Stat 991: Multivariate Analysis, Dimensionality Reduction, and Spectral Methods **Lecture: 22**

Spectral methods for learning HMMs

Instructor: Sham Kakade

1 The Transformed Representation

Assume M is invertible. For $x = 1, \ldots, d$, define

$$\widetilde{A}_x = M A_x M^{-1}$$

Also, as before, define $\widetilde{h}_t = M h_t$ and

$$\widetilde{g}_t = \mathbb{E}[\widetilde{h}_t | x_{<t}] = M \widetilde{g}_t$$

We now have the following updates.

Lemma 1.1. *In this representation, the HMM update rules are:*

$$\begin{aligned}
\widetilde{g}_1 &= M\pi_1 \\
\widetilde{g}_\infty^\top &= 1_m^\top M^{-1} \\
\widetilde{g}_{t+1} &= \frac{\widetilde{A}_{x_t} \widetilde{g}_t}{\widetilde{g}_\infty^\top \widetilde{A}_{x_t} \widetilde{g}_t} \\
\Pr[x_{t+1} | x_1, \ldots, x_t] &= \widetilde{g}_\infty^\top \widetilde{A}_{x_{t+1}} g_{t+1}
\end{aligned}$$

We also have that:

$$\Pr[x_1, \ldots, x_t] = \widetilde{g}_\infty^\top \widetilde{A}_{x_t} \ldots \widetilde{A}_{x_1} \widetilde{g}_1$$

Proof. The equations follow directly from the definitions and our HMM representation. The last equation (in the first claim) follows from $UM = O$. The second claim also follows from our previous expression (in the last lecture) of the joint probability $\Pr[x_1, \ldots, x_t]$. □

2 Learning

Assumption 1. *Assume that T and O are full rank. Also, assume that $\pi_1 > 0$.*

Define the following matrices:

$$\begin{aligned}
[P_1]_i &= \Pr(x_1 = i) \\
[P_{2,1}]_{i,j} &= \Pr(x_2 = i, x_1 = j) \\
[P_{3,x,1}]_{i,j} &= \Pr(x_3 = i, x_2 = x, x_1 = j)
\end{aligned}$$

Theorem 2.1. *Let the "thin" SVD of the cross correlation matrix at some timestep τ be $E[x_{\tau+1} x_\tau^\top] = UDV^\top$. Let $M = U^\top O$. Then M is invertible. Furthermore,*

$$\begin{aligned}
\widetilde{g}_1 &= = U^\top P_1 \\
\widetilde{g}_\infty &= \left(P_{2,1}^\top U\right)^+ P_1 \\
\widetilde{A}_x &= \left(U^\top P_{3,x,1}\right) \left(U^\top P_{2,1}\right)^+ \quad \forall x \in [d]\ .
\end{aligned}$$

Proof. For $\widetilde{g}_1$, we have that:

$$\widetilde{g}_1 = M\mathbb{E}[h_1] = U^\top E[x_1] = U^\top P_1$$

Now let us prove the equation for $\widetilde{A}_x$. Define:

$$[X]_{i,j} = \Pr(h_2 = i, x_1 = j)$$

and define:

$$[Y_x]_{i,j} = \Pr(h_3 = i, x_2 = x, x_1 = j)$$

We have that:

$$[Y_x]_{\cdot,j} = A_x[X]_{\cdot,j}$$

Hence,

$$Y_x = A_x X$$

and so:

$$U^\top(OY_x) = U^\top O A_x M^{-1} M X = \widetilde{A}_x U^\top(OX)$$

This and by definition of the P's,

$$U^\top P_{3,x,1} = \widetilde{A}_x U^\top P_{2,1}$$

which proves the result (using the rank conditions to argue that $U^\top P_{2,1}$ is rank k). For $\widetilde{g}_\infty$, first note that:

$$1^\top X = P_1^\top$$

and so:

$$1^\top M^{-1} U^\top O X = P_1^\top$$

Thus:

