

The author(s) shown below used Federal funds provided by the U.S. Department of Justice and prepared the following final report:

Document Title: Statistical Examination of Handwriting Characteristics using Automated Tools

Author(s): Sargur N. Srihari

Document No.: 241743

Date Received: April 2013

Award Number: 2010-DN-BX-K037

This report has not been published by the U.S. Department of Justice. To provide better customer service, NCJRS has made this Federally-funded grant report available electronically.

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

*Statistical Examination of Handwriting Characteristics using
Automated Tools*

FINAL TECHNICAL REPORT: NIJ Award Number: 2010-DN-BX-K037

Sargur N. Srihari
University at Buffalo, The State University of New York
Buffalo, New York 14260
Email: srihari@buffalo.edu

SUBMITTED TO:

U.S. Department of Justice, Office of Justice Programs
National Institute of Justice
810 Seventh Street N.W.
Washington, DC 20531

AWARDEE:

Research Foundation of the State University of New York

February 11, 2013

Abstract

In the examination of handwritten items, the questioned document (QD) examiner follows a sequence of steps in many of which there is a degree of uncertainty to be resolved by experience and power of recall. Some examples of such decisions are: determining type/comparability, whether there is an adequate quantity of information and determining whether a set of characteristics is individualizing or representative of a class. Statistical models can play a significant role in assisting the QD examiner in dealing with uncertainty. In particular, the need for a statistical description of handwriting characteristics has long been felt. Efforts have been limited due to lack of efficient methods for collecting the data, computational problems of dealing with the very large number of combinatorial possibilities and the lack of clear direction for use of such results by the QD examiner. This research developed new statistical methods and software tools to: (i) extract samples of commonly encountered letter forms from the handwriting of typical writers, (ii) determine characteristics that would be used by QD examiners to describe common letter forms, (iii) have QD examiners enter perceived characteristics of the samples with a user interface, (iv) determine the frequency of occurrence of combinations of handwriting characteristics, (v) use those frequencies to construct a probabilistic model while handling the combinatorial possibilities and sample requirements, and (vi) use such models to infer the probability of characteristics to determine whether they are individualizing and in forming an opinion. Previously collected samples of extended handwriting, whose writers were representative of the United States population, were used to extract snippets of common letter combinations. From these scanned images the words *th* and *and* were extracted. The word snippets of each writer were presented to QD examiners who entered values for several characteristics using an interactive tool developed for the purpose; the characteristics depended on writing type: cursive or hand-printed. From this data the frequencies of the characteristics and their combinations were evaluated. Since the combinations of characteristics is very large, exact statistical models are infeasible. Instead, probabilistic graphical models are used to model the joint distribution. Both directed and undirected graphical models were learnt from data using algorithms that use independence tests between pairs of variables and a global measure of the goodness. Methods for inferring useful probabilities from the models were developed, e.g., rarity as a measure of individualizing characteristics, and the probability of random correspondence of the observed sample among n writers. Using these methods, the probabilities of nearly 1,500 writing styles of *and* were determined and tabulated. An indication of how the developed techniques can be incorporated into the work-flow of the QD examiner is given.

Contents

| | | |
|----------|---|----------|
| 1 | Executive Summary | 2 |
| 2 | Research narrative | 8 |
| 2.1 | Introduction | 8 |
| 2.1.1 | Current practice | 9 |
| 2.1.2 | Statement of the problem | 12 |
| 2.1.3 | Literature review | 13 |
| 2.1.4 | Rationale for the research | 16 |
| 2.2 | Methods: Data Preparation | 16 |
| 2.2.1 | Handwriting Samples | 17 |
| 2.2.2 | Letter combinations | 17 |
| 2.2.3 | Characteristics | 19 |
| 2.2.4 | Extraction of Snippets | 19 |
| 2.2.5 | User Interface | 19 |
| 2.2.6 | Ground-truthing | 21 |
| 2.3 | Methods: Statistical Model Construction | 24 |
| 2.3.1 | Problem Complexity | 24 |
| 2.3.2 | Bayesian Networks | 25 |
| 2.3.3 | Markov Networks | 34 |
| 2.4 | Methods: Statistical Inference | 41 |
| 2.4.1 | Probability of Evidence | 41 |
| 2.4.2 | Probability of Identification | 45 |
| 2.4.3 | Opinion Scale | 48 |
| 2.5 | Methods: QD Work-flow | 54 |
| 2.5.1 | Standard Work-Flow | 54 |
| 2.5.2 | Use of Probabilistic Methods | 54 |
| 2.6 | Conclusions | 54 |
| 2.7 | Implications for policy and practice | 56 |
| 2.8 | Implications for further research | 56 |
| 2.9 | Dissemination | 56 |
| 2.9.1 | Publications | 56 |
| 2.9.2 | Presentations | 57 |
| 2.9.3 | Students | 57 |
| 2.10 | Acknowledgement | 58 |
| 2.11 | Appendix 1: Handwriting sample source | 63 |
| 2.11.1 | Source Document | 64 |
| 2.11.2 | Writer Population | 66 |
| 2.12 | Appendix 2: Tool for extracting image snippets | 68 |
| 2.13 | Appendix 3: Comparison of <i>th</i> marginals with previous work. | 69 |
| 2.13.1 | Chi-squared Test Short Description | 69 |
| 2.13.2 | Results of χ^2 Tests | 69 |

| | | |
|--------|---|----|
| 2.13.3 | Discussion of Results | 70 |
| 2.14 | Appendix 4: <i>and</i> examples | 71 |
| 2.14.1 | Cursive | 72 |
| 2.14.2 | Hand-print | 74 |
| 2.15 | Appendix 5: Type Determination | 76 |
| 2.15.1 | Characteristics of Type | 76 |
| 2.15.2 | Dataset | 78 |
| 2.15.3 | Type Distribution | 78 |
| 2.15.4 | Results | 78 |
| 2.15.5 | Conclusions | 79 |
| 2.16 | Appendix 6: Mapping Likelihood Ratio to Opinion Scale | 80 |

Chapter 1

Executive Summary

The most common task in Questioned Document (QD) examination deals with handwritten items. The task involves performing several decisions such as determining the type and comparability of the known and questioned items, determining whether there is an adequate amount of information and whether the observed characteristics are representative of a class or are individualizing. These decisions are usually based entirely on the examiner's experience and power of recall. The availability of statistical models in each of these steps offers the promise for providing a scientific basis for the choices made and the opinions expressed.

The project demonstrated the construction of probabilistic models for several of the steps in handwriting examination, with a particular focus on handwriting characteristics. Frequencies of characteristics, determined from samples collected over a representative population, are useful to determine a probability distribution from which several useful inferences can be made, e.g., whether observed characteristics are individualizing and in quantifying opinion.

The goal was to: (i) develop methods to extract samples of commonly encountered letter forms from extended handwriting samples of typical writers in the United States, (ii) prepare the appropriate format to present the samples to QD examiners who would then enter perceived characteristics with a user interface, (iii) determine the frequency of occurrence of combinations of handwriting characteristics, (iv) use those frequencies to construct a probabilistic model without the method being overwhelmed by the combinatorial possibilities and sample requirements, (v) develop methods to infer the probability of evidence from the model, and (vi) indicate where such methods can be used in the QD examiner's work-flow for examining handwritten items. The project tasks are divided into four parts: *data preparation*, *model construction*, *inference* and *QD work-flow*.

The first part was to prepare the data set of handwriting characteristics in a form suitable for statistical analysis. Previously created collections of handwriting samples that are representative of the United States population were used. The most common letter combinations in the English language were determined. Since the handwriting samples are of extended writing, the letter combinations of interest were isolated in them and the corresponding image snippets were extracted. QD examiners determined the characteristics appropriate for the letter combinations, with the characteristics being dependent on whether the writing was cursive or of hand-print. The samples of each writer were presented in an interactive manner to QD examiners who entered the characteristics that they observed using a pull-down menu. In particular, characteristics for the word *and* were determined considering the cursive and hand-print cases.

The second part began with the construction of probability models from the data. Since the frequencies of all combinations of characteristics cannot be exhaustively determined, probabilistic graphical models (PGMs) were used to model the joint distributions. The appropriate model is learnt from the data so that the joint distribution of the characteristics is captured by several conditional frequencies of the characteristics. New algorithms for constructing directed PGMs (Bayesian networks) and undirected PGMs (Markov networks) were proposed and their use demonstrated with handwriting data.

The third part consisted of methods of inference using the models. Methods were developed to determine: the probability of given evidence, probability of random correspondence (PRC), conditional PRC associated with a given sample and the probability of finding a similar one within tolerance in a database of given size.

The probabilities of nearly 1,500 writing styles of *and* were determined and tabulated in decreasing order of probability. Also, a method for evaluating the probability of identification, when evidence is compared with a known, is described. It involves taking into account two factors, one based on *distance* (or *similarity*) and the other on *rarity* (which is the reciprocal of probability).

The final part of the project is to indicate as to how QD examiners can incorporate the results of such analysis into their work-flow. They are used in choosing individualizing characteristics to compare between two handwritten items and in making a statement in the testimony as to how likely it is that a randomly selected person would have the characteristics observed in the evidence.

The project benefitted from the advise of several QD examiners, particularly Kirsten Singer of the Department of Veteran's Administration, Traci Moran of the Financial Management Service and Lisa Hanson of the Minnesota Criminal Apprehension Laboratory. The views expressed here are of the author alone and do not reflect the opinions of the QD examiners nor of the Department of Justice.

List of Figures

| | | |
|------|---|----|
| 2.1 | Samples of a handwritten word from eight different writers showing between-writer and within-writer variability. The word “referred” occurs in the third paragraph of the CEDAR letter described in Appendix 1. | 9 |
| 2.2 | Work Flow of Forensic Handwriting Examination. In steps 7 and 10 probabilistic analysis is useful. . | 11 |
| 2.3 | Samples of handwritten <i>th</i> of two writers showing two different writing styles as well as within-writer variability. | 17 |
| 2.4 | Samples of cursively written <i>and</i> of a single writer. Using the characteristics listed in Table 2.3 (a) this writing is encoded as 101122012 | 18 |
| 2.5 | Samples of hand-printed <i>and</i> of a single writer. Using the characteristics listed in Table 2.3 (b) this writing is encoded as 010110112 | 18 |
| 2.6 | GUI for determining the features for <i>th</i> exemplars of a given writer: values are assigned manually using pull-down menus for each feature. | 21 |
| 2.7 | GUI for ground-truthing of <i>and</i> : written cursively. | 22 |
| 2.8 | GUI for ground-truthing of <i>and</i> hand-printed. | 23 |
| 2.9 | Bayesian network of <i>th</i> : (a) manually constructed directed graph, (b) marginal distributions, (c)–(g) conditional probability distributions. | 26 |
| 2.10 | Bayesian Network (BN) structure learning, where edge (x_4, x_5) , with the next highest χ^2 value is being considered: (a) BN G^* before considering edge (x_4, x_5) ; (b) candidate BN G_{e_1} with edge $(x_4 \rightarrow x_5)$; (c) candidate BN G_{e_2} with edge $(x_5 \rightarrow x_4)$ | 27 |
| 2.11 | Heuristics of determining Structure in Algorithm BNSL | 28 |
| 2.12 | Evaluation of BN structures learnt from <i>th</i> data: (a) branch and bound algorithm, (b) human designed BN based on causality; (c) algorithm BNSL, and (d) performance metrics. | 30 |
| 2.13 | Bayesian networks for <i>and</i> data: (a) $BN_{cursive-and}$, (b) $BN_{handprint-and}$, (c) table of marginal probabilities for cursive, and (d) table of marginal probabilities for handprint. The necessary CPTs for (a) are given in Table 2.5 and for (b) in Table 2.6. | 31 |
| 2.14 | Dividing the number line into nine intervals for sampling. | 33 |
| 2.15 | Gain G_Q in the objective function, see expr. 2.13, for a given feature f with empirical probability E_S and expected value E_Q w.r.t. the probability distribution defined by MRF Q | 38 |
| 2.16 | Accuracy and construction time comparison of the greedy algorithm with L_1 -regularization on weights [47] (green dot line) and the designed fast greedy algorithm (blue solid line) for data sets with different number of variables. | 40 |
| 2.17 | Candidate MRFs: first two were manually constructed, the third – by the modified Chow-Liu algorithm, fourth MRF – by the original greedy algorithm and the fifth – by the proposed FGAM; and (f) the highest probability ‘ <i>th</i> ’ in data set and (g) a low probability ‘ <i>th</i> ’. | 41 |
| 2.18 | Examples of rarity evaluation: (a) the highest probability <i>th</i> in data set, and (b) a low probability <i>th</i> | 42 |
| 2.19 | Graphical models for determining random correspondence: (a) PRC, the probability of two samples having the same value within ϵ , where the <i>indicator</i> variable Z : $P(Z X_1, X_2)$ has the distribution shown in (b), (c) the probability of some pair of samples among n having the same value, n PRC, and (d) conditional n PRC, the probability of finding X_s among n samples and. | 43 |
| 2.20 | Probability of finding a random match for <i>and</i> among n writers for cursive and handprint. These plots of n PRC show the discriminative power of the characteristics, with cursive writing of <i>and</i> being more individualistic than hand-printing. | 44 |

