DD2434 Machine Learning, Advanced Course
Assignment 2

Salma El Alaoui Talibi December 23, 2015

1 Graphical Models

1.1 Qualitative effects in a Directed Graphical Model (DGM)

Question 1: *In which pairs is one value larger than the other? Explain your choices.*

- pair 1: $p(t^1|d^1) < p(t^1)$

$$\begin{aligned} p(t^1|d^1) &= p(t^1|d^1,c^0)p(c^0|d^1) + p(t^1|d^1,c^1)p(c^1|d^1)\{T \perp D|C\} \\ &= p(t^1|c^0)p(c^0|d^1) + p(t^1|c^1)p(c^1|d^1) \\ p(t^1|d^0) &= p(t^1|c^0)p(c^0|d^0) + p(t^1|c^1)p(c^1|d^0) \end{aligned}$$

We have two linear combinations of the same variables $p(t^1|c^0)$ and $p(t^1|c^1)$ with weights that sum to one:

$$\begin{aligned} p(c^0|d^1) + p(c^1|d^1) = 1 \\ p(c^0|d^0) + p(c^1|d^0) = 1 \end{aligned}$$

From the influences between the variables in the graphical model we have:

$$\begin{aligned} p(t^1|c^1) > p(t^1|c^0) \\ p(c^1|d^0) > p(c^1|d^1) \end{aligned}$$

Therefore:

$$p(t^1|d^0) > p(t^1|d^1) \tag{1}$$

By conditioning T by D:

$$p(t^1) = p(t^1|d^0)p(d^0) + p(t^1|d^1)p(d^1)$$

which is a linear combination of $p(t^1|d^0)$ and $p(t^1|d^1)$ with weights that sum to one:

$$p(d^0) + p(d^1) = 1$$

And therefore using 1 we obtain:

$$p(t^1|d^1) < p(t^1)$$

- pair 2: $p(d^1|t^0) > p(d^1)$

$$\begin{aligned} p(d^1|t^1) &= p(d^1|t^1,c^1)p(c^1|t^1) + p(d^1|t^1,c^0)p(c^0|t^1)\{T \perp D|C\} \\ &= p(d^1|c^1)p(c^1|t^1) + p(d^1|c^0)p(c^0|t^1) \\ p(d^1|t^0) &= p(d^1|t^0,c^1)p(c^1|t^0) + p(d^1|t^0,c^0)p(c^0|t^0)\{T \perp D|C\} \\ &= p(d^1|c^1)p(c^1|t^0) + p(d^1|c^0)p(c^0|t^0) \end{aligned}$$

We have two linear combinations of the same variables $p(d^1|c^1)$ and $p(d^1|c^0)$ with weights that sum to one:

$$p(c^1|t^1) + p(c^0|t^1) = 1$$
$$p(c^1|t^0) + p(c^0|t^0) = 1$$

From the influences between the variables in the graphical model we have:

$$p(d^1|c^1) < p(d^1|c^0)$$
$$p(c^1|t^1) > p(c^1|t^0) \Rightarrow p(c^0|t^1) < p(c^0|t^0)$$

Therefore:

$$p(d^1|t^0) > p(d^1|t^1) \tag{2}$$

By conditioning D by T:

$$p(d^1) = p(d^1|t^0)p(t^0) + p(d^1|t^1)p(t^1)$$

which is a linear combination of $p(d^1|t^0)$ and $p(t^1|d^1)$ with weights that sum to one:

$$p(t^0) + p(t^1) = 1$$

And therefore using 2 we obtain:

$$p(d^1|t^0) > p(d^1)$$

- pair 3: $p(h^1|e^1, f^1) > p(h^1|e^1)$

$$p(h^1|e^1) = p(h^1|e^1, f^1)p(f^1|e^1) + p(h^1|e^1, f^0)p(f^0|e^1)$$

We have a linear combination of $p(h^1|e^1, f^1)$ and $p(h^1|e^1, f^0)$ with weights that sum to one:

$$p(f^1|e^1) + p(f^0|e^1) = 1$$

From the graphical model, we can see that when one exercises(e^1) despite having little free time(f^1), then he is more likely to be health conscious(h^1) than when he exercises when having free time(f^0). Therefore:

$$p(h^1|e^1, f^1) > p(h^1|e^1, f^0)$$

which results in:

$$p(h^1|e^1, f^1) > p(h^1|e^1)$$

- pair 5: $p(c^1|h^0) > p(c^1)$
  There are 2 trails between C and H:
  - $C \leftarrow D \leftarrow H$
  - $C \leftarrow D \rightarrow W \leftarrow E \leftarrow H$

The second trail is blocked in W because in the v-structure $D \rightarrow W \leftarrow E$, W is not observed. Therefore we only consider the first trail.

$$\begin{aligned} p(c^1|h^0) &= p(c^1|h^0, d^0)p(d^0|h^0) + p(c^1|h^0, d^1)p(d^1|h^0)\{C \perp H|D\} \\ &= p(c^1|d^0)p(d^0|h^0) + p(c^1|d^1)p(d^1|h^0) \\ p(c^1|h^1) &= p(c^1|h^1, d^0)p(d^0|h^1) + p(c^1|h^1, d^1)p(d^1|h^1)\{C \perp H|D\} \\ &= p(c^1|d^0)p(d^=0|h^1) + p(c^1|d^1)p(d^1|h^1) \end{aligned}$$

We have two linear combinations of the same variables $p(c^1|d^0)$ and $p(c^1|d^1)$ that sum to one:

$$\begin{aligned} p(d^0|h^0) + p(d^1|h^0) = 1 \\ p(d^0|h^1) + p(d^1|h^1) = 1 \end{aligned}$$

From the influences between the variables in the graphical model we have:

$$p(d^1|h^1) > p(d^1|h^0) \Rightarrow p(d^0|h^1) < p(d^0|h^0)p(c^1|d^0) \qquad > p(c^1|d^1)$$

Therefore:

$$p(c^1|h^0) > p(c^1|h^1) \tag{3}$$

By conditioning C by H we obtain:

$$p(c^1) = p(c^1|h^0)p(h^0) + p(c^1|h^1)p(h^1)$$

which is a linear combination of $p(c^1|h^0)$ and $p(c^1|h^1)$ with weights that sum to one:

$$p(h^0) + p(h^1) = 1$$

And therefore using 3 we obtain:

$$p(c^1|h^0) > p(c^1)$$

Question 2: *Which pairs are equal? Explain your choices.*

- pair 4: $p(c^1|f^0) = p(c^1)$ There are 2 trails between C and F:
  - $C \leftarrow D \leftarrow H \rightarrow E \leftarrow F$
    This trail is not active because in the v-structure $H \rightarrow E \leftarrow F$, neither E nor its descendant W is in F, and therefore E and W are not observed.
  - $C \leftarrow D \rightarrow W \leftarrow E \leftarrow F$
    This trail is not active because in the v-structure $D \rightarrow W \leftarrow E$, W is not in F, and therefore it's not observed.

  Therefore all the trails between C and F are not active given F, which means:

$$p(c^1|f^0) = p(c^1)$$

- pair 6: $p(c^1|h^0, f^0 = p(c^1|h^0)$ There are 2 trails between C and F:

- $F \rightarrow E \leftarrow H \rightarrow D \rightarrow C$

  This trail is not active given H because in the v-structure $F \rightarrow E \leftarrow H$, neither E nor any of its descendants is in H, and therefore it's not observed.

- $F \rightarrow E \rightarrow W \leftarrow D \rightarrow C$

  This trail is not active given H because in the v-structure $E \rightarrow W \leftarrow D$, W is not in H, and therefore it's not observed.

Therefore all the trails between F and C are not active given H, so F and C are d-separated given H, which means that : $C \perp F|H$, and therefore:

$$p(c^1|h^0, f^0 = p(c^1|h^0)$$

- pair 7: $p(d^1|h^1, e^0) = p(c^1|h^0)$ There are 2 trails between D and E:

  - $D \leftarrow H \rightarrow E$

    This fork is not active given H because H is observed.

  - $E \rightarrow W \leftarrow D$

    This v-structure is not active given H because W is not in H, and therefore it's not observed.

  Therefore all the trails between D and E are not active given H, so D and E are d-separated given H, which means that: $D \perp E|H$, and therefore:

$$p(d^1|h^1, e^0) = p(c^1|h^0)$$

Question 3: *Which pairs are incomparable (i.e., the two values can not be compared based on the information available in the DAG.) Explain your choices.*

- pair 8: $p(d^1|e^1, f^0, w^1)$ and $p(d^1|e^1, f^0)$ are incomparable.

$$\begin{aligned} p(d^1|e^1, f^0) &= p(d^1|e^1, f^0, w^0)p(w^0|e^1, f^0) + p(d^1|e^1, f^0, w^1)p(w^1|e^, f^)\{W \perp F|E\} \\ &= p(d^1|e^1, f^0, w^0)p(w^0|e^1) + p(d^1|e^1, f^0, w^1)p(w^1|e^1) \end{aligned}$$

  We have a linear combination of $p(d^1|e^1, f^0, w^0)$ and $p(d^1|e^1, f^0, w^1)$ with weights that sum to one:

$$p(w^0|e^1) + p(w^1|e^1) = 1$$

  Therefore, in order to compare $p(d^1|e^1, f^0)$ and $p(d^1|e^1, f^0, w^1)$, we need to compare $p(d^1|e^1, f^0, w^0)$ and $p(d^1|e^1, f^0, w^1)$, which we can't do because we don't know anything about H.

