Probabilistic Graphical Models and Inference Methods for Handwriting

Danyang Chen
danyangc@buffalo.edu

Duc Thanh Anh Luong
ducthanh@buffalo.edu

Abstract

In this report, we present our approach for generating and learning probabilistic graphical models (PGMs) from a dataset of handwriting of word "*and*". We give a comparison between different PGMs of different generations involving in this dataset. We also use inference methods to determine the mean and entropy of each distribution and the relative entropy between distributions.

1 Introduction

In this report, we solve the problem of learning PGMs from the given data and perform inferences from the learned models. In particular, we constructed the Bayesian Networks and Markov Networks from the dataset of handwriting and performed inferences on Bayesian Networks. With Markov Networks, the task of computing the join probability is prohibitively expensive so that we left the task of doing inferences on Markov Networks(which involves computing join probability) as future works.

In section 2, we give a brief description about the dataset. We also give an explanation on dividing the dataset into different groups and draw a roadmap for further experiments in section 3 and 4. In section 3, we present our approach for constructing Bayesian Network, learning its parameters, and perform inferences on it. In section 4, we present our approach for constructing Markov Network and learning its parameters. In section 5, we give experimental results with Bayesian Network and Markov Network. Finally, in section 6, we conclude our works and discuss ideas for future works.

2 Dataset

We are given a dataset of children handwriting from a large number of students as they are learning (2nd grade) or have just learned (3rd and 4th grade) how to produce cursive and printed writing for 3 consecutive years 2012 - 2013 - 2014.

From the writing samples, the word "*and*" is extracted and features are generated from it. All features extracted from this process are discrete with maximum 5 distinct values. Figure 2 in project description gave a detailed explanation for each feature and its corresponding value. In short, there are totally 12 features for cursive handwriting and 12 features for handprint writing. In the dataset, there are some files with only 11 features which don't match the data description so we simply ignore those files.

In the dataset, there are missing and inconsistent values which are encoded with value -1 (missing value) and 99 (inconsistent). In our experiments when building the PGMs from dataset, we simply ignore the missing value whenever it appears and consider inconsistent value (99) as another value for that corresponding feature.

In order to compare the difference between handwriting for different generations, we group our dataset in the following:

- Cursive: for cursive handwriting, we construct 3 distinct groups of student which study in the same class in 2 consecutive years. The detail of each group is as follow:
  - Generation 1: consists of samples from students of 2011-2012 Grade 4 and 2012-2013 Grade 5 (471 samples)
  - Generation 2: consists of samples from students of 2011-2012 Grade 3 and 2012-2013 Grade 4 (559 samples)
  - Generation 3: consists of samples from students of 2012-2013 Grade 3 (176 samples)
- Handprint: for handprint handwriting, we construct 4 distinct groups of student which study in the same class in 3 consecutive years. The detail of each group is as follow:
  - Generation 1: consists of samples from students of 2011-2012 Grade 4 and 2012-2013 Grade 5 (503 samples)
  - Generation 2: consists of samples from students of 2011-2012 Grade 3, 2012-2013 Grade 4 and 2013-2014 Grade 5 (742 samples)
  - Generation 3: consists of samples from students of 2011-2012 Grade 2, 2012-2013 Grade 3 and 2013-2014 Grade 4 (1030 samples)
  - Generation 4: consists of samples from students of 2011-2012 Grade 1 (170 samples)

In section 3 and 4, we will construct PGMs for each generation mentioned above. We also construct PGMs for the whole dataset for both cursive and handprint data. With PGMs learned from the whole dataset, we can identify the common writing and rare writing in the dataset.

3 Bayesian Networks for handwriting dataset

3.1 Independence test

Since the Bayesian Network captures the set of independencies between variables, the first step that we need to do is to have an independence test, which we used Pearson's Chi-squared test [2]. The subroutine to perform this independence test can be found in the file *chi_square.m* of our code.

In brief, the Chi-squared test receive input including X_1, X_2 and X_3 where X_1 and X_2 are random variables while X_3 is a set of random variables. It checks whether $(X_1 \perp X_2|X_3)$ by using hypothesis testing. In particular, for each assignment x_3 of X_3, we perform the test $(X_1 \perp X_2|x_3)$. The formula for computing the chi-square value is as follow:

$$\chi^2 = \sum_{x_1} \sum_{x_2} \frac{[(O_{x_1,x_2}|x_3) - (E_{x_1,x_2}|x_3)]^2}{(E_{x_1,x_2}|x_3)} \tag{1}$$

where the first sum loops for all values x_1 of X_1, the second sum loops for all values x_2 of X_2, $(O_{x_1,x_2}|x_3)$ is the total number of occurrences of x_1, x_2, x_3 and $(E_{x_1,x_2}|x_3)$ is the expected value of number of occurrences of x_1, x_2, x_3 with an assumption that $X_1 \perp X_2|X_3$. In brief, the equation [1] computes the score of deviation from independence of our variables X_1, X_2 given values of X_3. If the score is too big, we then reject our hypothesis $(X_1 \perp X_2|x_3)$. More precisely, we convert the χ^2 score into probability and reject the hypothesis if the false rejection probability is more than 5%.

Chi-squared test accepts the hypothesis $(X_1 \perp X_2|X_3)$ when it doesn't reject any sub-hypotheses $(X_1 \perp X_2|x_3)$ for any value x_3.

3.2 Structure Learning

In this section, we explain our approach for constructing Bayesian Network from a given dataset. The approach that we chose follows closely the textbook [1, p. 78-92].

Before going to explain our algorithm, we define some terms that we will use in the subsequent sections.