| | | |
|------|--|----|
| 2.21 | Probability of finding a matching entry for th in a database of size n , called conditional n PRC: (a) exact match, and (b) match with one feature mismatch allowed. The two graphs in each figure correspond to X_s being the most common th and a rare th whose forms are shown in Figure 2.18. | 45 |
| 2.22 | Probability of identification is a sigmoid function of the log-likelihood ratio. | 46 |
| 2.23 | Distributions of sources for object and evidence: (a) Normally distributed sources, where each source s_i ($i = 1, 2, \dots$) is $\mathcal{N}(\theta_i, \sigma^2)$. Samples may come from a common source s_i or different sources s_i and s_j ($s_i \neq s_j$). (b) Source means $\{\theta_1, \theta_2, \dots\}$, are assumed to have distribution $\mathcal{N}(\mu, \tau^2)$, with $\tau \gg \sigma$. Samples coming from sources θ_1, θ_{10} are rarer (less frequent) than samples from θ_6 and θ_7 , suggesting that information about the distribution of source means is useful to assess strength of evidence. | 47 |
| 2.24 | Comparison of five methods of computing the likelihood ratio. For each method the average error rates are on the left (blue) and time per sample on the right (red). | 48 |
| 2.25 | Confidence intervals. Scores on normal distribution $\mathcal{N}(50, 25)$ (a-b) and uniform distribution $\mathcal{U}[0, 100]$ (c-d): i) Green solid lines – 95% confidence intervals for the log-likelihood ratio mean; ii) Blue circles – log-likelihood ratio; iii) Red dots – discounted weighted log-likelihood ratio score with trivial weights (eq. 2.31) and discount function d_1 (eq. 2.33) and threshold $m = 100$ | 51 |
| 2.26 | Handwriting Source: (a) document copied by writers includes all alphabets, and (b) a digitally scanned handwritten sample provided by a writer. | 65 |
| 2.27 | Transcript mapping function of CEDAR-FOX for extracting letter combinations from handwritten pages: (a) window or an input text file, and (b) truth super-imposed on handwriting image. | 68 |
| 2.28 | Examples of word type: (a) cursive: $f_1 = 0, f_2 = 2.5$, (b) hand-printed: $f_1 = 5, f_2 = 0.5$, and (c) predominantly cursive: $f_1 = 3, f_2 = 1.33$ | 76 |
| 2.29 | Determination of cursive (a) and hand print (b) within the CEDAR-FOX system. Screenshots in (c) and (d) show result on a continuous scale as predominantly cursive and predominantly hand-printed. | 77 |
| 2.30 | Histogram of binned feature f_1 (left) and representative Gamma distributions (right) with their thresholds. | 78 |

List of Tables

| | | |
|------|---|----|
| 2.1 | Most frequently occurring letter pairs (bigrams) with expected occurrences per 2000 letters. The bigrams are not permitted to span across consecutive words. | 17 |
| 2.2 | Characteristics of the <i>th</i> combination: six variables and their values [51]. | 19 |
| 2.3 | Characteristics of <i>and</i> as Specified by Document Examiners. | 20 |
| 2.4 | Evaluation of BN Models (log loss) for <i>and</i> data. | 30 |
| 2.5 | Conditional Probability Tables of characteristics of <i>and</i> for <i>cursive writing</i> needed in the BN shown in Figure 2.13 (a). | 32 |
| 2.6 | Conditional Probability Tables of characteristics of <i>and</i> for <i>handprint writing</i> needed in the BN shown in Figure 2.13 (b). | 33 |
| 2.7 | Examples of samples with highest and lowest joint probabilities: (a) Cursive- Highest (b) Cursive- Lowest (c) Handprint- Highest (d) Handprint- Lowest. The samples were obtained by Gibbs sampling of BNs. | 34 |
| 2.8 | Results of MRF structure learning with 3-fold cross-validation. | 41 |
| 2.9 | Probability of finding an identical match for a given <i>and</i> among 7000 writers. | 45 |
| 2.10 | CEDAR Letter Data Attributes: (a) positional frequency of letters in text, and (b) demographics of writers. | 64 |
| 2.11 | Comparison of two marginal distributions | 69 |
| 2.12 | Rules for obtaining an opinion on the 9 point scale | 81 |

List of Algorithms

| | | |
|---|---|----|
| 1 | BNSL: Bayesian Network Structure Learning | 28 |
| 2 | Gibbs Sampling | 30 |
| 3 | FGAM: A fast greedy algorithm for MRF structure learning | 37 |
| 4 | Constructing a set of features with high empirical probabilities | 38 |
| 5 | Constructing a set of features $\{f\}$ with high expected values $E_Q[f]$ with respect to the probability distribution defined by MRF Q | 39 |
| 6 | Comparison of handwritten items with statistical tools | 54 |

Chapter 2

Research narrative

2.1 Introduction

In the forensic sciences evidence is usually quantified by means of several characteristics. Frequencies of such characteristics from samples collected over a representative population are useful both in the investigative and prosecution phases of a case. Knowing the frequencies of characteristics allows the calculation of the rarity of a piece of evidence and thereby quantify its probative value. In the field of DNA identification, the evidence is characterized by a set of alleles. Knowing allele frequencies allows one to make a statement such as “the chance that a randomly selected person would have the same DNA pattern as that of the sample source and the suspect is 1 in 24,000,000”.

The importance of quantitative methods to characterize confidences in pattern-based forensic identification has been underscored by various court rulings and the recent report of the National Academy of Sciences [54]. Among the forensic sciences dealing with visual patterns is Questioned Document (QD) examination. An important component of QD examination is the analysis of handwritten items. The goal of such analysis is typically to compare a QD with known writing. The comparisons typically use several characteristics of handwriting. The aim of this project was to develop techniques to provide a statistical basis for the characteristics that document examiners use.

Since the examination of handwritten items involves human determination of characteristics, QD examiners have to create data sets of characteristics which can then be analyzed statistically. The resulting probabilities will be useful in identifying rare formations (known as individualizing characteristics) and in providing an opinion regarding the correspondence between evidence and known. This research is a first step towards developing the tools and methods necessary for the ambitious goal of creating a statistical foundation for the handwriting facet of QD examination.

The remainder of this report is organized as follows. Section 2.1.1 describes the terminology and procedure for examining handwritten items as described by SWGDOC (Scientific Working Group for Forensic Document Examination) and published as ASTM standards. Section 2.1.2 defines the problem statement for this project. Some relevant literature is described in Section 2.1.3 although most references are given in the the narrative. The project rationale is summarized in Section 2.1.4.

The discussion of research methods is divided into four parts: data preparation (Section 2.2), statistical model construction (Section 2.3), statistical inference (Section 2.4) and incorporating the methods into the QD examiner’s work-flow (Section 2.5.1).

There are six appendices as follows. The source of handwriting data is described in Appendix 1. A software tool for extracting snippets of data from extended handwriting is in Appendix 2. A comparison of marginal probabilities of *th* with earlier work is in Appendix 3. Samples of *and* used in constructing statistical models is in Appendix 4. A statistical method for determining handwriting type is in Appendix 5. A method for mapping a likelihood ratio into an opinion scale is described in Appendix 6.

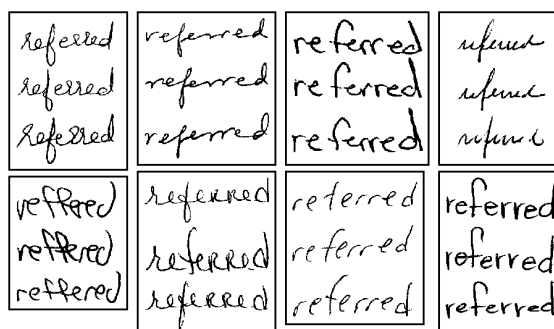


Figure 2.1: Samples of a handwritten word from eight different writers showing between-writer and within-writer variability. The word “referred” occurs in the third paragraph of the CEDAR letter described in Appendix 1.

2.1.1 Current practice

QD examination involves the comparison and analysis of documents, printing and writing instruments in order to identify or eliminate persons as the source. Handwriting comparison is based on the premise that no two persons write the same way, while considering the fact that the writing of each person has its own variabilities[58, 39]. Individuals write differently both because they were taught differently, e.g, Palmer and D’Nelian methods which are called class-characteristics, and due to individual habits known as individualizing characteristics. Examples of such variations are seen in Figure 2.1.

QD examiners specify handwriting characteristics (features) based on years of training [39]. QD examiner’s judgements are often based entirely on the examiner’s experience and power of recall. Statistical data concerning frequency of occurrence of forms and combinations would offer promise for providing a scientific basis for their opinions. QD examiners tend to assign probative values to specific handwriting characteristics and their combinations. These characteristics are termed the seven S’s, viz., *size, slant, spacing, shading, system, speed, strokes* [32].

The terminology and procedure for the examination of handwritten items by QD examiners are given in the ASTM documents *Standard Guide for Examination of Handwritten Items* [7] and *Standard Terminology for Expressing Conclusion of Forensic Document Examiners* [8] which we summarize below. This terminology is also used in the discussion below.

A. Questioned Document (QD) terminology

- *absent character*: present in one and not in the other
- *character*: language symbol: letter, numeral, punctuation
- *characteristic*: a feature, quality, attribute or property
- *class characteristics*: properties common to a group
- *comparable*: same types, also contemporaneous, instruments
- *distorted*: unnatural: disguise, simulation, involuntary
- *handwritten item*; cursive, hand-print or signatures
- *individualizing characteristics*: unique to individual
- *item*: object or material on which observations are made
- *known (K)*: of established origin in matter investigated
- *natural writing*: without attempt to control/alter execution

- *questioned (Q)*: source of question, e.g., common with *K*
- *range of variation*: deviations within a writer's repetitions
- *significant difference*: individualizing charac. outside range
- *significant similarity*: common individualizing characteristic
- *sufficient quantity*: volume required to assess writers' range
- *type of writing*: hand-print, cursive, numerals, signatures
- *variation*: deviations introduced by internal (illness, medication) and external (writing conditions, instrument)

B. Workflow

The workflow of examining handwritten items is given below; it summarizes steps in the ASTM document *Standard Guide for Examination of Handwritten Items* [7].

1. Determine if comparison is *Q v. Q*, *K v. K*, or *Q v. K*. The first when there are no suspects or to determine number of writers. The second to determine variation range. The third to confirm/repudiate writership.
2. Determine whether *Q* and *K* are original or copies. If not original, evaluate quality of best reproduction and check whether significant details are reproduced with sufficient clarity. If not discontinue procedure.
3. Determine whether *Q* and *K* are distorted.
4. Determine the type of writing. If more than one, separate into groups of single type.
5. Check for internal inconsistencies in groups. If inconsistencies suggest multiple writers, divide groups into consistent subgroups. For *K*, if there are unresolved inconsistencies, stop procedure and report accordingly.
6. Determine range of variation for each group/subgroup.
7. Detect presence/absence of individualizing characteristics.
8. Evaluate comparability of *Q* and *K*, e.g., both cursive or both hand-print. If not comparable request new *K* and repeat.
9. Compare bodies of writing.
10. Compare and analyze differences and similarities to form conclusion. The recommended nine-point terminology for expressing FDE conclusion is [8]:
 - (a) Identified as same
 - (b) Highly probable same
 - (c) Probably did
 - (d) Indications did
 - (e) No conclusion
 - (f) Indications did not
 - (g) Probably did not
 - (h) Highly probable did not
 - (i) Identified as Elimination

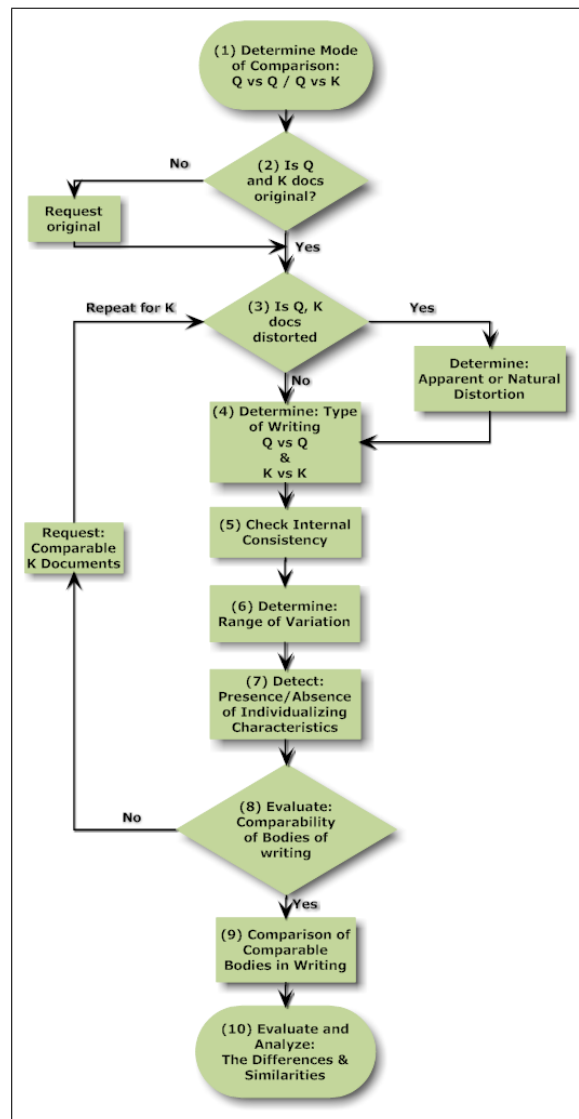


Figure 2.2: Work Flow of Forensic Handwriting Examination. In steps 7 and 10 probabilistic analysis is useful.

