Global Estimation of Hidden Markov Models using Interval Arithmetic

Tiago M. Montanher
Walter F. Mascarenhas
Outline

• The Problem
  • Statement
  • Objectives

• Interval arithmetic algorithm
  • Interval branch and bound
  • Evaluating the objective function
  • Discarding boxes

• Numerical experiments
  • Plotting solution sets
  • Basins of attraction
What is a Hidden Markov Model?

Dilma’s voting intention

<table>
<thead>
<tr>
<th>State</th>
<th>High</th>
<th>Low</th>
<th>Initial State</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>70%</td>
<td>30%</td>
<td>60%</td>
</tr>
<tr>
<td>Low</td>
<td>50%</td>
<td>50%</td>
<td>40%</td>
</tr>
</tbody>
</table>

Petrobras Stock value PETR4($)

<table>
<thead>
<tr>
<th>Performance</th>
<th>Increase</th>
<th>Unchanged</th>
<th>Decrease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>30%</td>
<td>50%</td>
<td>20%</td>
</tr>
<tr>
<td>Probability</td>
<td>60%</td>
<td>30%</td>
<td>10%</td>
</tr>
</tbody>
</table>

We observe only the Market performance (Increase, Decrease, Unchanged, Decrease,...)
Hidden Markov Models

- Let $q_t$ be a discrete time Markov chain assuming states $S_1, \ldots, S_N$.

- Let $O_t$ be a discrete time stochastic process assuming states $V_1, \ldots, V_M$ and satisfying

  \[ P(O_t = V_i | q_t = S_j) = b_{ij} \]

- A Hidden Markov Model is a triple $\theta = (A, B, \pi)$ where
  \[ \sum_{j=1}^{N} a_{ij} = 1, \quad \forall i = 1, \ldots, N. \quad a_{ij} \geq 0 \]
  \[ \sum_{i=1}^{M} b_{ij} = 1, \quad \forall j = 1, \ldots, N. \quad b_{ij} \geq 0 \]
  \[ \sum_{i=1}^{N} \pi_i = 1. \quad \pi_i \geq 0 \]
Problem Statement

We are given -

1. A set of observations $O_1, \ldots, O_T$.

2. The number of states $N$ for the hidden process.

3. The number of states $M$ for the observable process

$$\max z = \mathbb{P}(\theta|O)$$

$$e_N^T A_i = 1 \quad \forall i = 1, \ldots, N.$$  

$$e_M^T B^i = 1 \quad \forall i = 1, \ldots, N.$$  

$$e_N^T \pi = 1$$  

$$a_{ij}, b_{ij}, \pi_i \geq 0.$$
Computing Probabilities

- We evaluate probabilities through the formula
  \[ \mathbb{P}(O|\theta) = \pi^T d(B_{O_1}) Ad(B_{O_2}) \ldots Ad(B_{O_T}) \mathbf{1} \]

  where \( d(B_{O_t}) \) is the diagonal matrix given by line \( O_t \).
- Function \( \mathbb{P}(O|\theta) \) is non-convex.
- Since all parameters are non-negative, the objective function and their derivatives are non-negative.
- In order to make the evaluation easier we employ the backward recursion
  \[
  \beta_T = \mathbf{1} \\
  \beta_t = Ad(B_{O_{t+1}}) \beta_{t+1} \quad t = T - 1, \ldots, 1 \\
  \mathbb{P}(\theta|O) = \pi^T d(B_{O_1}) \beta_1
  \]
Objectives

State of the Art

- Baum-Welch algorithm is a local method widely used to maximize $\mathbb{P}(\theta|O)$.
- It is a good option if the number of observations is large.