Definition 3.1 *We say that a graph K is a perfect map (P-map) for a set of independencies I if we have that $I(K) = I$. We say that K is a perfect map for a distribution P if $I(K) = I(P)$.*

Definition 3.2 *If X and Y are not adjacent in G^* (an I-map of a distribution P) then we can find a set U such that $P \models (X \perp Y|U)$. We call this set U a witness of their independence.*

Definition 3.3 *A v-structure $X \rightarrow Z \leftarrow Y$ is an immorality if there is no direct edge between X and Y.*

Definition 3.4 *Let G be a DAG. A chain graph[1] K, is a class $PDAG$ of the equivalence class[2] of G if shares the same skeleton as G and contains a directed edge $X \rightarrow Y$ if and only if all G' that are I-equivalent to G contain the $X \rightarrow Y$.*

3.2.1 Building a skeleton for Bayesian Network

Assume that our distribution P in handwriting dataset has a perfect map. Let G^* be a perfect map for our distribution. The subroutine Build-PMap-Skeleton recovers the undirected skeleton for a distribution P that has a P-map. The pseudo-code for this subroutine can be written as follow:

Algorithm 1 Build-PMap-Skeleton

Input: $X = \{X_1, \cdots, X_n\}$ (set of random variables), P (distribution over X), d (bound on witness set)
Output: H (undirected graph over X), U (the set of witness)

```
Let H be a complete undirected graph over X
for X_i, X_j ∈ X do
    U_{X_i,X_j} = ∅
    for U ∈ Witness(X_i, X_j, H, d) do
        if P ⊨ (X_i ⊥ X_j|U) then
            U_{X_i,X_j} ← U
            Remove X_i − X_j from H
            break
        end if
    end for
end for
return H, {U_{X_i,X_j} : i, j ∈ {1, ··· , n}}
```

In the given pseudo-code of Algorithm 1, it sequentially detects an edge that is not in the skeleton and removes it by checking whether there exists a set U such that $(X_i \perp X_j|U)$.

One assumption that we make when running Build-PMap-Skeleton is that we limit the size of witness to d. Moreover, we can efficiently search for the set of witness U by enumerating it in increasing order of size.

3.2.2 Finding immoralities in skeleton

Having found the skeleton of P-map for distribution P, we now can add direction to some of the edges of the undirected graph returned in Build-PMap-Skeleton by checking the v-structure in the skeleton. Notice that, in a v-structure $X \rightarrow Z \leftarrow Y$, the information can flow from X to Y given Z. Therefore, by checking a path $X - Z - Y$ in the skeleton, if Z is not in witness set of X and Y, we can add the orientation for this v-structure as $X \rightarrow Z \leftarrow Y$. The formal description for above observation is given in the Algorithm 2

3.2.3 Build PDAG

Having constructed the skeleton of the perfect map G^*, and after adding the orientation for all immoralities that we have found in the skeleton, we can add some more orientation for this resulted graph by 3 rules given in Figure 1. A formal specification for this is given in Algorithm 3.

At this point, we have found the PDAG that captures all independencies of a distribution P. All remaining undirected edges can be oriented arbitrarily without losing any independencies in the

[1] a chain graph is graph which may have both directed and undirected edges but without any directed cycles
[2] 2 graphs that are in the same equivalent class if they capture the same set of independencies

Algorithm 2 Find-Immoralities

Input: $X = \{X_1, \cdots, X_n\}$ (set of random variables), S (Skeleton), $\{U_{X_i,X_j} : i, j \in \{1, \cdots, n\}\}$ (witness found by Build-PMap-Skeleton)
Output: K (Skeleton with edges involving in immoralities being directed)

$K \leftarrow S$
for X_i, X_j, X_k such that $X_i - X_j - X_k \in S$ and $X_i - X_k \notin S$ **do**
    //$X_i - X_j - X_k$ is a potential immorality
    if $X_j \notin U_{X_i,X_k}$ **then**
        Add orientation $X_i \rightarrow X_j$ and $X_j \leftarrow X_k$ to K
    end if
end for
return K

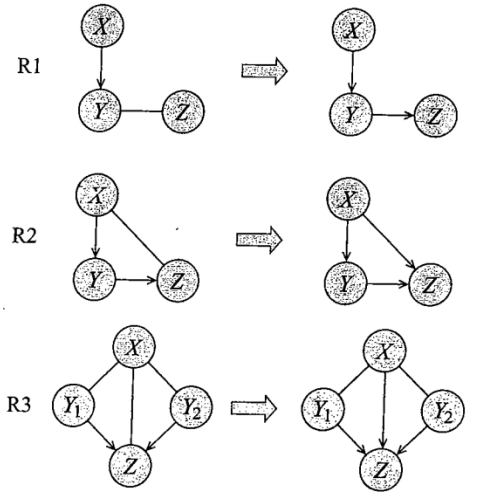


Figure 1: Rules for orienting edges in PDAG

Algorithm 3 Build-PDAG

Input: $X = \{X_1, \cdots, X_n\}$ (set of random variables), P (distribution of interest)
Output: K (PDAG of P)

$S, \{U_{X_i,X_j}\} \leftarrow$ Build-PMap-Skeleton(X, P)
$K \leftarrow$ Find-Immoralities$(X, S, \{U_{X_i,X_j}\})$
while not converged **do**
    Find a subgraph in K matching the left-hand side of a rule R1-R3
    Replace the subgraph with the right-hand side of the rule.
end while
return K

original distribution P. Theorem 3.10 in the textbook [1, p. 90] verifies the correctness of the algorithm.

3.3 Parameters Learning

The parameter set of Bayesian Network given its directed graph is the set of conditional probability distribution (CPD) which capture the local independencies of the distribution. In order to compute these CPDs, we follow the Maximum Likelihood Estimation (MLE) approach.

Let Gbe the Bayesian Network and θ be the set of parameters in G. For a dataset D with M samples $\{\xi[1], \cdots, \xi[M]\}$, the likelihood of parameter set θ given the dataset D is computed as follow:

$$
\begin{aligned}
L(\theta : D) &= \prod_{m=1}^{M} P_G(\xi[m] : \theta) && (2)\\
&= \prod_{m=1}^{M} \prod_{i} P(x_i[m]|pa_{X_i}[m] : \theta) && (3)\\
&= \prod_{i} \prod_{m=1}^{M} P(x_i[m]|pa_{X_i}[m] : \theta) && (4)\\
&= \prod_{i} L_i(\theta_{X_i|pa_{X_i}} : D) && (5)
\end{aligned}
$$

where $L_i(\theta_{X_i|pa_{X_i}} : D) = \prod_{m=1}^{M} P(x_i[m]|pa_{X_i}[m] : \theta)$ captures the local likelihood given the dataset.

At this point, our MLE formulation turns out to be the set of local MLE formulations which can be solved independently.

The local MLE is governed by a set of parameters $\{\theta_{x|u}$ for $x \in Val(X_i)$ and $u \in Val(pa_{X_i})\}$ of multinomial distribution. Formally, we can write the local likelihood as follow:

$$
\begin{aligned}
L_i(\theta_{X_i|pa_{X_i}} : D) &= \prod_{m=1}^{M} \theta_{x_i[m]|pa_{X_i}[m]} && (6)\\
&= \prod_{u \in Val(pa_{X_i})} \left[\prod_{x \in Val(X_i)} \theta_{x|u}^{M[u,x]} \right] && (7)
\end{aligned}
$$

where $M[u, x]$ is the occurrences of u and x in the dataset D.