- pair 9: $p(t^1|w^1, f^0)$ and $p(t^1|w^1)$ are incomparable.
  Knowing W activates the v-structure $H \rightarrow E \leftarrow F$, because W is a child of E, which means that H and F become dependent. Therefore we can't compare $p(t^1|w^1, f^0)$ and $p(t^1|w^1)$ without knowing anything about H.

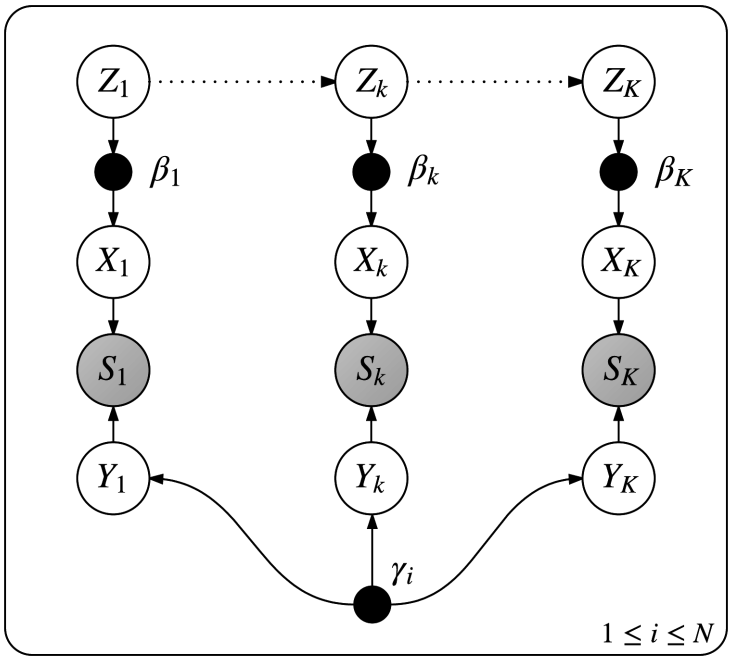


Figure 1: The casino model

1.2 The casino model

Question 4: *Provide a drawing of the graphical model θ.*

Figure 1 represents the graphical model θ.

- Z_k is a binary variable that indicates if table is primed or not.

$$Z_k = \begin{cases} 0 & \text{if the table is not primed} \\ 1 & \text{if the table is primed} \end{cases}$$

  It is governed by the following transition matrix in table 1:

Table 1: Transition matrix A

	$Z_{k+1} = 0$	$Z_{k+1} = 1$
$Z_k = 0$	$\frac{1}{4}$	$\frac{3}{4}$
$Z_k = 1$	$\frac{3}{4}$	$\frac{1}{4}$

- Y_k is the outcome of the player i's dice in the table k: $Y_k \sim Cat(\gamma_i)$
- X_k is the outcome of the table's dice in the table k: $X_k \sim Cat(\beta_k)$
- S_k is the sum of the two outcomes in table k

Question 5: *Provide an implementation of the model θ.*

```python
import numpy as np
import matplotlib.pyplot as plt

#generates a sequence of length K of the variable Z (a sequence of hidden states)
# Z = 1 if the table is primed and 0 if the table is not primed
def generateTables(K, initProb):
    sequence =[]
    #sample first table
    s = np.random.binomial(1, initProb, 1)
    #sample k-1 tables
    sequence.append(s[0])
    for k in range(K-1):
        if sequence[len(sequence)-1] == 0:
            s = np.random.binomial(1,3/4,1)
        else:
            s = np.random.binomial(1,1/4,1)
        sequence.append(s[0])
    return sequence

#Samples from the outcome of the table's dice
#primed and unPrimed are the categorical distributions for K primed and K unprimed Tables
def sampleTableDice(primed, unPrimed, tables):
    sequence = []
    for k in range(len(tables)):
        if tables[k] == 0:
            s = np.random.multinomial(1, unPrimed[k, :])
        else:
            s = np.random.multinomial(1, primed[k, :])
        outcome = np.argmax(s)+ 1
        sequence.append(outcome)
    return(sequence)

#Samples from the outcome of the player's dice
def samplePlayerDice(playerDice, tables):
    sequence = []
    for k in range(len(tables)):
        s =  np.random.multinomial(1, playerDice)
        outcome = np.argmax(s)+ 1
        sequence.append(outcome)
    return(sequence)

#Observations: sum of the two dices
def demo(nTables, primed, unPrimed, playerDice):
    tables = generateTables(nTables, 0.5)
    tableOutcome = np.asarray(sampleTableDice(primed, unPrimed, tables))
    playerOutcome = np.asarray(samplePlayerDice(playerDice, tables))
    sum = tableOutcome + playerOutcome
    return sum

nTables = 10
```

```
#initial state distribution
pi = [0.5, 0.5]
tables = generateTables(nTables, pi[1])
#unbiased dice
primed = np.ones((nTables,6))*1/6
unPrimed = np.ones((nTables,6))*1/6
playerDice = np.ones(6)*1/6
#observations
observations = demo(nTables, primed, unPrimed, playerDice)
```

Listing 1: Casino model implementation

Question 6: *Provide data generated using at least three different sets of categorical dice distributions – what does it look like for all perfect dice with uniform distributions, for example, or if all of them are perfect instead of one, or if all are bad in the same way?*

We plot the the histograms of 10000 sequences of observations of the casino model with one player, 10 tables and an equiprobable initial state distribution for the primed and non-primed tables, ie $\pi = [0.5, 0.5]^T$. Figure 2 shows the data generated under different sets of categorical dice distributions:

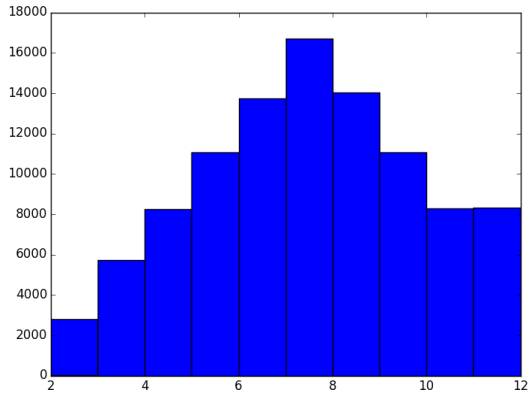
- in figure 2a, all the dice are fair.
- in figure 2b, all the tables have fair dice while the player's dice is unfair : $Y_k \sim Cat(\gamma_i = [0.5, 0.5, 0, 0, 0, 0]^T)$. We can see that $s_k \leq 8$.
- in figure 2c, all the non-primed tables have biased dice: $X_k \sim Cat(\beta_k = [0.5, 0.5, 0, 0, 0, 0]^T)$ if $Z_k = 0$. We can see that the bins where $s_k \geq 8$ have less samples than in figure 2a, where all the dice were unbiased.
- in figure 2d, all the dice have the same biased categorical distribution: $Cat(\beta_k = [0, 0, 0, 0, 0.5, 0.5]^T)$. We can see that $s_k \geq 10$.

1.3 Sampling tables given dice sum

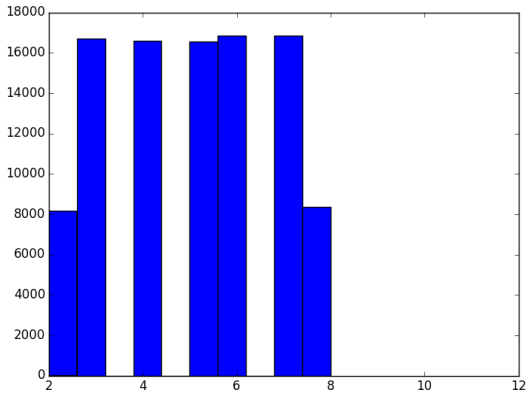
Question 7: *Describe an algorithm that, given (1) the parameters θ of the full casino model of Task 2.2 (so, θ is all the categorical distributions corresponding to all the dice), (2) a sequence of tables $z_1 ..., z_K$, and (3) an observation of dice sums $s_1, ..., s_K$, outputs $p(z_1, ..., z_K | s_1, ..., s_K, \theta)$.*

Using Bayes rule, we can write:

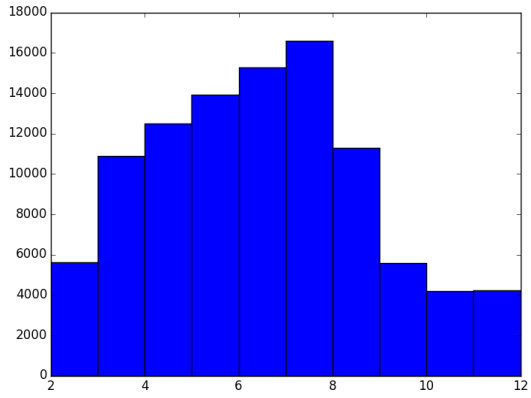
$$p(Z_{1:K}|s_{1:K}, \theta) = \frac{p(s_{1:K}|Z_{1:K}, \theta)p(Z_{1:K}|\theta)}{p(s_{1:K}|\theta)}$$



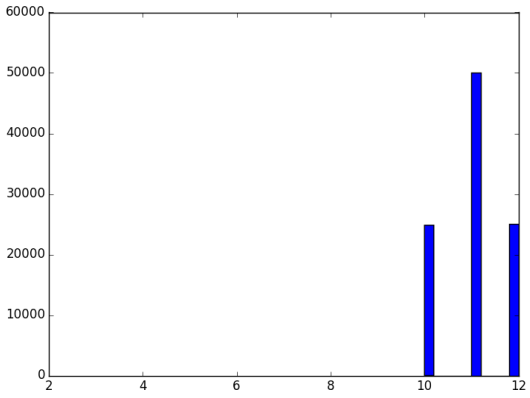
(a) unbiased distributions for all dice



(b) Biased dice for the player



(c) Biased dice for all non-primed tables



(d) biased distributions for all dice

Figure 2: Observations of the Casino model under different dice settings

We look at the three terms separately:

