CSE 674 Project 1: Determining Probabilities of Handwriting Formations using PGMs

Sargur N. Srihari
University at Buffalo, The State University of New York
Buffalo, New York 14260
Contact: 716-645-6162 (O), srihari@buffalo.edu

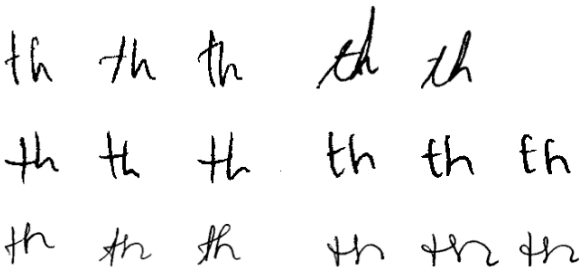
December 30, 2018

1 Objective

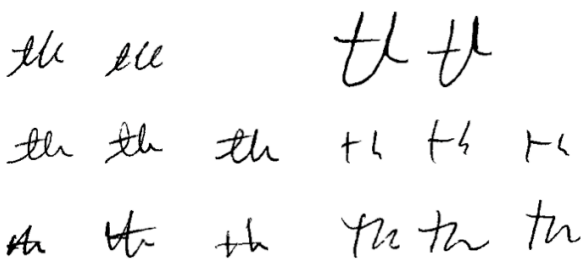
This project is to develop probabilistic graphical models (PGMs) to determine probabilities of observations which are described by several variables. We will work with handwriting patterns which are described by document examiners. They can be used to determine whether a particular handwriting sample is common or rare and which in turn can be useful to determine whether a particular handwriting sample was written by a certain individual. We consider only the letter pair *th* in this study. Since it is the most commonly encountered pair of letters (called a bigram) in English (see Fig. 1). Some examples of handwritten *th* are shown in Figure 2. They are in groups of samples of different writers.

Bigram	Count	Bigram	Count	Bigram	Count
th	50	at	25	st	20
er	40	en	25	io	18
on	39	es	25	le	18
an	38	of	25	is	17
re	38	or	25	ou	17
he	33	nt	24	ar	16
in	31	ea	22	as	16
ed	30	ti	22	de	16
nd	30	to	22	rt	16
ha	26	it	20	ve	16

Figure 1: Bigram frequencies in English.



(a)



(b)

Figure 2: Some examples of handwritten *th*. The examples are grouped by being written by different individuals. For instance the first three from the top left were written by the same person, next two by another, etc.

2 Feature definitions

A characterization of the structure of *th* as given by document examiners as shown in Table 1. In this characterization here are six random variables x_1-x_6 taking on the values a, b, c, d, e as indicated in parentheses.

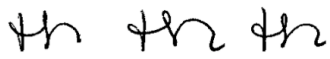
Table 1: Characteristics of *th* as provided by document examiners.

x_1 **(Height Relationship of *t* to *h*)**	x_2 **(Shape of Loop of *h*)**	x_3 **(Shape of Arch of *h*)**	x_4 **(Height of Cross on *t* staff)**	x_5 **(Baseline of *h*)**	x_6 **(Shape of *t*)**
t shorter than h (a)	retraced (a)	rounded arch (a)	upper half of staff (a)	slanting upward (a)	tented (a)
t even with h (b)	curved right side and straight left side (b)	pointed (b)	lower half of staff (b)	slanting downward (b)	single stroke (b)
t taller than h (c)	curved left side and straight right side (c)	no set pattern (c)	above staff (c)	baseline even (c)	looped (c)
no set pattern (d)	both sides curved (d)		no fixed pattern (d)	no set pattern (d)	closed (d)
	no fixed pattern (e)				mixture of shapes (e)

2.1 Examples of encoding

Two examples of images encoded in this way are given in Figure 3. Their probabilities can be determined from the graphical model constructed. How would you use such probabilities to determine whether a a particular handwriting sample was written by a given individual with known handwriting characteristics?