The work-flow is presented in the form of a flowchart in Figure 2.2.

The work-flow has at least two points in it where probabilistic analysis can play a useful role. In Step 7, the QD examiner selects individualizing characteristics. Such characteristics are those that are rare within the population, i.e., the letter/word formations have a low probability of occurrence. The second is in the last step of forming an opinion (Step 10), which is essentially a discretized probability of identification/exclusion.

C. Statistical Characterization

The role of probabilities of handwriting characteristics was recognized over a hundred years ago [58]. However their use in court testimony is not currently feasible due to the complexity of the problem. There have been very few efforts to characterize the statistical characteristics of such features, a notable one being [51]. On the other hand there have been efforts to compute features automatically– but the features tend to be gross approximations of the characteristics employed[60, 14] or the features do not correspond to human determined characteristics at all[74, 12]. While these automated methods perform well in objective tests they do not lend support to the document examiner in testimony. Much of automatic handwriting recognition is concerned with determining the identity of a given letter or combination of letters by learning from example data about different forms encountered. On the other hand the goal of forensic handwriting examination is to determine as to how unusual a given structure or formation is so that it can be used to identify the writer. While an unusual, or rare, handwriting formation is central to identifying the writer, it is of little consequence and even considered as noise in recognition.

Several types of probabilistic queries can be useful in the examination of handwriting evidence: (i) the probability of observed evidence, (ii) the probability of a particular feature observed in the evidence, (iii) the probability of finding the evidence in a representative database of handwriting exemplars. The probability of evidence can be used together with the probability of similarity to obtain a strength of opinion. As an example, in the field of DNA evidence a probabilistic statement can be made regarding the rarity of the observed profile by multiplying the probabilities of the observed allele frequencies. The strength of opinion can also be determined by a likelihood ratio[10, 86, 89]. In the case of fingerprints a similar statement can be made about the rarity of a particular minutiae pattern [81] and the strength of opinion by a likelihood ratio[76, 57, 56].

2.1.2 Statement of the problem

The goal is to develop methods, algorithms and software to make it possible to construct probabilistic models of handwriting characteristics so that the models can be used to answer queries of interest to the QD examiner. There are four parts. The first, to prepare the necessary data, the second to construct probability models, the third to perform useful inferences in answering queries and the fourth to indicate its use. In data preparation, the focus is on what letter formations are of interest, how to obtain them and how to assign characteristic values to them. Choice of probability models is important since all combinations cannot be exhaustively determined. Inference concerns answering probabilistic queries. An example query is the probability of a given set of characteristics. Another is the probability of random correspondence with one of n individuals. The end goal of the project is that the results of such analysis could be incorporated into the QD work-flow, e.g., in the choice of individualizing characteristics, in making a statement in the testimony as to how likely it is that a randomly selected person would have those same characteristics. These are further expanded as follows:

1. Data Preparation

- (a) Obtain samples of handwriting representative of the population of interest
- (b) Determine the letter combinations to be analyzed from language statistics
- (c) Have QD examiners specify characteristics specific to cursive writing and hand-print
- (d) Extract snippets of the desired letter combinations from the handwriting samples
- (e) Develop a software tool to assign characteristic values to extracted samples

- (f) Have QD examiners assign characteristic values to samples using the tool
 - (g) From the entered data create files for statistical analysis
2. Statistical Model Construction
 - (a) Determine the type of statistical model to be used, e.g., directed or undirected PGM
 - (b) Determine the structure and parameters of the model from the data
 3. Inference
 - (a) Determine the queries of interest, e.g., probability of finding the evidence among n writers
 - (b) Develop inference procedures to answer queries of interest from the model
 4. Incorporation into QD workflow
 - (a) Formalize QD examiner's work-flow for handwritten items
 - (b) Indicate in work-flow where methods can be incorporated

2.1.3 Literature review

The examination of handwritten items is the most common task in QD examination, also known as forensic document examination (FDE). The examiner has to deal with various aspects of documents, with writership being the central issue. Procedures for handwriting FDE have been described over the course of a century [58, 21, 33, 36, 39]. The requirements for handwriting examination include: (i) known exemplars are comparable to the disputed text, (ii) adequate in amount and (iii) timely or contemporaneous; abbreviated as CAT [35]. We discuss here the literature in the following five areas: (i) type/comparability, (ii) adequacy, (iii) handwriting characteristics, (iv) statistical analysis of characteristics and (v) QD work-flow for handwritten items.

A. Type and Comparability

Handwritten items may be all uppercase, all lower case or a hybrid of hand-print and cursive writing. The question of how the type of writing (hand-printed or cursive) affects results has arisen in the courts, e.g., some United States federal district court judges referred to a lack of information regarding proficiency in identifying writers of hand-printed documents. In the case of *US v. Jeffrey H. Feingold*¹, error rates in [41] on analyzing handwriting were specifically questioned regarding their applicability to hand printing. In that case, the court called a months-long break so that the results could be separated. The study demonstrated the superior proficiency of document examiners compared with laypeople in identifying writers of hand-printed writing.

Hand-printed writing poses some unique challenges and has been specifically explored for decades [20]. It has been noted in [6] that many writers are “less habituated to printing than to cursive writing,” and “in cursive script, connecting strokes within and between letters constitute a critical feature for identification”. Also [6] describes the importance of this problem as “handprinting is often the script of choice for writers of anonymous or disguised messages, apparently because many believe that handprinting is less identifiable than cursive script.”

The above issues and the proliferation of hand-printing in recent years emphasizes the need for the type of writing to be considered in statistical analysis.

¹9th Cir, April 2004, CR 02-0976-PHX-SMM

B. Adequacy

In [35] it is mentioned that “adequacy refers to the number of exemplars needed to allow the FDE to determine the writer’s range of variation. It is not possible to set a definite number of known exemplars. Factors such as the length of the questioned text, the writer’s cooperation in providing comparable exemplars, complexity of letter formations, and natural writing variation affect the number of known exemplars deemed adequate for the comparison to the disputed material. For example 20 to 30 repetitions of the letterforms in a robbery note may be adequate for examination in one case, while 2 or 3 pages of dictated text of a disputed 5-page letter may be adequate for that particular examination.”

The amount of information available can be naturally incorporated into statistical models of identification. In CEDAR-FOX the opinion scale factors-in the amount of information while mapping a log-likelihood ratio to an opinion scale [40] .

C. Handwriting Characteristics

We distinguish between two types of handwriting characteristics: those used by QD examiners in their analysis and those automatically computed.

1. *QD Examiner determined characteristics.* There are several books over the years describing class- and individualizing-characteristics of handwriting [36, 39]. Class-characteristics pertain to the writing characteristics of a group of individuals, e.g., those trained in a certain way. Individualizing characteristics pertain to the habits of an individual writer.

A study of frequencies of QD examiner characteristics of the letter pair *th* was done in 1977 by Muehlberger [51] and colleagues. They specified six characteristics for *th* without considering whether the writing was cursive or hand-printed, as is customary today. Based on the writing of 200 individuals they provided marginal distributions of the six variables and joint distributions of a few pairs of variables. No effort was made to specify the joint distribution of all the six variables together.

Since writing itself changes over the years, e.g., there is more hand-print than cursive writing today, new such evaluations are needed on contemporary handwriting samples. During our previous work on the individuality of handwriting we collected handwriting samples of 1500 individuals. Individuals who wrote the samples are stratified by gender and age and selected to be representative of the United States population. A detailed description of the data set is given in [73] and also summarized in Appendix 2. These samples are useful to obtain statistics of QD examiner characteristics.

2. *Automatically determined characteristics.* Several computational tools for FDE have been developed over the last two decades by the pattern analysis and machine intelligence community [63, 75]. Specific tools include FISH[34], CEDAR-FOX[73, 77], and FLASH-ID[66]. Such tools, which have the capability of extracting handwriting features for the purpose of side-by-side comparison, have been used to establish scientific foundations such as the individuality of handwriting [73, 66] and quantifying the strength of evidence as a likelihood ratio [74]. We summarize characteristics used in two systems and give a comparison:

- In the CEDAR-FOX system [74] characteristics include those specified by QD examiners and those that can be computed easily [78, 46]. These features have been tested extensively and found to work quite well. The system uses 13 macro features characterizing the entire document and micro features for characters used in the comparison. The number of characters used when comparing a pair of documents is variable since it depends on the content and recognition results on the documents, e.g., if there were N occurrences of character a in the first document and M occurrences of character a in the second document, then that results in $N \times M$ comparisons resulting in $N \times M$ features for the character a alone. The distributions of similarity of characteristics under the *same writer* and *different writer* scenarios are determined from the CEDAR data set thereby allowing a likelihood ratio to be computed between given evidence (handwriting sample) and known writing.

- A research group at George Mason University and Gannon Technologies, under funding from the FBI, developed the system known as Forensic Language-independent Analysis System for Handwriting IDentification (FLASH ID) [67]. The feature extraction method involves extracting graphs of characters from word segments, building a graph feature vector, and identifying the unknown character graph by matching against a database containing a set of known character graphs. The graphs known as isocodes are built considering nodes as the ends and cross-points of curves and the curves as the edges. The distribution of the isocodes represents the document which is then compared to the distribution of the known document using the Kullback-Liebler distance. It is clear that these features are quite different from those used by document examiners.
- CEDAR-FOX has a number of user interfaces for interaction as well for understanding the system's decision. For instance the transcript mapping function allows the user to associate characters with correct recognition decisions before a comparison is made. CEDAR-FOX can function when the questioned document has very little writing since it relies on past handwriting to obtain its statistics. FLASH-ID needs a significant amount of material in the questioned document in order for it to be able to compute a distribution of iso-codes. Finally, CEDAR-FOX can be downloaded freely for evaluation purposes.

It should be noted that the present research is about human QD examiners specifying the characteristics rather than using automatically determined characteristics. Thus we begin with QD examiner determined characteristics and focus only on the statistical inference done by computational methods.

D. Statistical Analysis of Characteristics

Two types of statistical analysis are pertinent to QD analysis. The first relates to the characteristics of a given document and the second to the comparison of two documents.

1. *Probability of Characteristics.* The evaluation of frequency of evidence was considered in [51]. In particular samples of th were obtained from 200 writers. The statistical analysis was limited to the extent of providing several conditional probability tables. The complexity of evaluating joint probabilities was not considered. However it was a good starting point for further research in this area.
2. *Probability of Identification.* Forensic identification concerns whether observed evidence arose from a known source. The probabilistic approach is to determine the likelihood ratio positive ($LR+$) [28, 1, 85, 56, 70] whose numerator is the joint probability of the evidence and source under the null, or *prosecution*, hypothesis that the evidence arises from the source and the denominator is the joint probability under the alternate, or *defense*, hypothesis that the evidence does not arise from the object. The probability of identification is readily obtained from $LR+$ as discussed in Section 2.4.2. Determining the joint probability has high data requirements, e.g., if evidence and object are both characterized by n binary features, the joint distribution requires 2^{2n} probabilities or parameters. Even for small n this requires extremely large data sets for estimating parameters. Furthermore in forensic applications data sets are usually small making the approach infeasible. There are two solutions:
 - (a) *Distance Method:* A solution is to use the probability of distance, or similarity, between the evidence and known instead of the joint probability [56, 76]. The distance between features of ensemble of pairs of same and different writer documents are modeled parametrically in CEDAR-FOX using gamma/Gaussian density functions. If p_i^s denotes the density function modeling the distance between *same* writer document pairs for the i^{th} feature and p_i^d denotes the density function modeling the distance between *different* writer pairs for the i^{th} feature, then the likelihood ratio between two documents with distances d_i between the features, is given by $LR = \prod_i \frac{p_i^s(d_i)}{p_i^d(d_i)}$.
 - (b) *Rarity Method:* The distance based method, which has a *constant* number of parameters, is simple to compute but there is a severe loss of information in going from a high-dimensional joint probability space to a one-dimensional distance space. This research considers a third method based on a result of Lindley [48] for univariate Gaussian samples which combines the probability of distance with the probability of the mean of evidence and object, called *rarity*. Computing rarity

exactly has the complexity of 2^n still, for which probabilistic graphical models (e.g. Bayesian networks) and mixture models are used to further simplify the computation.