By simple calculation, we can show that MLE parameters of multinomial distribution can be computed as follow:

$$
\hat{\theta_{x|u}} = \frac{M[u, x]}{M[u]} \tag{8}
$$

where $M[u]$ is the number of occurrences of u in the dataset D.

3.4 Sampling for Bayesian Network

3.4.1 Ancestral sampling

When sampling from Bayesian Network with ancestral sampling, we firstly order all random variables in topological order and sample each random variable in this order.

3.4.2 Gibbs sampling

When sampling Bayesian Network with Gibbs sampling, we first determine the order of random variables that we want to sample. The first sample can be obtained in some arbitrary initial distribution. The next sample will be obtained by sampling each random variable in the pre-determined order given the previous value of all other random variables. The formal description of Gibbs sampling can be given as follow:

Algorithm 4 Ancestral-Sample

Input: B (Bayesian Network over X - set of all random variables)
Output: $(x_1, \cdots, x_n)$
1: Let $X_1, \cdots, X_n$ be a topological ordering of X
2: **for** $i = 1, \cdots, n$ **do**
3: $u_i \leftarrow x(Pa_{X_i})$ //Assignment to Pa_{X_i} in $x_1, \cdots, x_{i-1}$
4: Sample x_i from $P(X_i|u_i)$
5: **end for**
6: **return** $(x_1, \cdots, x_n)$

Algorithm 5 Gibbs-Sample

Input: X (set of all random variables), Φ (set of factors defining P_Φ), $P^{(0)}(X)$ (initial state distribution), T (number of samples)
Output: $x^{(0)}, \cdots, x^{(T)}$
1: Sample $x^{(0)}$ from $P^{(0)}(X)$
2: **for** $t = 1, \cdots, T$ **do**
3: $x^{(t)} \leftarrow x^{(t-1)}$
4: **for** each $X_i \in X$ **do**
5: Sample $x_i^{(t)}$ from $P_\Phi(X_i|x_{-i})$ // Change X_i in $x^{(t)}$
6: **end for**
7: **end for**
8: **return** $x^{(0)}, \cdots, x^{(T)}$

3.5 Inference Methods

In inference for Bayesian Network, we mostly have to deal with 2 problems:

- Approximate the solution by using samples generated from the distribution of Bayesian Network, which we have explained 2 methods: ancestral sampling and Gibbs sampling.
- Compute the joint probability of a sample in Bayesian Network, which we will explain in the next section.

3.5.1 Compute the joint probability in Bayesian Network

Given a Bayesian Network B and a data sample x, the joint proability of x can be computed as follow:

$$P(x) = \prod P(x_i|Pa_{X_i} = Pa_{x_i}) \tag{9}$$

3.5.2 Mean of a distribution

The mean of a distribution in a Bayesian Network can be approximately computed with M samples $\{x_1, \cdots, x_M\}$ as follow:

$$\hat{E}[p(x)] = \frac{1}{M} \sum_{k=1}^{M} x_k \tag{10}$$

3.5.3 Entropy of a distribution

The entropy of a distribution in a Bayesian Network can be approximately computed with M samples $\{x_1, \cdots, x_M\}$ as follow:

$$\hat{H}[p(x)] = -\frac{1}{M} \sum_{k=1}^{M} \ln(p(x_k)) \tag{11}$$

3.5.4 Relative entropy between two distributions

The relative entropy (KL divergence) between 2 distributions can be approximately computed with M samples $\{x_1, \cdots, x_M\}$ as follow:

$$\hat{KL}[p||q] = -\frac{1}{M}\sum_{k=1}^{M}[\ln(q(x_k)) - \ln(p(x_k))] \tag{12}$$

4 Markov Networks for handwriting dataset

4.1 Structure Learning

For a distribution P to have a corresponding Markov Network, the distribution must be positive (no entry in the joint probability can be zero).

Assuming that our distribution P is positive, let H be the Markov Network corresponding to P. We also denote that $X = \{X_1, \cdots, X_n\}$ be the set of random variables in distribution P.

In order to construct Markov Network, there are 2 naive ways to do it:

- Pairwise independencies: if the edge $\{X_i, X_j\}$ is not in H, then X_i and X_j must be independent given all other nodes in the graph regardless of which other edges the graph contains. Thus, at the very least, to guarantee that H is an I-map, we must add direct edges between all pairs of nodes X_i and X_j such that

  $$P \not\models (X_i \perp X_j | X - \{X_i, X_j\}) \tag{13}$$

  We can now define H to include an edge $X_i - X_j$ for all X_i, X_j for which equation [13] holds.

- Markov blanket: in this approach, we use the local independencies and the notion of minimality. For each variable X, we define the neighbors of X to be a minimal set of Y that render X independent of the rest of the nodes. More precisely, we define:

  Definition 4.1 *A set U is a Markov blanket of Y in a distribution P if $Y \notin U$ and if U is a minimal set of nodes such that*

  $$(Y \perp X - \{Y\} - U | U) \in I(P) \tag{14}$$

  We then define a graph H by introducing an edge $\{Y, Z\}$ for all Y and all $Z \in MB_P(Y)$ where $MB_P(Y)$ is a set of Markov blanket of Y.

Although 2 methods above give an intuition how Markov Network can be constructed, they are not tractable because both involve the entire set of variables X and hence require measuring the probability of exponentially many events. Moreover, independencies that involve many variables lead to fragmentation of the data, and are much harder to evaluate without error. To estimate the distribution sufficiently well as to evaluate these independencies reliably, we would need exponentially many data points.

Suppose that in the Markov Network that we are looking for, the size of the set of Markov blanket for every node is bounded above by d. Given our assumption and perfect independence test, the Build-PMap-Skeleton procedure of algorithm 1 reconstructs the correct Markov structure H.

4.2 Parameters Learning

In order to find parameters for Markov Network, we need to have a finer representation for our Markov Network. In this section, we first explain the Log-linear model to represent Markov Network. Then, we define the set of features associate with our Markov Network. Finally, the weight of those features will be determined by gradient descent.