(a)



(b)

Figure 3: Example encodings of *th* images using characeristics described in Table 1: (a) Samples of Writer 1, which are jointly encoded as $x_1 = b, x_2 = a, x_3 = a, x_4 = d, x_5 = a, x_6 = b$, and (b) Samples of Writer 2, which are jointly encoded as $x_1 = b, x_2 = b, x_3 = a, x_4 = b, x_5 = a, x_6 = c$.

3 Marginal Probabilities

The marginal distributions of the six variables are given in Table 2 which are the same as listed in a more verbose form in Figure 3.

Table 2: Marginal distributions of the six features of *th*.

Values	x_1 (Relative height of t to h)	x_2 (Shape of h loop)	x_3 (Shape of h arch)	x_4 (Height of Cross of t)	x_5 (Baseline of h)	x_6 (Shape of t)
a	78%(156)	27.5%(55)	18%(36)	71.5%(143)	37.5%(75)	1.5%(3)
b	1.5%(3)	32%(64)	66%(132)	10.5%(21)	11%(22)	32%(64)
c	5.5%(11)	2.5% (5)	16%(32)	1%(2)	10.5%(21)	14%(28)
d	15%(30)	17%(34)		17%(34)	41%(82)	31.5%(63)
e		21%(42)				21%(42)

1. **Height relationship of the "t" to the "h":**
   a. **78% (156)** made "t" shorter than "h"
   b. **1.5% (3)** made "t" even with "h"
   c. **5.5% (11)** made "t" taller than "h"
   d. **15% (30)** no set pattern
2. Shape of the loop of the "h":
   a. **27.5% (55)** made retraced staff
   b. **32% (64)** made loop with curved right side and straight left
   c. **2.5% (5)** made loop with curved left side and straight right
   d. **17% (34)** made a loop with both sides curved
   e. **21% (42)** had no fixed pattern
3. Shape of the arch of the "h":
   a. **18% (36)** made rounded arch
   b. **66% (132)** made pointed arch
   c. **16% (32)** made arch with no set pattern
4. Height of cross on "t" staff:
   a. **71.5% (143)** made cross in upper half of staff
   b. **10.5% (21)** made cross in lower half of staff
   c. **1% (2)** made cross above staff
   d. **17% (34)** made cross with no fixed pattern
5. Baseline of the "h":
   a. **37.5% (75)** made baseline slanting upward
   b. **11% (22)** made baseline slanting downward
   c. **10.5% (21)** made baseline even
   d. **41% (82)** had no set pattern
6. Shape of the "t":
   a. **1.5% (3)** made tented "t"
   b. **32% (64)** made single stroke "t"
   c. **14% (28)** made looped "t"
   d. **31.5% (63)** made closed "t"
   e. **21% (42)** made a mixture of "t" shapes

Figure 4: Marginal probabilities (verbose). Based on samples from 200 individuals. From Muehlberger, et. al., "A Statistical Examination of Selected Handwriting Characteristics,", Journal of Forensic Sciences (1977), 205-211.

4 Conditional Probabilities

Several conditional distributions of the the six variables are given in Tables 3, 4, 5 and 6.

4.1 Distributions conditioned on x_1, height relationship of *t* to *h*

We have three conditional distributions available: (i) $p(x_2|x_1)$, where x_2 is the shape of the h loop, (ii) $p(x_4|x_1)$, where x_4 is the height of cross on t staff, and (iii) $p(x_6|x_1)$, where x_6 is the shape of t.

Table 3: Distributions conditioned on x_1, height relationship of t to h.

