# HW 6: PCA and Markov model

Anjali Krishna Prasad

November 2016

## 1 Principal Component Analysis

a)

$$J = \frac{1}{N}\sum_{i=1}^{N}(x_i - p_{i_1}e_1 - p_{i2}e_2)^T(x_i - p_{i_1}e_1 - p_{i2}e_2)$$

Differentiating the above equation wrt pi2

$$\frac{\partial J}{\partial p_{i2}} = \frac{2}{N}\sum_{i=1}^{N}(x_i - p_{i1}e_1 - p_{i2}e_2).(-e_2) = 0$$

$$-x_i e_2^T + e_2 e_1 p_{i1} + e_2^T e_2 p_{i2} = 0$$

as

$$e_2 e_1 = 0; e_2.e_2 = 1$$

$$-x_i e_2^T + p_{i2} = 0$$

$$p_{i2} = e_2^T x_i$$

b)

$$J = -e_2^T S e_2 + \lambda_2(e_2^T e_2 - 1) + \lambda_{12}(e_2^T e_1 - 0)$$

We know that:

$$e_2^T e_2 = 1; e_2^T e_1 = 0$$

Substituting these values in above equation we get:

$$J = -e_2^T S e_2 + 0 + 0 = -e_2^T S e_2$$

differentiating the above equation wrt e2

$$\frac{\partial J}{\partial e_2} = -(S + S^T)e_2 = 0$$

$$-Se_2 - S^T e_2 = 0$$

As matrix S is symmetric

$$S = S^T$$

$$-\lambda_2 e_2 - \lambda_2 e_2 = 0$$

$$-2\lambda_2 e_2 = 0$$

As lambda2 cannot be 0, e2 = 0 . therefore lambda2 minimizes e2 where lambda2 is second largest eigenvalue. Hence proved.

$$\lambda_1 = 1626.52, \qquad \lambda_2 = 128.99, \qquad \lambda_3 = 7.10$$

$$\vec{u}_1 = \begin{bmatrix} 0.22 \\ 0.41 \\ 0.88 \end{bmatrix}, \quad \vec{u}_2 = \begin{bmatrix} 0.25 \\ 0.85 \\ -0.46 \end{bmatrix}, \quad \vec{u}_3 = \begin{bmatrix} 0.94 \\ -0.32 \\ -0.08 \end{bmatrix}$$

## 2 A Real Example

a)Finding eigen values and eigen vectors of the matrix in the problem:

b)lambda1 is much larger than lambda2 and lambda3. Principal component u1 accounts for 92.28 percent of the variation in the data,

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = 98.28 Percent$$

and the second u2 accounts for 7.32 percent .The third u3 accounts for only 0.40 percent of data and is negligible compared to the first two. Therefore u3 orthonormal direction corresponding to eigen value lambda3 can be omitted without loosing much information.

c) lambda1 is much larger than lambda2 and lambda3. Principal component u1 accounts for 92.28 percent of the variation in the data,

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = 98.28 Percent$$

and the second u2 accounts for 7.32 percent .The third u3 accounts for only 0.40 percent. Therefore vector u1 corresponding to eigen value lambda1 contains the most of information regarding this data.

the third entry, weight, of u1 is the largest, so weight is the most significant. This means a change in one unit of weight tends to affect the size more so than a change in one unit of length or wingspan. The second entry of u1 is the next largest, which corresponds to wingspan. Thus, wingspan is the next most important factor in determining a bird's size

## 3 Hidden Markov Model

a)Probability of an observed sequence. Calculate

$$P(O; \theta)$$

We will use Forward Probability algorithm.We are given the following:

initail state probability

| | S1 | S2 |
|---|---|---|
| X | 0.6 | 0.4 |

transition probabilities

| | X2 | |
|---|---|---|
| | S1 | S2 |
| | 0.7 | 0.3 |
| X1 | 0.4 | 0.6 |

Emission probabiities

| | A | C | G | T |
|---|---|---|---|---|
| S1 | 0.4 | 0.2 | 0.3 | 0.1 |
| S2 | 0.2 | 0.4 | 0.1 | 0.3 |

Figure 1: The given data

Given Data:

$$\pi_1 = P(X = S_1) = 0.6; \pi_2 = P(X_1 = S_2) = 0.4$$

$$P(X_2 = S_1 | X_1 = S_1) = a_{11} = 0.7; P(X_2 = S_2 | X_1 = S_1) = a_{12} = 0.3; P(X_2 = S_1 | X_1 = S_2) = a_{21} = 0.4$$