Using the product rule and $S_i \perp S_{i+1}|Z_i$, we can write the first term in the numerator as:

$$\begin{aligned} p(s_{1:K}|Z_{1:K},\theta) &= \prod_{k=1}^{K} p(s_k|Z_k,\theta) \\ &= \prod_{k=1}^{K}\sum_{x,y}^{6} p(X_k=x|Z_k,\theta)p(Y_k=y|\theta)I(x+y=s_k) \\ &= \prod_{k=1}^{K}\sum_{x,y}^{6} I(x+y=s_k)Cat(x|\beta_k)Cat(y|\gamma) \\ &= \prod_{k=1}^{K}\sum_{x,y}^{6} I(x+y=s_k)\beta_{kx}\gamma_y \end{aligned} \tag{4}$$

Using the first order Markov property of the states, we can factorize the second term in the numerator as:

$$\begin{aligned} p(Z_{1:K}|\theta) &= p(Z_{1:K} = z_{1:K}|\theta) \\ &= \prod_{k=1}^{K-1} p(Z_{k+1} = z_{k+1}|Z_k = z_k)p(Z_1 = z_1) \\ &= A_{z_k z_{k+1}} \pi_{z_1} \end{aligned} \tag{5}$$

where A is the transition matrix and π is the initial state distribution vector.
Using the sum rule, we can write the denominator as:

$$p(s_{1:K}|\theta) = \sum_l p(s_{1:K}, Z_K = l|\theta) = \sum_l \alpha_K(l)$$

$$\begin{aligned} \alpha_K(l) = p(s_{1:K}, Z_K = l|\theta) &= p(s_{1:K-1}, s_K, Z_K = l|\theta) \\ &= p(s_K|Z_K = l, s_{1:K-1}, \theta)p(Z_k = l, s_{1:K-1}|\theta) \\ &= p(s_K|Z_K = l, \theta) \sum_i p(Z_K = l, Z_{K-1} = i, s_{1:K-1}|\theta) \\ &= p(s_K|Z_K = l, \theta) \sum_i p(Z_K = l|Z_{K-1} = i, s_{1:K-1}, \theta)p(Z_{K-1} = i, s_{1:K-1}|\theta) \\ &= p(s_K|Z_K = l, \theta) \sum_i p(Z_K = l|Z_{K-1} = i)\alpha_{K-1}(i) \end{aligned} \tag{6}$$

Therefore, we derive the following DP-algorithm to compute $\alpha_K(l)$:

Initialize:

$\alpha_1(l) = \pi_l \; p(s_1|Z_1 = l)$

for k = 2 to K **do**

$\alpha_k(l) = p(s_k|Z_k = l, \theta) \sum_i p(Z_k = l|Z_{k-1} = i)\alpha_{k-1}(i)$

end for

Where the emission probability is

$$p(s_k|Z_k = l, \theta) = \sum_{x,y}^{6} I(x + y = s_k)\beta_{kx}\gamma_y$$

Question 8: *You should also show how to sample $z_1, ..., z_k$ from $p(Z_1, ..., Z_K|s_1, ..., s_K, \theta)$ as well as implement and show test runs of this algorithm.*

$$p(Z_1 ... Z_K|s_1, ..., s_K, \theta) \tag{7}$$

$$= p(Z_{1:K} = z_{1:K}|s_{1:K}, \theta) \tag{8}$$

$$= p(Z_{1:K-1} = z_{1:K-1}|Z_K = z_K, s_{1:K}, \theta)p(Z_K = z_K|s_{1:K}, \theta) \tag{9}$$

Using Bayes' rule, the second term of equation 9 can be written as:

$$p(Z_K = z_K|s_{1:K}, \theta) = \frac{p(Z_K = z_K, s_{1:K}|\theta)}{p(s_{1:K}|\theta)}$$

We recognize the variable $\alpha_K(z_k)$ derived in equation 6, and therefore:

$$p(Z_k = z_k|s_{1:K}, \theta) = \frac{\alpha_K(z_K)}{\sum_t \alpha_K(t)}$$

We can sample from this distribution trivially by computing the probability for $p(Z_k = 0|s_{1:K}, \theta)$ and $p(Z_k = 1|s_{1:K}, \theta)$ and sampling from the resulting Bernoulli distribution.
By using the conditional independence of the Markov chain, the first term of equation 9 does not depend on s_K, and therefore:

$$p(Z_{1:K-1} = z_{1:K-1}|Z_K = z_K, s_{1:K}, \theta) \tag{10}$$
$$=p(Z_{1:K-1} = z_{1:K-1}|Z_K = z_K, s_{1:K-1}, \theta) \tag{11}$$
$$=p(Z_{1:K-2} = z_{1:K-2}|Z_{K-1} = z_{K-1}, Z_K = z_K, s_{1:K-1}, \theta)p(Z_{K-1} = z_{K-1}|Z_K = z_K, s_{1:K-1}, \theta) \tag{12}$$
$$=p(Z_{1:K-2} = z_{1:K-2}|Z_{K-1} = z_{K-1}, s_{1:K-1}, \theta)p(Z_{K-1} = z_{K-1}|Z_K = z_K, s_{1:K-1}, \theta) \tag{13}$$

The second term of equation 13 can be writen as:

$$p(Z_{K-1} = z_{K-1}|Z_K = z_K, s_{1:K-1}, \theta) \tag{14}$$
$$=\frac{p(Z_K = z_k|Z_{K-1} = z_{K-1}, s_{1:K-1}, \theta)p(Z_{K-1} = z_{K-1}, s_{1:K-1}, \theta)}{p(Z_K = z_k|s_{1:K-1}, \theta)} \tag{15}$$
$$=\frac{A_{z_{k-1}z_k}\alpha_{K-1}(z_{K-1})}{\sum_t A_{tz_k}\alpha_{K-1}(t)} \tag{16}$$

while the first term of equation 13 is of the same form as the first term of equation 9. So we can sample trivially from 16, and recurse backwards over timesteps $T-2, T-3, ..., 1$. We have then produced a sample from the posterior distribution in equation 7.

We use a casino model with one player, 10 tables and an equiprobable initial state distribution for the primed and non-primed tables, ie $\pi = [0.5, 0.5]^T$. We use fair dice for the player and the primed tables, while the non-primed tables all have biased dice: $X_k \sim Cat(\beta_k = [0.5, 0.5, 0, 0, 0, 0]^T)$ if $Z_k = 0$.
Table 2 shows some samples from the posterior distribution in equation 7, and the observations used to produce them. We observe that for all outcomes $s_k \geq 9$, the corresponding table in the sample is primed,

Table 2: Samples from the posterior distribution of the casino model

Observations	**Posterior**
$[4, 7, 6, 12, 7, 4, 9, 7, 7, 3]$	$[0, 0, 1, 1, 0, 0, 1, 1, 0, 0]$
$[4, 5, 9, 4, 7, 7, 6, 5, 11, 7]$	$[1, 0, 1, 1, 1, 0, 0, 0, 1, 1]$
$[8, 8, 6, 7, 9, 4, 4, 11, 6, 9]$	$[1, 0, 1, 0, 1, 1, 0, 1, 1, 1]$
$[10, 2, 10, 5, 10, 7, 7, 7, 3, 7]$	$[1, 0, 1, 0, 1, 0, 1, 0, 1, 0]$

ie $z_k = 1$. It means that the posterior distribution respects the constraint that we have defined.

1.4 Expectation-Maximization (EM)

Question 9: *Present the algorithm written down in a formal manner (using both text and mathematical notation, but not pseudo code).*

The complete data is:

$$D = \{(x_{1:K}^{(1)}, z_{1:K}^{(1)}, s_{1:K}^{(1)}), ..., (x_{1:K}^{(N)}, z_{1:K}^{(N)}, s_{1:K}^{(N)})\}$$

The likelihood of the data can be written as:

$$\begin{aligned}
p(D|\theta) &= \prod_{n=1}^{N} p(x_{1:K}^{(n)}, z_{1:K}^{(n)}, s_{1:K}^{(n)}|\theta) \\
&= \prod_{n=1}^{N} p(s_{1:K}^{(n)}|x_{1:K}^{(n)}, z_{1:K}^{(n)}, \theta) p(x_{1:K}^{(n)}, z_{1:K}^{(n)}|\theta) \\
&= \prod_{n=1}^{N} p(s_{1:K}^{(n)}|x_{1:K}^{(n)}, z_{1:K}^{(n)}, \theta) p(x_{1:K}^{(n)}|\theta) p(z_{1:K}^{(n)}|\theta) \\
&= \prod_{n=1}^{N} \prod_{k=1}^{K} p(s_{k}^{(n)}|x_{k}^{(n)}, z_{k}^{(n)}, \theta) p(x_{k}^{(n)}|\theta) p(z_{k}^{(n)}|\theta) \\
&= \prod_{n=1}^{N} \prod_{k=1}^{K} I(s_{k}^{(n)} = x_{k}^{(n)} + z_{k}^{(n)}) p(x_{k}^{(n)}|\theta) p(z_{k}^{(n)}|\theta) \\
&= \prod_{n=1}^{N} \prod_{k=1}^{K} p(x_{k}^{(n)}|\theta) p(z_{k}^{(n)}|\theta) \\
&= \prod_{n=1}^{N} \prod_{k=1}^{K} \prod_{m=1}^{6} \prod_{l=1}^{6} \big(p(X_k = l|\theta) p(Z^n = m|\theta)\big)^{I(x_k^{(n)}=l, z_k^{(n)}=m)}
\end{aligned}$$

Let $p(X_k = l|\theta) = \alpha_{k,l}$ and $p(Z^n = m|\theta) = \phi_{n,m}$.