$x_1 =$	a (t shorter than h)	b (t even with h)	c (t taller than h)	d (No set pattern)
Total	78%(156)	1.5%(3)	5.5%(11)	15%(30)
$x_2 = a$ (retraced staff)	23.1%(36)	66.6%(2)	45.5%(5)	40%(12)
$x_2 = b$ (curved right)	36.5%(57)	0%(0)	9.1%(1)	20%(6)
$x_2 = c$ (curved left)	2.6%(4)	0%(0)	0%(0)	3.3%(1)
$x_2 = d$ (both curved)	17.3%(27)	0%(0)	18.2%(2)	16.7%(5)
$x_2 = e$ (no pattern)	20.5%(32)	33.3%(1)	27.3%(3)	20%(6)
$x_4 = a$ (upper staff)	73.7%(115)	100%(3)	72.7%(8)	56.7%(17)
$x_4 = b$ (lower staff)	7.7%(12)	0%(0)	27.3%(3)	20%(6)
$x_4 = c$ (above staff)	1.3%(2)	0%(0)	0%(0)	0%(0)
$x_4 = d$ (no pattern)	17.3%(27)	0%(0)	0%(0)	23.3%(7)
$x_6 = a$ (tented t)	1.9%(3)	0%(0)	0%(0)	0%(0)
$x_6 = b$ (single stroke t)	28.2%(44)	66.6%(2)	54.5%(6)	40%(12)
$x_6 = c$ (looped t)	12.8%(20)	33.3%(1)	9.1%(1)	20%(6)
$x_6 = d$ (closed t)	35.2%(55)	0%(0)	18.2%(2)	20%(6)
$x_6 = e$(mixed shapes)	21.8%(34)	0%(0)	18.2%(2)	20%(6)

	Samples	2a	2b	2c	2d	2e	4a	4b	4c	4d	6a	6b	6c	6d	6e
Incidence of "t" shorter than "h" (1a)															
Number	156	36	57	4	27	32	115	12	2	27	3	44	20	55	34
Percentage	78	23.1	36.5	2.6	17.3	20.5	73.7	7.7	1.3	17.3	1.9	28.2	12.8	35.2	21.8
Incidence of "t" even with "h" (1b)															
Number	3	2	0	0	0	1	3	0	0	0	0	2	1	0	0
Percentage	1.5	66.6	0	0	0	33.3	100	0	0	0	0	66.6	33.3	0	0
Incidence of "t" taller than "h" (1c)															
Number	11	5	1	0	2	3	8	3	0	0	0	6	1	2	2
Percentage	5.5	45.5	9.1	0	18.2	27.3	72.7	27.3	0	0	0	54.5	9.1	18.2	18.2
Incidence of no set pattern (1d)															
Number	30	12	6	1	5	6	17	6	0	7	0	12	6	6	6
Percentage	15	40	20	3.3	16.7	20	56.7	20	0	23.3	0	40	20	20	20

Figure 5: Conditional probabilities given x_1 (height relationship of t and h) in verbose form. Distributions of x_2, x_4 and x_6 are shown. A sample reading of the table is, of the 11 people making t taller than h, 8 crossed t in the upper half of staff (72.7%) and 6 made a single stroke t (54.5%).

4.2 Distributions conditioned on x_2, shape of loop of *h*.

We have two conditional distributions available: (i) $p(x_3|x_2)$, where x_3 is the shape of the h arch, and (ii) $p(x_5|x_2)$, where x_5 is the baseline of h.

Table 4: Distributions conditioned on x_2, shape of loop of h.

$x_2 =$	a(retraced staff)	b(curved right)	c(curved left)	d(both curved)	e (no pattern)
Total	27.5%(55)	32%(64)	2.5%(5)	17%(34)	21%(42)
$x_3 = a$ (rounded arch)	12.7%(7)	26.6%(17)	20%(1)	17.6%(6)	11.9%(5)
$x_3 = b$ (pointed arch)	74.5%(41)	65.6%(42)	80%(4)	70.6%(24)	50%(21)
$x_3 = c$ (no pattern)	12.7%(7)	7.8%(5)	0%(0)	11.8%(4)	38.1%(16)
$x_5 = a$ (upward)	41.8%(23)	33.4%(22)	60%(3)	38.2%(13)	33.4%(14)
$x_5 = b$ (downward)	7.3%(4)	10.9%(7)	40%(2)	14.7%(5)	9.5%(4)
$x_5 = c$ (even)	10.9%(6)	12.5%(8)	0%(0)	11.8%(4)	7.1%(3)
$x_5 = d$ (no pattern)	40%(22)	42.2%(27)	0%(0)	35.3%(12)	50%(21)

