Grammatical Inference as a Principal Component Analysis

Raphaël BAILLY, François DENIS, Liva RALAIVOLA

LIF, Marseille
CNRS, Aix-Marseille Université

ICML 2009

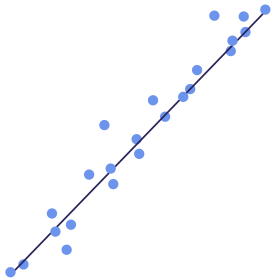
Strings from Σ^*

$S =$ACGTGACTGGTA, GTAACTGACGTGACTGACTG, CCGTACCT, GTACCTGATCT-TAACCGATCTGAC,...

$\Downarrow$

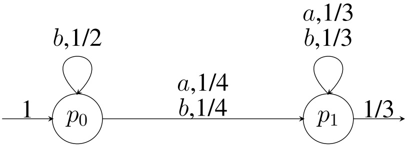
points of $l^2(\Sigma^*) \subset \mathbb{R}^{\Sigma^*}$

$p_S, \dot{A}p_S, \dot{C}p_S, \dot{G}p_S, \dot{T}p_S, \ldots$



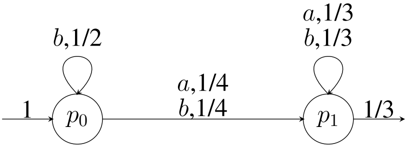
Grammatical Inference $\Leftrightarrow$ Finding the d-dimensional vector subspace wich minimizes the distance to the set of points

Probabilistic Automata (PA) $\simeq$ (HMM)



- starts on state p_0 with probability 1
- moves to state p_1 emitting symbol a with probability $1/4$
- stops on state p_1 with probability $1/3$

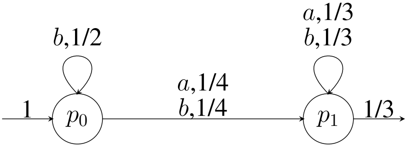
Probabilistic Automata



$$I = \begin{pmatrix} 1 \\ 0 \end{pmatrix} T = \begin{pmatrix} 0 \\ 1/3 \end{pmatrix} M_a = \begin{pmatrix} 0 & 1/4 \\ 0 & 1/3 \end{pmatrix} M_b = \begin{pmatrix} 1/2 & 1/4 \\ 0 & 1/3 \end{pmatrix}$$

- $p(ba) = I \times M_b \times M_a \times T \sim 0,069$

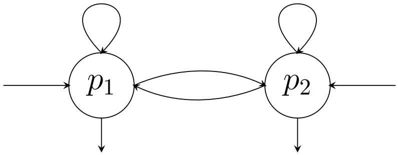
Probabilistic Grammatical Inference



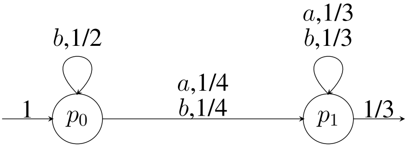
From a sample, find an automaton wich computes a probability distribution close to the underlying sample distribution

Algorithm: Baum-Welch [Baum et al. 1970]

- Structure of automaton known a priori (authorized states and transition)
- Sets coefficients to maximize likelihood of a training sample



Weighted Automata



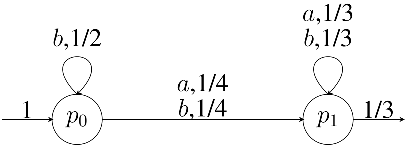
- Coefficients in $\mathbb{R}$
- $p(a_0 \dots a_n) = I \times M_{a_0} \cdots \times M_{a_n} \times T$

Residuals

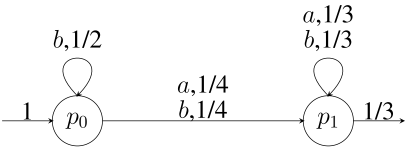
- $\dot{u} : \mathbb{R}^{\Sigma^*} \mapsto \mathbb{R}^{\Sigma^*}$ for $u \in \Sigma^*$
- $\dot{u}r(w) = r(uw)$

- Residuals of r: linear combination of $\dot{u}r$
- Residual space of r: vector space spanned by the residuals of r

- A mapping r is computed by a *WA* (i.e is a rational series) if and only if its Residual space has a finite dimension