$$P(X_2 = S_2 | X_1 = S_2) = a_{22} = 0.6$$

$$b_{1A} = P(X_t =' A' | X_t = S_1) = 0.4; b_{1C} = P(X_t =' C' | X_t = S_1) = 0.2; b_{1G} = P(X_t =' G' | X_t = S_1) = 0.3$$

$$b_{1T} = P(X_t =' T' | X_t = S_1) = 0.1; b_{2A} = P(X_t =' A' | X_t = S_2) = 0.2; b_{2C} = P(X_t =' C' | X_t = S_2) = 0.4$$

$$b_{2G} = P(X_t =' G'|X_t = S_2) = 0.1; b_{2T} = P(X_t =' T'|X_t = S_2) = 0.3$$

The following equation shows the joint probability of X being a state and y having a sequence from 1 to t

$$\alpha_t(j) = P(X_t = s_j, y_{i:t})$$

according to forward algorithm we have the following base case(as 'A' is the 1st character in the sequence:

$$\alpha_1(j) = P(y_1|X_1 = s_j)P(X_1 = s_j) = \pi_j P(y_1 =' A'|X_1 = s_j)$$

$$\alpha_1(1) = \pi_1 P(y_1 =' A'|X_1 = S_1) = 0.6 * 0.4 = 0.24$$

$$\alpha_1(2) = \pi_2 P(y_1 =' A'|X_1 = S_2) = 0.4 * 0.2 = 0.08$$

Now we have our base case. Now we use recursion to get other alpha's:

$$\alpha_t(j) = P(y_t|X_t = S_j) \sum_i a_{ij}\alpha_{t-1}(i)$$

$$\alpha_2(1) = P(y_2 =' C'|X_t = S_1) \sum_i a_{i1}\alpha_1(i) = 0.2 * ((0.7 * 0.24) + (0.4 * 0.08)) = 0.04$$

$$\alpha_2(2) = P(y_2 =' C'|X_t = S_2) \sum_i a_{i2}\alpha_1(i) = 0.4 * ((0.3 * 0.24) + (0.6 * 0.08)) = 0.048$$

$$\alpha_3(1) = P(y_3 =' C'|X_t = S_1) \sum_i a_{i1}\alpha_2(i) = 0.2 * ((0.7 * 0.04) + (0.4 * 0.048)) = 0.00944$$

$$\alpha_3(2) = P(y_3 =' C'|X_t = S_2) \sum_i a_{i2}\alpha_2(i) = 0.4 * ((0.3 * 0.04) + (0.6 * 0.048)) = 0.01632$$

$$\alpha_4(1) = P(y_4 =' G'|X_t = S_1) \sum_i a_{i1}\alpha_3(i) = 0.3 * ((0.7 * 0.00944) + (0.4 * 0.01632)) = 0.0039408$$

$$\alpha_4(2) = P(y_4 =' G'|X_t = S_2) \sum_i a_{i2}\alpha_3(i) = 0.1 * ((0.3 * 0.00944) + (0.6 * 0.01632)) = 0.0012624$$

$$\alpha_5(1) = P(y_5 =' T'|X_t = S_1) \sum_i a_{i1}\alpha_4(i) = 0.1 * ((0.7 * 0.0039408) + (0.4 * 0.0012624)) = 0.000326$$

$$\alpha_5(2) = P(y_5 =' G'|X_t = S_2) \sum_i a_{i2}\alpha_4(i) = 0.3 * ((0.3 * 0.0039408) + (0.6 * 0.0012624)) = 0.000582$$

$$\alpha_6(1) = P(y_6 =' A'|X_t = S_1) \sum_i a_{i1}\alpha_5(i) = 0.4 * ((0.7 * 0.000326) + (0.4 * 0.000582)) = 0.000184$$

$$\alpha_6(2) = P(y_6 =' G'|X_t = S_2) \sum_i a_{i2}\alpha_5(i) = 0.2 * ((0.3 * 0.000326) + (0.6 * 0.000582)) = 0.0000894$$