We can write the complete log-likelihood as:

$$l(s^{1:N}; x^{1:N}; z^{1:N}; \theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{m=1}^{6} \sum_{l=1}^{6} I(x_k^{(n)} = l, z_k^{(n)} = m) \Big(log(\alpha_{k,l}) + log(\phi_{n,m}) \Big)$$

And therefore the expected complete log-likelihood is:

$$\begin{aligned}
Q(\theta, \theta') &= \sum_{n=1}^{N} E_{p(x_{1:K}^{(n)}, z_{1:K}^{(n)}|s_{1:K}^{(n)}, \theta)} l(s_{1:K}^{(n)}; x_{1:K}^{(n)}, z_{1:K}^{(n)}; \theta') \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{m=1}^{6} \sum_{l=1}^{6} p(x_k^{(n)} = l, z_k^{(n)} = m|s_k^{(n)}, \theta) \Big(log(\alpha'_{k,l}) + log(\phi'_{n,m}) \Big)
\end{aligned}$$

Let $p(x_k^{(n)} = l, z_k^{(n)} = m|s_k^{(n)}, \theta) = N_{k,n,m,l}$. Therefore:

$$\begin{aligned}
Q(\theta, \theta') &= \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{m=1}^{6} \sum_{l=1}^{6} N_{k,n,m,l} \Big(log(\alpha'_{k,l}) + log(\phi'_{n,m}) \Big) \\
&= \sum_{k=1}^{K} \sum_{l=1}^{6} \Big[\sum_{n=1}^{N} \sum_{m=1}^{6} N_{k,n,m,l} \Big] log(\alpha'_{k,l}) + \sum_{n=1}^{N} \sum_{m=1}^{6} \Big[\sum_{k=1}^{K} \sum_{l=1}^{6} N_{k,n,m,l} \Big] log(\phi'_{n,m})
\end{aligned}$$

The two terms can be maximised separately by :

$$\widehat{\alpha'_{k,l}} = \frac{\sum_{n=1}^{N} \sum_{m=1}^{6} N_{k,n,m,l}}{\sum_{l=1}^{6} \sum_{n=1}^{N} \sum_{m=1}^{6} N_{k,n,m,l}}$$

$$\widehat{\phi'_{n,m}} = \frac{\sum_{k=1}^{K} \sum_{l=1}^{6} N_{k,n,m,l}}{\sum_{m=1}^{6} \sum_{k=1}^{K} \sum_{l=1}^{6} N_{k,n,m,l}}$$

Finally, we derive $N_{m,l}$:

$$\begin{aligned} N_{k,n,m,l} &= p(x_k^{(n)} = l, z_k^{(n)} = m | s_k^{(n)}, \theta) \\ &= \frac{p(x_k^{(n)} = l, z_k^{(n)} = m | s_k^{(n)}, \theta)}{p(s_k^{(n)} | \theta)} \\ &= \frac{\alpha_{k,l}\phi_{n,m} I(l + m = s_k^{(n)})}{\sum_{i=1}^{6}\sum_{j=1}^{6} \alpha_{k,i}\phi_{n,i} I(i + i = s_k^{(n)})} \end{aligned}$$

2 Non-Gaussian Latent Representations

2.1 Independent Component Analysis (ICA)

Question 12: *First, whiten the data as described in Section 5 of Hyvarinen and Oja. Show plots that illustrate both the two obtained eigenvectors and their eigenvalues, and a plot of the whitened pointset* $\{\widetilde{x}_i\}$*, and describe (using both text and mathematical notation, but not pseudo code) how you obtained it.*

The data set is X, an $2 \times n$ matrix where n is the number of observations. The rows contains the 2-D representation of the linear combination of the 2 independent signals.
We first center X, by subtracting off the mean for each row X_i, so that the row-wise zero empirical mean becomes 0.
The second step is to whiten X, which means to apply a linear transformation to obtain a new vector $\widetilde{X}$, such that its components are uncorrelated and their variances equal 1. In other words, the covariance matrix of the new vector $\widetilde{X}$ equals the identity matrix:

$$E(\widetilde{X}\widetilde{X}^T) = I$$

Since the covariance matrix $E(XX^T)$ is symmetric and real, it is diagonalizable and can be factorized as:

$$E(XX^T) = EDE^T$$

Where E is the orthogonal matrix of eigenvectors and D is the diagonal matrix of eigenvalues. Since forming XX^T to perform its eigen-value decomposition can cause loss of precision, we use singular value decomposition instead.
Lets define a new matrix Y as an $n \times 2$ matrix:

$$Y = \frac{1}{\sqrt{n-1}} X^T$$

Let's analyze Y^TY :

$$\begin{aligned} Y^TY &= \left(\frac{1}{\sqrt{n-1}} X^T\right)^T \frac{1}{\sqrt{n-1}} X^T \\ &= \frac{1}{n-1} XX^T \\ &= E(XX^T) \end{aligned}$$

By construction, Y^TY equals the covariance matrix of X. Applying singular value decomposition to Y gives:

$$Y = U\Sigma V^T$$

where U and V are orthogonal matrices and Σ is a diagonal matrix. Let's rewrite Y^TY:

$$\begin{aligned} Y^TY &= (U\Sigma V^T)^T U\Sigma V^T \\ &= V\Sigma U^T U\Sigma V^T \end{aligned}$$

Since U is orthogonal, $U^TU = I$ and we can therefore write:

$$E(XX^T) = Y^TY = V\Sigma^2 V^T = EDE^T$$

We can therefore see the correspondence with the eigen-value decomposition of $E(XX^T)$:

$$\begin{aligned} E &= V \\ D &= \Sigma^2 \end{aligned}$$

The eigenvectors (columns of E) are represented in figure 3 in centered data space, and are scaled by the square root of the corresponding eigenvalue (diagonal of D).
We can now write the white data as:

$$\begin{aligned} \widetilde{X} &= ED^{-\frac{1}{2}}E^TX \\ \widetilde{X}\widetilde{X}^T &= ED^{-\frac{1}{2}}E^TXX^TED^{-\frac{1}{2}}E^T \end{aligned}$$

By using $XX^T = (n-1)E(XX^T) = (n-1)EDE^T$, we get:

$$\widetilde{X}\widetilde{X}^T = (n-1)ED^{-\frac{1}{2}}E^TEDE^TED^{-\frac{1}{2}}E^T$$

Since E is orthogonal, $E^TE = EE^T = I$, which results in:

$$\widetilde{X}\widetilde{X}^T = (n-1)I$$

And therefore, the covariance of the new vector $\widetilde{X}$ is indeed equal to identity:

$$E(\widetilde{X}\widetilde{X}^T) = \frac{n-1}{n-1}\, I = I$$

The whitened data $\widetilde{X}$ is represented in figure 4.

Question 13: *Then, describe (using both text and mathematical notation, but not pseudo code) why a PPCA transform can not recover the independent components in the data.*

In PPCA, the latent variable z are assumed to be Gaussian, and its prior distribution is therefore:

$$p(z) = \mathcal{N}(Z|0, I)$$

The observed variable x is defined by a linear transformation of the latent variable z plus additive Gaussian noise, ie:

$$x = Wz + \mu + \epsilon \qquad \text{where} \qquad \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

And therefore the marginal density of x is given by:

$$p(x) = \int p(x|z)p(z)dz = \mathcal{N}(x|\mu, C)$$

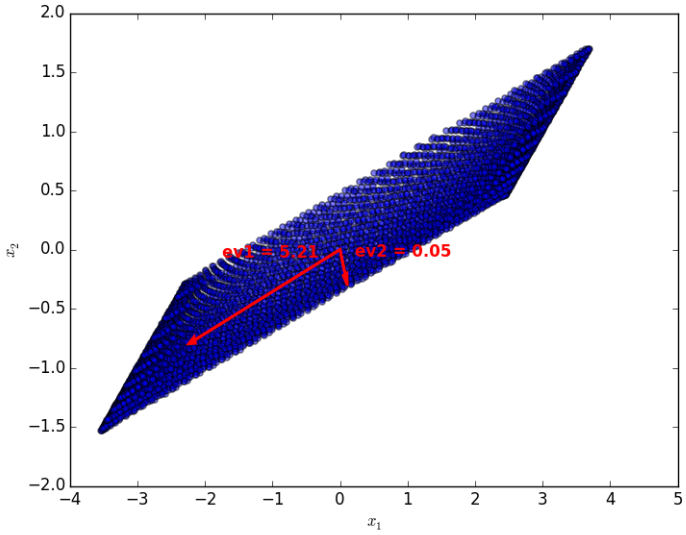


Figure 3: The eigenvectors of the covariance matrix $E(XX^T)$

Where $C = WW^T + \sigma^2$.
We can see that if we make the transformation $W \rightarrow WR$, where R is an orthogonal matrix, the marginal density of x, and therefore the likelihood function are unchanged because the covariance matrix C is invariant. This means that PPCA cannot distinguish between two choices for the latent variable if they differ by a rotation in the latent space. If we were to apply PPCA here, we would only be able to estimate the mixing matrix A up to a rotation, which means that it wouldn't be identifiable. We can recover the best linear subspace in which the signals lie, but cannot uniquely recover the signals themselves.

