q > 1$ so-called clusters. One of its most prominent toolbox is spectral clustering, with such a success that its recent developments have been qualified elsewhere as a "gold rush" in classification [1–9] (and many others), pioneered by works ranging from spectral graph theory [10] to image segmentation [11]. Roughly speaking, spectral clustering consists of finding some principal axes of a similarity matrix. The subspace they span, onto which the data are projected, may yield clusters optimizing a criterion that takes into account both the maximization of the within-cluster similarity, and the minimization of the between-clusters similarity. The papers that have so far investigated spectral clustering have two common points. First, they consider a hard membership assignment of data: the clusters induce a partition of the set of objects. It is widely known that soft membership, that assigns a fraction of each object to each cluster, is sometimes preferable to improve the solution, or for the problem at hand. This is clearly the case of our text classification task (Section 6), as words may belong to more than one language cluster. In fact, this is also the case for the probabilistic (density estimation) approaches to clustering, pioneered by the popular expectation maximization method [12]. Their second common point is linked to the first: the solution of clustering is obtained after thresholding the spectral clustering output. This is crucial because in most (if not all) cases, the optimization of the clustering quality criterion is *NP-Hard* for the hard membership assignment [11]. To be more precise, the principal axes yield the polynomial time *optimal* solution to an optimization problem whose criterion is the *same* as that of hard membership (modulo a constant factor), *but* whose domain is unconstrained. Hard membership makes it necessary to fit (threshold) this optimal solution to a constrained domain. Little is known about the quality of this approximation [13], except for the NP-hardness of the task. A recent paper has stressed the need to extend the criteria used on spectral clustering to soft membership [5]. The authors propose to extend the normalized cut criterion from Ref. [11]. The extension they propose departs from mainstream probabilistic interpretations of spectral clustering for two reasons. The first is the criterion used: their criterion aggregates local conductances between clusters, rather than global measures, as earlier used in the Markov chain interpretation of normalized cuts [6], and even earlier in the mixing properties of Markov chains [14]. Second, their algorithm does not directly work on the criterion: it prefers a relaxed (i.e., modified) criterion better suited to the optimization technique chosen. Third and last, this technique is not spectral relaxation (eigen decomposition), but an iterative bound optimization scheme, which usually converges to a local optimum.

* Corresponding author. Fax: +33 596 72 74 03.
E-mail addresses: Richard.Nock@martinique.univ-ag.fr (R. Nock), Pascal.Vaillant@guyane.univ-ag.fr (P. Vaillant), Claudia.Henry@martinique.univ-ag.fr (C. Henry), nielsen@lix.polytechnique.fr (F. Nielsen)
URLs: http://www.univ-ag.fr/~rnock/Articles/PR07/ (R. Nock), http://www.univ-ag.fr/~rnock (R. Nock), http://www.lix.polytechnique.fr/~nielsen (F. Nielsen).

0031-3203/$30.00 © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.06.024

Our paper, which focuses on spectral clustering, departs from the mainstream for the following reasons and contributions. First, we introduce a new interpretation of spectral clustering for soft membership assignment, whose solution *is* spectral relaxation. In this extended framework, we prove that soft memberships still enjoy the links with stochastic processes that were previously known for hard membership [6], which brings a wealth of probabilistic justifications to this method in terms of stability and conductance of the clusters. The soft decomposition of clusters is located "on top" of those usually obtained (through the correlations with the axes). Roughly speaking, soft memberships are distributions that meet a particular orthogonality condition, and this set of distributions puts emphasis on two classical components of some (ergodic) Markov chain:

- percolation probabilities between clusters are encoded in its eigenvalues,
- the first cluster is always its stationary distribution.

Second, we provide an application of soft spectral clustering on a particularly appropriate and challenging problem: cluster words on corpora whose languages may have permeable borders, i.e., in which each word may belong to more than one language. Among the very few attempts to cast spectral clustering to text classification, one of the first builds the similarity matrix via the computation of cosines between vector-based representations of words, and then builds a normalized graph Laplacian out of this matrix to find out the principal axes [2]. Motivated by the relationships between spectral clustering and stochastic processes, we prefer possibly more natural approach that fits this matrix to a Markovian stochastic process following a popular bigram model [15]. Our approach involves however a novel construction of a maximum likelihood Markov chain that satisfies two essential properties: it is always suited to spectral decomposition this is not the case for arbitrary Markov chains), while it removes highly undesirable assumptions about the stochastic process for our task at hand. Thus, the scope of this result goes beyond the scope of this paper, as it may be useful for hard spectral clustering as well. Third and last, we provide experiments with a readily available program, that clearly display the potential of this method for visual data mining when speaking of text classification.

Independently of the interest in extending the scope of spectral clustering, we feel that such results may be interesting because they tackle the interpretation of the *tractable* part of spectral clustering, avoiding the complexity gap that follows after hard membership. Section 2 summarizes related works on hard spectral clustering; Section 3 presents soft spectral clustering; Section 4 discusses the theory; Section 5 gives the theory for its application to text classification; Sections 6 and 7 describe experiments and give a final summary and conclusion. For the sake of readability of the paper, only the proofs of the main results are in the paper's body. The remaining ones have been postponed to Appendix A.

2. Hard spectral clustering

This section provides a synthesis of related spectral clustering works. First we give some definitions. In this paper, calligraphic faces such as $\mathscr{X}$ denote sets and blackboard faces such as $\mathbb{S}$ denote subsets of $\mathbb{R}$, the set of real numbers; whenever applicable, indexed lower cases such as x_i $(i=1,2,\ldots)$ enumerate the elements of $\mathscr{X}$. Upper cases like M denote matrices, with $m_{i,j} \in \mathbb{R}$ being the entry in row i, column j of M; $M^\top$ is the transpose of M, $\mathrm{tr}(M)$ the trace of M, and $diag(M)$ is the vector $\boldsymbol{m}$ of its diagonal elements. Boldfaces such as $\boldsymbol{x}$ denote column vectors, with x_i being the ith element of $\boldsymbol{x}$, and $\langle .,. \rangle$ denotes the inner product for real-valued vector spaces. We are given a set $\mathscr{V}$ of size $|\mathscr{V}| = v$ ($|.|$ denotes the cardinal), together with a symmetric matrix $W_{v\times v}$ with $w_{i,j} \geqslant 0$. We define $D_{v\times v}$ the diagonal matrix with $d_{i,i} = d_i = \sum_j w_{i,j}$. Fix $q > 1$ some user-fixed integer that represents the number of clusters to find. The ideal objective would be to find a mapping $Z : \mathscr{V} \to \mathbb{S}^q$, with $\mathbb{S} = \{0,1\}$, mapping that we represent by a matrix $Z = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_q] \in \mathbb{S}^{v\times q}$.

Under appropriate constraints, the mapping should minimize a multiway normalized cuts (MNC) criterion, that is a quantity representing the sum, for all clusters defined by the mapping Z, of some ratio of cluster cohesion to cluster size. To define the MNC we make use of the following symbols:

$$\kappa_k(Z) = \sum_{i,j=1}^{v} w_{i,j}(z_{i,k} - z_{j,k})^2 \quad (1)$$

Here (1), $\kappa_k(Z)$ measures the cut that Z defines betweenthe inside and the outside of cluster k (the sum of the weight of the crossing edges).

$$\alpha_k(Z) = \sum_{i=1}^{v} z_{i,k}^2 d_i \quad (2)$$

Here (2), $\alpha_k(Z)$ defines a measure of cluster size (the sum of the weight of all edges outgoing from a point in the cluster).

With the following notations, we define the MNC as:

$$\begin{aligned} \arg\min_{Z\in\mathbb{S}^{v\times q}} \quad & \mu(Z) = \sum_{k=1}^{q} \kappa_k(Z)/\alpha_k(Z) \\ \text{s.t.} \quad & Z^\top Z \text{ positive diagonal} \\ \text{s.t.} \quad & \mathrm{tr}(Z^\top Z) = v. \end{aligned} \quad (3)$$

Since this does not change the value of $\mu(Z)$, we suppose without loss of generality that $w_{i,i} = 0, \forall 1 \leqslant i \leqslant v$. The columns of Z are pairwise orthogonal, and Z defines a partition of $\mathscr{V}$ into q clusters $\mathscr{V}_k$ (where $1 \leqslant k \leqslant q$). The clusters that follow from this hard membership assignment are naturally

$$\forall 1 \leqslant k \leqslant q, \quad \mathscr{V}_k = \{v_i : z_{i,k} = 1\}. \quad (4)$$

There is at least one reason why clustering gets better as MNC in Eq. (3) is minimized. Define $P_{v\times v}$ with

$$P = D^{-1}W. \quad (5)$$

Then P is row stochastic: $p_{i,j} \geqslant 0 (1 \leqslant i,j \leqslant v)$ and $\sum_{j=1}^{v} p_{i,j} = 1 (1 \leqslant i \leqslant v)$. We can define a (first order) Markov chain $\mathscr{M}$, with state space $\mathscr{V}$, and transition probability matrix P. Suppose $\mathscr{M}$ is ergodic, regardless of the initial distribution, $\mathscr{M}$ settles down over time to a single stationary distribution $\boldsymbol{\pi}$, the solution of $P^\top \boldsymbol{\pi} = \boldsymbol{\pi}$. Suppose we start (at $t = 0$) a random walk with $\mathscr{M}$, from distribution $\boldsymbol{\pi}$. Let $[\mathscr{V}_k]_t$ be the event that the Markov chain is in cluster k at time $t \geqslant 1$. We have the following important theorem [6].