	Samples	3a	3b	3c	5a	5b	5c	5d
Incidence of retraced "h" loop (2a)								
Number	55	7	41	7	23	4	6	22
Percentage	27.5	12.7	74.5	12.7	41.8	7.3	10.9	40
Incidence of curved right side, straight left (2b)								
Number	64	17	42	5	22	7	8	27
Percentage	32	26.6	65.6	7.8	34.4	10.9	12.5	42.2
Incidence of curved left side, straight right (2c)								
Number	5	1	4	0	3	2	0	0
Percentage	2.5	20	80	0	60	40	0	0
Incidence of both sides curved (2d)								
Number	34	6	24	4	13	5	4	12
Percentage	17	17.6	70.6	11.8	38.2	14.7	11.8	35.3
Incidence of no set pattern (2e)								
Number	42	5	21	16	14	4	3	21
Percentage	21	11.9	50	38.1	33.4	9.5	7.1	50

Figure 6: Conditional probabilities given x_2 (shape of loop of h). Distributions of x_3 and x_5. A sample reading of the table is, of the 34 people making both sides of th h loop curved, 24 made pointed arch of h (70.6%) and 13 made baseline of h slanting upwards (38.2%) .

4.3 Distributions conditioned on x_3, shape of arch of h

We have three conditional distributions available: (i) $p(x_2|x_3)$, where x_2 is the shape of the h loop, (ii) $p(x_5|x_3)$, where x_5 is the baseline of h, and (iii) $p(x_6|x_3)$, where x_6 is the shape of t.

Table 5: Distributions conditioned on x_3, shape of arch of h.

$x_3 =$	a (rounded arch)	b (pointed arch)	c (no pattern)
Total	18%(36)	66%(132)	16%(32)
$x_2 = a$ (retraced staff)	19.4%(7)	31.1%(41)	21.9%(7)
$x_2 = b$ (curved right)	47.2%(17)	31.8%(42)	15.6%(5)
$x_2 = c$ (curved left)	2.8%(1)	3.03%(4)	0%(0)
$x_2 = d$ (both curved)	16.7%(6)	18.2%(24)	12.5%(4)
$x_2 = e$ (no pattern)	13.9%(5)	15.9%(5)	50%(16)
$x_5 = a$ (upward)	36.1%(13)	39.4%(52)	31.3%(10)
$x_5 = b$ (downward)	8.3%(3)	11.4%(15)	12.5%(4)
$x_5 = c$ (even)	22.2%(8)	9.1%(12)	3.1%(1)
$x_5 = d$ (no pattern)	33.3%(12)	40.2%(53)	53.1%(17)
$x_6 = a$ (tented t)	0%(0)	2.3%(3)	0%(0)
$x_6 = b$ (single stroke t)	38.9%(14)	31.8%(42)	25%(8)
$x_6 = c$ (looped t)	8.3%(3)	15.2%(20)	15.6%(5)
$x_6 = d$ (closed t)	36.1%(13)	30.3%(40)	20.3%(10)
$x_6 = e$ (mixed t shapes)	16.7%(6)	20.4%(27)	28.1%(9)

	Samples	2a	2b	2c	2d	2e	5a	5b	5c	5d	6a	6b	6c	6d	6e
Incidence of rounded arch of "h" (3a)															
Number	36	7	17	1	6	5	13	3	8	12	0	14	3	13	6
Percentage	18	19.4	47.2	2.8	16.7	13.9	36.1	8.3	22.2	33.3	0	38.9	8.3	36.1	16.7
Incidence of pointed arch of "h" (3b)															
Number	132	41	42	4	24	21	52	15	12	53	3	42	20	40	
Percentage	66	31.1	31.8	3.03	18.2	15.9	39.4	11.4	9.1	40.2	2.3	31.8	15.2	30.3	0.5
Incidence of no set pattern (3c)															
Number	32	7	5	0	4	16	10	4	1	17	0	8	5	10	9
Percentage	16	21.9	15.6	0	12.5	50	31.3	12.5	3.1	53.1	0	25	15.6	[illegible]3	28.1

Figure 7: Conditional probabilities given x_3 (shape of arch of h). Distributions of x_2, x_5 and x_6. A sample reading of the table is, of the 36 people making a rounded arch of the h, 14 made a single stroke t and 13 made a closed t (36.1%).