Question 14: *Finally, recover the two independent components using FastICA as described in Section 6 of Hyvarinen and Oja. Show plots that illustrate both the two obtained mixing vectors, and a plot of the decorrelated pointset $\{s_i\}$, and describe (using both text and mathematical notation, but not pseudo code) how you obtained it.*

To extract a single component, we need to find a unit vector w such that the projection $w^T x$ maximises nongaussianity, which is here measured by the approximation of negentropy $J(w^T x)$, for which we use the following nonquadratic functions:

$$g(u) = tanh(u)$$
$$g'(u) = 1 - tanh^2(u)$$

We extract the weight vector w for a single component by following these steps:

1. Choose an initial random weight vector w
2. $w^+ = E\{xg(w^T x)\} - E\{g'(w^T x)\}w$
3. $w = \frac{w^+}{\|w^+\|}$
4. if not converged, go back to 2.

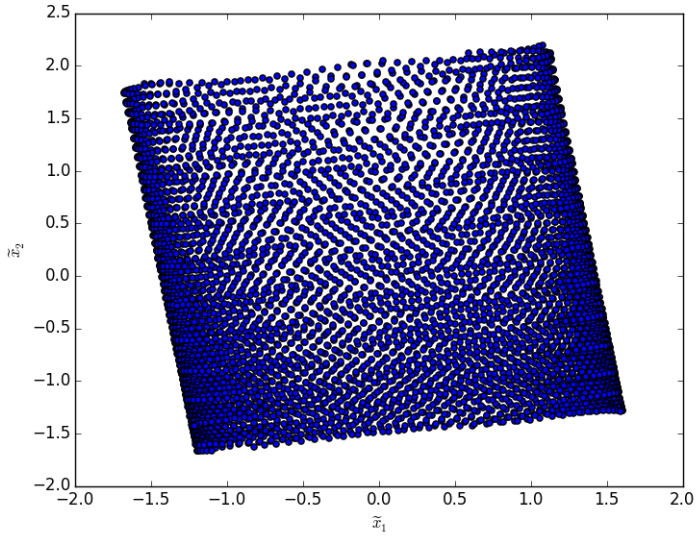


Figure 4: The joint distribution of the whitened mixtures

Where x is the whitened data matrix and the average is over all column-vectors of x.
Since we want to estimate two independent components, we have to run the previous fast ICA algorithm using two units : w_1 and w_2. We need to prevent the different vectors from converging to the same maxima, and for that we must decorrelate $w_1^T x$ and $w_2^T x$. For that we use the Gram-Schmidt-like decorrelation described in *Hyvärinen and Oja*: to estimate a new component p, we run the one-unit fixed-point algorithm for w_p, and after every iteration step subtract from w_p the projections of the previously estimated $p-1$ vectors, and then renormalise w_p:

$$w_p = w_p - \sum_{j=1}^{p-1} w_j w_p^T w_j$$
$$w_p = \frac{w_p}{\|w_p\|}$$

After obtaining the un-mixing matrix W where each row projects X (the whitened data matrix) onto an independent component, we obtain the independent component matrix S by:

$$S = WX$$

And the mixing matrix A by:

$$A = W^{-1}$$

Figure 5 represents the two independent components. We note that because both S and A are unknown, we can only recover s_1 and s_2 up to a scaling factor: we cannot determine the variances of the independent components. We cannot determine their order either.
Figure 6 represents the two mixing vectors A plotted on the whitened data.

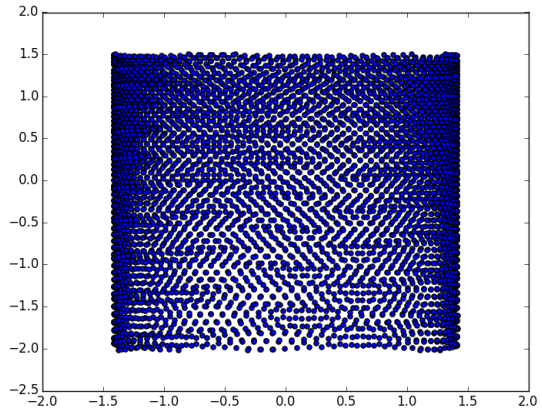


Figure 5: The unmixed 2D distribution of the signals

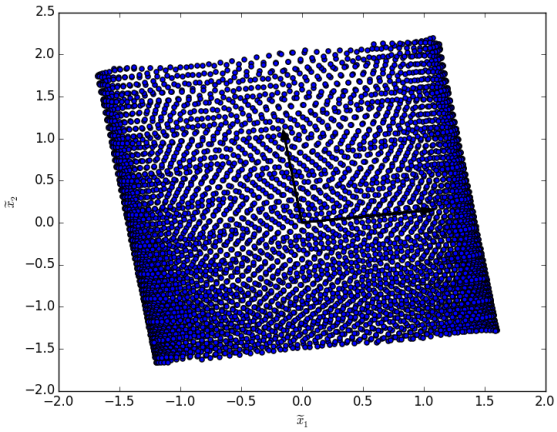


Figure 6: Mixing vectors

2.2 Implementation of Latent Dirichlet Allocation (LDA)

Question 15: *Implement LDA with Gibbs sampling and run it with the given R3 data. Vary the settings of the three parameters K, α and σ. What effect do they have on the learned topic space? (In particular, you should try K = 3, i.e., one topic for each class.) In the report, you should show a list of the 20 most common words in each topic k, along with their weights in the learned per topic word distribution β_k. You should also, for one of the training documents m, show the latent per document topic distribution θ_m for that document. Print out the words of the document m in the report and explain, with examples from the document, the reasons for this topic distribution θ_m.*

For $K = 3$ and $K = 10$, we ran the sampler for 1000 iterations, and print out the 30 most common words in each topic with their weights in the learned per-topic distribution β_k.

- $K = 3$: The topics correspond to the three classes of documents. We can see in table 3 that many of the most common words for each topic are in fact common prepositions, conjunctions and verbs (stop-words), and they have similar weights in the 3 topics. That is due to the fact that these words occur very often in all documents. However some of the most common words (colored in red) still allow us to identify each topic: topic 0 corresponds to the class crude, topic 1 corresponds to the class fx-money, and topic 2 corresponds to the class trade.
  We also print out document 19, and for each word specify its number of occurrences in the document and its topic assignment in this document.

  ('to', 4.0, 0) ('and', 2.0, 0) ('the', 5.0, 0) ('of', 5.0, 0) ('prices', 1.0, 0) ('last', 1.0, 0) ('in', 3.0, 2) ('s', 1.0, 2) ('said', 1.0, 2) ('by', 1.0, 0) ('be', 1.0, 0) ('reuter', 1.0, 0) ('crude', 2.0, 0) ('today', 1.0, 0) ('oil', 5.0, 0) ('market', 1.0, 0) ('opec', 2.0, 0) ('meet', 2.0, 0) ('agreement', 1.0, 2) ('some', 1.0, 0) ('other', 1.0, 0) ('would', 1.0, 0) ('will', 1.0, 2) ('their', 1.0, 2) ('six', 1.0, 0) ('petroleum', 1.0, 0) ('are', 1.0, 0) ('from', 1.0, 0) ('kuwait', 1.0, 0) ('members', 1.0, 0) ('united', 1.0, 0) ('arab', 3.0, 0) ('emirates', 2.0, 0) ('official', 2.0, 2) ('countries', 1.0, 0) ('marketing', 1.0, 0) ('gulf', 3.0, 0) ('qatar', 1.0, 0) ('deputy', 2.0, 0) ('officials', 1.0, 0) ('states', 2.0, 2) ('help', 1.0, 0) ('saudi', 1.0, 0) ('sunday', 1.0, 0) ('each', 1.0, 0) ('cooperation', 1.0, 2) ('council', 1.0, 2) ('gcc', 2.0, 0) ('bahrain', 2.0, 0) ('arabia',

1.0, 0) ('uae', 1.0, 0) ('agency', 1.0, 0) ('four', 1.0, 0) ('news', 1.0, 0) ('ministers', 3.0, 2) ('discuss', 1.0, 0) ('coordination', 1.0, 2) ('wam', 2.0, 2) ('reported', 1.0, 0) ('discussing', 1.0, 0) ('implementation', 1.0, 2) ('doha', 1.0, 0) ('organisation', 1.0, 0) ('exporting', 1.0, 0) ('face', 1.0, 2) ('stiff', 1.0, 0) ('buyer', 1.0, 2) ('resistance', 1.0, 0) .

We can see that the majority of the words in this document are assigned to the topic crude (**opec, crude, oil, exporting, saudi, qatar, emirates**), many of which are among the most common in topic crude. A few words are assigned to the topic trade (**buyer, cooperation, agreement**), and no word is assigned to the topic fx-money. Indeed, the topic distribution for this document is:

$$\theta_{19} = \begin{pmatrix} 0.759 \\ 0.004 \\ 0.236 \end{pmatrix}$$

- $K = 10$: Here the topics are the latent low-dimensional representation of the document. We can see in tables 4 and 5 that conjunctions, prepositions and stop-words no longer appear among the most-common words for some topics (topics 1, 1, 4, 5, 6, 7 ,8) while they still do for others (topics 2, 3, 9). We can see that the words in some topics have semantical similarity: bank, currency, monetary, exchange, dollar, banks, rate, securities in topic 1, bill, means, democrat, laws, chairman, congressional in topic 8. On the other hand, other topics have words that are not semantically close, but must have other types of associations.