E. QD Work-flow

The work-flow for the examination of handwritten items by QD examiners are given in the SWGDOC document *Standard Guide for Examination of Handwritten Items* [7]. This forms a good starting point for the inclusion of computational methods.

Handwriting examination practice continues to be a largely manual intensive effort based on FDE training. The situation is not dissimilar to expert systems where automation is only a part of the process, e.g., medical diagnosis, where the stakes are high. Thus there is a need to systematize human procedures so that they can be better understood, validated and improved. Such procedure specification has been referred to as computational thinking [90]. The need for validation is also vital to the forensic sciences [54]. Applying computational thinking to forensic procedures is computational forensics[71].

2.1.4 Rationale for the research

In nearly every step of the QD examination process there is uncertainty. For instance in determining type/comparability and whether there is an adequate quantity of information. Also in the critical step of determining whether a set of characteristics is individualizing or representative of a class. Thus statistical models can play a significant role in assisting the QD examiner. In particular, the need for a statistical description of handwriting characteristics has long been felt, but it has so far not been possible due to lack of efficient methods for collecting the data, computational problems of dealing the very large number of combinatorial possibilities and lack of clear direction for use of such results by the QD examiner.

Existing methods to extract handwriting characteristics automatically have fallen short, since: (i) they are not as discriminative as expert human perception, and (ii) since they do not correspond to human intuition they do not lend support the testimony of the QD examiner. The focus of this effort is quite different from previous methods involving automation in that it involves a cooperative effort between the QD examiner who determines the values of the characteristics and automated methods are only used to build sophisticated probabilistic models.

The goal is to demonstrate how the characteristics of handwritten letter combinations can be captured, and how their statistics can be useful for QD investigation and testimony. This involves data preparation, statistical model construction, statistical inference and incorporation of methods into the QD work-flow.

2.2 Methods: Data Preparation

Preparation of the data for subsequent statistical analysis consisted of the following tasks:

1. *Handwriting Samples*: Determine the collection of handwriting samples from which frequencies will be extracted
2. *Letter Combinations*: Select the letter combinations for which statistical characteristics will be determined
3. *Characteristics*: Determine for each letter combination the set of characteristics that are used by QD examiners
4. *Extraction of snippets* of letter combination images
5. *User Interface*: Preparing a user interface for data entry
6. *Ground-truthing* the letter combinations by QD examiners

Details of each of the six steps are given below.

Table 2.1: Most frequently occurring letter pairs (bigrams) with expected occurrences per 2000 letters. The bigrams are not permitted to span across consecutive words.

| 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 |

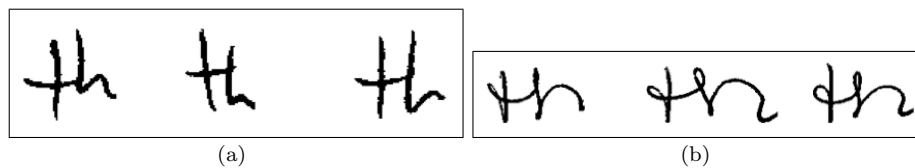


Figure 2.3: Samples of handwritten *th* of two writers showing two different writing styles as well as within-writer variability.

2.2.1 Handwriting Samples

The handwriting samples used in this project are derived from the *CEDAR dataset* consists of writing samples of over 1,500 individuals representing the United States. The population was stratified over gender, age, ethnicity, education, and handedness. Each individual copied a document that contains all possible letters in the English language, including many common letter pairs such as "th", "an", "he" and "nd". Each document is scanned at a resolution of 300 pixels per inch. A description of this data is given in Appendix 1.

2.2.2 Letter combinations

The choice of letter combinations was based on frequency in the English language. Their availability in the handwriting samples was also a consideration. The longer the word string, the higher is the discriminatory power of the formation. Adequacy of handwriting for comparison is relevant for QD examination. Thus we decided to consider pairs and triples of letters rather than a single letter since it is likely to be more individualistic. The most common letter bigrams in the English language are listed in Table 2.1.

The most frequently occurring letter pair is *th* which has been studied in the QD literature [51]. Thus we decided to begin our analysis with *th* whose examples are shown in Figure 2.3. No distinction was made between cursive and hand-print as was done subsequently with the letter triple *and* which includes both the fourth most frequently occurring letter pair *an* and the ninth most frequent pair *nd*. There are nine instances of *th* in the CEDAR letter, five of which are initial, three in the middle (or end) of a word and one with an uppercase "T". There are five instances of *and* in the CEDAR letter, all of which are individual words rather than part of words. Since there are 3 samples of writing of each page per writer, it potentially gives us 27 samples of "th" and fifteen samples of "and" per individual. Some examples of cursively written *and* of a single writer are given in Figure 2.4. Examples of hand-printed *and* of a writer are in Figure 2.5.

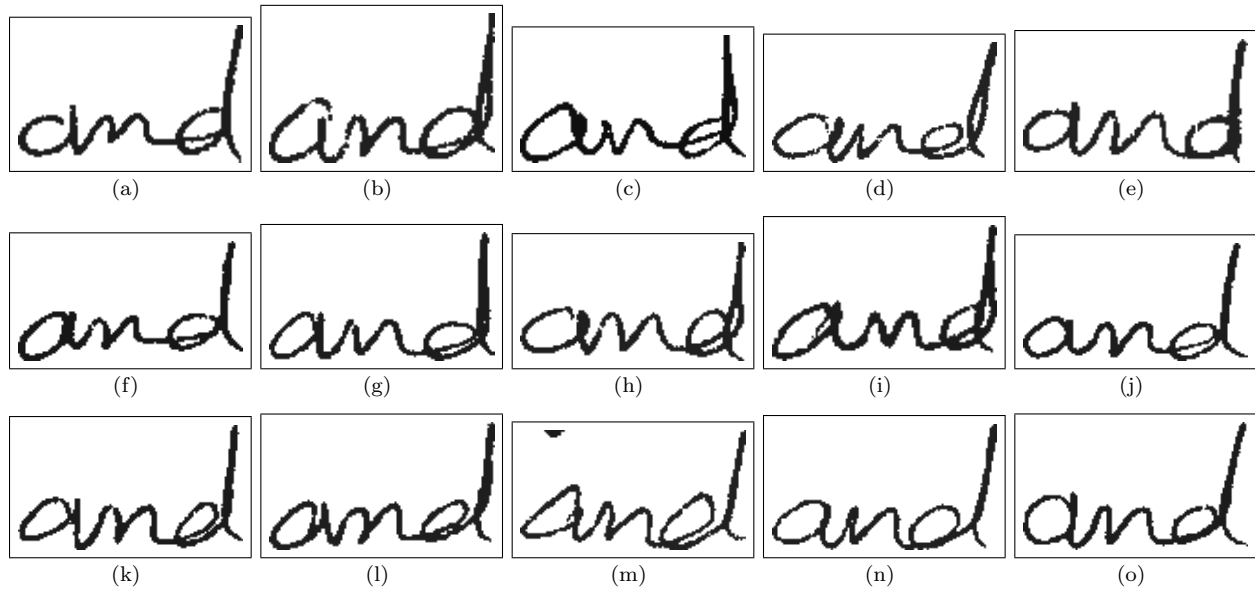


Figure 2.4: Samples of cursively written *and* of a single writer. Using the characteristics listed in Table 2.3 (a) this writing is encoded as 101122012

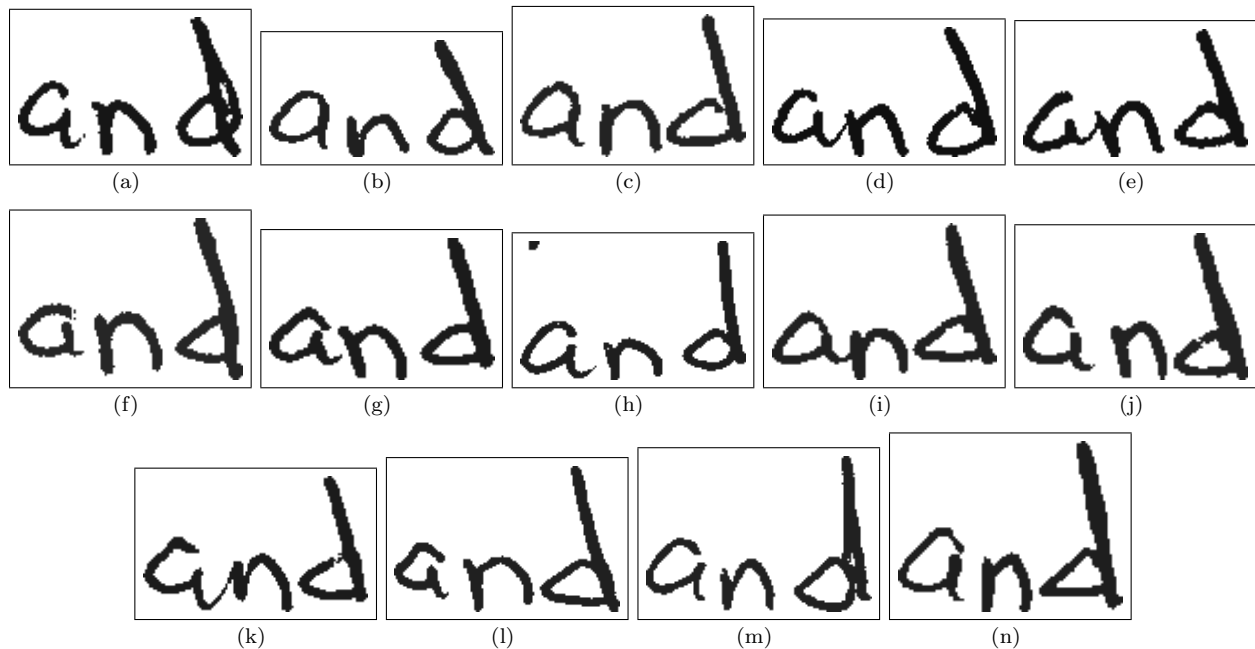


Figure 2.5: Samples of hand-printed *and* of a single writer. Using the characteristics listed in Table 2.3 (b) this writing is encoded as 010110112

Table 2.2: Characteristics of the th combination: six variables and their values [51].

| $R =$ Height Relationship of t to h | $L =$ Shape of Loop of h | $A =$ Shape of Arch of h | $C =$ Height of Cross on t staff | $B =$ Baseline of h | $S =$ Shape of t |
|---|--|----------------------------|------------------------------------|---------------------------|---------------------------|
| $r^0 = t$ shorter than h | $l^0 =$ retraced | $a^0 =$ rounded arch | $c^0 =$ upper half of staff | $b^0 =$ slanting upward | $s^0 =$ tented |
| $r^1 = t$ even with h | $l^1 =$ curved right side and straight left side | $a^1 =$ pointed | $c^1 =$ lower half of staff | $b^1 =$ slanting downward | $s^1 =$ single stroke |
| $r^2 = t$ taller than h | $l^2 =$ curved left side and straight right side | $a^2 =$ no set pattern | $c^2 =$ above staff | $b^2 =$ baseline even | $s^2 =$ looped |
| $r^3 =$ no set pattern | $l^3 =$ both sides curved | | $c^3 =$ no fixed pattern | $b^3 =$ no set pattern | $s^3 =$ closed |
| | $l^4 =$ no fixed pattern | | | | $s^4 =$ mixture of shapes |

2.2.3 Characteristics

The essential starting point for the statistical study was for QD examiners to provide a list of characteristics for letter combinations. It was necessary for them to specify for each letter combination, such as for those listed in Table 2.1, the characteristics that they would use for discriminating between writers. Some of this work previously done. For instance, the characteristics for the most frequent letter combination th were provided in [51] which are reproduced in Table 2.2; perhaps because it was an early experiment, [51] did not differentiate between handprint and cursive.