Theorem 1. *We have* $\mu(Z) = 2\sum_{k=1}^{q} \mathbf{Pr}([\overline{\mathscr{V}_k}]_{t+1} | [\mathscr{V}_k]_t)$ *for the partition defined in Eq.* (4).

(The paper [6] actually gives the proof for $q = 2$, yet its extension is immediate.) Thus, $\mu(Z)$ sums the probabilities of escaping a cluster given that the random walk is located inside the cluster minimizing $\mu(Z)$ amounts to partitioning $\mathscr{V}$ into "stable" components with respect to $\mathscr{M}$. It is interesting to note here that the criterion proposed by Dempster et al. to extend $\mu(Z)$ to $q > 2$ clusters is different, as it is proportional to $\sum_{k\neq k'} \mathbf{Pr}([\mathscr{V}_{k'}]_{t+1} | [\mathscr{V}_k]_t)$.

Unfortunately, the minimization of MNC is NP-hard, already when $q = 2$ [11]. To approximate this problem, one relaxes the output and

Please cite this article as: R. Nock, et al., Soft memberships for spectral clustering, with application to permeable language distinction, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.06.024

rewrites the goal as seek [1]:

$$\arg\min_{Y\in\mathbb{R}^{v\times q}} \nu(Y)=\sum_{k=1}^{q}\kappa_k(Y) \quad \text{s.t.} \quad Y^\top DY=I \tag{6}$$

We say that Y is *clusterwise constant* iff its rows come from a set of at most q distinct row vectors. In this case, without loss of generality, we suppose that identical row vectors in Y are contiguous. The following theorem says that minimizing (3) or (6) *constrained* to clusterwise constant matrices are equivalent. For a proof, we refer e.g., to Refs. [1,9].

Theorem 2. *For any clusterwise constant $Y\in\mathbb{R}^{v\times q}$ satisfying the constraint of Eq. (6), there exists $Z\in\mathbb{S}^{v\times q}$ satisfying the constraints of Eq. (3) such that $\mu(Z)=(1/2)\nu(Y)$, and reciprocally.*

This theorem is important as it explains that the hard clustering solution naturally arises from the rows of Y (hence our terminology). In addition, it shows that solving (6) restricted to the subset of clusterwise constant matrices is NP-hard; fortunately, the unconstrained problem (6) is tractable via the following spectral decomposition of $\mathscr{M}$ [1–3,5,6,8,11]: Y is the set of the q column eigenvectors associated with the smallest eigenvalues of the generalized eigenproblem $(\forall 1\leqslant k\leqslant q)$: $D-W\boldsymbol{y}_k=\lambda_k D\boldsymbol{y}_k$, and it follows that

$$\nu(Y)=2\sum_{k=1}^{q}\lambda_k. \tag{7}$$

If we suppose, without loss of generality, that eigenvalues are ordered, $\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_q$, then it easily follows that $\lambda_1=0$, associated with a constant eigenvector $\boldsymbol{y}_1$. People usually discard this first eigenvector, and keep the following ones to compute Z. The map Z is obtained either by thresholding Y (sometimes, recursively), or by running a hard membership clustering algorithm such as q-means in a subspace spanned by columns of Y. The hope is obviously that $\mu(Z)$ is not too large compared to $(\frac{1}{2})\nu(Y)$. The point is that spectral relaxation finds Y in polynomial time, $O(qv^3)$ without algorithmic sophistication. This is one advocacy for an alternative interpretation of the output Y.

3. Soft spectral clustering

We now suppose that each object may belong to *all* clusters in varying proportions. Define matrix $\tilde{Y}$ by

$$\tilde{y}_{i,k}=d_i y_{i,k}^2. \tag{8}$$

The following property is immediate: because of Eq. (6), each column vector $\tilde{\boldsymbol{y}}_k$ of $\tilde{Y}$ defines a *probability distribution* over $\mathscr{V}$. Since $\tilde{\boldsymbol{y}}_k$ is associated with principal axis k, it seems natural to define it as the probability to draw v_i given that we are in $\mathscr{V}_k$, the cluster associated with the axis. Following the notations of Theorem 1, we thus let

$$\tilde{y}_{i,k}=\mathbf{Pr}([v_i]_t|[\mathscr{V}_k]_t) \tag{9}$$

be the probability to pick v_i, given that we are in cluster k, at time t. For all $1\leqslant k\leqslant q$, define matrix $P^{(k)}$ by

$$p_{i,j}^{(k)}=(w_{i,j}y_{j,k})/(d_i y_{i,k}) \tag{10}$$

Lemma 1. *For all $1\leqslant k\leqslant q$, we have $P\boldsymbol{y}_k=(1-\lambda_k)\boldsymbol{y}_k$, $\tilde{\boldsymbol{y}}_k^\top P^{(k)}=(1-\lambda_k)\tilde{\boldsymbol{y}}_k^\top$, and $P^{(k)}\mathbf{1}=(1-\lambda_k)\mathbf{1}$.*

(proof straightforward). Lemma 1 shows that $P^{(k)}$ is not so far from the transition matrix of some Markov chain $\mathscr{M}^{(k)}$: the sum over each row does not depend on the row and it is "almost" 1, and distribution $\tilde{y}_k$ is "almost" the stationary distribution for $P^{(k)}$. The gap of λ_k to reach constraints satisfied by a Markov chain, and the fact that the entries in $P^{(k)}$ are not all positive, indicate altogether that $P^{(k)}$ may enclose a little bit more than the Markov chain itself. Let us make the assumption that it encodes the *difference* between transition probabilities akin to those of Markov chains. Suppose that $p_{i,j}^{(k)}$ is the difference between the probabilities of reaching, respectively, $\mathscr{V}_k$ and $\overline{\mathscr{V}_k}$ in j, *given* that the random walk is located on i $(\forall t\geqslant 0)$

$$p_{i,j}^{(k)}=\mathbf{Pr}([v_j\wedge\mathscr{V}_k]_{t+1}|[v_i]_t)-\mathbf{Pr}([v_j\wedge\overline{\mathscr{V}_k}]_{t+1}|[v_i]_t) \tag{11}$$

We now make use of the following assumption, which we call (**A**), which states that reaching an object outside cluster k at time $t+1$ does not depend on the starting point at time t.

(**A**) For all $1\leqslant i,j\leqslant v$,

$$\mathbf{Pr}([v_j\wedge\overline{\mathscr{V}_k}]_{t+1}|[v_i\wedge\mathscr{V}_k]_t)=\mathbf{Pr}([v_j\wedge\overline{\mathscr{V}_k}]_{t+1}|[v_i]_t) \tag{12}$$

$$=\mathbf{Pr}([v_j\wedge\overline{\mathscr{V}_k}]_{t+1}|[\mathscr{V}_k]_t) \tag{13}$$

We show that, under (**A**), the probabilistic interpretations given in Eqs. (9) and (11) to the terms defined in Eqs. (8) and (10) (respectively), allow to give a probabilistic interpretation to the function $\nu(Y)$ too. This extends the probabilistic result of Theorem 1 for $\mu(Z)$ under hard clustering to $\nu(Y)$ for *soft* clustering as well.

Theorem 3. *Eqs. (9) and (11) yield under (**A**): $\nu(Y)=4\sum_{k=1}^{q}\mathbf{Pr}([\overline{\mathscr{V}_k}]_{t+1}|[\mathscr{V}_k]_t)$.*

Proof. For all $1\leqslant k\leqslant q$, we first show that $\lambda_k/2=\mathbf{Pr}([\overline{\mathscr{V}_k}]_{t+1}|[\mathscr{V}_k]_t)$. Bayes rule yields $\mathbf{Pr}([\overline{\mathscr{V}_k}]_{t+1}|[\mathscr{V}_k]_t)=\mathbf{Pr}([\overline{\mathscr{V}_k}]_{t+1}\wedge[\mathscr{V}_k]_t)/\mathbf{Pr}([\mathscr{V}_k]_t)$. Now, we have

$$\mathbf{Pr}([\overline{\mathscr{V}_k}]_{t+1}\wedge[\mathscr{V}_k]_t)=\sum_{i,j}\mathbf{Pr}([v_j\wedge\overline{\mathscr{V}_k}]_{t+1}\wedge[v_i\wedge\mathscr{V}_k]_t)$$

$$=\sum_{i,j}\mathbf{Pr}([v_j\wedge\overline{\mathscr{V}_k}]_{t+1}|[v_i\wedge\mathscr{V}_k]_t)\mathbf{Pr}([v_i\wedge\mathscr{V}_k]_t),$$

from which we get

$$\mathbf{Pr}([\overline{\mathscr{V}_k}]_{t+1}|[\mathscr{V}_k]_t)=\sum_{i,j}\mathbf{Pr}([v_j\wedge\overline{\mathscr{V}_k}]_{t+1}|[v_i\wedge\mathscr{V}_k]_t)\mathbf{Pr}([v_i|\mathscr{V}_k]_t)$$

$$=\sum_{i,j}\mathbf{Pr}([v_j\wedge\overline{\mathscr{V}_k}]_{t+1}|[v_i]_t)\tilde{y}_{i,k}. \tag{14}$$

Here, we have made use of Eq. (12) in (**A**). Now, the axiom of total probabilities yields

$$p_{i,j}=\mathbf{Pr}([v_j\wedge\mathscr{V}_k]_{t+1}|[v_i]_t)+\mathbf{Pr}([v_j\wedge\overline{\mathscr{V}_k}]_{t+1}|[v_i]_t)$$

This, and Eq. (11), bring altogether $\mathbf{Pr}([v_j\wedge\overline{\mathscr{V}_k}]_{t+1}|[v_i]_t)=(1/2)(p_{i,j}-p_{i,j}^{(k)})$. Finally, we obtain

$$\mathbf{Pr}([\overline{\mathscr{V}_k}]_{t+1}|[\mathscr{V}_k]_t)=(1/2)\sum_{i,j}d_i y_{i,k}^2\left(\frac{w_{i,j}}{d_i}-\frac{w_{i,j}y_{j,k}}{d_i y_{i,k}}\right)$$

$$=(1/2)\sum_{i,j}w_{i,j}y_{i,k}(y_{i,k}-y_{j,k})$$

$$=(1/2)\boldsymbol{y}_k^\top(D-W)\boldsymbol{y}_k=\lambda_k/2 \tag{15}$$

There remains to sum for all k, and use Eq. (7) to get the statement of the theorem. □

Please cite this article as: R. Nock, et al., Soft memberships for spectral clustering, with application to permeable language distinction, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.06.024

An immediate corollary to the theorem is the following one.