  We print out the words in the previous document with their counts and their new topic assignments:

  ('to', 4.0, 9), ('and', 2.0, 9), ('the', 5.0, 9), ('of', 5.0, 9), ('prices', 1.0, 9), ('last', 1.0, 9), ('in', 3.0, 9), ('s', 1.0, 2), ('said', 1.0, 9), ('by', 1.0, 9), ('be', 1.0, 9), ('reuter', 1.0, 9), ('crude', 2.0, 9), ('today', 1.0, 9), ('oil', 5.0, 9), ('market', 1.0, 9), ('opec', 2.0, 9), ('meet', 2.0, 2), ('agreement', 1.0, 9), ('some', 1.0, 3), ('other', 1.0, 9), ('would', 1.0, 9), ('will', 1.0, 9), ('their', 1.0, 9), ('six', 1.0, 2), ('petroleum', 1.0, 9), ('are', 1.0, 9), ('from', 1.0, 9), ('kuwait', 1.0, 9), ('members', 1.0, 9), ('united', 1.0, 2), ('arab', 3.0, 9), ('emirates', 2.0, 9), ('official', 2.0, 9), ('countries', 1.0, 2), ('marketing', 1.0, 9), ('gulf', 3.0, 9), ('qatar', 1.0, 9), ('deputy', 2.0, 2), ('officials', 1.0, 2), ('states', 2.0, 9), ('help', 1.0, 9), ('saudi', 1.0, 9), ('sunday', 1.0, 9), ('each', 1.0, 9), ('cooperation', 1.0, 9), ('council', 1.0, 2), ('gcc', 2.0, 9), ('bahrain', 2.0, 9), ('arabia', 1.0, 9), ('uae', 1.0, 9), ('agency', 1.0, 9), ('four', 1.0, 2), ('news', 1.0, 2), ('ministers', 3.0, 9), ('discuss', 1.0, 2), ('coordination', 1.0, 2), ('wam', 2.0, 9), ('reported', 1.0, 9), ('discussing', 1.0, 9), ('implementation', 1.0, 2), ('doha', 1.0, 9), ('organisation', 1.0, 9), ('exporting', 1.0, 9), ('face', 1.0, 9), ('stiff', 1.0, 9), ('buyer', 1.0, 9), ('resistance', 1.0, 9).

The new topic distribution θ_{19} is :

$$\begin{pmatrix} 0.0042 \\ 0.0042 \\ 0.1873 \\ 0.0183 \\ 0.0183 \\ 0.0042 \\ 0.0042 \\ 0.0042 \\ 0.00422 \\ 0.7647 \end{pmatrix}$$

We can see that all topics have equally low weights, except topic 2 (0.1873) and topic 9 (0.7647). Topic 9 has many stop-words, as well as some words associated with the class crude among its most common words: oil, dlrs, pct, barrels, prices, crude.

Question 16: *Now use a K > 3, for example K = 10 or K = 15. For each test document mtest, infer the topic distribution, using the per topic word distributions learned in the training phase. Classify each θ_{mtest} with kNN and the training document representations θ_m. Study a couple of correctly classified documents, and a couple of wrongly classified documents. Are the correctly classified documents more typical for their class?*

We run the sampler with $K = 10$ for 1000 iterations. After inferring the topic distribution $\theta_{m_{test}}$ for each document, we classify them with KNN (10 neighbours) using the training document topic distributions θ_m. We get an accuracy of 89.39%. The confusion matrix is:

$$\begin{pmatrix} 121 & 0 & 0 \\ 24 & 49 & 2 \\ 4 & 0 & 83 \end{pmatrix}$$

As we can see, all the documents from the crude class were correctly predicted, as well as most documents from the money-fx class. However, many documents from the class trade were predicted as belonging to the class crude. We compare document 4 which is labeled from the class trade and has been classified in the class crude, and document 8 which was correctly classified in class trade. Document 4 seems less typical because it mixes international trade vocabulary (japan, sanctions, cooperation, official, trade) to words from the economic jargon that appear frequently in crude documents (economy, barriers, export, markets), while no such words appear in document 8. In general, the classes crude and trade both have a strong international and political semantic component which could explain why it appears difficult to correctly separate them.

Another interesting example is document 46 which has the label money-fx but has been classified as belonging to the class crude. This document has many occurences of the word kenya and the word shilling, which is a currency that never appeared in the training documents, and other words like dollar which appears frequently in both classes. The class crude has the lower precision of all classes, which explains why this document has been assigned to this class.

2.3 Derivation of Gibbs sampling for Latent Dirichlet Allocation (LDA)

Question 17: *Please derive Equation (3) from the dependencies represented in Figure 3. We require small steps where only one type of mathematic operation is allowed in each step. Comments should be added between each step.*

In this question we use the same notations as in the paper *Gibbs sampling in the generative model of Latent Dirichlet Allocation*, by *Tom Griffiths*:

- For each document, the multinomial distribution over T topics:

$$z_i|\theta^{(d_i)} \sim Discrete(\theta^{(d_i)})$$
$$\theta \sim Dirichlet(\alpha)$$

- For each topic, the the multinomial distribution over W words in the vocabulary:

$$w_i|z_i, \phi^{(z_i)} \sim Discrete(\phi^{(z_i)})$$
$$\phi \sim Dirichlet(\beta)$$

We want to calculate the full conditional probability for z_i :

$$\begin{aligned}
p(z_i = j|z_{-i}, w) &= p(z_i = j|z_{-i}, w_i, w_{-i}) && (17)\\
&= \{Bayes'rule\} && (18)\\
&= \frac{p(z_i = j, w_i, w_{-i}|z_{-i}}{p(w|z_{-i})} && (19)\\
&\propto p(z_i = j, w_i|z_{-i})p(w_i|z_i = j, w_{-i}, z_{-i}) && (20)\\
&= \{z_i = j \perp w_{-i}|z_{-i}\} && (21)\\
&= p(z_i = j|z_{-i})p(w_{-i}|z_{-i})p(w_i|z_i = j, w_{-i}, z_{-i}) && (22)\\
&\propto p(w_i|z_i = j, w_{-i}, z_{-i})p(z_i = j|z_{-i}) && (23)
\end{aligned}$$

By integrating over ϕ in the first term of equation 23:

$$\begin{aligned}
p(w_i|z_i = j, w_{-i}, z_{-i}) &= \int p(w_i, \phi^{(j)}|z_i = j, w_{-i}, z_{-i})d\phi^{(j)} && (24)\\
&= \int p(w_i|\phi^{(j)}, z_i = j, w_{-i}, z_{-i})p(\phi^{(j)}|z_i = j, w_{-i}, z_{-i})d\phi^{(j)} && (25)\\
&= \{w_i \perp (w_{-i}, z_{-i})|\phi^{(j)}, z_i = j\}\ \{\phi^{(j)} \perp Z_i = j|w_{-i}, z_{-i}\} && (26)\\
&= \int p(w_i|\phi^{(j)}, z_i = j)p(\phi^{(j)}|z_{-i}, w_{-i})d\phi^{(j)} && (27)
\end{aligned}$$

We look at the rightmost term of equation 27:

$$\begin{aligned}
p(\phi^{(j)}|z_{-i}, w_{-i}) &= \frac{p(\phi^{(j)}, w_{-i}|z_{-i})}{p(w_{-i}|z_{-i}} && (28)\\
&\propto p(\phi^{(j)}, w_{-i}|z_{-i}) && (29)\\
&= p(w_{-i}|\phi^{(j)}, z_{-i})p(\phi^{(j)}|z_{-i}) && (30)\\
&= \{\phi^{(j)} \perp z_{-i}|v - structure, w \text{ unknown}\} && (31)\\
&= p(w_{-i}|\phi^{(j)}, z_{-i})p(\phi^{(j)}) && (32)
\end{aligned}$$

Since $p(\phi^{(j)})$ is $Dirichelet(\beta)$ and conjugate to $p(w_{-i}|\phi^{(j)}, z_{-i})$ which is $Discrete(\phi^{(j)})$, we know that the posterior $p(\phi^{(j)}|z_{-i}, w_{-i})$ will be $Dirichlet$, and we find its parameter:

$$p(\phi^{(j)}) = \frac{1}{B(\beta)} \prod_{x=1}^{W} (\phi_x^{(j)})^{\beta-1} \propto \prod_{x=1}^{W} (\phi_x^{(j)})^{\beta-1} \tag{33}$$

$$\begin{aligned} p(w_{-i}|\phi^{(j)}, z_{-i}) &= \prod_{p\neq i} \prod_{x=1}^{W} \phi_{v_x}^{(j) I(z_p=j, w_p=v_x)} \\ &= \prod_{x=1}^{W} \phi_{v_x}^{(j) \sum_{p\neq i} I(z_p=j, w_p=v_x)} \\ &= \prod_{x=1}^{W} \phi_{v_x}^{(j) n_{-i,j}^{(v_x)}} \end{aligned} \tag{34}$$

where $n_{-i,j}^{(v_x)}$ is the number of instances of word v_x assigned to topic j, not including the current word. By replacing the two terms in 32 by their expressions in equations 33 and 34, we obtain:

$$\begin{aligned} p(\phi^{(j)}|z_{-i}, w_{-i}) &\propto \prod_{x=1}^{W} \phi_{v_x}^{(j) n_{-i,j}^{(v_x)} + \beta - 1} \\ &\sim Dirichlet(\beta + n_{-i,j}^{(v_x)}) \end{aligned} \tag{35}$$

The leftmost term of equation 27 is simply :

$$p(w_i|z_i = j, \phi^{(j)}) = \phi_{w_i}^{(j)}$$

We can now replace the two terms in the integral in equation 27:

$$\begin{aligned} p(w_i|z_i = j, w_{-i}, z_{-i}) &= \int p(w_i|\phi^{(j)}, z_i = j) p(\phi^{(j)}|z_{-i}, w_{-i}) d\phi^{(j)} \\ &= \int \phi_{w_i}^{(j)} \frac{1}{B(\beta + n_{-i,j}^{(v)})} \prod_{x=1}^{W} \phi_{v_x}^{(j) n_{-i,j}^{(v_x)} + \beta - 1} d\phi^{(j)} \\ &= \frac{1}{B(\beta + n_{-i,j}^{(v)})} \int \prod_{x=1}^{W} \phi_{v_x}^{(j) I(v_x=w_i) + n_{-i,j}^{(v)} + \beta - 1} d\phi^{(j)} \end{aligned}$$