4.4 Distributions conditioned on x_4, height of cross on t staff

We have three conditional distributions available: (i) $p(x_1|x_4)$, where x_1 is the relationship of t to h (ii) $p(x_2|x_4)$, where x_2 is the shape of the h loop, and (iii) $p(x_6|x_4)$, where x_6 is the shape of t.

Table 6: Distributions conditioned on x_4, height of cross on t staff.

$x_4 =$	a (upper staff)	b (lower staff)	c (above staff)	d (no pattern)
Total	71.5%(143)	10.5%(21)	1%(2)	17%(34)
$x_1 = a$ (t shorter than h)	80.4%(115)	57.1%(12)	100%(2)	79.4%(27)
$x_1 = b$ (t even with h)	2.1%(3)	0%(0)	0%(0)	0%(0)
$x_1 = c$ (t taller than h)	5.6%(8)	14.3%(3)	0%(0)	0%(0)
$x_1 = d$ (no set pattern)	11.9%(17)	28.6%(6)	0%(0)	20.6%(7)
$x_2 = a$ (retraced staff)	30.8%(44)	23.8%(5)	0%(0)	17.6%(6)
$x_2 = b$ (curved right)	32.2%(46)	28.6%(6)	100%(2)	32.3%(11)
$x_2 = c$ (curved left)	2.8%(4)	0%(0)	0%(0)	2.9%(1)
$x_2 = d$ (both curved)	15.4%(22)	19%(4)	0%(0)	23.5%(8)
$x_2 = e$ (no pattern)	19.6%(28)	28.6%(6)	0%(0)	23.5%(8)
$x_6 = a$ (tented t)	2.1%(3)	0%(0)	0%(0)	0%(0)
$x_6 = b$ (single stroke t)	28%(40)	57.1%(12)	0%(0)	35.3%(12)
$x_6 = c$ (looped t)	15.4%(22)	14.3%(3)	0%(0)	8.8%(3)
$x_6 = d$ (closed t)	32.9%(47)	19%(4)	50%(1)	32.3%(11)
$x_6 = e$ (mixed t)	21.7%(31)	9.5%(2)	50%(1)	23.5%(8)

	Samples	1a	1b	1c	1d	2a	2b	2c	2d	2e	6a	6b	6c	6d	6e
Incidence of cross in upper half (4a)															
Number	143	115	3	8	17	44	46	4	22	28	3	40	22	47	31
Percentage	71.5	80.4	2.1	5.6	11.9	30.8	32.2	2.8	15.4	19.6	2.1	28	15.4	32.9	21.7
Incidence of cross in lower half (4b)															
Number	21	12	0	3	6	5	6	0	4	6	0	12	3	4	2
Percentage	10.5	57.1	0	14.3	28.6	23.8	28.6	0	19	28.6	0	57.1	14.3	19	9.5
Incidence of cross above staff (4c)															
Number	2	2	0	0	0	0	2	0	0	0	0	0	0	1	1
Percentage	1	100	0	0	0	0	100	0	0	0	0	0	0	50	50
Incidence of no set pattern (4d)															
Number	34	27	0	0	7	6	11	1	8	8	0	12	3	11	8
Percentage	17	79.4	0	0	20.6	17.6	32.3	2.9	23.5	23.5	0	35.3	8.8	32.3	23.5

Figure 8: Conditional probabilities given x_4 (height on the cross of t shaft). Distributions of x_1, x_2 and x_6. A sample reading of the table is, of the 21 people crossing t in lower half, 12 made t shorter than h (57.1%), none made a tented t and 12 made a single stroke t (57.1%).

4.5 Distributions conditioned on x_5, baseline of *h*.

We have two conditional distributions available: (i) $p(x_2|x_5)$, where x_2 is the shape of h , and (ii) $p(x_3|x_5)$, where x_3 is the shape of the h loop.