Corollary 1. *We have* $1-\lambda_k = \mathbf{Pr}([\mathscr{V}_k]_{t+1}|[\mathscr{V}_k]_t) - \mathbf{Pr}([\overline{\mathscr{V}_k}]_{t+1}|[\mathscr{V}_k]_t)$.

Proof. We use Eq. (15), and obtain $1-\lambda_k = 1-2\mathbf{Pr}([\overline{\mathscr{V}_k}]_{t+1}|[\mathscr{V}_k]_t)$, $\forall t \geqslant 0$. This brings the statement of the corollary. □

This is consistent with the fact that $1-\lambda_k \in [-1,1]$ [6]. Lemma 1 still holds under Eqs. (8), (11) and Corollary 1. Theorem 3 allows us to extend to soft membership the links between spectral clustering and Markov chains that were coined in Ref. [6]: we seek soft clusters having high stability (there is also a link with the conductance of clusters, see Ref. [6]). Finally, we remark that the soft membership solution is significantly different from the hard membership solution, as each cluster is now built from the *columns* of $\tilde{Y}$, and not from the rows of Y (Theorem 2).

4. Discussion

The case $k=1$ is particular, not only because $\boldsymbol{y}_1$ is constant [1,3,6,8,11]. It can also be shown from Lemma 1 that $\tilde{\boldsymbol{y}}_1 = \pi$, the stationary distribution of an ergodic Markov chain $\mathscr{M}$ whose transitions are modeled by Eq. (5).[1] This is natural, as this distribution is the one that best explains the data. On the other hand, there is no percolation possible from $\mathscr{V}_1$ to $\overline{\mathscr{V}}_1$, which is explained by the fact that $\lambda_1 = 0$ in Corollary 1, and by the fact that $p_{i,j}^{(k)} = p_{i,j} = \mathbf{Pr}([v_j \wedge \mathscr{V}_k]_{t+1}|[v_i]_t)$. We call soft cluster $\mathscr{V}_1$ the *stationary cluster*, as it only carries the ergodicity information about $\mathscr{M}$. In previous literature, Ref. [2] obtained some spectral clustering results on distributions observed from a text corpus; they made a 2D plot on the second and third principal axes, *after* having made a prior selection of the most frequent words to be plotted. Our results on $k=1$bring a justification to this, as it amounts to making a selection of words according to the *first* cluster (principal axis). Furthermore, since the first axis encodes the word frequencies, the next axes, that are 2-2 orthogonal, are not "affected" anymore by these frequencies.

So far, our probabilistic interpretation of spectral clustering, which appears in Eq. (8), does not fully integrate the constraint of Eq. (6), namely $Y^\top DY = I$. More precisely, $diag(Y^\top DY) = \mathbf{1}$ is a subset of the constraint that makes it possible to define our distributions in Eq. (8). The zero elements outside the diagonal also imply that the distributions have "zero correlation". This means that, if any two distributions of the v dimensional probability simplex, say $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{y}}'$, fit to the constraint, then there exists $\Sigma \in \{-1,+1\}^{v\times 2}$ such that

$$\sum_{i=1}^{v} \sqrt{\tilde{y}_i \tilde{y}'_i}\, \sigma_{i,1}\sigma_{i,2} = 0. \tag{16}$$

Clearly, this constraint does not enforce different distributions. For example, consider q uniform distributions, and v a power of 2; using for Σ any subset of $q \leqslant v$ columns of an Hadamard matrix easily yields zero correlation (16) among any two pairs of these distributions. In practice of course, the numerical approximations in computing the eigenvectors may yield non-zero correlations, so this constraint actually does not really hold. However, for the sake of completeness, we have wondered to what extent Eq. (16) is hard to satisfy theoretically. Here is an answer.

[1] Actually, when the data considered is a text corpus, this stationary distribution is the overall frequency distribution of the vocabulary items in the corpus, i.e., the word frequencies: in fact, since all $y_{i,1}$ are constant in the first eigenvector of P, all $\tilde{y}_{i,1}$ are proportional to d_i (the number of occurrences of a word type ω_i) with a normalization factor. This point is developed later (Section 5).

Theorem 4. *Let D be defined as in* Section 2, *and $Y \in (\mathbb{R}^+)^{v\times q}$ such that* $diag(Y^\top DY) = \mathbf{1}$. *Then, there exists $\Sigma \in \{-1,+1\}^{v\times q}$ such that*

$$\forall 1 \leqslant k \neq l \leqslant q, |f_{k,l}(\Sigma)| \leqslant 2\sqrt{\langle \tilde{\boldsymbol{y}}_k, \tilde{\boldsymbol{y}}_l \rangle \log(2q^2)} \tag{17}$$

with $f_{k,l}(\Sigma) = \sum_{i=1}^{v} \sqrt{\tilde{y}_{i,k}\tilde{y}_{i,l}}\,\sigma_{i,k}\sigma_{i,l}$.

(the proof is in Appendix A, Section A.1). This bound is better when the inner product is small. Due to Eq. (6), the distributions in $\tilde{Y}$ generally have very small inner products: either the distributions are very different from each other, or they are not but in that case, they are well spread over $\mathscr{V}$ (recall that $\tilde{\boldsymbol{y}}_1 = \boldsymbol{\pi}$). To see why the bound is small in that latter case, consider two uniform distributions $\tilde{\boldsymbol{y}}_k, \tilde{\boldsymbol{y}}_l$: we obtain $|f_{k,l}(\Sigma)| = \mathcal{O}(\sqrt{(\log q)/v})$, a tiny real since in general $q \ll v$.

Finally, when $k>1$, there may be masking problems from our analysis, as the estimators $p_{i,j}^{(k)}$ are not necessarily confined to the interval $[-p_{i,j}, p_{i,j}]$. Rather than a limit of spectral clustering, we think that this either follows from the nature of the criterion optimized in Eq. (6) which may bring such masking problems [16], or it accounts for a deeper analysis of the problem. At least, our analysis demonstrates that there is indeed something to drill down from classical spectral clustering analyses, to bring a probabilistic account to the tractable part of this powerful method. We note that these eventual problems are purely theoretical, and have absolutely no experimental impact, as we do not depict the components of $P^{(k)}$, our representations rely only on $\tilde{Y}$.

5. Maximum likelihood ring text generation

We begin with more definitions. A corpus $\mathscr{C}$ is a set of texts, $\{\mathscr{T}_1, \mathscr{T}_2, \ldots, \mathscr{T}_m\}$, with m the length of the corpus. For all $1 \leqslant k \leqslant m$, text $\mathscr{T}_k$ is a string of tokens (occurrences of words or punctuation marks), $\mathscr{T}_k = \omega_{k,1}\omega_{k,2}, \ldots, \omega_{k,|\mathscr{T}_k|}$, of size/length $|\mathscr{T}_k|$. The *size* of the corpus, $|\mathscr{C}| = n$, is the sum of the length of the texts: $n = \sum_{i=1}^{m} |\mathscr{T}_i|$. The size of a corpus is implicitly measured in words (number of word occurrences), but it may contain punctuation marks as well. The *vocabulary* of $\mathscr{C}$, $\mathscr{V}$, is the set of distinct linguistically relevant words or punctuation marks, the tokens of which are contained in the texts of $\mathscr{C}$. The size of the vocabulary is denoted $v = |\mathscr{V}|$. The elements of $\mathscr{V} = \{v_1, v_2, \ldots, v_v\}$ are *types*: each one is unique and appears only once in $\mathscr{V}$. For all $i,j \in \{1,2,\ldots,v\}$, we let n_i denote the number of occurrences of type i in $\mathscr{C}$, and $n_{i,j}$ denote the number of times a word of type i immediately precedes (left) a word of type j in $\mathscr{C}$.