$$1^\top M^{-1} U^\top P_{2,1} = P_1^\top$$

which proves the result. □

3 General observation events

We can consider arbitrary past observation events, in vector representation denoted by $X_{t,p}$ (which an events vector determined by $x_{t-1}, x_{t-2}, \ldots x_\infty$), as opposed to just singleton observations x_1; and arbitrary future observation events in vector representation $X_{t,f}$ (an events vector determined by $x_t, x_{t+1}, \ldots$) as opposed to just singleton observations. Let the set of some past events be $\{1, \ldots, m_p\}$, which is represented as an m_p-dimensional vector; and let the set of some future events be $\{1, \ldots, m_f\}$, which is represented as an m_f-dimensional vector.

Let us assume that $E[h_1]$ is the stationary distribution (and that time goes back to $-\infty$).. Define the event matrix $\widetilde{O}^p \in \mathbb{R}^{m_p \times K}$ by

$$\widetilde{O}^p_{\cdot,j} = \mathbb{E}\,[X_{t,p} | h_t = j].$$

which is not time varying as we have assumed the chain starts at the stationary distribution. Similarly, define $\widetilde{O}^f \in \mathbb{R}^{m_f \times K}$ by

$$\widetilde{O}^f_{\cdot,j} = \mathbb{E}\,[X_{t,f} | h_t = j].$$

which is again not time varying.

Define the matrix $\widetilde{P}_{2,1} \in \mathbb{R}^{m_f \times m_p}$ by

$$\widetilde{P}_{2,1} = \mathbb{E}\; X_{2,f} X_{1,p}^\top.$$

Then

$$\widetilde{P}_{2,1} = \widetilde{O}^f T \operatorname{diag}(\pi) \widetilde{O}^{p\top}$$

where T is our usual transition matrix, taking us from h_1 to h_2.

Lemma 3.1. *Assume that the HMM representation is "minimal" — that there is no HMM, with a fewer number of hidden states, which has identical probabilities for observable sequences. Define the range of the process to be* $\mathrm{span}\{\mathbb{E}[x_t|x < t]|x_{<t}\}$*, and the dimension of the process to be the dimension of this range.*

There exist a set of past and future events such that the rank of $\widetilde{P}_{2,1}$ *is the dimension of the process. Furthermore, let the thin SVD* $\widetilde{P}_{2,p} = U\Sigma V^\top$ *and let* $M = U^\top O^f$*, there exists an* $\widetilde{M}$ *such that* $\widetilde{M}M$ *acts as the identity on any belief state (this* $\widetilde{M}$ *may not be the pseudo-inverse, but it acts as the inverse on the belief states).*

Define $\widetilde{P}_{3,x,p} \in \mathbb{R}^{m_f \times m_p}$ by

$$\widetilde{P}_{3,x,p} = \Pr[x_2 = x]\mathbb{E}\left[X_{3,f}X_p^\top | x_2 = x\right].$$

Then

$$\begin{aligned} U^\top \widetilde{P}_{3,x,p} &= U^\top \widetilde{O}^f A_x T \,\mathrm{diag}(\pi)\widetilde{O}^{p\top} \\ &= U^\top \widetilde{O}^f A_x (U^\top \widetilde{O}^f)^{-1}(U^\top \widetilde{O}^f) T \,\mathrm{diag}(\pi)\widetilde{O}^{p\top} \\ &= (U^\top \widetilde{O}^f) A_x (U^\top \widetilde{O}^f)^{-1}(U^\top \widetilde{P}_{2,p}) \end{aligned}$$

so

$$B_x = (U^\top \widetilde{O}^f) A_x (U^\top \widetilde{O}^f)^{-1} = (U^\top \widetilde{P}_{3,x,p})(U^\top \widetilde{P}_{2,p})^+$$

Acknowledgements

Part of these notes were modified from Daniel Hsu's notes (on the general observation events).