In [51] the writing of th is characterized by a set of six features $X = \{R, L, A, C, B, S\}$ where R takes on four possible values indicated by lower-case letters superscripted as r^0, r^1, r^2, r^3 and so on. The value is assigned to a particular writing sample, which can consist of several instances of th , as shown in Figures 2.3 and 2.6. For instance the three samples in Figure 2.3(a) will be jointly encoded as $r^1, l^0, a^0, c^3, b^1, s^2$ and the samples in Figure 2.3(b) as $r^2, l^2, a^0, c^1, b^0, s^2$.

For the purpose of this project QD examiners provided a characterization for the word (letter triple) *and*. The characteristics vary depending on whether the writing is cursive and hand-print as shown in Table 2.3. This is also consistent with the comparability issue in Step 8 of the QD work-flow (Figure 2.2).

2.2.4 Extraction of Snippets

Since we are dealing with a large number of handwriting samples (over 4,500 pages) an automatic tool is useful to extract the letter combinations of interest. They were extracted largely automatically from the scanned handwriting images using the transcript mapping function of CEDAR-FOX [37]. Given the typed transcript of the handwritten document, this function maps words to images, as shown in Appendix 3 (Section 2.12). Since some of the mapping may be erroneous, there is provision for manually correcting the results. This process provides the location of letter combination of interest within a page of handwriting from which the image snippets can be extracted.

2.2.5 User Interface

An interface was constructed for ground-truthing image snippets. The graphics interface for truthing the th samples is given in Figure 2.6. The display of image snippets is on the left consisting of all snippets found in the document. Each of the features has a pull-down menu for values that can be entered by the user.

In the case of *and* the interface has two choices: cursive and hand-print, one of which the user must select. The screen for the cursive choice is given in Figure 2.7, and the screen for the hand-print choice is given in Figure 2.8. Handwritten QD examination requires that the letter forms compared in two samples be of the same type [36]. The characteristics depend on whether the writing is cursive or hand-print. Thus the user interface has to facilitate classifying the input into whether the writing is cursive or hand-print. One solution is to automatically perform this classification; which will allow the appropriate characteristics

Table 2.3: Characteristics of *and* as Specified by Document Examiners.

(a) Cursive

| Initial stroke of formation of <i>a</i> (x_1) | Formation of staff of <i>a</i> (x_2) | Number of arches of <i>n</i> (x_3) | Shape of arches of <i>n</i> (x_4) | Location of mid-point of <i>n</i> (x_5) | Formation of staff of <i>d</i> (x_6) | Formation of initial stroke of <i>d</i> (x_7) | Formation of terminal stroke of <i>d</i> (x_8) | Symbol in place of the word <i>and</i> (x_9) |
|---|--|--|---------------------------------------|---|--|---|--|--|
| Right of staff (0) | Tented (0) | One (0) | Pointed (0) | Above baseline (0) | Tented (0) | Overhand (0) | Curved up (0) | Formation (0) |
| Left of staff (1) | Retraced (1) | Two (1) | Rounded (1) | Below baseline (1) | Retraced (1) | Underhand (1) | Straight across (1) | Symbol (1) |
| Center of staff (2) | Looped (2) | No fixed pattern (2) | Retraced (2) | At baseline (2) | Looped (2) | Straight across (2) | Curved down (2) | None (2) |
| No fixed pattern (3) | No staff (3) | | Combination (3) | No fixed pattern (3) | No fixed pattern (3) | No fixed pattern (3) | No obvious ending stroke (3) | |
| | No fixed pattern (4) | | No fixed pattern (4) | | | | No fixed pattern (4) | |

(b) handprint

| Number of strokes of formation of <i>a</i> (x_1) | Formation of staff of <i>a</i> (x_2) | Number of strokes of formation of <i>n</i> (x_3) | Formation of staff of <i>n</i> (x_4) | Shape of arch of <i>n</i> (x_5) | Number of strokes of formation of <i>d</i> (x_6) | Formation of staff of <i>d</i> (x_7) | Initial stroke of <i>d</i> (x_8) | Unusual formation (x_9) |
|--|--|--|--|-------------------------------------|--|--|--------------------------------------|-----------------------------|
| One continuous (0) | Tented (0) | One continuous (0) | Tented (0) | Pointed (0) | One continuous (0) | Tented (0) | Top of staff (0) | Formation (0) |
| Two strokes (1) | Retraced (1) | Two strokes (1) | Retraced (1) | Rounded (1) | Two strokes (1) | Retraced (1) | Bulb (1) | Symbol (1) |
| Three strokes (2) | Looped (2) | Three strokes (2) | Looped (2) | No fixed pattern (2) | Three strokes (2) | Looped (2) | No fixed pattern (2) | None (2) |
| Upper case (3) | No staff (3) | Upper case (3) | No staff (3) | | Upper Case (3) | Single down (3) | Undetermined (3) | |
| No fixed pattern (4) | Single line down (4) | No fixed pattern (4) | No fixed pattern (4) | | No fixed pattern (4) | Single up (4) | | |
| | No fixed pattern (5) | | | | | No fixed pattern (5) | | |

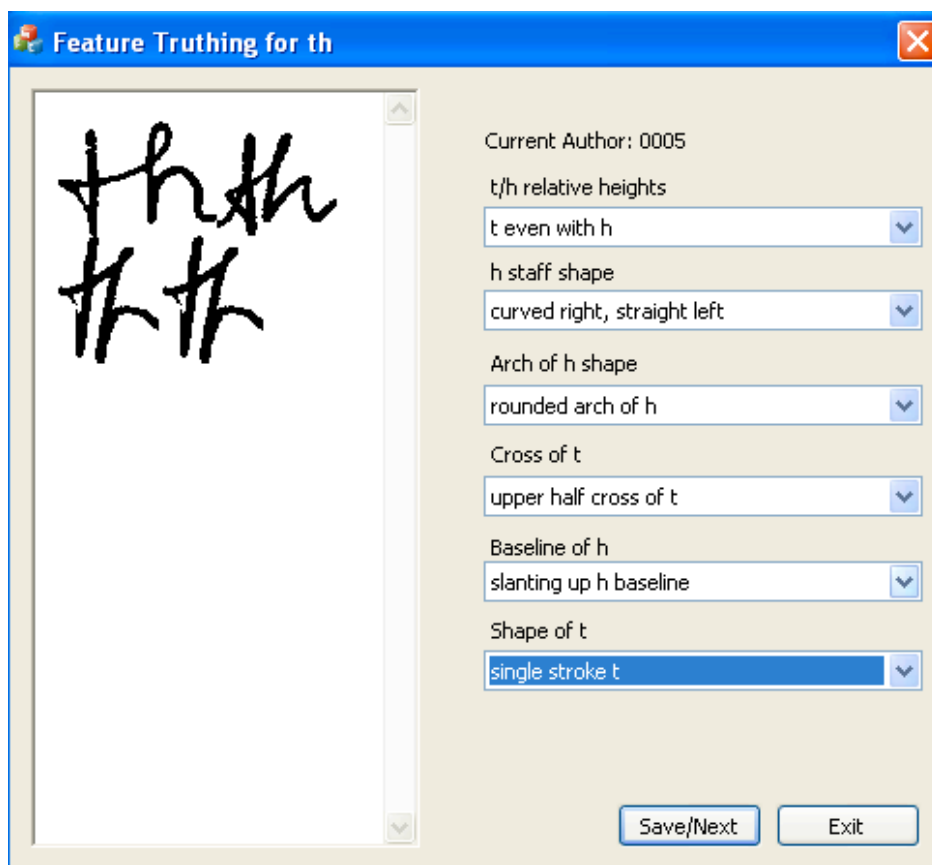


Figure 2.6: GUI for determining the features for *th* exemplars of a given writer: values are assigned manually using pull-down menus for each feature.

menu to be automatically brought-up (see Appendix 5). The other is to have a user enter this information manually. For the present, we used manual input.

When the data for a given set of images has been entered by the user, the next set of snippets is displayed. The user can save work and resume later if necessary. The frequencies of different combinations of features are accumulated so that different marginal and conditional probabilities can be computed from them.

2.2.6 Ground-truthing

Ground-truthing is the task of assigning values for each of the characteristics for given samples of a handwritten word. It is best done by QD examiners who are trained to observe handwriting characteristics. For the very first test effort we undertook, the data entry for *th* was done by lay people (computer science students).

In the case of *and* the data entry interface was used by QD examiners to enter the values for the characteristics. They first enter whether the samples are cursive or handprint. This provides the appropriate pull-down menus for data entry.

The work-load was shared primarily between two QD examiners. A small set at the end was entered by a third QD examiner. The end-result is a valuable resource for characterizing the distribution of *and* in U.S. handwriting.

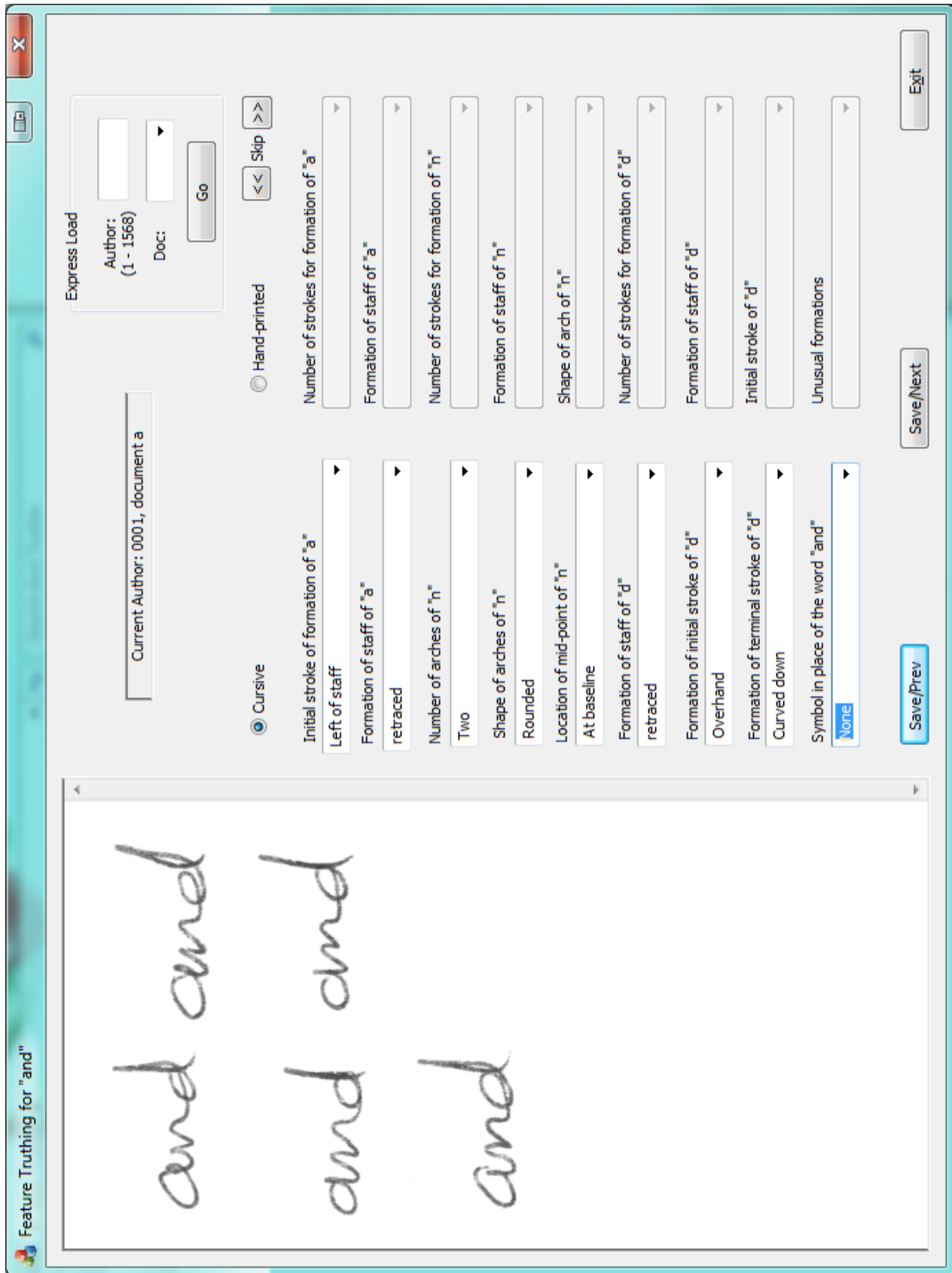


Figure 2.7: GUI for ground-truthing of *and*: written cursively.