Suppose that $\mathscr{C}$ is generated from a random walk of a Markov chain $\mathscr{M}$. Perhaps the most intuitive way to compute P is to perform conventional maximum likelihood out of a popular bigram model [15], which would yield $p_{i,j} = n_{i,j}/n_i$ (folklore). However, there are at least two reasons not to remain as general. The first is technical: the eigenvalues of the spectral decomposition of P may be *complex*, preventing their soft spectral interpretation. The second is linguistic. This computation of P is convenient *if* we make the assumption that a text is written from the left to the right. This corresponds to our *a priori* intuition of speakers of European languages, who have been taught to read and write in languages where the graphical transcription of the linearity of speech is done from left to right. However, a more thorough reflection on the empirical nature of the problem has led us to question this approach. The method being developed should be able to work on any type of written language, making no assumption on its transcription conventions. Some languages (including important literary languages like Hebrew or Arabic) have a tradition of writing from right to left, and this sometimes goes down to having the actual stream of bytes in the file also going "from right to left" (in the file access sense). The new Unicode standard for specifying language directionality circumvents this, by allowing the file to always be coded in the logical order, and managing the visual rendering so

Please cite this article as: R. Nock, et al., Soft memberships for spectral clustering, with application to permeable language distinction, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.06.024

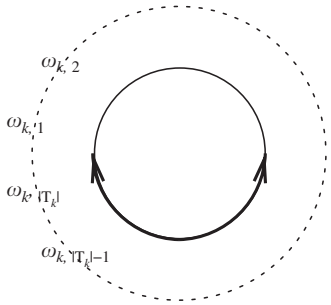

Fig. 1. A "circular" generation of a text $\mathscr{T}_k$ makes it possible to eliminate both the direction for writing $\mathscr{C}$ *and* the choice of the first word written.

that it suits the language conventions, even in the case of mixed-language texts (i.e., English texts with Hebrew quotes); but large corpora are still encoded in the old way, and the program should not be sensitive to this, making no more postulates than necessary.

We have found a convenient approach to eliminate this directionality dependence. It also has the benefit of removing the dependence in the choice of the first word to write down a text. What we actually are considering is the likelihood of $\mathscr{C}$ as if all of its texts were read and written in a direction-insensitive way. Let us illustrate this by supposing that any text observed in the corpus $\mathscr{C}$ can be represented as a circular object (Fig. 1), where two contiguous words in the text are represented by two contiguous points on the circle, and the end of the text loops to the beginning. The unidirectional reading process consists of walking around the circle in a monotonous direction, starting from a given point (the first word); its linear projection is the text that has actually been observed. A direction-insensitive reading process consists of walking around the circle, starting from any point, and then jumping at every step to a contiguous point, but in an unspecified direction. Its linear projection *may* yield the text that actually was observed, but it also may yield a great number of other possibilities. What we are doing is, assuming that we are dealing with the second type of process, and that the text we have observed is but one of the possible outputs of a direction-insensitive random walk process.

The following theorem gives the new max-likelihood transition matrix P.

Theorem 5. *With the circular writing approach, the maximum likelihood transition matrix P is defined by $p_{i,j} = (n_{i,j} + n_{j,i})/(2n_i)$, with $1 \leqslant i,j \leqslant v$.*

(proof in Appendix A, Section A.2). Now, if we define $W_{v\times v}$ with $w_{i,j} = (n_{i,j} + n_{j,i})/2$, and $D_{v\times v}$ the diagonal matrix with $d_{i,i} = d_i = n_i$, it follows that P is the product of these two symmetric matrices, and it satisfies (5). Its spectral decomposition has only real eigenvalues, and P fits well to spectral clustering. Furthermore, the circular way to write down the texts of $\mathscr{C}$ has another advantage: $\mathscr{M}$ is irreducible. Let us make the assumption that $\mathscr{M}$ is also aperiodic. This is a mild assumption: our way to model the corpus implies that there exists loops of length 2 for any type (since after any step leading from word ω_i to word ω_j, the process may go back to ω_i). Assume that there exists at least one type $v_\star$ with two special occurrences separated by an even number of other words in the text. This yields a direct loop $\ell_\star$ of odd length for $v_\star$. Now, for any other word ω, we can generate a loop of odd length, by going from ω to whichever special occurrence of $v_\star$, then following $\ell_\star$, and finally returning to ω by the same path. Thus aperiodicity follows from a simple assumption about one type, and it is all the more likely to happen as the corpus is big. Taken together, irreducibility and aperiodicity now imply that $\mathscr{M}$ is ergodic, and it is easy to check that its stationary distribution satisfies $\pi_i = n_i/n$. Whenever the state space $\mathscr{V}$ of $\mathscr{M}$ comes from more than a single language, soft spectral clustering may provide a basis for their smooth discrimination.

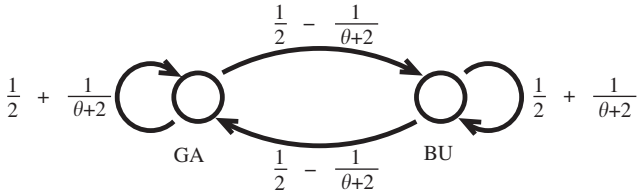

Fig. 2. A toy maximum likelihood Markov chain $\mathscr{M}$ for language "GABU".

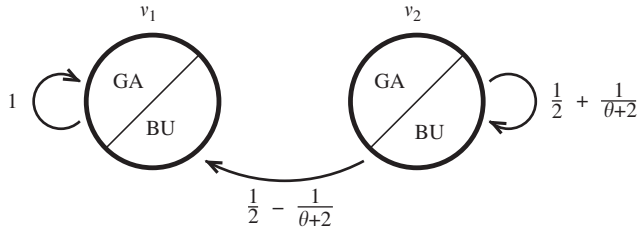

Fig. 3. The two clusters induced by our spectral analysis of $\mathscr{M}$ in Fig. 2. Disks display the relative proportions of words in each cluster, and arrows denote the transition probabilities computed from Corollary 1. Note that transitions confirm the ergodicity of $\mathscr{M}$.

Before embarking into experiments displaying how this discrimination occurs, the theory developed here for text generation may be used to obtain an appealing interpretation of the material of Section 3.

Consider a simple toy language, called "GABU", with vocabulary $\mathscr{V} = \{\text{GA}, \text{BU}\}$, whose transition matrix P is parameterized by a real "temperature" parameter $\theta \in \mathbb{R}_+$:

$$P = \begin{bmatrix} \frac{1}{2} + \frac{1}{2+\theta} & \frac{1}{2} - \frac{1}{2+\theta} \\ \frac{1}{2} - \frac{1}{2+\theta} & \frac{1}{2} + \frac{1}{2+\theta} \end{bmatrix}. \quad (18)$$

Fig. 2 displays the associated Markov chain $\mathscr{M}$. We assume the same number of GA's and BU's in $\mathscr{T}$, so that $D = (n/2)I$. One can check that:

$$\boldsymbol{y}_1 = [1/\sqrt{n}, 1/\sqrt{n}]^\top, \quad \lambda_1 = 0, \quad (19)$$

$$\boldsymbol{y}_2 = [1/\sqrt{n}, -1/\sqrt{n}]^\top, \quad \lambda_2 = 1 - 2/(2+\theta) \quad (20)$$

Corollary 1 tells us from Eq. (20) that

$$\mathbf{Pr}([\mathscr{V}_2]_{t+1} | [\mathscr{V}_2]_t) = \frac{1}{2} + \frac{1}{2+\theta}, \quad (21)$$

and Eqs. (19)–(20) yield $\tilde{\boldsymbol{y}}_1 = \tilde{\boldsymbol{y}}_2 = [\frac{1}{2}, \frac{1}{2}]^\top$, i.e., points are equally distributed inside each cluster, which is not surprising because of the tiny vocabulary size, the even distribution of GA's and BU's in D, and the fact that P is symmetric (P is doubly stochastic). As a simple matter of fact, because of our way to write a text, P is always symmetric when there are only two types in the vocabulary, but this is not the case for larger vocabularies. The fact that distributions are identical in our toy example is also a consequence of thevocabulary size. Words in natural language tend to follow a Zipf-Mandelbrot distribution.In a real-world corpus, such a highly non-uniform distribution in $\tilde{\boldsymbol{y}}_1$ would hardly spread to other clusters: the orthogonality constraint in Eq. (6) would necessitate to have distinct subsets with identical sum of square-root of probabilities.

Fig. 3 presents these two clusters. As seen from this figure, parameter θ controls the percolation from cluster 2 to the stationary cluster—because of the ergodicity of $\mathscr{M}$, it can only be one-way. When $\theta \to 0$, the two clusters are well separated, which goes along with the

Please cite this article as: R. Nock, et al., Soft memberships for spectral clustering, with application to permeable language distinction, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.06.024