- States ⇔ Residuals (Minimal Case: base of the Residual space)
- Coefficients: linear relations between residuals
- $\dot{b}p_0 = \frac{1}{2}p_0 + \frac{1}{4}p_1$



- $\dot{I}$: p in the base (p_0, p_1)
- $p = 1 \times p_0 + 0 \times p_1$
- matrix M_a: matrix of $\dot{a}$ in the base (p_0, p_1)
- $\dot{a}p_0 = \frac{1}{4}p_1$, $\dot{a}p_1 = \frac{1}{3}p_1$

$$I = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

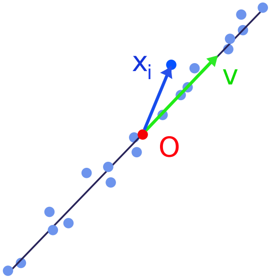
$$M_a = \begin{pmatrix} 0 & 1/4 \\ 0 & 1/3 \end{pmatrix}$$

Consequences

- B a base of the Residual space of r (dimension d) $\Leftrightarrow$ Transition matrices of a d-state automaton wich computes r
- I = coordinates of r in this base
- T = empty word probability of the base residuals

Principal Component Analysis

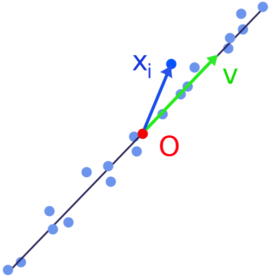
- $\{x_i\}$ a set of points in a vector space E with a distance
- For a given dimension d, one looks for a vector subspace F_d of E wich minimizes the sum of the squares of the distances from x_i to F_d (Reconstruction Error)



PCA- Dot product

If E is equipped with a dot product, F_d is spanned by $v_1 \dots v_d$, eigenvectors associated to the d first eigenvalues of $M=$ variance matrix of $\{x_i\}$

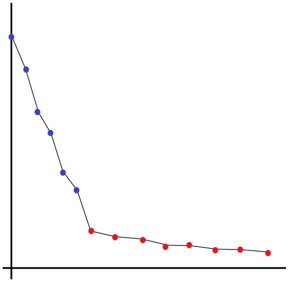
The sum of the remainig eigenvalues is equal to the reconstruction error



Elbow and Dimension

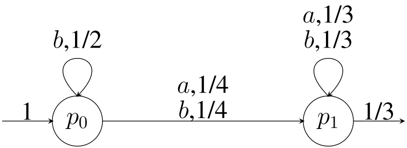
After the eigenvalue "elbow", the eigenvectors are meaningless.

Here, only the vectors associated to the blue eiegnvalues will be kept.

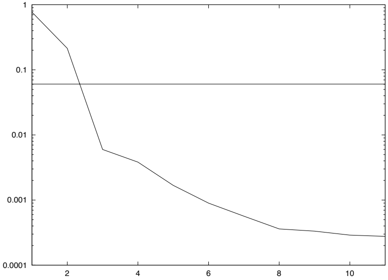


Finding the automaton rank

- S a sample, p_S the empirical distribution, $N = \{\dot{w}p_S, w \in \Sigma^*\}$
- Perform a PCA on N
- Use upper bound of the reconstruction error to find a lower bound of the dimension
- Find the elbow on the eigenvalues curve greater than this bound



Automate A, S i.i.d w.r.t p_A, $|S| = 1000$



Finding the parameters of the Automaton

The dimension d is given.

- PCA on the residuals: base $\{w_1 \ldots w_d\}$ of eigenvectors, spanning V_d
- Π_{V_d} is the projection upon V_d. $\dot{a}$ is the linear mapping: $r \in \Sigma^*, r \to \dot{a}r$
- Given $x \in \Sigma$, the matrix $M_x =$ matrix of $\Pi_{V_d} \circ \dot{x}$ in the base $\{w_1 \ldots w_d\}$
- $I =$ cordinates of $\Pi_{V_d}(p_S)$ in the base $\{w_1 \ldots w_d\}$
- $T = (w_1(\epsilon), \ldots, w_d(\epsilon))$