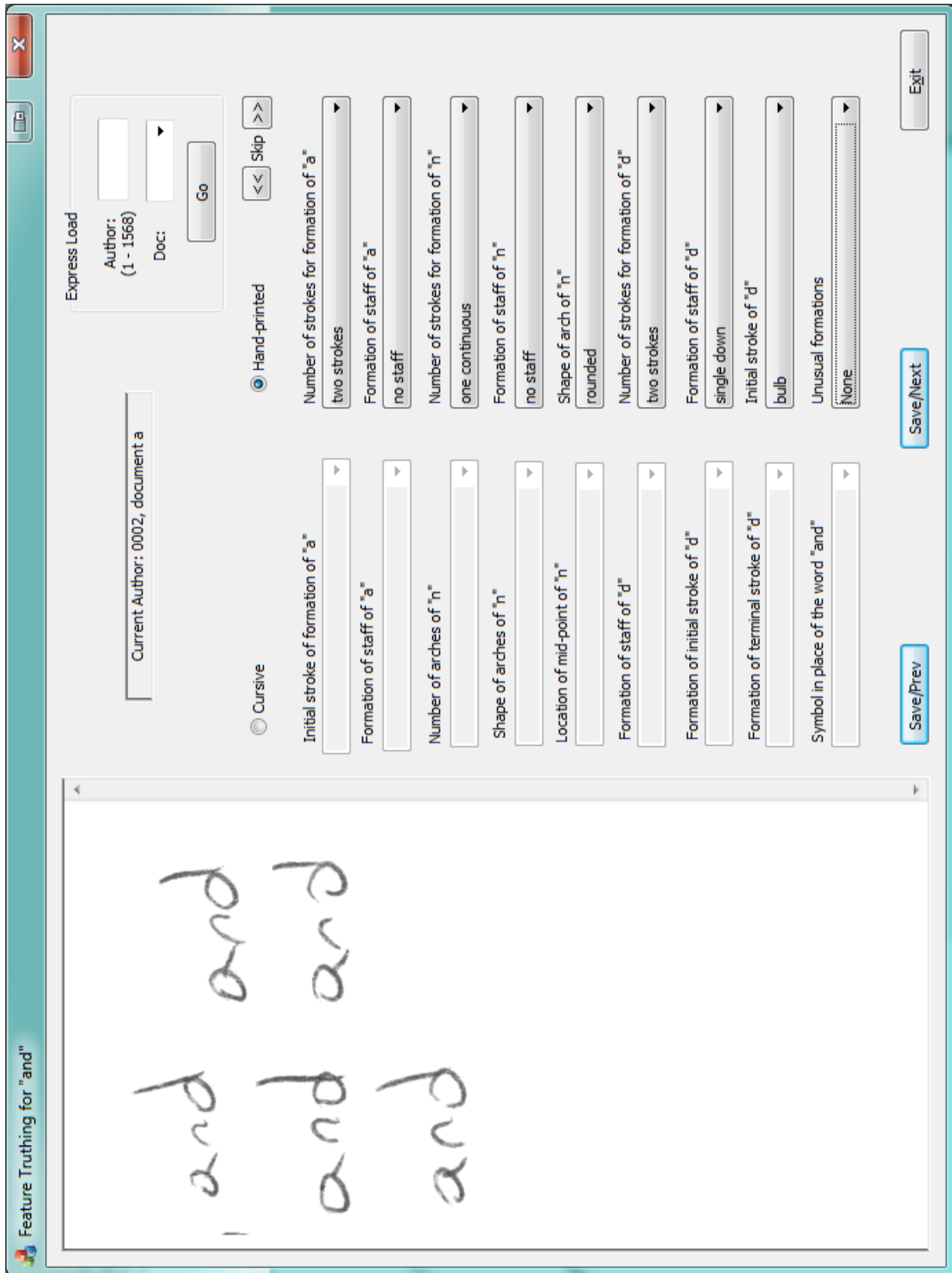


Figure 2.8: GUI for ground-truthing of *and* hand-printed.

2.3 Methods: Statistical Model Construction

Once the frequency data for handwriting characteristics is available, using methods described in Section 2.2, the next step is to construct a statistical model. The model, in the form of a joint distribution, can have several uses:

- The probability of a given handwritten item, say a word represented by a set of characteristics, can be determined. This probability is a measure of whether the characteristics belong to a class (when the probability is high) or are individualizing (low probability).
- Marginal probabilities of individual variables, joint probabilities of pairs (or more) of variables and conditional probabilities, which are useful to refer to characteristic(s) in isolation, can be inferred.
- Generate samples for analysis, e.g., approximate inference. While the samples generated will be in the form of characteristic values, images can be artificially synthesized for visualization, testing, etc.

In constructing models, there are several fundamental issues to be faced. The number of combinatorial possibilities of feature combinations increases exponentially with the number of characteristics. The number of samples needed also increases proportionally.

2.3.1 Problem Complexity

To illustrate the complexity of constructing a joint distribution, consider the example of *th* whose characteristics as given by QD examiners [51] is given in Table 2.2. Thus the writing of *th* is characterized by a set of six features $X = \{R, L, A, C, B, S\}$ where R takes on four possible values indicated by lower-case letters superscripted as r^0, r^1, r^2, r^3 and so on. The value is assigned to a particular writing sample, which can consist of several instances of *th*, as shown in Figures 2.3 and 2.6. For instance the three samples in Figure 2.3(a) will be jointly encoded as $r^1, l^0, a^0, c^3, b^1, s^2$ and the samples in Figure 2.3(b) as $r^2, l^2, a^0, c^1, b^0, s^2$.

In the probabilistic formulation each characteristic is considered to be a random variable. These six variables each have multinomial distributions with 4, 5, 3, 4, 4 and 5 possible values. If we assume that the variables are independent then the number of independent probabilities (parameters) to be estimated is $3 + 4 + 2 + 3 + 3 + 4 = 19$. On the other hand if we allow all dependencies, the number of parameters needed is $4 \times 5 \times 3 \times 4 \times 4 \times 5 - 1 = 4,799$. The complexity of joint probability distribution increases exponentially with the number of variables. However it is unsatisfactory to assume full independence between all variables.

Consider now one of the most common words in the English language *and*. Characteristics for this word specified by QD examiners is given in Table 2.3 for cursive and handprint writing. Thus the writing of *and* is characterized by a set of nine features $X = X_1, X_2, \dots, X_9$ where X_i takes up to 5 values for the cursive dataset and 6 values for the handprint dataset. Each sample with a set of feature values can represent several instances of *and*. In the probabilistic formulation each feature is considered to be a random variable. The nine features each have multinomial distributions with 4, 5, 3, 5, 4, 4, 4, 5 and 3 possible values for cursive data and 5, 6, 5, 5, 3, 5, 6, 4, and 3 possible values for handprint data. If we assume that all variables are dependent on every other variable, the number of parameters needed for cursive data is $4 \times 5 \times 3 \times 5 \times 4 \times 4 \times 4 \times 5 \times 3 - 1 = 287,999$ and for handprint data is $5 \times 6 \times 5 \times 5 \times 3 \times 5 \times 6 \times 4 \times 3 - 1 = 809,999$.

Just moving from the two letter word *th* to a three letter word *and* increases the number of parameters needed from about 5,000 to 288,000 or more. Given a full *London letter* [58], which has a dozen or more words, it would be impossible to characterize the probability distribution with a full set of parameters.

The computational complexity and the need for samples can be managed by exploiting statistical independencies that exist between some variables but without resorting to the assumption that all characteristics are statistically independent of each other. Probabilistic graphical models are useful to express such independencies [44]. We can use either directed graphical models, known as Bayesian networks, or undirected graphical models, known as Markov networks. In the rest of this section we discuss methods to construct both types of models.

2.3.2 Bayesian Networks

A Bayesian network (BN) is a representation of a joint probability distribution of several random variables. It is represented in the form of a directed acyclic graph (DAG) together with associated conditional probability distributions. It is essentially a collection of *directed dependencies* (or *causality*) between variables.

A BN for the six variables in Table 2.2, BN_{th} , is given in Figure 2.9(a). It incorporates causality such as: the shape of t (S) influences the shape of h loop (L), the shape of h -arch (A) influences the baseline of h (B), etc. Together with the conditional probability distributions (CPDs), the Bayesian network represents the full joint distribution of the six variables. BN_{th} factorizes the distribution of th into component CPDs as

$$P(X) = P(R)P(L|S)P(A|L)P(C|S)P(B|R, A)P(S|R). \quad (2.1)$$

The CPD values are derived from data obtained from the ground-truthing described previously. The number of independent parameters needed to specify BN_{th} is $3 + 16 + 10 + 20 + 15 + 36 = 100$ which is far fewer than 4,799 to directly specify the distribution. The marginal distributions of the variables are in Figure 2.9(b) and the conditional probability tables (CPTs) for Eq. 2.1 are given in Figure 2.9(c-g). The method of estimating the parameters is discussed after we introduce *Bayesian Network Structure Learning*.

Note that given the BN, the marginal probability of any single characteristic, or the joint probability of any combination of characteristics can be determined. For example, if we are interested in the joint probability $P(R, L)$, it can be determined using the sum rule of probability as $P(R, L) = \sum_{A,C,B,S} P(X)$.

Bayesian Network Structure Learning

Manually specifying causality between variables is not an easy task. Thus automatic methods are useful. The goal of BN structure learning is to find a BN that gives the best representation of all directed dependencies between variables. However it is a computationally intractable problem that is NP-complete problem. Learning the structure of Bayesian networks (BNs) that approximate the joint distribution of a large number of variables is an important problem in machine learning and data mining [18, 43, 44, 82].

Existing methods for BN structure learning can be divided into three types: *constraint* based, *score* based and *Bayesian model averaging* methods. By viewing the BN as a representation of dependencies, constraint based methods attempt to find a network structure that best explains dependencies. But it is sensitive to errors in testing single dependencies [44]. Score based methods view learning as a *model selection* problem; by defining a scoring function which assesses the fitness of each model, it searches for a high-scoring network structure. As the search space is super-exponential, enumerating scores for all models is often *NP-hard*. Therefore it has to resort to heuristic search. Examples are the K2 algorithm [22], and the optimized branch and bound algorithm [26]. The third type of methods, Bayesian model averaging, make predictions by averaging across all possible structures. But the disadvantage is that some may not have closed form. Recently Peters *et al.* [61] proposed an algorithm that infers the causal structure of discrete variables using additive noise models. It also points out that the limitation of the χ^2 test on small data sets.

We use both constraints and a score. Deviance from independence between pairs of variables are constraints and the structure is scored using log-loss. Instead of exhaustively searching the entire solution space, the deviances are used as guidance to search for a structure with a low log-loss. Constraints and score are defined as follows:

1. *Constraint.* The chi-squared (χ^2) independence test (also known as *Pearson's chi-squared test*.) is used to test for independence of two categorical variables [31, 88]. It provides a measure of deviance from the null hypothesis of independence [44].

Let X and Y be two multinomial variables governed by distributions $P(X = x)$ and $P(Y = y)$. Consider a data set \mathcal{D} with a total of M samples, where $M[x, y]$ is the *observed* count for each *joint assignment* of $X = x$ and $Y = y$. Given the null hypothesis that X and Y are independent, the *expected* count for (x, y) is $E[x, y] = M \cdot \hat{P}(x) \cdot \hat{P}(y)$ where \hat{P} indicates the estimate of P from \mathcal{D} . Then the