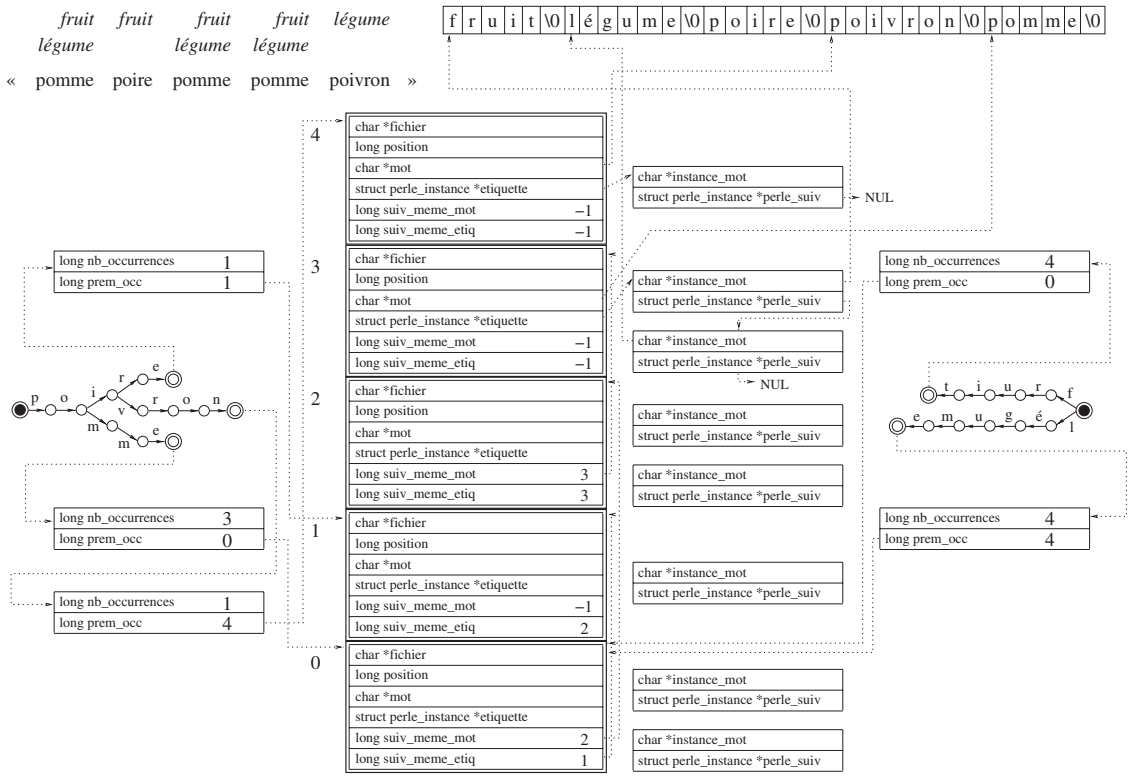

Fig. 4. The specialized index table format developed for this application. At the center, the table of token occurrences (used to compute matrix W by moving a contextual window). At the left, a trie (lexical tree) indexing the words of the vocabulary. The table may also be used for tagging, hence the use of the "etiquette" labels and of the second trie on the right (useful to index the tags of the tagset). The figure exemplifies the result of indexing a text containing the words "pomme poire pomme pomme poivron".

fact that $\mathscr{M}$ is not connected anymore. As θ increases, the transitions become uniform on $\mathscr{M}$ and the chances to hop onto the stationary cluster increase.

6. Experiments

6.1. Implementation of the system

A computer program (MOTS[2]) has been developed in C on an Intel computer running a Debian GNU/Linux operating system. The program makes use of various functions of the GNU C library (glibc). For the algebraic computing, it relies on the ATLAS optimization system for BLAS (basic linear algebra subprograms);[3] and for solving eigensystems, on the LAPACK library,[4] written in FORTRAN. Overall, MOTS contains 16,000 lines of code; when statically linked, it yields a 1.2 MB executable file.

The program takes a text of arbitrarily long size as input.[5] The main processing chain of the program works in five steps:

1 it automatically detects the text format and encoding, and converts everything to raw text encoded in Unicode UTF-8,
2 it performs a stage of tokenization, i.e., it segments the raw stream of bytes into tokens of words, figures or typographical signs.
3 it builds an index table suited for fast access to word type information (designed on the lexical tree, or *trie*, model). Fig. 4 gives a schema of the index table format that has been implemented.
4 it computes the bigram transition matrix $P = D^{-1}W$, by moving a contextual window along the tokens put in their text order, and incrementing n_{ij} for every seen occurrence of a transition (ω_i, ω_j), where ω_i and ω_j are two given words.
5 it calls SGEEV, a function of the LAPACK library, to compute the eigenvalues and eigenvectors of the matrix.

The time-consuming step is (5), which relies on a non-optimized FORTRAN reference implementation. For a corpus with a vocabulary of 13,000 distinct words, the system (an average 2005 Pentium chip running at 2 GHz) stayed more than 24 h in the SGEEV function. Probable improvements would be gained by working with a more performant workstation, but time is not a very critical factor for our purpose.

6.2. Data sources

Experiments were made on several examples of multilingual corpora. We have used different sources of publicly available texts, such as online digital libraries projects (like the Project Gutenberg, the Wikisource online library, Projekt Gutenberg-DE for German texts, ABU Association des Bibliophiles Universels for French texts …); online text repositories supported by established institutions (like the Gallica digital library of the Bibliothèque Nationale de France, the Runeberg project at the University of Linköping, the Etext center at the University of Virginia, or the ATHENA site at the University of

[2] Our system is available from the URL: http://www.univ-ag.fr/~pvaillan/mots/.

[3] ATLAS was developed and made available to the community by R. Clint Whaley, University of Texas at San Antonio; ATLAS web page: http://math-atlas.sourceforge.net/.

[4] LAPACK was developed over years by a team of researchers, mainly located at the University of Tennessee at Knoxville; LAPACK web page: http://www.netlib.org/lapack/.

[5] There actually is a practical limit, which is caused by the 2 GB (2^{32} bytes) file size limit that the system imposes on the file storing the matrix of floating numbers!

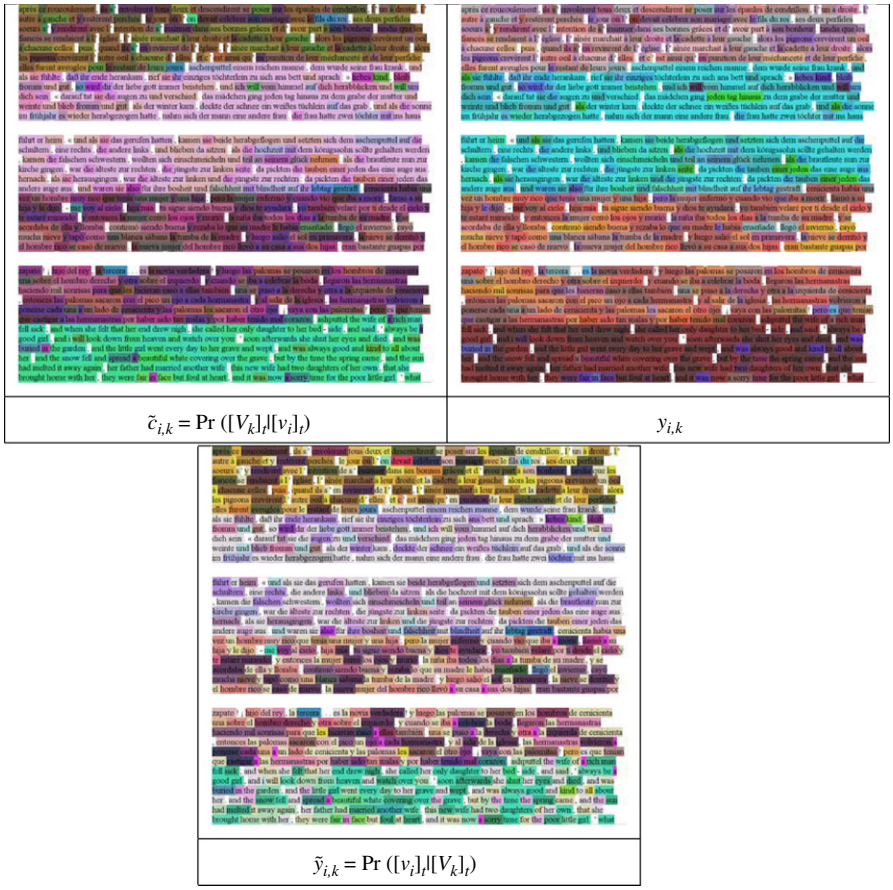

Fig. 5. Experiments on multilingual passages of *Cinderella*. Each row crops a borderline between two languages (from the top to the bottom): French/German, German/Spanish, Spanish/English. Bottom row = quantities that are represented by RGB colors in each column; each color level is associated with a principal axis $k \in 2, 3, 4$ (see text).

Geneva…); sources of legal texts or international treaties and conventions (like EUR-Lex, the multilingual legal website of the European Union); or various other sources, such as movie transcripts databases. For texts in Creole languages (see below, p. 18), the resources are scarcer. However, for French-based Creoles, some collections of short digital texts are available from a few websites: Potomitan,[6] Krakemanto,[7] or M.-C. Hazaël-Massieux's Creole studies website at the Université de Provence.[8]

6.3. RGB representation

Plotting a text according to the soft spectral clustering interpretation described above is quite simple. We can represent each word with a RGB color, where each color level is associated with some principal axis k, and scales the component of the plotted vector,say $\boldsymbol{u}$, for each word. More precisely, we display $\chi = 5$ different color levels on each axis, and fit each level to contain approximately the same number of points ($\approx v/5$). The χ corresponding intervals of values of $\boldsymbol{u}$ do not necessarily have the same width, but we have maximal visual contrast. There are actually *three* kinds of $\boldsymbol{u}$ that we plot. The first two are naturally $\boldsymbol{y}_k$ and $\tilde{\boldsymbol{y}}_k$. But the most interesting plot to make is perhaps not $\mathbf{Pr}([v_i]_t|[\mathscr{V}_k]_t) = \tilde{y}_{i,k}$. Since we plot colors for each word, it is much more interesting to plot the probability of being in a cluster *given* that we observe some word, i.e.,

$$\mathbf{Pr}([\mathscr{V}_k]_t|[v_i]_t) = \tilde{c}_{i,k} = \mathbf{Pr}([v_i]_t|[\mathscr{V}_k]_t)\mathbf{Pr}([\mathscr{V}_k]_t)/\mathbf{Pr}([v_i]_t), \quad (22)$$

and we let $\tilde{C}$ denote the matrix containing the $\tilde{\boldsymbol{c}}_k$ as column vectors (while $\tilde{Y}$ is column stochastic, $\tilde{C}$ is row stochastic). Since we have $\mathbf{Pr}([v_i]_t) = d_i/d$, the only unknown to compute this probability is $\mathbf{Pr}([\mathscr{V}_k]_t)$ (noted $p^\star_k$ for short); given any $i = 1, 2, \ldots, v$, summing Bayes rules over $k = 1, 2, \ldots, v$ yields $\sum_{k=1}^{v} \mathbf{Pr}([v_i]_t|[\mathscr{V}_k]_t)\mathbf{Pr}([\mathscr{V}_k]_t) = \mathbf{Pr}([v_i]_t)$, i.e. $\mathbf{p}^\star$ satisfies:

$$\mathbf{p}^\star = \tilde{Y}^{-1}\pi \quad (23)$$

Solutions to Eq. (23) exist only when $\tilde{Y}$ is invertible, which necessitates that clusters be distinct from each other. As discussed in Section 5, such situations typically arise for tiny vocabularies, such as for our toy GABU example. In practice, our vocabularies and corpora were far large enough to prevent such a situation, and we never met inversion problems.