To Find the final answer i.e probability of sequence we use the following formula:

| alpha | S(j)=S1 | S(j)=S2 |
|---|---|---|
| alpha_1(j) | 0.24 | 0.08 |
| alpha_2(j) | 0.04 | 0.048 |
| alpha_3(j) | 0.00944 | 0.01632 |
| alpha_4(j) | 0.0039408 | 0.0012624 |
| alpha_5(j) | 0.000326 | 0.000582 |
| alpha_6(j) | 0.000184 | 0.0000894 |

Figure 2: The alphas

$$P(y_{1:T}) = \sum_j \alpha_T(j)$$

$$P(y_{1:6}) = \sum_j \alpha_6(j) = 0.000184 + 0.0000894 = 0.0002734$$

b)Filtering. Calculate

$$P(X_6 = S_j|O; Q); j = 1, 2$$

We will find the beta valuses using backpropogation algorithm so that we will be able to find the probability in question.

$$\beta_t(j) = P(y_{t+1:T}|X_T = S_j)$$

We will start with the base case of backward propogation algorithm:

$$\beta_T(j) = 1$$

therefore:

$$\beta_6(1) = 1; \beta_6(2) = 1$$

Now we come to the recursive step of backward propagation:

$$\beta_{t-1}(i) = \sum_j \beta_t(j) a_{ij} P(y_t|X_t = S_j)$$

$$\beta_5(1) = \sum_j \beta_6(j) a_{1j} P(y_6 =' A'|X_6 = S_j) = [1 * 0.7 * 0.4] + [1 * 0.3 * 0.3] = 0.34$$

$$\beta_5(2) = \sum_j \beta_6(j) a_{2j} P(y_6 =' A'|X_6 = S_j) = [1 * 0.4 * 0.4] + [1 * 0.6 * 0.2] = 0.28$$

$$\beta_4(1) = \sum_j \beta_5(j) a_{1j} P(y_5 =' T'|X_5 = S_j) = [0.34 * 0.7 * 0.1] + [0.28 * 0.3 * 0.3] = 0.049$$

$$\beta_4(2) = \sum_j \beta_5(j) a_{2j} P(y_5 =' T'|X_5 = S_j) = [0.34 * 0.4 * 0.1] + [0.28 * 0.6 * 0.3] = 0.064$$

$$\beta_3(1) = \sum_j \beta_4(j) a_{1j} P(y_4 =' G'|X_4 = S_j) = [0.049 * 0.7 * 0.3] + [0.064 * 0.3 * 0.1] = 0.01221$$

$$\beta_3(2) = \sum_j \beta_4(j) a_{2j} P(y_4 =' G'|X_4 = S_j) = [0.049 * 0.4 * 0.3] + [0.064 * 0.6 * 0.1] = 0.00972$$

$$\beta_2(1) = \sum_j \beta_3(j) a_{1j} P(y_3 =' C'|X_3 = S_j) = [0.01221 * 0.7 * 0.2] + [0.00972 * 0.3 * 0.4] = 0.002876$$

$$\beta_2(2) = \sum_j \beta_3(j) a_{2j} P(y_3 =' C'|X_3 = S_j) = [0.01221 * 0.4 * 0.2] + [0.00972 * 0.6 * 0.4] = 0.00331$$

$$\beta_1(1) = \sum_j \beta_2(j) a_{1j} P(y_2 =' C'|X_2 = S_j) = [0.002876 * 0.7 * 0.2] + [0.00331 * 0.3 * 0.4] = 0.000799$$

$$\beta_1(2) = \sum_j \beta_2(j) a_{2j} P(y_2 =' C'|X_2 = S_j) = [0.002876 * 0.4 * 0.2] + [0.00331 * 0.6 * 0.4] = 0.001024$$

Now we have all the beta values from back propagation algorithm. we know that :

| beta | S(j)=S1 | S(j)=S2 |
|---|---|---|
| beta_1(j) | 0.000799 | 0.001024 |
| beta_2(j) | 0.002876 | 0.00331 |
| beta_3(j) | 0.01221 | 0.00972 |
| beta_4(j) | 0.049 | 0.064 |
| beta_5(j) | 0.34 | 0.28 |
| beta_6(j) | 1 | 1 |