TABLE 6—*Characteristic 5: baseline of the "h" for a total of 200 samples.*[a]

	Samples	2a	2b	2c	2d	2e	3a	3b	3c
Incidence of baseline slanting upwards (5a)									
Number	75	23	22	3	13	14	13	52	10
Percentage	37.5	30.7	29.3	4	17.3	18.7	17.3	69.3	13.3
Incidence of baseline slanting downwards (5b)									
Number	22	4	7	2	5	4	3	15	4
Percentage	11	18.2	31.8	9.1	22.7	18.2	13.6	68.2	18.2
Incidence of even baseline (5c)									
Number	21	6	8	0	4	3	8	12	1
Percentage	10.5	28.6	38.1	0	19	14.3	38.1	57.1	4.8
Incidence of no fixed pattern (5d)									
Number	82	22	27	0	12	21	12	53	17
Percentage	41	26.8	32.9	0	14.6	25.6	14.6	64.6	20.7

[a] Code numbers and letters are explained in the text. A sample reading of the table is this: of the 22 people making baseline slanting downwards, 15 made a pointed arch of the "h" (68.2%).

Figure 9: Conditional probabilities given x_5 (Baseline of h). Distributions of x_2 and x_3. A sample reading of the table is, of the 22 people making baseline slanting down-wards, 15 made a pointed arch of the h (68.2%)

4.6 Distributions conditioned on x_6, shape of *t*.

We have four conditional distributions available: (i) $p(x_1|x_6)$, where x_2 is the relationship of t to h, (ii) $p(x_2|x_6)$, where x_2 is the shape of the h loop, (iii) $p(x_3|x_6)$, where x_3 is the shape of the h loop, and (iv) $p(x_4|x_6)$, where x_4 is the shape of the h loop.

TABLE 7.—*Characteristic 6: shape of the "t" for a total of 200 samples.*"

	Samples	1a	1b	1c	1d	2a	2b	2c	2d	2e	3a	3b	3c	4a	4b	4c	4d
Incidence of tented "t" (6a)																	
Number	3	3	0	0	0	0	1	0	2	0	0	3	0	3	0	0	0
Percentage	1.5	100	0	0	0	0	33.3	0	66.6	0	0	100	0	100	0	0	0
Incidence of single-stroke "t" (6b)																	
Number	64	44	2	6	12	18	19	0	15	12	14	42	8	40	12	0	12
Percentage	32	68.7	3.1	9.4	18.7	28.1	29.6	0	23.4	18.7	21.8	65.6	12.5	62.5	18.7	0	18.7
Incidence of looped "t" (6c)																	
Number	28	20	1	1	6	6	11	0	4	7	3	20	5	22	3	0	3
Percentage	14	71.4	3.6	3.6	21.4	21.4	39.2	0	14.2	25	10.7	71.4	17.9	78.6	10.7	0	10.7
Incidence of closed "t" (6d)																	
Number	63	55	0	2	6	20	20	5	6	12	13	40	10	47	4	1	11
Percentage	31.5	87.3	0	3.2	9.5	31.7	31.7	7.9	9.5	19	20.6	63.5	15.9	74.6	6.3	1.6	17.5
Incidence of mixture (6e)																	
Number	42	34	0	2	6	11	13	0	7	11	6	27	9	31	2	1	8
Percentage	21	80.9	0	4.8	14.3	26.2	30.9	0	16.7	26.2	14.3	64.3	21.4	73.8	4.8	2.4	19

"Code numbers and letters are explained in the text. A sample reading of the table is this: of the 63 people making closed "t," 55 made "t" shorter than "h" (87.3%); 20 made retraced "h" loop (31.7%); 47 crossed "t" in upper half (74.6%); and 40 made pointed "h" arch (63.5%). Of the 3 people making tented "t," 3 made "t" shorter than "h" (100%).