Fig. 5 presents such an experiment on a 1 Mb text, containing four versions of the same tale (Cinderella, from the Grimm Brothers), in four languages: French, German, Spanish, and English. While we can remark that the plot of $\tilde{C}$ displays perhaps the sharpest distinction between the languages, it also "orders" them in some sense. From the average color levels of each language, we can say that Red is principally German, Green is principally English, and Blue is principally Spanish. French is somewhere in between all of them. It is interesting to notice that the results are in accordance with what we know

[6] Potomitan collection of tales: http://www.potomitan.info/atelier/contes/; other texts in various Creole languages (mainly Martinique and Guadeloupe) available from other sections of the website.

[7] Krakémantò tales of French Guiana: http://www.krakemanto.gf/.

[8] M.-C. Hazaël-Massieux's creole website: http://creoles.free.fr/Cours/.

Please cite this article as: R. Nock, et al., Soft memberships for spectral clustering, with application to permeable language distinction, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.06.024

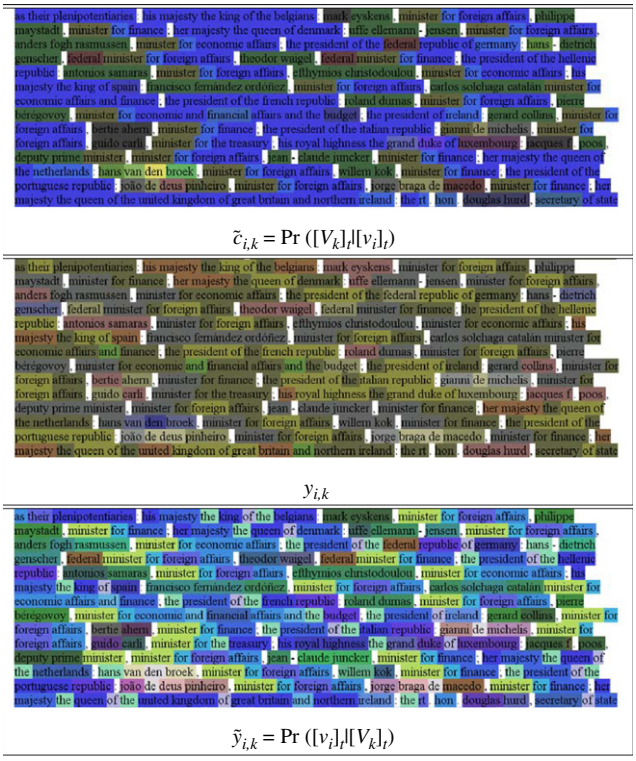

Fig. 6. More experiments (four languages compilation of the Maastricht treaty). Conventions follow Fig. 5.

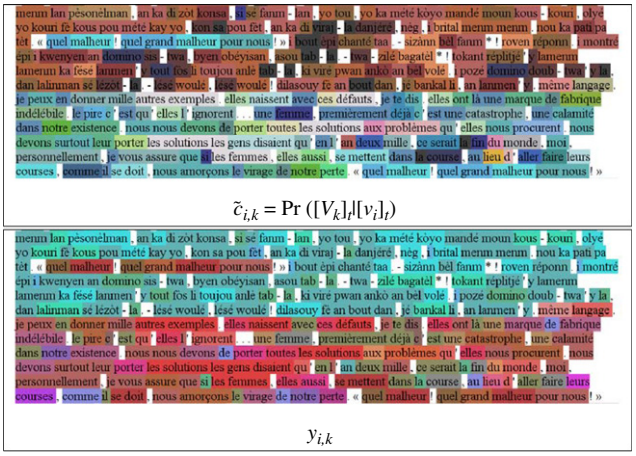

Fig. 7. Crop of an extract of *Lavwa egal*, where French and Creole are intertwined. Conventions follow Fig. 5.

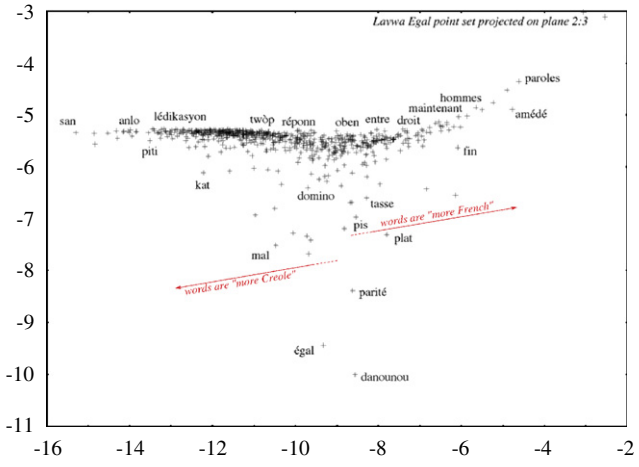

Fig. 8. Plot of $\tilde{\mathbf{c}}_3(y)$ as a function of $\tilde{\mathbf{c}}_2(x)$ for the significant words of *Lavwa egal* (see text for details).

about the genetic roots of these four languages, a fact which is of course unknown to the computer program. Similar conclusions were borne out from experiments on these four languages for the Maastricht treaty (see Fig. 6). Moreover, results of the treaty make the names of each country's plenipotentiaries appear (Fig. 6 displays the English crop for this part); it was quite surprising to see that, while Y gives uniform results by language, $\tilde{C}$ clearly makes those names appear (mostly dark green over dark blue for English).

An even more interesting experiment consisted of trying the program on texts where languages are more intricately mixed. This is quite typically so in literature from multilingual regions, like in the case of the Creole-speaking communities throughout the wide Caribbean area and USA. In at least some cases, the Creole language has remained in contact with its "lexifier" European language (none of those has in the meantime become extinct), in sociolinguistic situations which have sometimes been coined as "diglossic": this has especially been the case for English-based Creoles like Jamaican or Gullah spoken in the states of South Carolina and Georgia, and French-based Creoles spoken in the territories of Haiti, Guadeloupe, Martinique and French Guiana. In a diglossic situation, the European language is still in use as the official and prestige language, while the Creole language is the vernacular. This leads to very frequent code-switching and intermingling of languages in several domains. This is clearly an extreme situation of choice for soft clustering. We have processed a 200 kb extract (12,200 occurrences of 2,400 vocabulary items) of a French-Martinican Creole bilingual novel, *Lavwa egal* by Térèz Léotin,[9] where segments in French and Creole are strongly intertwined. While both languages share many words, the results display quite surprising contrasts, and these are actually sharper for $\tilde{C}$. Fig. 7 displays a crop in which the program has even managed to extract a short French sentence (*quel malheur*, *quel grand malheur pour nous*) out of a Creole segment.

6.4. Representation using $\tilde{C}$

The most conventional plot for spectral clustering results would consists of projecting the words on two selected eigenvectors. We have carried out such a representation for our experiments on the Maastricht treaty and *Lavwa egal*, but for matrix $\tilde{C}$ instead of Y, to further test its ability for smooth separation of languages.

Fig. 8 displays a two-dimensional distribution of the most significant units in the "Lavwa egal" corpus on the plane defined by $\tilde{\mathbf{c}}_3$ and $\tilde{\mathbf{c}}_2$. The probability is shown on logarithmic scale (numbers appearing on the axes are powers of 10). By "significant words" we mean words occurring more than four times in the corpus—16% of the word types in this particular corpus. Note that including all words does not change the overall shape of the data cloud, but tends to scale down the most significant part of the diagram, by including outlying points. Only a few labels have been shown (20 out of 660) for the sake of readability, but they are consistent with the global distribution.

Fig. 9 shows the same data projected on the plane defined by $\tilde{\mathbf{c}}_4$ and $\tilde{\mathbf{c}}_3$, respectively. It resembles a plot one would obtain by folding Fig. 8 around a separating axis for French and Creole, and axis that

[9] Térèz Léotin, *Lavwa egal—La voix égale*, published by Ibis Rouge, French Guiana, 2003.

Please cite this article as: R. Nock, et al., Soft memberships for spectral clustering, with application to permeable language distinction, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.06.024

would be roughly parallel to the y axis. Since separation is very smooth between languages, we have chosen not to sketch the axis in the figure but it is quite intuitive.

A look at the two diagrams shows that units clearly tend to cluster along lines that significantly form in the 2:3:4 eigenvector space. In Fig. 8, Creole words are grouped on the left of the diagram, and French words on the right. In Fig. 9, the clusters are even more visible. Interestingly, this shows that on non-trivially small vocabularies (i.e. when the number of distinct words exceeds few units—see discussion in Section 5 about the toy GABU example), the information necessary for a soft distinction between language clusters can be nicely retrieved from row-stochastic matrix $\tilde{C}$, and not only on the signed coordinates of the direct solution to spectral decomposition, i.e., matrix Y. As witnessed by our experiments in Section 6.3, sometimes, it is also sharper.