4.2.1 Log-linear models

First, we give a formal definition of log-linear model which we will use later on to compute the parameters of Markov Network.

Definition 4.2 *A distribution P is a log-linear model over a Markov network H if it is associated with:*

- *a set of features $F = \{f_1(D_1), \cdots, f_k(D_k)\}$ where each D_i is a complete subgraph in H*
- *a set of weights $\theta_1, \cdots, \theta_k$ such that*

$$P(X_1, \cdots, X_n) = \frac{1}{Z} \exp\left[-\sum_{i=1}^{k} \theta_i f_i(D_i)\right]$$

4.2.2 Choosing set of features

From above definition, given a Markov Network structure, for every clique in the Markov Network, there is at least a feature that includes all variables in that clique. A question is how many features for each clique and how to define them. One easy approach that we choose is binary indicator feature. For every clique, the number of features will be the number of possible assignments for all random variables in that clique and each feature can be computed by using indicator function.

For example, if we have a clique of 2 random variables A and B, where $Val(A) = \{a_0, a_1\}$ and $Val(B) = \{b_0, b_1\}$. So the number of features for this clique is 4 and each feature can be defined as follow:

$$f_{a_i,b_j}(A, B) = 1(A = a_i)1(B = b_j) \text{ for } i, j \in \{0, 1\}$$

4.2.3 Gradient descent

Finally, we need to find the set of weight $\theta = \{\theta_1, \cdots, \theta_k\}$ for our distribution associated with log-linear model.

Let D be a set of samples $\{\xi[1], \cdots, \xi[M]\}$. The likelihood of our parameter θ given the dataset D is:

$$L(\theta : D) = \prod_{m=1}^{M} \frac{1}{Z(\theta)} \exp\left[-\sum_{i=1}^{k} \theta_i f_i(\xi[m])\right] \tag{15}$$

The log-likelihood can be written as follow:

$$\log L(\theta : D) = \sum_i \theta_i \left(\sum_{m=1}^{M} f_i(\xi[m])\right) - M \ln Z(\theta) \tag{16}$$

In order to maximize the above log-likelihood, we can use the gradient descent to find the optimal parameter θ. In gradient descent, the derivative with respect to parameters can be shown as follow:

$$\frac{\partial}{\partial \theta_i} \frac{1}{M} \log L(\theta : D) = E_D[f_i] - E_\theta[f_i] \tag{17}$$

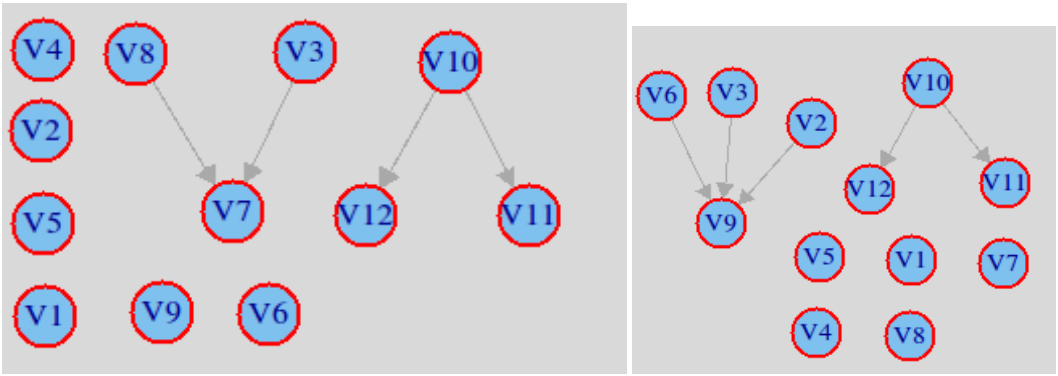
5 Experimental results

All the code that is used to perform experiments in this section can be found in the following link: `https://github.com/luongthanhanhduc/cse674`

5.1 Bayesian Networks

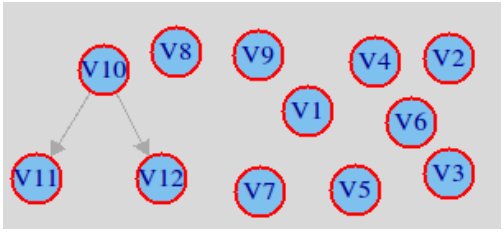
We have provided the theoretical background for Bayesian Network in section 3. We now move to the experiment with Bayesian Network. For Bayesian Network, we implemented the following functions:

- *BN_build_skeleton.m*: build the skeleton for Bayesian Network



(a) Generation 1

(b) Generation 2



(c) Generation 3

Figure 2: Bayesian Network for each generation in cursive dataset

- *BN_mark_immoralities.m*: mark all immoralities in the skeleton (adding direction for them)
- *BN_build_PDAG.m*: build PDAG of the distribution for Bayesian Network
- *BN_add_direction_PDAG.m*: add the direction arbitrarily for undirected edges in PDAG
- *BN_parameters.m*: compute parameters for Bayesian Network using MLE
- *BN_ancestral_sample.m*: ancestral sampling for Bayesian Network
- *BN_Gibbs.m*: Gibbs sampling for Bayesian Network
- *BN_compute_mean.m*: compute the approximate mean of the distribution given in Bayesian Network
- *BN_compute_entropy.m*: compute the approximate entropy of the distribution given in Bayesian Network
- *BN_compute_KL_divergence.m*: compute the approximate relative entropy between 2 distributions of 2 Bayesian Networks.
- *BN_common_writing.m*: identify the most common writing in the dataset
- *BN_rare_writing.m*: identify the rarest writing in the dataset