Figure 3: The betas

$$P(X_6 = S_j|O; \theta); j = 1, 2$$

is given by

$$P(X_6 = S_j|y_{i:T}) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{j'} \alpha_t(j')\beta_t(j')}$$

therefore:

$$P(X_6 = S_1|y_{1:T}) = \frac{\alpha_6(1)\beta_6(1)}{[\alpha_6(1)\beta_6(1)] + [\alpha_6(2)\beta_6(2)]}$$

$$P(X_6 = S_1|y_{1:T}) = \frac{0.000184 * 1}{[0.000184 * 1] + [0.0000894 * 1]} = 0.673007$$

for j=1 we get probability 0.673007

$$P(X_6 = S_2|y_{1:T}) = \frac{\alpha_6(2)\beta_6(2)}{[\alpha_6(1)\beta_6(1)] + [\alpha_6(2)\beta_6(2)]}$$

$$P(X_6 = S_2|y_{1:T}) = \frac{0.0000894 * 1}{[0.000184 * 1] + [0.0000894 * 1]} = 0.32699$$

for j=2 we get probability 0.32699

c) Similarly

$$P(X_4 = S_1|y_{1:T}) = \frac{\alpha_4(1)\beta_4(1)}{[\alpha_4(1)\beta_4(1)] + [\alpha_4(2)\beta_4(2)]}$$

$$P(X_4 = S_1|y_{1:T}) = \frac{0.0039408 * 0.049}{[0.0039408 * 0.049] + [0.064 * 0.0012624]} = 0.705017$$

for j=1 we get probability 0.705017

$$P(X_4 = S_2|y_{1:T}) = \frac{\alpha_4(2)\beta_4(2)}{[\alpha_4(2)\beta_4(2)] + [\alpha_4(1)\beta_4(1)]}$$

$$P(X_4 = S_2|y_{1:T}) = \frac{0.064 * 0.0012624}{[0.0039408 * 0.049] + [0.064 * 0.0012624]} = 0.294983$$

for j=2 we get probability 0.294983

d)Compute

$$X = \overline{X1X2....X6} = argmax_X P(X|O;\theta).$$

We will use viterbi algorithm to find the probability in question and sequence. Define the most likely path ending with j at time t

Base case :

$$\delta_1(j) = \pi_j P(y_1 =' A'|X_1 = S_j)$$

$$\delta_1(1) = \pi_1 P(y_1 =' A'|X_1 = S_1) = 0.6 * 0.4 = 0.24$$

$$\delta_1(2) = \pi_2 P(y_1 =' A'|X_1 = S_2) = 0.4 * 0.2 = 0.08$$

Lets define recursive part of the algorithm now:

$$\delta_t(j) = max_i \delta_{t-1}(i) a_{ij} P(y_t|X_t = S_j)$$

following recursion we get:

$$\delta_2(1) = 0.24 * 0.7 * 0.2 = 0.0336; i = 1$$

$$\delta_2(1) = 0.08 * 0.4 * 0.2 = 0.0.0064; i = 2$$

max out of above 2 is 0.0336 for j=1.Record this observation.

$$\delta_2(2) = 0.24 * 0.3 * 0.4 = 0.0.0288; i = 1$$

$$\delta_2(2) = 0.08 * 0.6 * 0.4 = 0.0.0192; i = 2$$

max out of above 2 is 0.0288 for j=2. Record this observation.

$$\delta_3(1) = 0.0336 * 0.7 * 0.2 = 0.004704; i = 1$$

$$\delta_3(1) = 0.0288 * 0.4 * 0.2 = 0.002304; i = 2$$

max out of above 2 is 0.004704 for j=1.Record this observation.

$$\delta_3(1) = 0.0336 * 0.3 * 0.4 = 0.0.004032; i = 1$$

$$\delta_3(1) = 0.0288 * 0.6 * 0.4 = 0.0.006912; i = 2$$

max out of above 2 is 0.006912 for j=2.Record this observation.

$$\delta_4(1) = 0.004704 * 0.7 * 0.3 = 0.000988; i = 1$$

$$\delta_4(1) = 0.0006912 * 0.4 * 0.3 = 0.000829; i = 2$$

max out of above 2 is 0.000988 for j=1.Record this observation.

$$\delta_4(2) = 0.004704 * 0.3 * 0.1 = 0.000141; i = 1$$

$$\delta_4(2) = 0.0006912 * 0.6 * 0.1 = 0.000415; i = 2$$

max out of above 2 is 0.000415 for j=2.Record this observation. observation.

$$\delta_5(1) = 0.000988 * 0.7 * 0.1 = 0.00006916; i = 1$$

$$\delta_5(1) = 0.000415 * 0.4 * 0.1 = 0.0000166; i = 2$$

max out of above 2 is 0.00006916 for j=1.Record this observation.