Our Aims

- To develop a global optimization algorithm able to find all isolated global maxima of $\mathbb{P}(\theta|O)$.
- To study the convergence properties of Baum-Welch algorithm.
Baum-Welch Algorithm

**Input** – An observation set $O$ and an initial estimate $\theta^0$.

**Output** – A (local) approximation for the maximum likelihood estimator $\hat{\theta}$.

\[
\begin{align*}
    a_{ij}^{t+1} &= 
    \frac{a_{ij}^t \frac{\delta \mathbb{P}(\theta | O)}{\delta a_{ij}}(\theta^t)}{
    \sum_{k=1}^N a_{ik}^t \frac{\delta \mathbb{P}(\theta | O)}{\delta a_{ik}}(\theta^t)} \\
    b_{ij}^{t+1} &= 
    \frac{b_{ij}^t \frac{\delta \mathbb{P}(\theta | O)}{\delta b_{ij}}(\theta^t)}{
    \sum_{k=1}^M b_{kj}^t \frac{\delta \mathbb{P}(\theta | O)}{\delta b_{kj}}(\theta^t)} \\
    \pi_i^{t+1} &= 
    \frac{\pi_i^t \frac{\delta \mathbb{P}(\theta | O)}{\delta \pi_i}(\theta^t)}{
    \sum_{k=1}^N \pi_k^t \frac{\delta \mathbb{P}(\theta | O)}{\delta \pi_k}(\theta^t)}
\end{align*}
\]
Interval Branch and Bound

**Input** – An interval extension $\mathbb{P}(\theta|O)$ of $\mathbb{P}(\theta|O)$.

**Output** – A list of boxes which contains all global maximum likelihood estimator

![Algorithm Diagram]

End of Algorithm  \[\text{List is Empty}\]  \[\text{Box Selection}\]

\[\text{Delete Box?}\]

\[\text{N}\]  \[\text{Upper Bound Update}\]

\[\text{Delete Boxes without glob. min}\]

\[\text{Newton Reduction}\]

\[\text{Save as a Candidate}\]

\[\text{Candidate?}\]

\[\text{N}\]  \[\text{Box Splitting}\]
• If we just replace the floating point operations by interval operations we have a natural extension

\[ P(\theta | O) = \pi^T d(B_{O_1})/\beta_1 \]

• For example, if \( N = 2 \) and \( M = 3 \) HMM with \( O = \{1, 3\} \) and a box \( x = [0.25, 0.75]^{12} \) then

\[ P(\theta | O) = [0.015625, 1.26562] \]

• We propose an extension based on the simplices structure of the problem. Let \( c = d(\beta_{O_{t+1}})/\beta_{t+1} \) then

\[ \beta_t(i) = \begin{bmatrix} \min c^T A_i & \max c^T A_i \\ e^T A_i = 1, & e^T A_i = 1 \\ a_{ij} \in a_{ij} & a_{ij} \in a_{ij} \end{bmatrix} \]

\[ P(\theta | O) = [0.0625, 0.5625] \]
Computing probabilities – Interval Arithmetic

\[
([a_1, a_1], \ldots, [a_N, a_N]) \quad \left[\begin{array}{c}
[b_1, b_1] \\
\vdots \\
[b_N, b_N]
\end{array}\right]
\]

4 * N + 3 rounding mode switching per backward iteration

Since we have only non-negative coefficients

\[
\left(\begin{array}{c}
a_1 \\
a_N \\
\vdots \\
a_1 \\
a_N
\end{array}\right) \quad \left(\begin{array}{c}
a_1 \\
a_N \\
\vdots \\
a_1 \\
a_N
\end{array}\right)
\]

3 rounding mode switching to evaluate the objective function and their derivatives
KKT Conditions

- If $\hat{\theta}$ is a solution then there exists $(\hat{\lambda}, \hat{\mu})$ such that

$$\frac{\delta \mathbb{P}(\theta|O)}{\delta a_{ij}}(\hat{\theta}) - \hat{\lambda}_{A_{i}} - \hat{\mu}_{a_{ij}} = 0 \quad \forall i, j = 1, \ldots, N.$$  

$$\frac{\delta \mathbb{P}(\theta|O)}{\delta b_{ij}}(\hat{\theta}) - \hat{\lambda}_{B_{i}} - \hat{\mu}_{b_{ij}} = 0 \quad \forall i = 1, \ldots, N \text{ and } j = 1, \ldots, M.$$  