5.1.1 Structure learning

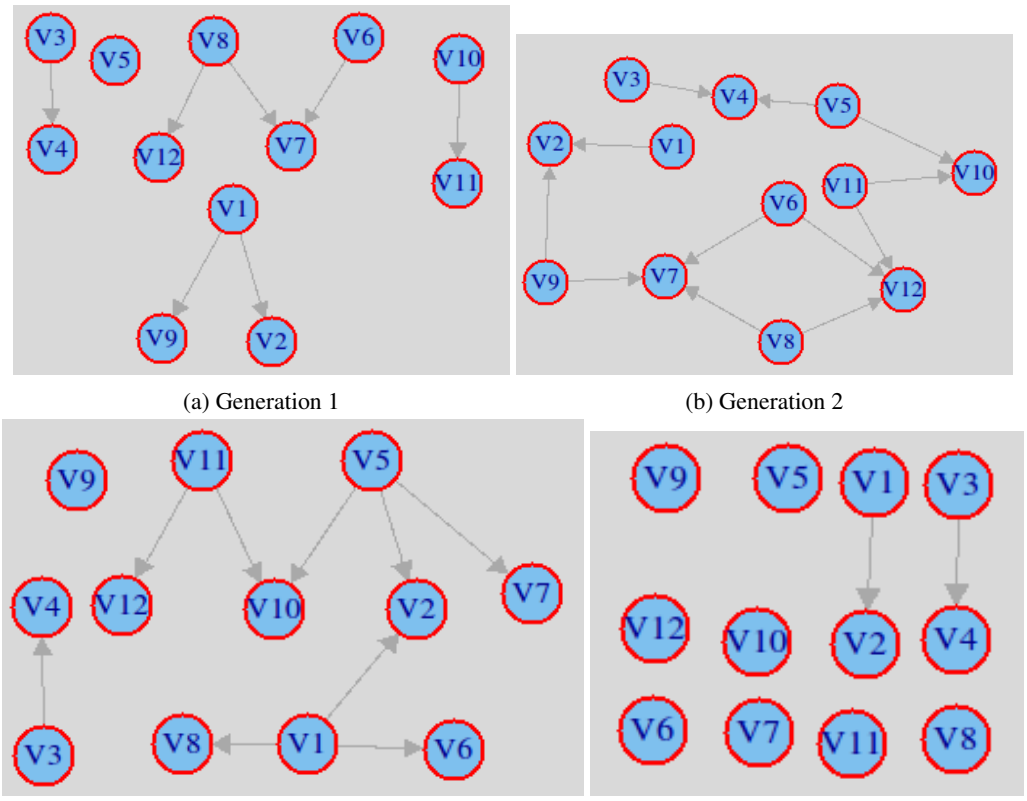
Figure 2 and 3 show the Bayesian Network for different generations in cursive and handprint dataset.

Figure 4 and 5 show the Bayesian Network for entire cursive and handprint dataset.

5.1.2 Mean of a distribution

For the mean of the distribution, we compute the approximate mean using formula given in section 3.5.2. We compute the mean from data set, ancestral sampling (10000 samples) and Gibbs sampling (10000 samples).

Table 1, 2 and 3 show the mean value for each random variables and give comparison between different mean values obtained from data set, ancestral sampling and Gibbs sampling in cursive dataset.



(a) Generation 1

(b) Generation 2

(c) Generation 3

(d) Generation 4

Figure 3: Bayesian Network for each generation in handprint dataset

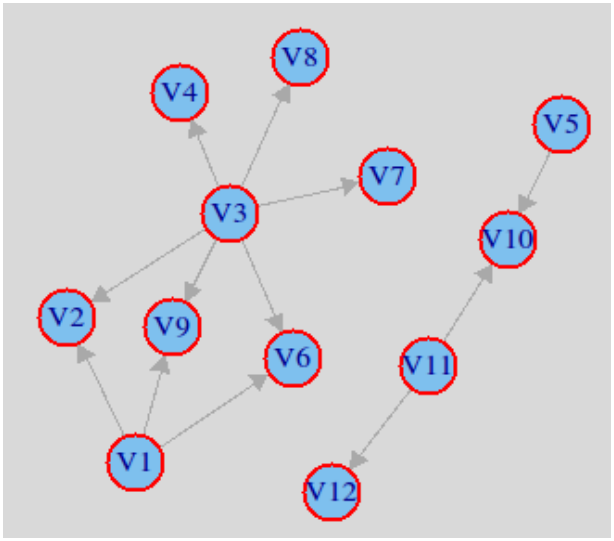


Figure 4: Bayesian Network for the entire cursive dataset

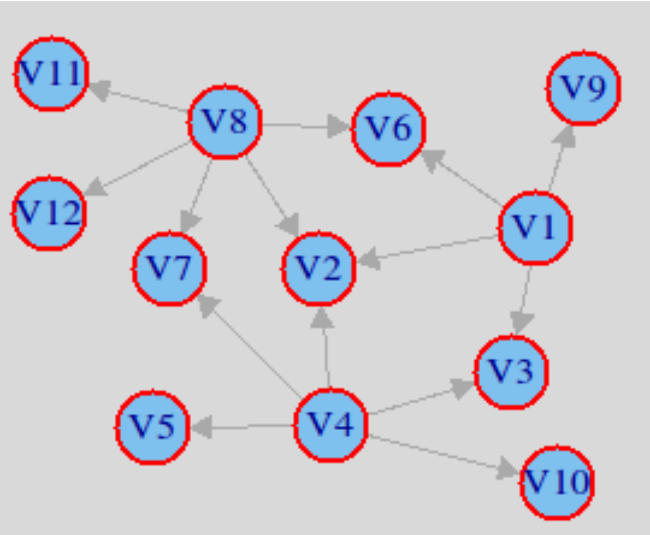


Figure 5: Bayesian Network for the entire handprint dataset

Table 5, 6, 7 and 8 show the mean value for each random variables and give comparison between different mean values obtained from data set, ancestral sampling and Gibbs sampling in handprint dataset.

Table 4 and 9 show the mean value for each random variables and give comparison between different mean values obtained from data set, ancestral sampling and Gibbs sampling in the entire dataset of cursive data and handprint one.

Table 1: Mean of distribution of Bayesian Network obtained from cursive dataset - generation 1

Data Samples	Ancestral Sampling	Gibbs Sampling
1.5615	1.5448	1.5749
1.047	1.0459	1.0482
0.81432	0.8081	0.8164
1.1946	1.1947	1.2015
0.86801	0.8737	0.8641
1.4013	1.4025	1.4081
0.45638	0.4496	0.4628
0.59955	0.5849	0.595
1.7808	1.7808	1.7878
0.96053	0.9433	0.9709
0.73465	0.7202	0.7326
0.85746	0.8649	0.8504

5.1.3 Entropy of a distribution

For the entropy of the distribution, we compute the approximate entropy using formula given in section 3.5.3. We compute the entropy of distribution from data set, ancestral sampling (10000 samples) and Gibbs sampling (10000 samples).

Table 10, 11 and 12 show the entropy value for each random variables and give comparison between different entropy values obtained from data set, ancestral sampling and Gibbs sampling in cursive dataset.