$$\delta_5(2) = 0.000988 * 0.3 * 0.3 = 0.00008892; i = 1$$

$$\delta_5(2) = 0.000415 * 0.6 * 0.3 = 0.0000747; i = 2$$

max out of above 2 is 0.00008892 for j=2.Record this observation.

$$\delta_6(1) = 0.00006916 * 0.7 * 0.4 = 0.00001736; i = 1$$

$$\delta_6(1) = 0.00008892 * 0.4 * 0.4 = 0.00001423; i = 2$$

max out of above 2 is 0.00001736 for j=1.Record this observation.

$$\delta_6(1) = 0.00006916 * 0.3 * 0.2 = 0.00000415; i = 1$$

$$\delta_6(1) = 0.00008892 * 0.6 * 0.2 = 0.0000107; i = 2$$

max out of above 2 is 0.0000107 for j=2.Record this observation.

$$argmax_j \delta_6(j) = 0.00001736$$

sequence =

$$S_1, S_1, S_2, S_1, S_2, S_1$$

| V_t(j) | S(j)=S1 | S(j)=S2 | Sequence |
|---|---|---|---|
| V_1(j) | 0.24 | 0.08 | S1 |
| V_2(j) | 0.0336 | 0.0288 | S1 |
| V_3(j) | 0.004704 | 0.006912 | S2 |
| V_4(j) | 0.000988 | 0.000415 | S1 |
| V_5(j) | 0.00006916 | 0.00008892 | S2 |
| V_6(j) | 0.00001736 | 0.0000107 | S1 |

Figure 4: The deltas and sequence

e) We will find probability of every alphabet being 7th in the sequence:

| | S(j)=S1 | S(j)=S2 | S1+S2 |
|---|---|---|---|
| A | 0.000066 | 0.000022 | 0.000088 |
| C | 0.000033 | 0.000044 | 0.000077 |
| G | 0.000049 | 0.000011 | 0.00006 |
| T | 0.000016 | 0.000033 | 0.000049 |

Figure 5: Best option for s7

A:

$$\alpha_7(1) = P(y_7 =' A'|X_7 = S_1)\sum_i a_{i1}\alpha_{t-1}(i) = 0.4((0.7 * 0.000184) + (0.4 * 0.0000894)) = 0.000066$$

$$\alpha_7(2) = P(y_7 =' A'|X_7 = S_2)\sum_i a_{i2}\alpha_{t-1}(i) = 0.2((0.3 * 0.000184) + (0.6 * 0.0000894)) = 0.000022$$

sum of probabilities the j=1 and j=2 is 0.000088

C:

$$\alpha_7(1) = P(y_7 =' C'|X_7 = S_1)\sum_i a_{i1}\alpha_{t-1}(i) = 0.2((0.7 * 0.000184) + (0.4 * 0.0000894)) = 0.000033$$

$$\alpha_7(2) = P(y_7 =' C'|X_7 = S_2)\sum_i a_{i2}\alpha_{t-1}(i) = 0.4((0.3 * 0.000184) + (0.6 * 0.0000894)) = 0.000044$$

sum of probabilities the j=1 and j=2 is 0.000077

G:

$$\alpha_7(1) = P(y_7 =' G'|X_7 = S_1)\sum_i a_{i1}\alpha_{t-1}(i) = 0.3((0.7 * 0.000184) + (0.4 * 0.0000894)) = 0.000049$$

$$\alpha_7(2) = P(y_7 =' C'|X_7 = S_2)\sum_i a_{i2}\alpha_{t-1}(i) = 0.1((0.3 * 0.000184) + (0.6 * 0.0000894)) = 0.000011$$

sum of probabilities the j=1 and j=2 is 0.000060

T:

$$\alpha_7(1) = P(y_7 =' T'|X_7 = S_1)\sum_i a_{i1}\alpha_{t-1}(i) = 0.1((0.7 * 0.000184) + (0.4 * 0.0000894)) = 0.000016$$

$$\alpha_7(2) = P(y_7 =' T'|X_7 = S_2)\sum_i a_{i2}\alpha_{t-1}(i) = 0.3((0.3 * 0.000184) + (0.6 * 0.0000894)) = 0.000033$$

sum of probabilities the j=1 and j=2 is 0.000049

Highest probability is the probability of 'A'