$$\frac{\delta \mathbb{P}(\theta|O)}{\delta \pi_{i}}(\hat{\theta}) - \hat{\lambda}_{\pi} - \hat{\mu}_{\pi_{i}} = 0 \quad \forall i, j = 1, \ldots, N.$$  

$$\epsilon^{T} \hat{\theta}_{A_{i}} = 1 \quad \forall i = 1, \ldots, N.$$  

$$\epsilon^{T} \hat{\theta}_{B_{i}} = 1 \quad \forall i = 1, \ldots, N.$$  

$$\epsilon^{T} \hat{\theta}_{\pi} = 1$$  

$$\mu_{a_{ij}} \hat{\theta}_{a_{ij}} = 0 \quad \forall i, j = 1, \ldots, N.$$  

$$\mu_{b_{ij}} \hat{\theta}_{b_{ij}} = 0 \quad \forall i = 1, \ldots, N \text{ and } j = 1, \ldots, M.$$  

$$\mu_{\pi_{i}} \hat{\theta}_{\pi_{i}} = 0 \quad \forall i, j = 1, \ldots, N.$$  

$$\mu_{a_{ij}}, \mu_{b_{ij}}, \mu_{\pi_{i}} \geq 0$$
Discarding Boxes

- From KKT conditions, we have

\[
\frac{\delta \mathbb{P}(\theta \mid O)}{\delta a_{ij}}(\hat{\theta}) - \hat{\lambda}_{Ai} - \hat{\mu}_{aij} = 0
\]

\[
\frac{\delta \mathbb{P}(\theta \mid O)}{\delta a_{ik}}(\hat{\theta}) - \hat{\lambda}_{Ai} - \hat{\mu}_{aik} = 0
\]

- Let \( g = \frac{\delta f}{\delta a_{ij}}(\theta) - \frac{\delta f}{\delta a_{ik}}(\theta) \), by complementary conditions:

1. \( 0 < g \Leftrightarrow \hat{\mu}_{aij} > 0 \) and \( \theta_{aij} = 0 \)

2. \( 0 > g \Leftrightarrow \hat{\mu}_{aik} > 0 \) e \( \theta_{aik} = 0 \).
Discarding Boxes

- We can re-label the observations in order that the most frequent is labeled by 1 the second most frequent 2, ... 

- After re-labeling we have that

\[ b_{11} \geq b_{1j} \quad j = 2, \ldots, N. \]

\[ b_{11} \geq \frac{1}{N}. \]

And, in general

\[ b_{ii} \geq b_{ij} \quad j = i + 1, \ldots, N. \]

- We discard any box that do not satisfy the constraints above.

- We also apply constraint propagation using these inequalities in order to reduce boxes.
Implementation details

- Algorithm in C++11.
- We developed our own interval arithmetic library based on cfenv.h library.
- Avoid round mode switching. Only three per function or gradient evaluation.
- To avoid underflow/overflow we use the functions frexp and scalbln at each vector operation.
- Interval analysis methods -
  - Midpoint test,
  - Constraints propagation(linear and Taylor based),
  - Taylor evaluation of objective function.
Solution set for all instances with $N = 2$, $M = 2$ and $T = 8$ (254 instances)
Numerical Experiments

Solution set for all instances with $N = 2$, $M = 2$ and $T = 9$ (510 instances)
Numerical Experiments

Solution set for all instances with $N = 2$, $M = 2$ and $T = 10$ (1022 instances)
Basins of Attraction – $N = 2$, $M = 2$ and $O = 0000011001$
Basins of Attraction – $N = 2$, $M = 2$ and $O = 1001000100001001000101111$
• We provide an alternative to estimate parameters of Hidden Markov Models.

• We derive a new interval extension and discarding tests based on the structure of the problem.

• Only few models can be correctly predicted when we have a small set of observations.

• The Baum Welch algorithm does not find a global maximum likelihood estimator for more than 50% of initial points.
Questions?