Table 14, 15, 16 and 17 show the entropy value for each random variables and give comparison between different entropy values obtained from data set, ancestral sampling and Gibbs sampling in handprint dataset.

Table 2: Mean of distribution of Bayesian Network obtained from cursive dataset - generation 2

Data Samples	Ancestral Sampling	Gibbs Sampling
1.3872	1.4026	1.3816
1.144	1.1483	1.1479
0.90078	0.899	0.9007
1.2665	1.2714	1.2618
0.79961	0.7948	0.793
1.4027	1.4032	1.3986
0.393	0.3868	0.3953
0.35603	0.3473	0.3606
1.7354	1.7426	1.738
0.74419	0.7328	0.7317
0.59213	0.6016	0.5934
0.91413	0.9028	0.9078

Table 3: Mean of distribution of Bayesian Network obtained from cursive dataset - generation 3

Data Samples	Ancestral Sampling	Gibbs Sampling
1.4602	1.4589	1.4695
0.92045	0.9218	0.9209
0.77841	0.774	0.7786
1.125	1.1284	1.1121
1.2898	1.2823	1.2779
1.2784	1.2655	1.2802
0.41477	0.4095	0.4145
0.63068	0.6391	0.6404
1.75	1.7562	1.7486
1.0114	1.0247	1.0095
0.85795	0.8598	0.8569
0.86932	0.886	0.8669

Table 4: Mean of distribution of Bayesian Network obtained from entire cursive dataset

Data Samples	Ancestral Sampling	Gibbs Sampling
1.467	1.475	1.4658
1.0712	1.0836	1.0946
0.84785	0.8493	0.8475
1.2164	1.2248	1.2135
0.90237	0.8988	0.9081
1.3829	1.4206	1.4232
0.42128	0.4187	0.4216
0.49428	0.494	0.4922
1.7555	1.7947	1.8031
0.8665	0.8771	0.8801
0.68598	0.6897	0.6802
0.88581	0.8907	0.8889

Table 5: Mean of distribution of Bayesian Network obtained from handprint dataset - generation 1

Data Samples	Ancestral Sampling	Gibbs Sampling
0.206	0.212	0.2087
1.1904	1.1931	1.1915
0.038	0.0338	0.0384
1.462	1.4332	1.4663
0.83166	0.8322	0.8269
0.17836	0.1786	0.1841
1.618	1.614	1.6254
0.44088	0.4423	0.4351
1.7725	1.7652	1.7699
0.70717	0.7118	0.7185
0.22754	0.2445	0.2302
0.56574	0.5708	0.5513

Table 6: Mean of distribution of Bayesian Network obtained from handprint dataset - generation 2

Data Samples	Ancestral Sampling	Gibbs Sampling
0.19315	0.2005	0.2029
1.0904	1.0768	1.1029
0.045205	0.0462	0.043
1.3877	1.3991	1.391
0.89589	0.8999	0.8981
0.19041	0.1913	0.1811
1.7671	1.7902	1.7804
0.42055	0.422	0.4292
1.8959	1.8982	1.8936
0.7965	0.8188	0.7926
0.21968	0.2283	0.2178
0.42645	0.4827	0.4681

Table 7: Mean of distribution of Bayesian Network obtained from handprint dataset - generation 3

Data Samples	Ancestral Sampling	Gibbs Sampling
0.22913	0.2281	0.2206
1.1854	1.1872	1.1903
0.066019	0.0634	0.07
1.6961	1.6778	1.6912
0.87379	0.8699	0.8724
0.25584	0.2563	0.2676
1.8922	1.8762	1.9014
0.44455	0.4446	0.4446
1.9252	1.9233	1.927
0.64757	0.6495	0.6593
0.32816	0.3217	0.3361
0.64078	0.626	0.642

Table 8: Mean of distribution of Bayesian Network obtained from handprint dataset - generation 4

Data Samples	Ancestral Sampling	Gibbs Sampling
0.29586	0.3035	0.2966
1.2647	1.2909	1.2528
0.15882	0.1577	0.1552
2.2765	2.2735	2.258
0.83432	0.8324	0.828
0.3787	0.37	0.3827
2.1479	2.1529	2.1474
0.3432	0.3464	0.3357
0.95266	0.9494	0.9555
0.75882	0.7538	0.7637
0.48824	0.5079	0.4856
0.85882	0.854	0.8415

Table 9: Mean of distribution of Bayesian Network obtained from entire handprint dataset

Data Samples	Ancestral Sampling	Gibbs Sampling
0.2182	0.2168	0.2307
1.1634	1.1836	1.1792
0.060494	0.0662	0.0697
1.5959	1.6	1.5859
0.86903	0.8678	0.8657
0.22877	0.2585	0.2563
1.816	1.8091	1.8449
0.42951	0.4255	0.429
1.8835	1.8888	1.8793
0.71277	0.7048	0.7249
0.28571	0.2798	0.2889
0.57552	0.5615	0.5749

Table 13 and 18 show the entropy value for each random variables and give comparison between different entropy values obtained from data set, ancestral sampling and Gibbs sampling in the entire dataset of cursive data and handprint one.

Table 10: Entropy of distribution of Bayesian Network obtained from cursive dataset - generation 1

Data Samples	Ancestral Sampling	Gibbs Sampling
9.6408	10.2484	9.6642

Table 11: Entropy of distribution of Bayesian Network obtained from cursive dataset - generation 2

Data Samples	Ancestral Sampling	Gibbs Sampling
9.3936	10.2776	9.3277

Table 12: Entropy of distribution of Bayesian Network obtained from cursive dataset - generation 3

Data Samples	Ancestral Sampling	Gibbs Sampling
9.7506	10.6981	9.7420

5.1.4 Relative entropy between 2 distributions

For the relative entropy between 2 distributions, we compute the approximate relative entropy using formula given in section 3.5.4. We compute the relative entropy between different distribution using data given from data set. In this experiment, we don't use the samples generated using Gibbs sampling and ancestral sampling because data generated from a particular distribution will bias the result of KL divergence. Table 19 gives the KL divergence score between different Bayesian Network in cursive data set. Table 20 gives the KL divergence score between different Bayesian Network in handprint dataset.