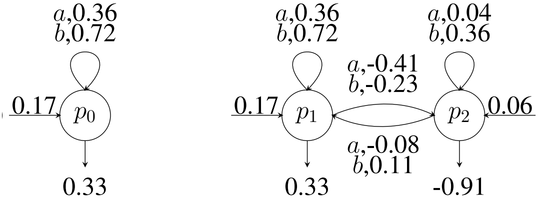


Figure: Computed automata for $d = 1$ (A_1) and $d = 2$ $(A_2)(|S| = 1000)$

	ε	a	b	aa	ab	ba	bb
p_A	0.0	0.083	0.083	0.028	0.028	0.069	0.069
$p_{r_{A_2}}$	0.000	0.10	0.086	0.028	0.030	0.077	0.072

Properties

- Identification in the limite of the rank (Number of states)
- Convergence of the automaton's coefficients towards those of the target in $O(1/n^{1/2})$

Consequence:

- l_1-convergence of the estimated distribution to the target

Toy examples

- 500 randomly generated automata with 4 states on a 2 letters alphabet
- Building automata for several number of states
- Rank selection with several criteria: distance minimization (l_1, l_2 ou KL), eigenvalues curve

$\|S\| = 100000$	$\|\|\|\|_1$	$\|\|\|\|_2$	KL-divergence	Eigenvalue curve
Correct rank	48%	29%	13%	60%

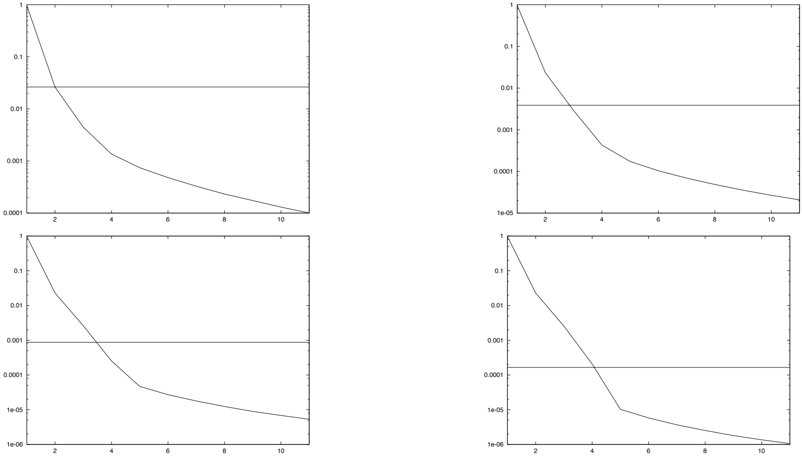


Figure: Eigenvalues for sample size of 1000, 5000, 20000 and 100000.

Biological data

- Data: DNA sequences of a promoter (C.Jejuni)
- Learning sample: 140 strings of 122 bases, Test sample: 35 strings
- HMM Structure (based on a priori biological knowledge): 11 states [Petersen et al. 03], 10 states [Won et al. 04]
- Comparison between Baum-Welch on HMM, and boosted PCA

Results

- 7-state Weighted Automaton
- Improved likelihood performances on the test sample with PCA method

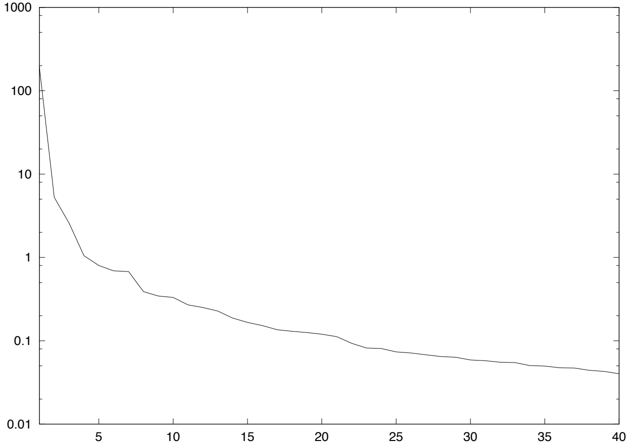


Figure: Eigenvalues curve for biological data.

Conclusion

- Probabilistic Grammatical Inference method with convergence theoretical results
- Good performances compared to generally used methods
- Inner product-based method: one can extend to kernel metrics, akin to Kernel PCA [Schölkopf Smola Müller 99], and embedding distribution in an RKHS [Smola Gretton Song Schölkopf 07]