Figure 10: Conditional probabilities given x_6 (Shape of t). Distributions of x_1, x_2, x_3 and x_4. A sample reading of the table is, of the 63 people making a closed t, 55 made t shorter than h (87.3%), 20 made a retraced h loop (31.7%), 47 crossed t in upper half (74.6%) and 40 made pointed h arch (63.5%). Of the 3 people making tented t, 3 made t shorter than h (100%).

5 PGMs

We would like to create a probabilistic graphical model for the given features so that we can evaluate the probability of any given *th*.

As an illustration, one possible probabilistic graphical model is the Bayesian network (not necessarily optimal) shown in Figure 11. The graph factorizes the joint distribution of the six variables into $p(x_1, x_2, x_3, x_4, x_5, x_6) = p(x_1)p(x_6|x_1)p(x_4|x_1)p(x_2|x_6)p(x_4|x_6)p(x_3|x_2)p(x_5|x_2)$. Since the CPDs for all of the factors are available, we can calculate the probability of any given assignment of values to the features.

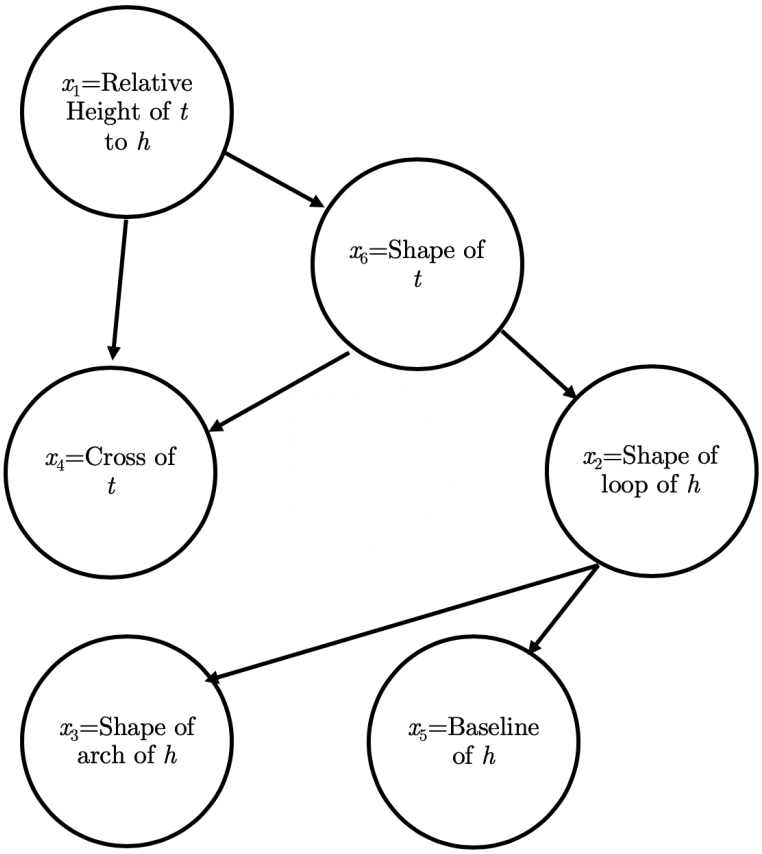


Figure 11: A possible Bayesian Network for *th*.

6 Tasks

1. Evaluate independences and conditional independences that exist in the data. Note that you can determine whether x_i and x_j are independent by testing if $p(x_i, x_j) \approx p(x_i)p(x_j)$. You can determine the joint probability as $p(x_i, x_j) = p(x_i|x_j)p(x_j)$.

2. Construct a Bayesian network with the fewest number of edges that maximizes the likelihood. You can construct several Bayesian networks and determine the likelihood it assigns to samples generated (using ancestral sampling). Describe some high probability *th* formations.

3. Convert the Bayesian network into a Markov network using moralization. Compare inferences using Bayesian network and the Markov network.

7 Deliverables

Write a report describing your methods. Divide it into three parts describing results obtained for each of the three tasks.

8 Data Files

The files provided to you are as follows:

1. th.rar - 3125 automatically extracted cases of 'th'

2. th-dataset (thdb-20110113-thonly) - 33 manually extracted cases of 'th' from the 'the' dataset