We recognize the form of a Dirichlet in the equation above, so we multiply and divide by the normalizing constant to complete the integral:

$$p(w_i|z_i = j, w_{-i}, z_{-i}) \tag{36}$$

$$= \frac{B(I(v = w_i) + n_{-i,j}^{(v)} + \beta)}{B(\beta + n_{-i,j}^{(v)})} \int \frac{1}{B(I(v = w_i) + n_{-i,j}^{(v)} + \beta)} \prod_{x=1}^{W} \phi_{v_x}^{(j)I(v_x=w_i)+n_{-i,j}^{(v_x)}+\beta-1} d\phi^{(j)} \tag{37}$$

$$= \frac{\Gamma(\sum_{x=1}^{W} \beta + n_{-i,j}^{(v_x)})}{\prod_{x=1}^{W} \Gamma(\beta + n_{-i,j}^{(v_x)})} \frac{\prod_{x=1}^{W} \Gamma(\beta + n_{-i,j}^{(v_x)} + I(v_x = w_i))}{\Gamma(\sum_{x=1}^{W} (\beta + n_{-i,j}^{(v_x)} + I(v_x = w_i)))} \tag{38}$$

$$= \frac{\Gamma(\sum_{x=1}^{W} \beta + n_{-i,j}^{(v_x)})}{\prod_{x=1}^{W} \Gamma(\beta + n_{-i,j}^{(v_x)})} \frac{\prod_{x=1}^{W} \Gamma(\beta + n_{-i,j}^{(v_x)} + I(v_x = w_i))}{\Gamma(\sum_{x=1}^{W} (\beta + n_{-i,j}^{(v_x)}) + 1)} \tag{39}$$

$$= \{using\ \Gamma(x+1) = x\Gamma(x)\} \tag{40}$$

$$= \frac{\Gamma(\sum_{x=1}^{W} \beta + n_{-i,j}^{(v_x)})}{\prod_{x=1}^{W} \Gamma(\beta + n_{-i,j}^{(v_x)})} \frac{\prod_{x=1}^{W} \Gamma(\beta + n_{-i,j}^{(v_x)} + I(v_x = w_i))}{\Gamma(\sum_{x=1}^{W} (\beta + n_{-i,j}^{(v_x)})) \sum_{x=1}^{W} (\beta + n_{-i,j}^{(v_x)})} \tag{41}$$

$$= \frac{\Gamma(\beta + n_{-i,j}^{(w_i)} + 1) \prod_{x=1}^{W} (\Gamma(\beta + n_{-i,j}^{(v_x)} + I(v_x = w_i)))^{I(v_x \neq w_i)}}{\Gamma(\beta + n_{-i,j}^{(w_i)}) \prod_{x=1}^{W} (\Gamma(\beta + n_{-i,j}^{(v_x)}))^{I(v_x \neq w_i)}} \frac{1}{\sum_{x=1}^{W} (\beta + n_{-i,j}^{(v_x)})} \tag{42}$$

$$= \frac{(\beta + n_{-i,j}^{(w_i)}) \Gamma(\beta + n_{-i,j}^{(w_i)})}{\Gamma(\beta + n_{-i,j}^{(w_i)})} \frac{1}{\sum_{x=1}^{W} (\beta + n_{-i,j}^{(v_x)})} \tag{43}$$

$$= \frac{\beta + n_{-i,j}^{(w_i)}}{\sum_{x=1}^{W} (\beta + n_{-i,j}^{(v_x)})} \tag{44}$$

By integrating over θ in the second term of equation 23:

$$p(z_i = j|z_{-i}) = \int p(z_i = j, \theta^{(d_i)}|z_{-i}) d\theta^{(d_i)} \tag{45}$$

$$= \int p(z_i = j|\theta^{(d_i)}, z_{-i}) p(\theta^{(d_i)}|z_{-i}) d\theta^{(d_i)} \tag{46}$$

$$= \{z_i \perp z_{-i}|\theta^{(d_i)}\} \tag{47}$$

$$= \int p(z_i = j|\theta^{(d_i)}) p(\theta^{(d_i)}|z_{-i}) d\theta^{(d_i)} \tag{48}$$

We look at the rightmost term of equation 48:

$$p(\theta^{(d_i)}|z_{-i}) = \frac{p(\theta^{(d_i)}, z_{-i})}{p(z_i)} \propto p(z_{-i}|\theta^{(d_i)}) p(\theta^{(d_i)})$$

Since $p(\theta^{(d_i)})$ is $Dirichlet(\alpha)$ and conjugate to $p(z_{-i}|\theta^{(d_i)})$ which is $Discrete(\theta^{(d_i)})$, the posterior $p(\theta^{(d_i)}|z_{-i})$ will be $Dirichlet$. We find its parameter:

$$p(\theta^{(d_i)}) = \frac{1}{B(\alpha)} \prod_{t=1}^{T} (\theta_t^{(d_i)})^{\alpha-1}$$

$$p(z_{-i}|\theta^{(d_i)}) = \prod_{p \neq i} \prod_{t=1}^{T} (\theta_t^{(d_i)})^{I(z_p=t, d_p=d_i}$$

$$= \prod_{t=1}^{T} (\theta_t^{(d_i)})^{n_{i,t}^{(d_i)}}$$

And therefore, the posterior is:

$$p(\theta^{(d_i)}|z_{-i}) \propto \prod_{t=1}^{T}(\theta_t^{(d_i)})^{\alpha+n_{-i,t}^{(d_i)}-1}$$
$$\sim Dirichlet(\alpha + n_{-i,t}^{(d_i)})$$

The leftmost term of equation 48 is simply:

$$p(z_i = j|\theta^{(d_i)}) = \theta^{(d_i)}$$

We can now replace the two terms in the integral in equation 48:

$$\begin{aligned} p(z_i = j|z_{-i}) &= \int p(z_i = j|\theta^{(d_i)})p(\theta^{(d_i)}|z_{-i})d\theta^{(d_i)} \\ &= \int \theta^{(d_i)} \frac{1}{B(\alpha + n_{-i,(.)}^{(d_i)})} \prod_{t=1}^{T}(\theta_t^{(d_i)})^{\alpha+n_{-i,(.)}^{(d_i)}-1} d\theta^{(d_i)} \\ &= \frac{1}{B(\alpha + n_{-i,(.)}^{(d_i)})} \int \prod_{t=1}^{T}(\theta_t^{(d_i)})^{I(t=j)\alpha+n_{-i,t}^{(d_i)}-1} d\theta^{(d_i)} \end{aligned}$$

We recognize the form of a Dirichlet inside the integral, so we multiply and divide by the normalizing constant to complete the integral:

$$p(z_i = j|z_{-i}) \tag{49}$$
$$= \frac{B(I(t=j) + \alpha + n_{-i,(.)}^{(d_i)})}{B(\alpha + n_{i,(.)}^{(d_i)})} \int \frac{1}{B(I(t=j) + \alpha + n_{i,(.)}^{(d_i)})} \prod_{t=1}^{T}(\theta_t^{(d_i)})^{I(t=j)\alpha+n_{-i,t}^{(d_i)}-1} d\theta^{(d_i)} \tag{50}$$
$$= \frac{\Gamma(\sum_{t=1}^{T}(\alpha + n_{-i,t}^{(d_i)}))}{\prod_{t=1}^{T}\Gamma(\alpha + n_{i,t}^{(d_i)})} \frac{\prod_{t=1}^{T}\Gamma(\alpha + n_{-i,t}^{(d_i)} + I(t=j))}{\Gamma(\sum_{t=1}^{T}(\alpha + n_{-i,t}^{(d_i)} + I(t=j)))} \tag{51}$$
$$= \frac{\Gamma(\sum_{t=1}^{T}(\alpha + n_{-i,t}^{(d_i)}))}{\prod_{t=1}^{T}\Gamma(\alpha + n_{-i,t}^{(d_i)})} \frac{\prod_{t=1}^{T}\Gamma(\alpha + n_{-i,t}^{(d_i)} + I(t=j))}{\Gamma(\sum_{t=1}^{T}(\alpha + n_{-i,t}^{(d_i)}) + 1)} \tag{52}$$
$$= \{\Gamma(x+1) = x\Gamma(x)\} \tag{53}$$
$$= \frac{\Gamma(\sum_{t=1}^{T}(\alpha + n_{-i,t}^{(d_i)}))}{\prod_{t=1}^{T}\Gamma(\alpha + n_{i,t}^{(d_i)})} \frac{\prod_{t=1}^{T}\Gamma(\alpha + n_{-i,t}^{(d_i)} + I(t=j))}{\sum_{t=1}^{T}(\alpha + n_{-i,t}^{(d_i)})\Gamma(\sum_{t=1}^{T}(\alpha + n_{-i,t}^{(d_i)}))} \tag{54}$$
$$= \frac{\Gamma(\alpha + n_{-i,t}^{(d_i)} + 1)\prod_{t=1}^{T}(\Gamma(\alpha + n_{-i,t}^{(d_i)} + I(t=j)))^{I(t\neq j)}}{\Gamma(\alpha + n_{-i,t}^{(d_i)})\prod_{t=1}^{T}(\Gamma(\alpha + n_{-i,t}^{(d_i)}))^{I(t\neq j)}} \frac{1}{\sum_{t=1}^{T}(\alpha + n_{-i,t}^{(d_i)})} \tag{55}$$
$$= \frac{(\alpha + n_{-i,t}^{(d_i)})\Gamma(\alpha + n_{-i,t}^{(d_i)})}{\Gamma(\alpha + n_{-i,t}^{(d_i)})} \frac{1}{\sum_{t=1}^{T}(\alpha + n_{-i,t}^{(d_i)})} \tag{56}$$
$$= \frac{\alpha + n_{-i,t}^{(d_i)}}{\sum_{t=1}^{T}(\alpha + n_{-i,t}^{(d_i)})} \tag{57}$$