5.1.5 Common writing

The common writing in the dataset is defined as the most similar to the mean, which can be computed using the following formula for dissimilarity measure:

$$d(X,Y) = \frac{1}{D}\sum_{i=1}^{D}\frac{|X_i - Y_i|}{max(X_i)} \tag{18}$$

In our experiment, the most similar to the mean sample for cursive dataset is describe in Table 21 and the most similar to the mean sample for handprint dataset is describe in Table 22

5.1.6 Rare writing

The most unusual sample in the dataset is identified by computing the joint probability for each sample in the learned Bayesian Network and choose the sample with lowest probability. In our experiment, the most unusual sample for cursive dataset is describe in Table 23 and the most unusual sample for handprint dataset is describe in Table 24

5.2 Markov Network

5.2.1 Structure Learning and Parameter Learning

With the approach presented in earlier section, from the given dataset, we can construct the Markov Network and learn its parameters using gradient descent.

Figure 8 and 9 show the Markov Network for the entire cursive and handprint dataset.

Figure 6 and 7 show the Markov Network for each generation in each dataset (handprint and cursive).

As mentioned above, with appropriate assumption, the code for generating Markov Network is exactly the same with Bayesian Network, which has been implemented in the file *BN_build_skeleton.m*.

Table 13: Entropy of distribution of Bayesian Network obtained from entire cursive dataset

Data Samples	Ancestral Sampling	Gibbs Sampling
9.5120	10.9654	9.4824

Table 14: Entropy of distribution of Bayesian Network obtained from handprint dataset - generation 1

Data Samples	Ancestral Sampling	Gibbs Sampling
6.8677	7.2222	6.8752

Table 15: Entropy of distribution of Bayesian Network obtained from handprint dataset - generation 2

Data Samples	Ancestral Sampling	Gibbs Sampling
6.7094	7.6344	6.5811

Table 16: Entropy of distribution of Bayesian Network obtained from handprint dataset - generation 3

Data Samples	Ancestral Sampling	Gibbs Sampling
7.7421	9.0796	7.7469

Table 17: Entropy of distribution of Bayesian Network obtained from handprint dataset - generation 4

Data Samples	Ancestral Sampling	Gibbs Sampling
8.6482	8.9793	8.6248

Table 18: Entropy of distribution of Bayesian Network obtained from entire handprint dataset

Data Samples	Ancestral Sampling	Gibbs Sampling
7.6640	8.1616	7.6523

Table 19: Relative entropy between 2 distributions in cursive dataset

	Generation 1	Generation 2	Generation 3	Entire dataset
Generation 1	0	-0.4471	0.0551	-0.4381
Generation 2	0.4471	0	0.5022	0.0090
Generation 3	-0.0551	-0.5022	0	-0.4932
Entire dataset	0.4381	-0.0090	0.4932	0

Table 20: Relative entropy between 2 distributions in handprint dataset

	Generation 1	Generation 2	Generation 3	Generation 4	Entire dataset
Generation 1	0	-1.1353	-0.0079	0.1938	0.0606
Generation 2	1.1353	0	1.1273	1.3291	1.1959
Generation 3	0.0079	-1.1273	0	0.2018	0.0685
Generation 4	-0.1938	-1.3291	-0.2018	0	-0.1332
Entire dataset	-0.0606	-1.1959	-0.0685	0.1332	0

Table 21: The most similar to the mean sample of cursive dataset

Dissimilarity	0.0047
Intial stroke of “a”	staff left
Formation of “a” staff	retraced
Number of “n” arches	two
Shape of “n” arches	rounded
Location of “n” mid	at base
Formation of “d” staff	retraced
Formation of “d” initial	overhand
Formation of “d” terminal	curved up
Symbol	unusual
a-n relationship	a equal
a-d relationship	a equal
n-d relationship	n equal

Table 22: The most similar to the mean sample of handprint dataset

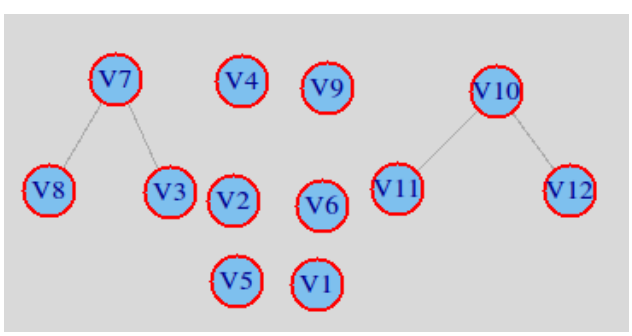
Dissimilarity	0.0043
# strokes in “a”	one
formation of “a” staff	retraced
# strokes in “n”	one
formation of “n” staff	retraced
shape of arch of “n”	rounded
# strokes in “d”	one
formation of “d” staff	retraced
initial stroke of “d”	bulb
unusual formation	and (2nd value)
a-n relationship	a equal
a-d relationship	a equal
n-d relationship	n equal

Table 23: The most unusual sample of cursive dataset

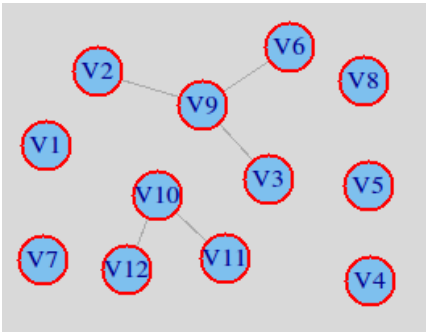
Probability	2.9604e-10
Intial stroke of “a”	staff left
Formation of “a” staff	retraced
Number of “n” arches	two
Shape of “n” arches	combination
Location of “n” mid	above base
Formation of “d” staff	tented
Formation of “d” initial	underhand
Formation of “d” terminal	no obvious end stroke
Symbol	unusual
a-n relationship	a equal
a-d relationship	a smaller
n-d relationship	n smaller

Table 24: The most unusual sample of handprint dataset

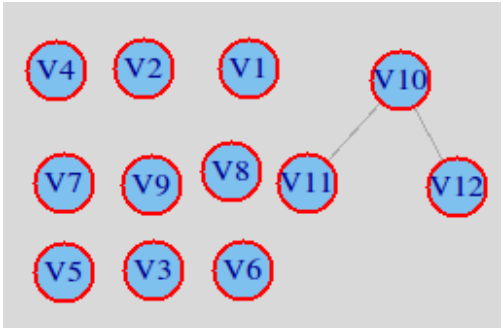
Probability	4.0677e-12
# strokes in “a”	one
formation of “a” staff	single down
# strokes in “n”	one
formation of “n” staff	retraced
shape of arch of “n”	inconsistent
# strokes in “d”	inconsistent
formation of “d” staff	no staff
initial stroke of “d”	staff top
unusual formation	formation
a-n relationship	a small
a-d relationship	inconsistent
n-d relationship	inconsistent