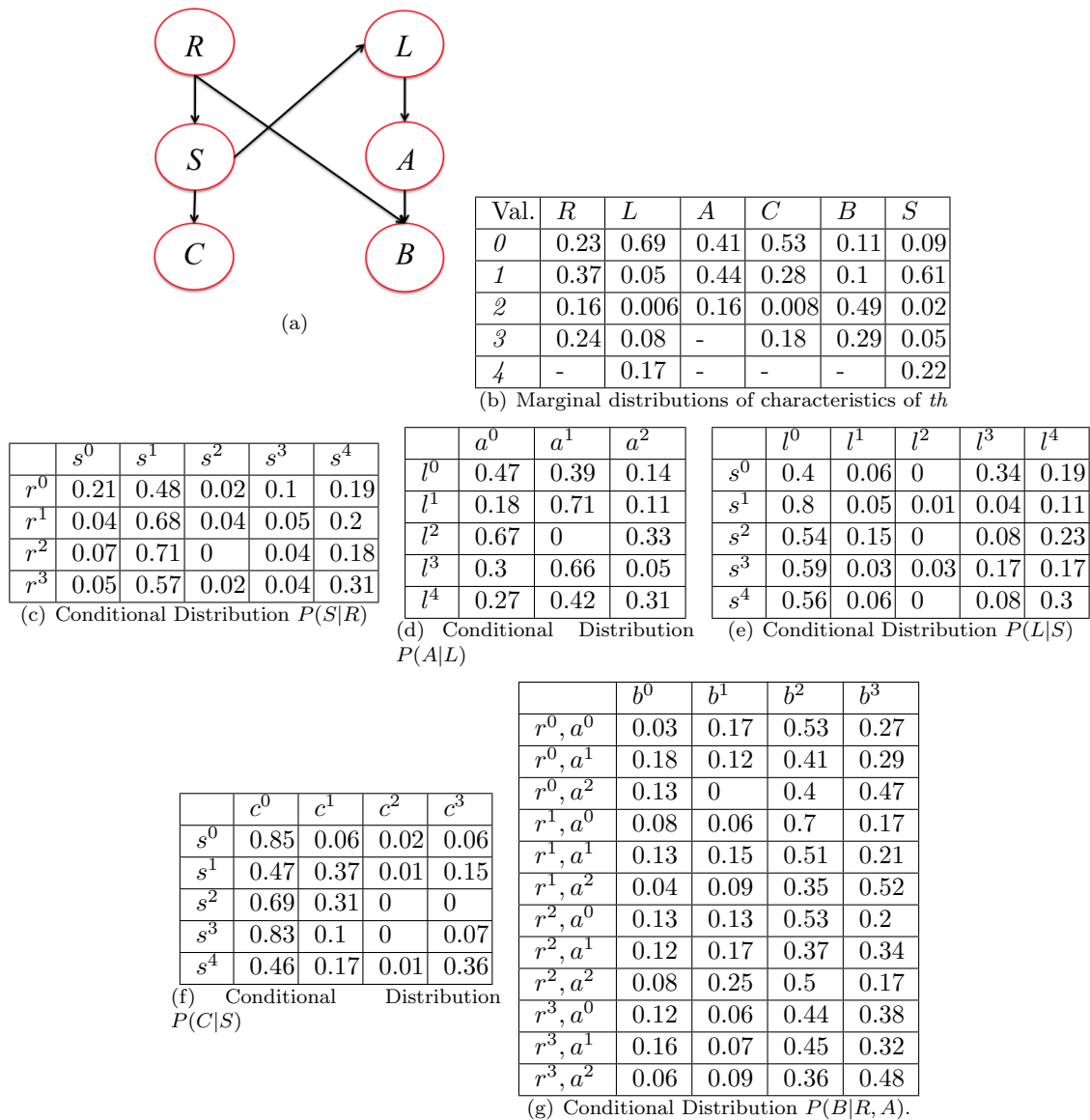


Figure 2.9: Bayesian network of th : (a) manually constructed directed graph, (b) marginal distributions, (c)–(g) conditional probability distributions.

deviance measure is defined by

$$d_{\chi^2}(\mathcal{D}) = \sum_{x,y} \frac{(M[x,y] - E[x,y])^2}{E[x,y]}. \quad (2.2)$$

2. *Score.* For a BN G with n variables $\{X_1, \dots, X_n\}$, its log-loss on a data set \mathcal{D} with M i.i.d samples is the negative log-likelihood given by

$$s(\mathcal{D}|G) = - \sum_{i=1}^n \sum_{m=1}^M \log P(x_i[m]|\text{pa}_{X_i}[m]), \quad (2.3)$$

where $x_i[m]$ is the value of the i^{th} variable (characteristic) in the m^{th} sample and $\text{pa}_{X_i}[m]$ are the values of all the parent variables of X_i in the m^{th} sample.

Structure Learning Algorithm

In Algorithm 1 the pairwise deviances between all possible pairs of nodes (features) are first calculated and stored in a set E_p in non-increasing order. Starting from a model without any edges but containing all the vertices, one edge (x_i, x_j) at a time is examined to determine whether or not the edge should be added and which direction should be used by comparing the three BNs: the BN before considering the edge (G^*), the BN with $(x_i \rightarrow x_j)$ added (G_{c_1}), and the BN with $(x_j \rightarrow x_i)$ added (G_{c_2}), as shown in Fig. 2.10. The goodness of each BN is measured by its log-loss defined in Eq. 2.3. If either edges do not decrease the log-loss of the model or do not create a DAG, no edge is added to G . The algorithm chooses the best model G^* so far with the lowest s among (G_c, G_{c_1}, G_{c_2}) , set current graph $G_c = G^*$, remove (v_i, v_j) out of E_p . Then it repeats the previous procedure, until all node pairs have been examined, i.e. $E_p = \emptyset$. This process is repeated until the set E_p became empty. The parameters of G_c, G_{c_1} , and G_{c_2} are estimated using either maximum likelihood or Bayesian approaches as described in the next section. The function $\text{isDAG}(G)$ returns a Boolean value indicating whether G is a *directed acyclic graph* (DAG).

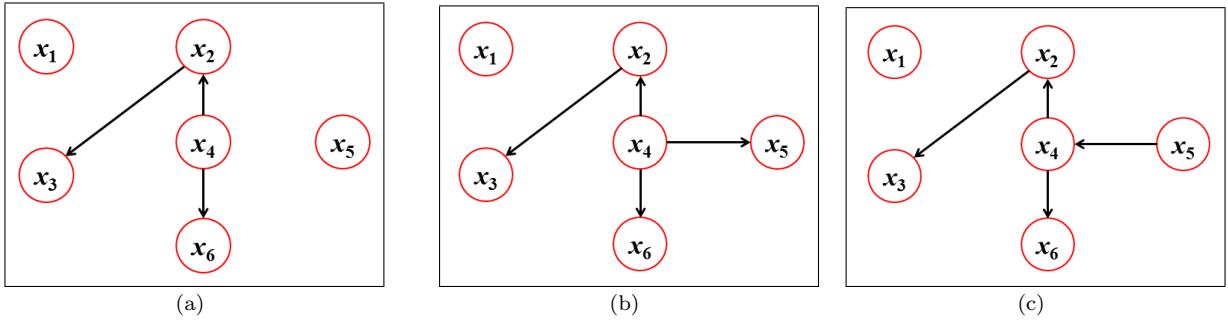


Figure 2.10: Bayesian Network (BN) structure learning, where edge (x_4, x_5) , with the next highest χ^2 value is being considered: (a) BN G^* before considering edge (x_4, x_5) ; (b) candidate BN G_{c_1} with edge $(x_4 \rightarrow x_5)$; (c) candidate BN G_{c_2} with edge $(x_5 \rightarrow x_4)$.

Since determining the log-loss to determine the addition of each edge takes exponential-time, while constructing the models, only up to two parents per node was considered. As shown in Figure 2.11(a) and 2.11(b), if the new node 'b' to be added will become the parent of node 'a', the edge is added to our structure as long as 'a' has no parent or 'a' has only a single parent. As shown in Figure 2.11(c), if 'a' already has 2 parents 'c' and 'd', the new edge is not added to the graph whether or not it decreases the log loss.

Algorithm Complexity. Step 3 takes time $O(n^2)$, step 12 takes $O(n^2 \log n)$ and $O(n \cdot 2^d)$, where d is the maximum number of parents per node, and the time required in the loop is bounded by $O(n^2)$, so the running time is $O(n^2 \log)$.

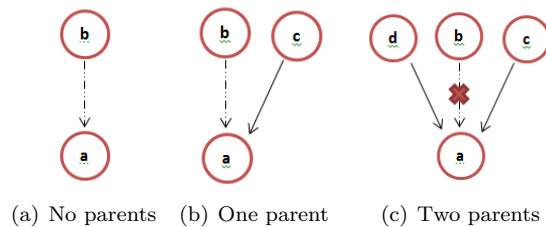


Figure 2.11: Heuristics of determining Structure in Algorithm BNSL

Algorithm 1 BNSL: Bayesian Network Structure Learning

-
- 1: **Input:** set $S = \{v_1, \dots, v_n\}$ of n random variables, training data \mathcal{D}_r and validation data \mathcal{D}_v
 - 2: **Output:** an optimal graph $G^* = \{V, E^*\}$
 - 3: Compute pairwise deviations, $d_{\chi^2}(v_i, v_j)$
 - 4: Construct ordered set $E_p = \{(v_i, v_j) | i \neq j\}$ in descending order of deviations
 - 5: Set $G^* = \{V, E^*\}$, where $V = S$, $E^* = \emptyset$
 - 6: Set $k = 1$
 - 7: **repeat**
 - 8: Pick the first pair (v_i, v_j) from E_p
 - 9: Create $G_{c_1} = (V, E_{c_1}), E_{c_1} = E^* + \{v_i \rightarrow v_j\}$
 - 10: Create $G_{c_2} = (V, E_{c_2}), E_{c_2} = E^* + \{v_j \rightarrow v_i\}$
 - 11: Compute s_{k-1}, s_{c_1} for G^* and G_{c_1} from \mathcal{D}_r
 - 12: **if** $(s_{c_1} > s_{k-1})$ & $\text{isDAG}(G_{c_1})$ **then**
 - 13: $G^* = G_{c_1}$
 - 14: **end if**
 - 15: Compute s_{k-1}, s_{c_2} for G^* and G_{c_2} from \mathcal{D}_r
 - 16: **if** $(s_{c_2} > s_{k-1})$ & $\text{isDAG}(G_{c_2})$ **then**
 - 17: $G^* = G_{c_2}$
 - 18: **end if**
 - 19: Increment k by 1
 - 20: $E_p = E_p - \{(v_i, v_j)\}$
 - 21: **until** $E_p = \emptyset$
 - 22: **return** G^*
-

Parameter Estimation

To estimate the parameters of a BN one can use either maximum likelihood or Bayesian estimation. If there are an insufficient number of samples then a maximum likelihood estimate, may yield several zero probabilities. Zero probabilities are dangerous since we can never recover from a product involving such a probability as a factor, There can also be a problem in inference where a division by that probability is necessary. A smoothing can be performed over maximum likelihood estimates or a Bayesian estimate can be used as described below.

1. **Maximum likelihood estimation.** Assume a multinomial variable X with k possible values $\{x^1, \dots, x^k\}$. The maximum likelihood estimate for $X = x^i$ ($i = 1, \dots, k$) is simply the fraction of the number samples taking on that value, $M[i]$, to the total number of samples M . In order to avoid zero probabilities, or to avoid cases of division by zero, additive smoothing is used:

$$\theta_i = \hat{P}(X = x^i) = \frac{M[i] + \alpha}{M + \alpha k}, \quad (2.4)$$

where, $\alpha = 1$ for add-one smoothing.

2. **Bayesian estimation.** We place a uniform Dirichlet prior on its parameters $\theta = \{\theta_1, \dots, \theta_k\}$, i.e. $\theta \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k)$, where α_i are called hyper-parameters with $\sum_i \alpha_i = \alpha$. Now consider a data set \mathcal{D} of independent observations where the number of observations of $x = x^i$ is $M[i]$. Then the *sufficient statistics* $\{M[1], \dots, M[k]\}$ is distributed by $\{M[1], \dots, M[k]\} \sim \text{Multinomial}(\theta_1, \dots, \theta_k)$. Because of the conjugacy between the Dirichlet and the multinomial distribution, the posterior distribution of θ given \mathcal{D} , is also Dirichlet, $\theta|\mathcal{D} \sim \text{Dirichlet}(\alpha'_1, \dots, \alpha'_k)$, where $\alpha'_i = \alpha_i + M[i], \forall i = 1, \dots, k$. The result as a prediction is quite similar to that with maximum likelihood parameters:

$$\hat{P}(X[M+1] = x^i) = \frac{M[i] + \alpha_i}{M + \alpha}, \quad (2.5)$$

With $\alpha_1 = \dots = \alpha_k = 1$ the Bayesian estimate reduces to the case of maximum likelihood estimate with smoothing. We describe next the parameters obtained for the BNs for the *th* and *and* handwritten data sets.

1. **Parameters for BN_{th} .** There were 3,125 samples from 1,254 documents written by 499 writers. The probabilities required by the BN in Figure 2.9(a) were estimated using the maximum likelihood approach, where no smoothing was necessary since there were few zero probabilities. The marginal distributions are given in Figure 2.9(b).

A comparison of the marginal distributions of the characteristics given in Figure 2.9(b) with those given in [51] was performed. The results are described in Appendix 3. The marginal probabilities of the variables have some similarity with those given in [51]: four of the six variables are accepted as having similar distributions and the other two rejected. That the correlation is not stronger can be attributed to several reasons: in [51] there is no mention that the handwriting samples are representative of any population whereas an effort was made in ours, handwriting samples can change over a period of a few decades, the proportions of cursive and hand-print in the two data sets may be quite different (since the *th* characteristics are not tuned to type), and some characteristic values are ambiguous as we found in our ground-truthing.

2. **Parameters for BN_{and} .** We used 10,111 samples with 1,555 writers. The parameters of the models were determined using maximum likelihood with smoothing and evaluating the Conditional Probability Tables (CPTs) for only the factors in the joint probability. BNs learned from the *and* data are given in Figure 2.13 and the necessary CPTs are given in Tables 2.5 and 2.6. The CPTs were calculated using maximum likelihood estimates with smoothing. The number of independent parameters needed for the BNs are as follows:

Cursive: $2 + 4 + 12 + 15 + 20 + 15 + 10 + 12 + 9 = 99$.