Figs. 10 and 11 show comparable results on a four-language corpus, consisting of versions of the European treaty of Maastricht in German, French, Spanish and English. The three angles of the plot in Fig. 10 roughly cluster the languages as follows: mostly German (upper right), French + Spanish (left) and English (bottom). But the most striking phenomenon does not arise from these two Figures. A comparison with the results of *Lavwa egal* in Figs. 8 and 9 displays surprising similar structures if we take figures two by two. Indeed, Fig. 10 relies on the same structure as Fig. 8: a dense point cloud on the top, elongated along $\tilde{c}_2$, with virtually the same shape on the two plots. In Fig. 8, this axis is separating French from Creole. The discrimination in Fig. 10 is slightly more complicated as it involves three languages (German, Spanish, French), certainly because there are more distinct languages to cluster. In Fig. 10, English plays the role of the language that creates a less dense cloud on the bottom of the figure, a cluster that would probably have appeared for *Lavwa egal* should it have mixed another language (e.g., Haitian Creole).

The situation is similar if we compare Figs. 9 and 11. The bend that appears in Fig. 9 with a clear boundary line also appears in Fig. 11. In the same way as it involves a distinction between French and Creole in Fig. 9, it makes a distinction between German and

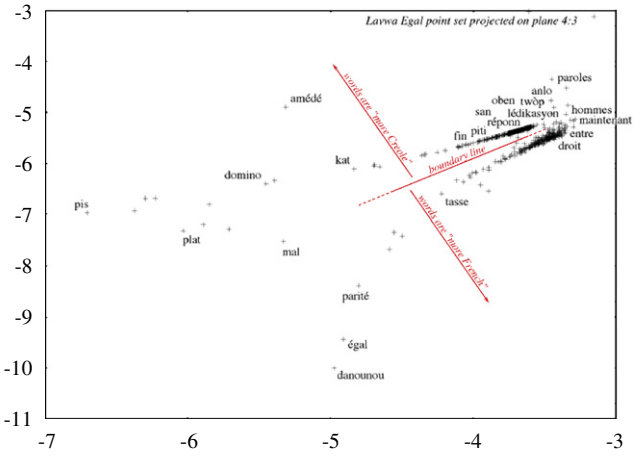

Fig. 9. Plot of $\tilde{\boldsymbol{C}}_3(y)$ as a function of $\tilde{\boldsymbol{C}}_2(x)$ for the significant words of Lavwa egal (see text for details).

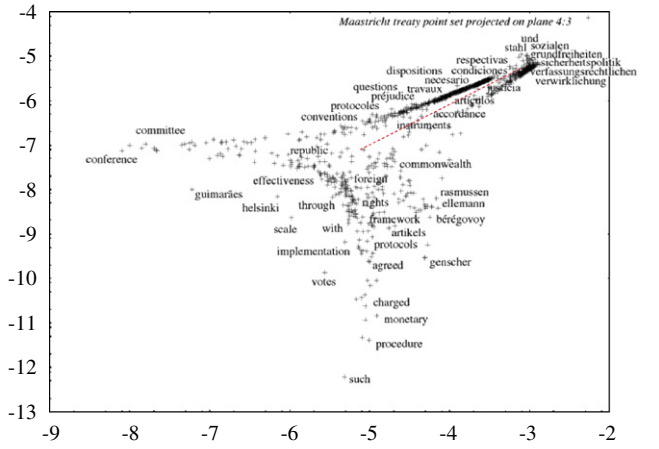

Fig. 11. Plot of $\tilde{\boldsymbol{C}}_3(y)$ as a function of $\tilde{\boldsymbol{C}}_4(x)$ for the significant words of the Maastricht treaty (see text for details about the red dashed line).

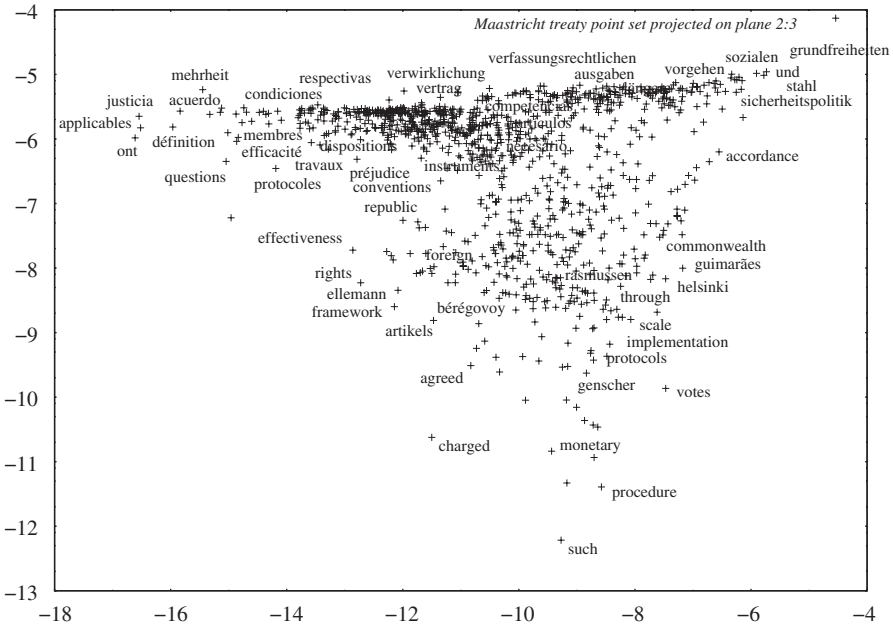

Fig. 10. Plot of $\tilde{\boldsymbol{C}}_3(y)$ as a function of $\tilde{\boldsymbol{C}}_2(x)$ for the significant words of the Maastricht treaty (see text for details).

Please cite this article as: R. Nock, et al., Soft memberships for spectral clustering, with application to permeable language distinction, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.06.024

(French + Spanish) in Fig. 11 (the red dashed line). Again, English plays the role of the "extra" language, in the same sense as it does for Figs. 8 and 10, since it fills in Fig. 11 the space occupied by very few words in Fig. 9 (*plat*, *mal*, *parité*, *égal*, *danounou*).

The fact that the same structures—with the same role, the same interpretation—appear in the plots of two extremely different corpora tends to make us think that they are not fortuitous. Drilling down in their properties is the matter of making more tests on more languages, and this shall be the subject of future works.

7. Summary and conclusion

We have described a soft clustering method for sets of elements that can be interpreted as states of a Markov chain, and an example application of the method to a problem in information extraction: identifying different languages in multilingual corpora, when those languages are not known in advance, and when their properties are not defined a priori. In such a setting, states are words, which can belong in any proportion to distinct languages, and transition probabilities are modeled from some maximum likelihood writing of texts in any direction.

More precisely, our technique draws its roots in the early interpretations of spectral cut methods in terms of Markov chains and the conductance of their clusters [6]. We fundamentally depart from these works in the clustering problem addressed: where the usual spectral approaches seek the approximation of a *hard* clustering, our spectral approach is the solution of the *soft* clustering problem.

We propose a relaxation of the constraints used to define classic "hard" clustering (Eq. (6) and following text). This allows us to find solutions which are computationally tractable, while keeping valid the framework that explains clustering in terms of clusters' conductance in Markov chains. This was an important objective, as this framework is of probabilistic nature, and is thus a good candidate for a natural explanation of soft clustering. Our solution yields the probabilistic memberships of points in clusters Section 3.

The domain to which we have applied this technique, with a readily available code, is particularly challenging: the classification of words in multilingual corpora, in the case where (1) the language models are not given a priori; and (2) the different languages present in the corpora share a number of common words or word sequences. Our method allows us to point to homogeneous portions of texts, and to mixed portions of texts. In the case of code mixing between related languages, it also allows us to find the words belonging clearly to one language or the other, and to find the words which may belong to both, to a mixed degree.

The main advantage of the method we propose is that it gives an easy way to compute interpretation of the cluster structure information inherently contained in the distribution of data. It can be used in cases where the data can be viewed as the result of a stochastic process, and when a soft clustering is more meaningful than a hard clustering.

Acknowledgements

The authors wish to thank the reviewers for useful comments and suggestions that helped to improve the quality of this manuscript. R. Nock and P. Vaillant acknowledge support from Grant ANR JC9009 (ANR "Programme Jeunes Chercheurs"); R. Nock and F. Nielsen acknowledge support from Grant ANR-07-BLAN-0328-1 (ANR "Programme Blanc").

Some Creole texts we have used—like the novel *Lavwa egal*, from Térèz Léotin, used in many figures in this paper—are not free of rights, and have been entrusted to us for research purposes by their authors or collectors, whom we are pleased to thank here. Our thanks go to authors: Raphaël Confiant, Marie-Denise Grangenois, Térèz Léotin, Georges Mauvois, Manuel Norvat (Martinique), Vincent Morin (France); to linguists: Ralph Ludwig (Germany), Mikael Parkvall (Sweden), Stefan Pfänder (Germany); and to students who collected and transcribed oral corpora: Christelle Lengrai (Marie-Galante) and Juliette Moustin (Martinique).

Appendix A.

