Grammatical Inference as a Principal Component Analysis

Raphaël BAILLY, François DENIS, Liva RALAIVOLA

LIF, Marseille
CNRS, Aix-Marseille Université

ICML 2009
Overview

Automata

Residuals

PCA

Algorithm

Results

Experiments

Conclusion and Future works

Raphaël BAILLY, François DENIS, Liva RALAIVOLA

LIF, Marseille CNRS, Aix-Marseille Université

Grammatical Inference as a Principal Component Analysis
Strings from $\Sigma^*$

$S = $ACGTGACTGGTA, GTAACCTGACGTGACTG, CCGTACCT, GTACCTGATCTCTAACCGATCTGAC,...

⇓

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

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

Grammatical Inference $\Leftrightarrow$
Finding the $d$-dimensional vector subspace which minimizes the distance to the set of points
Probabilistic Automata (PA) \(\sim\) (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

\[ I = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad T = \begin{pmatrix} 0 & 0 \\ 1/3 & 1/3 \end{pmatrix}, \quad M_a = \begin{pmatrix} 0 & 1/4 \\ 0 & 1/3 \end{pmatrix}, \quad M_b = \begin{pmatrix} 1/2 & 1/4 \\ 0 & 1/3 \end{pmatrix} \]

\[ p(ba) = I \times M_b \times M_a \times T \sim 0.069 \]
Probabilistic Grammatical Inference

From a sample, find an automaton which 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 $\mathbb{R}$
- $p(a_0 \ldots a_n) = I \times M_{a_0} \cdots \times M_{a_n} \times T$
<table>
<thead>
<tr>
<th>Overview</th>
<th>Automata</th>
<th>Residuals</th>
<th>PCA</th>
<th>Algorithm</th>
<th>Results</th>
<th>Experiments</th>
<th>Conclusion and Future works</th>
</tr>
</thead>
</table>

**Overview**

**Automata**

**Residuals**

**PCA**

**Algorithm**

**Results**

**Experiments**

**Conclusion and Future works**
Residuals

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

- Residuals of $r$: linear combination of $\dot{ur}$
- 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 $\Leftrightarrow$ Residuals (Minimal Case: base of the Residual space)

- Coefficients: linear relations between residuals
- $bp_0 = \frac{1}{2}p_0 + \frac{1}{4}p_1$
Grammatical Inference as a PCA Problem

Algorithm 2

Let \( S \) be a sample i.i.d. according to a rational stochastic language. Then \( \rho_{\text{MA}} \) converges to the target as the size of the sample.

Let \( \lambda_i \) be the first \( d \) eigenvalues of the matrix \( M_a \) of \( \dot{a} \) in the base \((p_0, p_1)\).

\[ l = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \]

\[ M_a = \begin{pmatrix} 0 & 1/4 \\ 0 & 1/3 \end{pmatrix} \]
Consequences

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

Automata

Residuals

PCA

Algorithm

Results

Experiments

Conclusion and Future works

Raphaël BAILLY, François DENIS, Liva RALAIVOLA

LIF, Marseille CNRS, Aix-Marseille Université

Grammatical Inference as a Principal Component Analysis
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 $E$ which minimizes the sum of the squares of the distances from $x_i$ to $F_d$ (Reconstruction Error)
If $E$ is equipped with a dot product, $F_d$ is spanned by $v_1 \ldots v_d$, eigenvectors associated to the $d$ first eigenvalues of $M = \text{variance matrix of } \{x_i\}$

The sum of the remaining 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 eigenvalues will be kept.
Overview

Automata

Residuals

PCA

Algorithm

Results

Experiments

Conclusion and Future works
Finding the automaton rank

- $S$ a sample, $p_S$ the empirical distribution, $N = \{ \hat{w} p_S, w \in \Sigma^* \}$
- 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 $V_d$

- $\Pi_{V_d}$ is the projection upon $V_d$. $\dot{a}$ is the linear mapping:
  $r \in \Sigma^*$, $r \rightarrow \dot{a}r$

- Given $x \in \Sigma$, the matrix $M_x = \text{matrix of } \Pi_{V_d} \circ \dot{x}$ in the base $\{w_1 \ldots w_d\}$

- $I = \text{coordinates of } \Pi_{V_d}(p_S)$ 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 $d = 2$ ($A_2$) ($|S| = 1000$)

<table>
<thead>
<tr>
<th></th>
<th>$\varepsilon$</th>
<th>$a$</th>
<th>$b$</th>
<th>$aa$</th>
<th>$ab$</th>
<th>$ba$</th>
<th>$bb$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_A$</td>
<td>0.0</td>
<td>0.083</td>
<td>0.083</td>
<td>0.028</td>
<td>0.028</td>
<td>0.069</td>
<td>0.069</td>
</tr>
<tr>
<td>$p_{A_2}$</td>
<td>0.000</td>
<td>0.10</td>
<td>0.086</td>
<td>0.028</td>
<td>0.030</td>
<td>0.077</td>
<td>0.072</td>
</tr>
</tbody>
</table>
Overview

Automata

Residuals

PCA

Algorithm

Results

Experiments

Conclusion and Future works

Grammatical Inference as a Principal Component Analysis
Properties

- Identification in the limit 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
Overview

Automata

Residuals

PCA

Algorithm

Results

Experiments

Conclusion and Future works

Grammatical Inference as a Principal Component Analysis
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.
<table>
<thead>
<tr>
<th>Overview</th>
<th>Automata</th>
<th>Residuals</th>
<th>PCA</th>
<th>Algorithm</th>
<th>Results</th>
<th>Experiments</th>
<th>Conclusion and Future works</th>
</tr>
</thead>
</table>

**Overview**

**Automata**

**Residuals**

**PCA**

**Algorithm**

**Results**

**Experiments**

**Conclusion and Future works**

Raphaël BAILLY, François DENIS, Liva RALAIVOLA

LIF, Marseille CNRS, Aix-Marseille Université

Grammatical Inference as a Principal Component Analysis
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]