(a) Generation 1

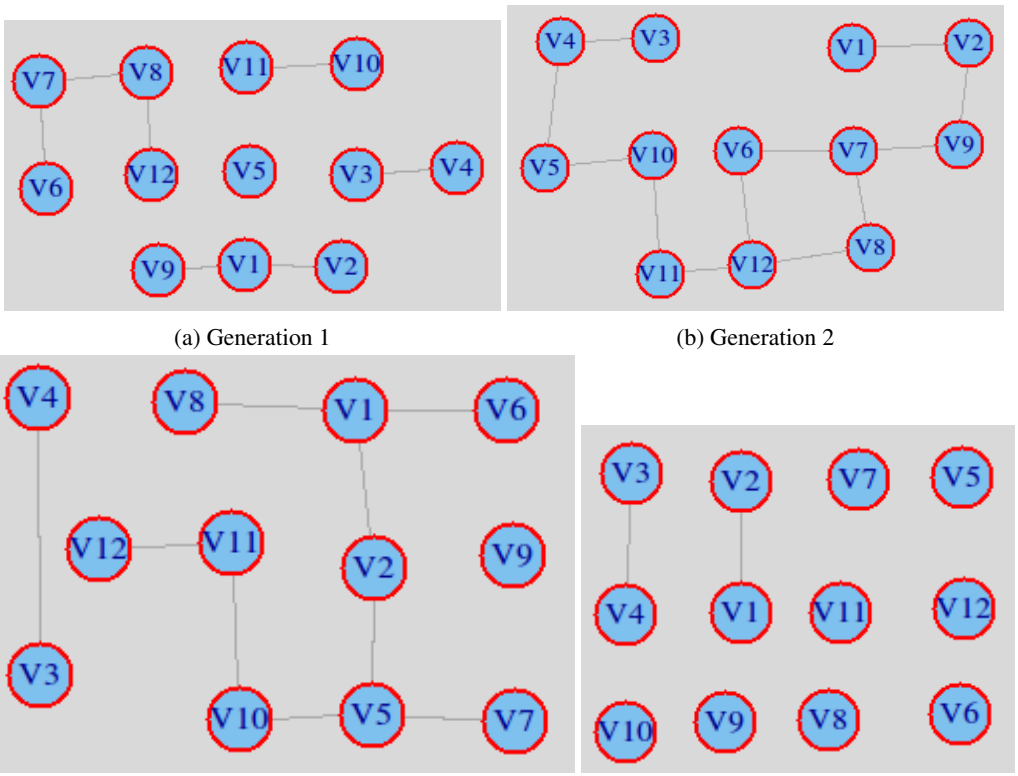


(b) Generation 2



(c) Generation 3

Figure 6: Markov Network for each generation in cursive dataset



(a) Generation 1

(b) Generation 2

(c) Generation 3

(d) Generation 4

Figure 7: Markov Network for each generation in handprint dataset

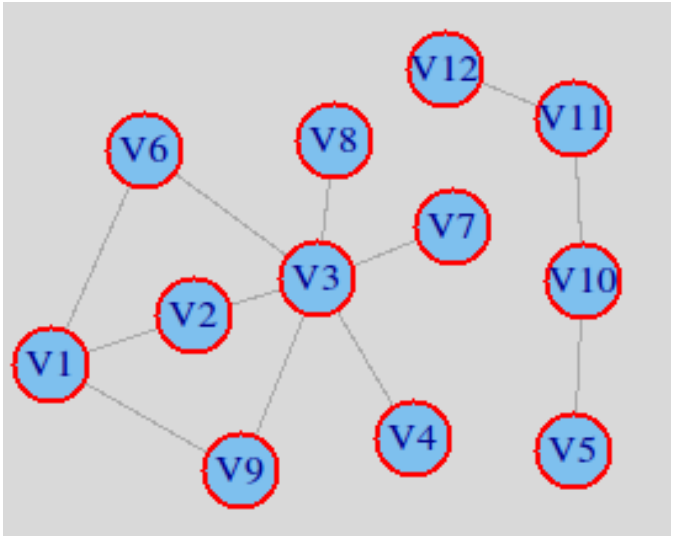


Figure 8: Markov Network for the entire cursive dataset

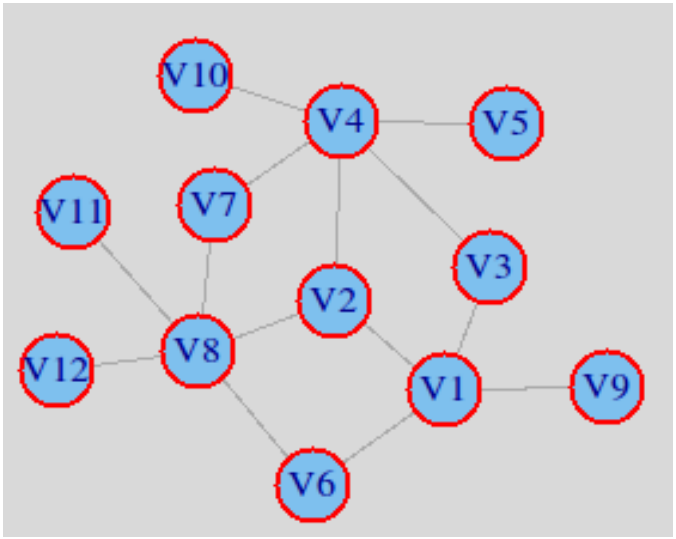


Figure 9: Markov Network for the entire handprint dataset

The code for implementing gradient descent to learn parameters for Markov Network can be found in files *Markov_learn_params.m*. Moreover, data can also sample from Markov Network using Gibbs sampling, and the detail of our implementation can be found in file *Markov_Gibbs.m*.

6 Conclusion

In this project, we have applied the Bayesian Network learning into the dataset of children handwriting. In particular, we constructed the structure of Bayesian Network from given dataset. We also use Maximum Likelihood approach to find parameters for our Bayesian Network. With the obtained Bayesian Network, we perform different inferences and answer many important questions involving children handwriting such as identifying rare writing as well as common writing.

References

[1] Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

[2] Pearson, K. (1900). *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50(302), 157-175.*