A.1. Proof of Theorem 4

We use the probabilistic method. Fix k,l such that $1 \leqslant k \neq l \leqslant q$. Suppose $\Sigma \in \{-1,+1\}^{\nu \times q}$ given, and $\Sigma' \in \{-1,+1\}^{\nu \times q}$ which differs from Σ by a single $\sigma_{i,l} \neq \sigma'_{i,l}$. One obtains

$$|f_{k,l}(\Sigma) - f_{k,l}(\Sigma')| \leqslant 2\sqrt{\tilde{y}_{i,k}\tilde{y}_{i,l}}, \tag{24}$$

and the same would hold whether the modification was made on column k. Now, provided Σ is picked uniformly at random, we have

$$\mathbf{E}_{\Sigma}(f_{k,l}(\Sigma)) = 0 \tag{25}$$

Together with Eqs. (24) and (25), the independent bounded difference inequality [17] on the random choices of Σ brings $\forall t_{k,l} \geqslant 0$:

$$\mathbf{Pr}(|f_{k,l}(\Sigma)| \geqslant t_{k,l}) \leqslant 2\exp\left(-\frac{t_{k,l}^2}{4\langle \tilde{\mathbf{y}}_k, \tilde{\mathbf{y}}_l \rangle}\right) \tag{26}$$

Fixing $t_{k,l} = 2\sqrt{\langle \tilde{\mathbf{y}}_k, \tilde{\mathbf{y}}_l \rangle \log(2q^2)}$ yields the upperbound $1/q^2$ on the probability. Thus, the probability that *some* couple among the $q(q-1)/2$ violates (26) is no more than $(q-1)/(2q) < 1/2$. Finally, with probability $> 1/2$ over the random choice of Σ, all possible couples (k,l) get concentrated following (26), and this brings the statement of the Theorem.

A.2. Proof of Theorem 5

For all $1 \leqslant k \leqslant m$, the circular likelihood of a text $\mathscr{T}_k$ of $\mathscr{C}$ is the following, under a bigram model:

$$\ell(\mathscr{T}_k) = \sum_{j=1}^{|\mathscr{T}_k|} \mathbf{Pr}(\omega_{k,j}) \prod_{l=1}^{|\mathscr{T}_k|-1} (\mathbf{Pr}(\omega_{k,l+1}|\omega_{k,l}) \times \mathbf{Pr}(\omega_{k,l}|\omega_{k,l+1})),$$

notice that the product does not depend on j since the circular writing of the text is made in the same way regardless of the first word chosen. The likelihood of $\mathscr{C}$ is

$$\ell(\mathscr{C}) = \prod_{k=1}^{m} \ell(\mathscr{T}_k) = z \times \prod_{i,j} (p_{i,j})^{n_{i,j}+n_{j,i}}, \tag{27}$$

where $z = \prod_{k=1}^{m} \sum_{j=1}^{|\mathscr{T}_k|} \mathbf{Pr}(\omega_{k,j})$ does not depend on P. The maximization of $\ell(\mathscr{C})$ under the ν constraints $\sum_{j=1}^{\nu} p_{i,j} = 1$ (with $1 \leqslant i \leqslant \nu$) may be obtained via the Lagrangian:

$$l\left(\mathscr{C}, \bigcup_{i=1}^{c} \lambda_i\right) = \ell(\mathscr{C}) + \sum_{i=1}^{c} \lambda_i \left(1 - \sum_{j=1}^{c} p_{i,j}\right).$$

Fix $1 \leqslant i \leqslant \nu$. Differentiating the Lagrangian with respect to $p_{i,j}$ and using Eq. (27), yields the following stationarity conditions ($\forall 1 \leqslant j \leqslant \nu$):

$$\frac{\partial l\left(\mathscr{C}, \bigcup_{i=1}^{\nu} \lambda_i\right)}{\partial p_{i,j}} = z s_{i,j} (n_{i,j} + n_{j,i}) (p_{i,j})^{n_{i,j}+n_{j,i}-1} - \lambda_i = 0, \tag{28}$$

with $s_{i,j} = \prod_{(k,l) \neq (i,j)} (p_{k,l})^{n_{k,l}+n_{l,k}}$. Eq. (28) yields ν expressions for λ_i, and if we equate two of them for $1 \leqslant j \neq j' \leqslant \nu$, we obtain $z s_{i,j}(n_{i,j} +$

Please cite this article as: R. Nock, et al., Soft memberships for spectral clustering, with application to permeable language distinction, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.06.024

$n_{j,i})(p_{i,j})^{n_{i,j}+n_{j,i}-1} = z s_{i,j'}(n_{i,j'} + n_{j',i})(p_{i,j'})^{n_{i,j'}+n_{j',i}-1}$, which, after simplification of $s_{i,\cdot}$, yields

$$(p_{i,j'})^{n_{i,j'}+n_{j',i}}(n_{i,j} + n_{j,i})(p_{i,j})^{n_{i,j}+n_{j,i}-1} = (p_{i,j})^{n_{i,j}+n_{j,i}}(n_{i,j'} + n_{j',i})(p_{i,j'})^{n_{i,j'}+n_{j',i}-1},$$

that is

$$p_{i,j'} = p_{i,j}(n_{i,j'} + n_{j',i})/(n_{i,j} + n_{j,i}). \quad (29)$$

Summing for $j' \neq j$ yields $1 - p_{i,j} = p_{i,j}(2n_i - n_{i,j} - n_{j,i})/(n_{i,j} + n_{j,i})$, i.e., $p_{i,j} = (n_{i,j} + n_{j,i})/(2n_i)$. It is easy to check that the stationary point found is the global maximum of the likelihood, as claimed.

References

[1] F.R. Bach, M.I. Jordan, Learning spectral clustering, in: S. Thrun, L. Saul, B. Schoelkopf (Eds.), NIPS* 16, MIT Press, Cambridge, MA, 2003.
[2] M. Belkhin, J. Goldsmith, in: Proceedings of the ACL'02 Workshop on Morphological and Phonological Learning, 2002.
[3] M. Belkhin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.), NIPS* 14, MIT Press, Cambridge, MA, 2001.
[4] C. Ding, A Tutorial on Spectral Clustering, in: Tutorials of the 21th ICML, 2004.
[5] R. Jin, C. Ding, F. Kang, A probabilistic approach for optimizing spectral clustering, in: NIPS* 18. MIT Press, Cambridge, MA, 2005.
[6] M. Meilă, J. Shi, S. Becker, Z. Ghahramani, Learning segmentation by random walks, in: T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.), NIPS* 14, MIT Press, Cambridge, MA, 2001.
[7] B. Mohar, Some applications of Laplace eigenvalues of graphs, in: G. Hann, G. Sabidussi (Eds.), Graph Symmetry: Algebraic Methods and Applications, NATO ASI Series, 1997, pp. 225–275.
[8] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm, in: T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.), NIPS* 14, MIT Press, Cambridge, MA, 2001.
[9] H. Zha, X. He, C. Ding, M. Gu, H. Simon, Spectral relaxation for k-means clustering, in: NIPS* 14, 2001, pp. 1057–1064.
[10] F.R.K. Chung, Spectral Graph Theory, vol. 92. Regional Conference Series in Mathematics, American Mathematical Society, Providence, RI, 1997.
[11] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE TPAMI 22 (2000) 888–905.
[12] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Stat. Soc. B 39 (1977) 1–38.
[13] R. Kannan, S. Vempala, A. Vetta, On clusterings: good, bad and spectral, J. ACM 51 (2004) 497–515.
[14] A. Sinclair, M. Jerrum, Approximate counting, uniform generation and rapidly mixing Markov chains, Inf. Comput. 82 (1989) 93–133.
[15] F. Peng, D. Schuurmans, Combining naive bayes and n-gram language models for text classification, in: Proceedings of the 25th ECIR, 2003.
[16] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, Annals of Statistics 28 (2000) 337–374.
[17] C. McDiarmid, Concentration, in: M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, B. Reed (Eds.), Probabilistic Methods for Algorithmic Discrete Mathematics, Springer, Berlin, 1998, pp. 1–54.

About the Author—RICHARD NOCK received the Agronomical Engineering degree from the Ecole Nationale Superieure Agronomique de Montpellier, France (1993), the PhD degree in Computer Science (1998), and an accreditation to lead research (HDR, 2002) from the University of Montpellier II, France. Since 1998, he has been a Faculty Member at the Universite Antilles-Guyane in Guadeloupe and in Martinique, where he is actually Full Time Professor of computer science. His primary research interests include machine learning, data mining, computational complexity, and image processing.

About the Author—PASCAL VAILLANT received the Engineering degree from the Institut National des Télécommunications, France (1992) and the PhD degree in Computer Science from the University of Paris-Orsay, France (1997). He is actually Assistant Professor in Linguistics and Computer Science at the Universite Antilles-Guyane, where his research interests range from machine learning and classification to semiotics.

About the Author—CLAUDIA HENRY received the Engineering degree from the Ecole Nationale Supérieure d'Ingénieurs en Mathématiques Appliquées de Grenoble, France (2001). She is actually PhD student in Computer Science, focusing on machine learning and classification.

About the Author—FRANK NIELSEN received the BSc and MSc degrees from Ecole Normale Superieure (ENS) of Lyon (France) in 1992 and 1994, respectively. He defended his PhD thesis on adaptive computational geometry prepared at INRIA Sophia-Antipolis (France) under the supervision of Professor J.-D. Boissonnat in 1996. As a Civil Servant of the University of Nice (France), he gave lectures at the engineering schools ESSI and ISIA (Ecole des Mines). In 1997, he served in the army as a Scientific member in the Computer Science laboratory of Ecole Polytechnique. In 1998, he joined Sony Computer Science Laboratories Inc., Tokyo (Japan), as a researcher. Now Senior Researcher, his current research interests include geometry, vision, graphics, learning, and optimization.

Please cite this article as: R. Nock, et al., Soft memberships for spectral clustering, with application to permeable language distinction, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.06.024