Therefore, from 23, 44 and 57 we have:

$$p(z_i = j|z_{-i}, w) = \frac{\beta + n_{-i,j}^{(w_i)}}{\sum_{x=1}^{W}(\beta + n_{-i,j}^{(v_x)})} \frac{\alpha + n_{-i,t}^{(d_i)}}{\sum_{t=1}^{T}(\alpha + n_{-i,t}^{(d_i)})}$$

Table 3: Most common words in each topic with their weights β_{kw}, $K = 3$

Topic 0	Topic 1	Topic 2
oil, 0.010	the, 0.0115	on, 0.0064
to, 0.0084	of, 0.0113	that, 0.0059
the, 0.0083	reuter, 0.0109	trade, 0.0054
reuter, 0.0082	to, 0.0108	said, 0.0054
a, 0.0080	in, 0.0103	in, 0.0054
said, 0.0077	said, 0.0097	he, 0.0053
of, 0.0074	a, 0.0087	for, 0.0053
and, 0.0071	and, 0.0087	a' 0.0052
for, 0.0065	bank, 0.0079	and, 0.0052
in, 0.0061	s, 0.0077	is, 0.0051
s, 0.0060	it, 0.0069	to, 0.0051
it, 0.0059	from, 0.0068	would, 0.0050
prices, 0.0057	market, 0.0067	of, 0.0049
crude, 0.0056	with, 0.0064	reuter, 0.0049
its, 0.0053	at, 0.0062	be, 0.0048
will, 0.0051	exchange, 0.0060	the, 0.0047
on, 0.00506	billion, 0.0056	as, 0.0046
by, 0.00502	dollar, 0.00518	not, 0.0043
mln, 0.00489	pct, 0.0050	but, 0.0041
dlrs, 0.0048	this, 0.0047	with, 0.0040
was, 0.0044	mln, 0.0047	by, 0.0040
is, 0.0044	currency, 0.0042	are, 0.0040
that, 0.0044	money, 0.0042	told, 0.0040
an, 0.0043	rate, 0.0041	was, 0.0037
as, 0.0042	had, 0.0040	have, 0.0037
at, 0.0042	today, 0.0040	it, 0.0036
barrels, 0.0042	treasury, 0.0039	an, 0.0035
be, 0.0040	was, 0.0038	japan, 0.0032
petroleum, 0.0039	rates, 0.0037	states, 0.0032
day, 0.0039	by, 0.0037	countries, 0.0032

Table 4: Most common words in each topic with their weights β_{kw}, $K = 10$ (I)

Topic 0	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**
head, 0.0025	bank, 0.0074	to, 0.0075	said, 0.013	dlrs, 0.0083
vital, 0.0019	currency, 0.0070	the, 0.0068	market, 0.0125	crude, 0.0081
day, 0.0011	monetary, 0.00648	reuter, 0.0067	of', 0.0122	today, 0.0061
whose, 0.0011	exchange, 0.0063	in, 0.0067	the, 0.0111	west, 0.0059
favour, 0.0011	dollar, 0.0052	and, 0.0064	bank, 0.0110	cts, 0.0057
turkish, 0.0011	banks, 0.0050	a, 0.0064	reuter, 0.0108	bbl, 0.0057
go, 0.0011	rate, 0.0042	of, 0.0063	it, 0.0107	effective, 0.0053
northwest, 0.0008	central, 0.0041	said, 0.0060	mln, 0.0107	canada, 0.0051
activities, 0.0008	currencies, 0.0039	on, 0.0060	money, 0.0102	texas, 0.0051
adviser, 0.0008	securities, 0.0034	trade, 0.0058	today, 0.0099	one, 0.0047
thus, 0.0008	fund, 0.0034	he, 0.0058	stg, 0.0093	raises, 0.0047
arrangement, 0.0008	fed, 0.0033	that, 0.0058	in, 0.0091	barrel, 0.0043
holding, 0.0008	dealers, 0.0030	for, 0.0056	a, 0.0088	oil, 0.0041
study, 0.0008	funds, 0.0030	with, 0.0052	england, 0.0086	posted, 0.0041
priority, 0.0008	dollars, 0.0028	would, 0.0051	to, 0.0083	unit, 0.0040
cross, 0.0008	financial, 0.0026	be, 0.0050	and, 0.0074	company, 0.0040
produced, 0.0008	mark, 0.0026	as, 0.0049	shortage, 0.0064	intermediate, 0.0038
formerly, 0.0008	term, 0.0026	is, 0.0048	assistance, 0.0061	postings, 0.0038
groups, 0.0008	german, 0.0026	not, 0.0047	at, 0.0061	prices, 0.0038
nigel, 0.0008	interest, 0.0025	it, 0.0046	forecast, 0.0060	co, 0.0036
stronger, 0.0008	system, 0.0025	by, 0.0044	bills, 0.00600	raised, 0.00341
become, 0.0008	economists, 0.0025	was, 0.0044	its, 0.0058	light, 0.0034
general, 0.0008	rates, 0.0025	has, 0.0040	around, 0.0056	canadian, 0.0032
seeking, 0.0008	intervention, 0.0024	have, 0.0039	treasury, 0.0050	to, 0.0032
had, 0.0008	capital, 0.0022	will, 0.0038	this, 0.0050	grade, 0.0032
reshuffle, 0.0008	fixed, 0.0022	told, 0.0038	revised, 0.0049	price, 0.0032
valued, 0.0008	banking, 0.0022	are, 0.0037	with, 0.0047	contract, 0.0030
court, 0.0008	french, 0.0022	japan, 0.0036	from, 0.0045	subsidiary, 0.0030
his, 0.0008	marks, 0.0021	an, 0.0036	system, 0.0045	grades, 0.0028
appointed, 0.0008	rise, 0.0021	this, 0.0036	billion, 0.0044	british, 0.0026

Table 5: Most common words in each topic with their weights β_{kw}, $K = 10$ (II)

Topic 5	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**
businessmen, 0.0023	earthquake, 0.0050	memory, 0.0016	bill, 0.0048	the, 0.0108
soviet, 0.0020847	venezuela, 0.0048	below, 0.00137	means, 0.0044	reuter, 0.0108
diplomats, 0.0015	pipeline, 0.00480	beginning, 0.0013	democrat, 0.0029	of, 0.0105
paper, 0.0013	pacific, 0.0043	start, 0.0013	laws, 0.0029	said, 0.0104
restrictions, 0.00133	ecuador, 0.00344	introduction, 0.0011	house, 0.0027	a, 0.0104
favourable, 0.001	mines, 0.00344	opening, 0.0011	force, 0.0027	and, 0.0099
chinese, 0.0013	night, 0.00299	read, 0.0011	chairman, 0.0027	in, 0.0096
preventing, 0.0010	jungle, 0.0027	economics, 0.0011	ways, 0.0025	oil, 0.00839
state, 0.0010	deputy, 0.0025	dry, 0.0011	congressional, 0.0025	for, 0.0079
easing, 0.0010	balao, 0.0025	office, 0.0011	richard, 0.0025	from, 0.0067
commercial, 0.0010	damaged, 0.0025	unveiled, 0.0011	authority, 0.0023	at, 0.0067
embassy, 0.0010	accounts, 0.0023	performance, 0.0011	relief, 0.0021	it, 0.0066
moved, 0.0010	venezuelan, 0.0023	counter, 0.0011	rostenkowski, 0.0021	on, 0.0065
others, 0.0010	five, 0.0023	quarter, 0.0011	measure, 0.0021	by, 0.0065
described, 0.0010	grisanti, 0.002	effects, 0.0008	subcommittee, 0.0021	was, 0.0062
control, 0.0010	hernandez, 0.0023	compliance, 0.0008	technology, 0.0021	is, 0.0059
peking, 0.0010	lend, 0.00209	ambassador, 0.0008	leaders, 0.0019	mln, 0.00586
applied, 0.0008	santos, 0.0020	industries, 0.0008	surpluses, 0.0019	its, 0.0055
labour, 0.0008	damage, 0.0020	extend, 0.0008	unfair, 0.0019	that, 0.0053
wanted, 0.0008	repair, 0.0018	sydney, 0.0008	retaliate, 0.0019	dlrs, 0.0053
deal, 0.0008	arturo, 0.0018	continuous, 0.0008	amendment, 0.0017	will, 0.0053
newspaper, 0.0008	quake, 0.0018	amount, 0.0008	proposal, 0.0017	pct, 0.0052
further, 0.0008	port, 0.0018	clients, 0.00085	texas, 0.0017	an, 0.0052
opportunities, 0.0008	colombia, 0.0018	distributors, 0.0008	toughen, 0.0017	year, 0.0046
agriculture, 0.0008	fernando, 0.0018	lower, 0.0008	provisions, 0.0017	were, 0.0046
encourage, 0.0008	coast, 0.0016	environmental, 0.0008	speaker, 0.0015	prices, 0.0045
noir, 0.00082	who, 0.0016	increase, 0.0008	unanimously, 0.0015	day, 0.0042
application, 0.0008	alvite, 0.0016	ltd, 0.0008	aide, 0.0015	barrels, 0.0039
expand, 0.00082	linking, 0.0016	working, 0.0008	passage, 0.0015	per, 0.0037
expectations, 0.00082	today, 0.0016	reversed, 0.0008	wright, 0.0015	crude, 0.00348