Handprint: $5 + 4 + 5 + 2 + 4 + 15 + 20 + 10 + 12 = 77$.

Evaluation of BN Construction Algorithm

The final issue in BN structure learning is as to how good are the resulting structures. Since the true joint distribution is unknown, we can compare with structures obtained using other algorithms, hand-constructed ones and the simplest structure that assumes all characteristics are independent.

1. **Goodness of BN_{th} .** To evaluate the performance of Algorithm 1 in constructing a BN for the *th* data, we used a baseline algorithm known as branch-and-bound (B & B) [26, 25] as well as the human-designed BN in Figure 2.9(a). Three evaluation metrics were used: log-loss s , average learning time \bar{T} , and sensitivity to training data σ_s . To study algorithm sensitivity, we ran validation tests 1000 times for all candidate graphs on *randomized* data. For each run, we randomly selected $\frac{2}{3}$ of the whole data for training, and the rest for testing. Sensitivity is defined by the *standard deviation* of the log-loss from multiple runs.

The corresponding three BN structures are shown in Fig. 2.12(a-c). Here we used maximum likelihood parameter estimation with smoothing. As can be seen, the log-losses in Figure 2.9(d) are somewhat similar, although BNSL is marginally better.

2. **Goodness of BN_{and} .** Log-loss values were calculated for the BNs and a model that assumes all variables are independent. As seen in Table 2.4, the log loss is higher with the independence assumption

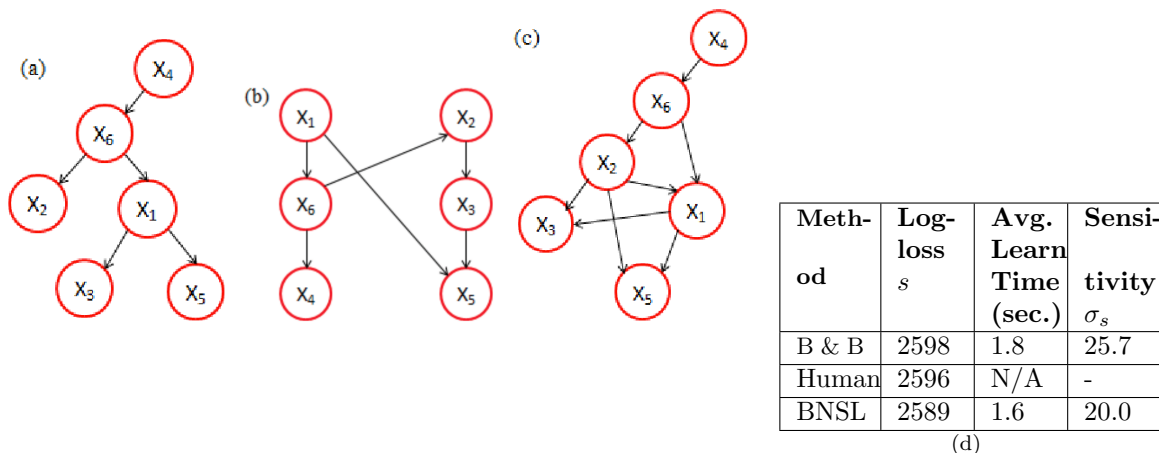


Figure 2.12: Evaluation of BN structures learnt from *th* data: (a) branch and bound algorithm, (b) human designed BN based on causality; (c) algorithm BNSL, and (d) performance metrics.

for both the cursive and hand-print data sets, thereby indicating that the learnt BN structure is the better model.

Table 2.4: Evaluation of BN Models (log loss) for *and* data.

| | Bayesian Network | Independent Variables |
|------------------|---------------------|--------------------------|
| Cursive | 25329 | 25994 |
| Handprint | 7898 | 8059 |

Sampling from Bayesian Networks

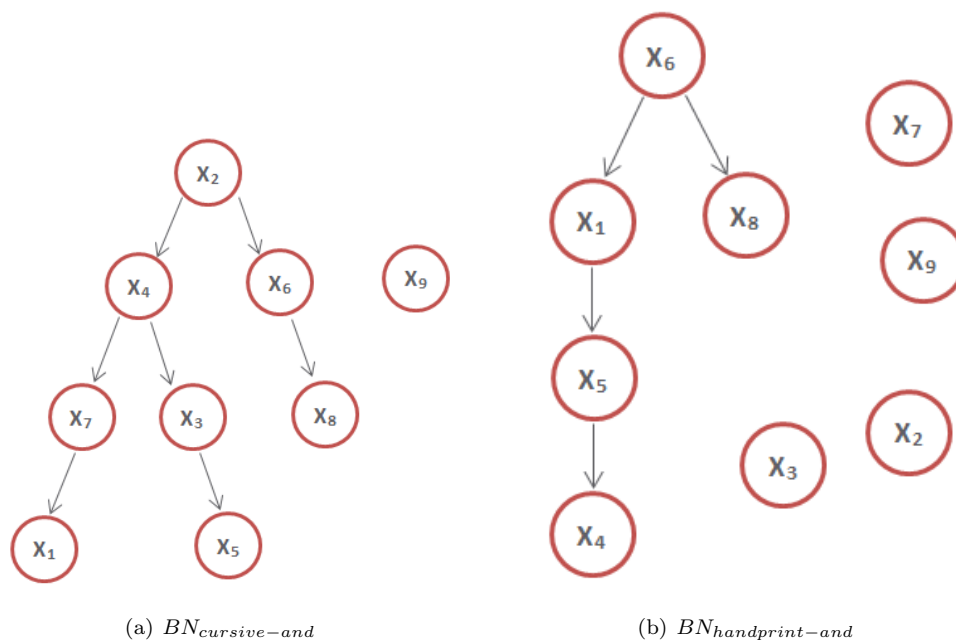
Bayesian networks allow the generation of samples satisfying the distribution modeled in a simple way. Gibbs sampling was used to generate a few samples whose probability is then evaluated. The algorithm for Gibbs sampling [15] is given in Algorithm 2.

Algorithm 2 Gibbs Sampling

- 1: Initialize $\{x_i : i = 1, \dots, M\}$
 - 2: **for** $\tau = 1, \dots, T$ **do**
 - 3: Sample $x_1^{(\tau+1)} \sim p(x_1 | x_2^{(\tau)}, x_3^{(\tau)}, \dots, x_M^{(\tau)})$
 - 4: Sample $x_2^{(\tau+1)} \sim p(x_2 | x_1^{(\tau+1)}, x_3^{(\tau)}, \dots, x_M^{(\tau)})$
 - 5: .
 - 6: .
 - 7: Sample $x_j^{(\tau+1)} \sim p(x_j | x_1^{(\tau+1)}, \dots, x_{j-1}^{(\tau+1)}, x_{j+1}^{(\tau)}, \dots, x_M^{(\tau)})$
 - 8: .
 - 9: .
 - 10: Sample $x_M^{(\tau+1)} \sim p(x_M | x_1^{(\tau+1)}, x_2^{(\tau+1)}, \dots, x_{M-1}^{(\tau+1)})$
 - 11: **end for**
-

Where, in our case, $M = 9$, as we have 9 features in each sample. $\tau = \text{Burn-in} + \text{Number of samples required}$. The Burn-in iterations are required to ensure that the initial values that we initialize the first value of X , has no effect on the samples generated.

The values of the conditional probabilities needed for sampling are calculated as follows.



(c) Marginal probabilities of cursive characteristics

| | X_1 | X_2 | X_3 | X_4 | X_5 | X_6 | X_7 | X_8 | X_9 |
|----------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|----------------------|----------------------|----------------------|----------------------|
| 0 | 1.34×10^{-1} | 1.36×10^{-1} | 9.75×10^{-2} | 3.79×10^{-1} | 1.99×10^{-1} | $1.69 \cdot 10^{-2}$ | $4.66 \cdot 10^{-1}$ | $1.96 \cdot 10^{-1}$ | $3.25 \cdot 10^{-4}$ |
| 1 | 3.58×10^{-1} | 6.12×10^{-1} | 8.79×10^{-1} | 2.22×10^{-1} | 2.21×10^{-2} | $2.26 \cdot 10^{-1}$ | $3.58 \cdot 10^{-1}$ | $2.61 \cdot 10^{-1}$ | $3.25 \cdot 10^{-4}$ |
| 2 | 3.34×10^{-1} | 1.21×10^{-1} | 2.34×10^{-2} | 7.73×10^{-2} | $7.61 \cdot 10^{-1}$ | $6.40 \cdot 10^{-1}$ | $1.17 \cdot 10^{-1}$ | $3.31 \cdot 10^{-1}$ | $9.99 \cdot 10^{-1}$ |
| 3 | 1.74×10^{-1} | 7.14×10^{-3} | | 3.06×10^{-1} | 1.82×10^{-2} | $1.17 \cdot 10^{-1}$ | $5.91 \cdot 10^{-2}$ | $1.31 \cdot 10^{-1}$ | |
| 4 | | $1.23 \cdot 10^{-1}$ | | $1.66 \cdot 10^{-2}$ | | | | $8.18 \cdot 10^{-2}$ | |

(d) Marginal probabilities of handprint characteristics

| | X_1 | X_2 | X_3 | X_4 | X_5 | X_6 | X_7 | X_8 | X_9 |
|----------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|----------------------|----------------------|----------------------|----------------------|
| 0 | 9.02×10^{-1} | 3.16×10^{-1} | 9.25×10^{-1} | 2.68×10^{-1} | 1.44×10^{-1} | $7.96 \cdot 10^{-1}$ | $2.89 \cdot 10^{-2}$ | $3.51 \cdot 10^{-1}$ | $4.04 \cdot 10^{-2}$ |
| 1 | 3.07×10^{-2} | 4.09×10^{-1} | 6.14×10^{-3} | 5.44×10^{-1} | 7.63×10^{-1} | $1.44 \cdot 10^{-1}$ | $2.62 \cdot 10^{-1}$ | $5.90 \cdot 10^{-1}$ | $8.79 \cdot 10^{-4}$ |
| 2 | 8.77×10^{-4} | 5.70×10^{-2} | 8.77×10^{-4} | 9.65×10^{-3} | 9.31×10^{-2} | $8.77 \cdot 10^{-4}$ | $1.32 \cdot 10^{-1}$ | $1.49 \cdot 10^{-2}$ | $9.59 \cdot 10^{-1}$ |
| 3 | 5.61×10^{-2} | 4.29×10^{-2} | 6.67×10^{-2} | 6.49×10^{-2} | | $3.77 \cdot 10^{-2}$ | $4.34 \cdot 10^{-1}$ | $4.39 \cdot 10^{-2}$ | |
| 4 | 1.05×10^{-2} | 2.89×10^{-2} | 1.75×10^{-3} | 1.13×10^{-1} | | $2.19 \cdot 10^{-2}$ | $5.08 \cdot 10^{-2}$ | | |
| 5 | | $1.45 \cdot 10^{-1}$ | | | | | $9.20 \cdot 10^{-2}$ | | |

Figure 2.13: Bayesian networks for *and* data: (a) $BN_{cursive-and}$, (b) $BN_{handprint-and}$, (c) table of marginal probabilities for cursive, and (d) table of marginal probabilities for handprint. The necessary CPTs for (a) are given in Table 2.5 and for (b) in Table 2.6.

(a) $X_4|X_2$

| | | | | | |
|-----------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| X_4/X_2 | 0 | 1 | 2 | 3 | 4 |
| 0 | 3.83×10^{-1} | 3.79×10^{-1} | 4.50×10^{-1} | 1.54×10^{-1} | 3.08×10^{-1} |
| 1 | 3.03×10^{-1} | 2.21×10^{-1} | 1.56×10^{-1} | 3.85×10^{-1} | 1.88×10^{-1} |
| 2 | 3.31×10^{-2} | 9.52×10^{-2} | 7.41×10^{-2} | 3.85×10^{-2} | 4.96×10^{-2} |
| 3 | 2.39×10^{-1} | 2.95×10^{-1} | 3.10×10^{-1} | 3.85×10^{-1} | 4.18×10^{-1} |
| 4 | 4.26×10^{-2} | 9.52×10^{-3} | 1.06×10^{-2} | 3.85×10^{-2} | 3.66×10^{-2} |

(b) $X_6|X_2$

| | | | | | |
|-----------|-----------------------|-----------------------|-----------------------|----------------------|-----------------------|
| X_6/X_4 | 0 | 1 | 2 | 3 | 4 |
| 0 | 4.50×10^{-2} | 1.32×10^{-2} | 1.33×10^{-2} | $4. \times 10^{-2}$ | 1.57×10^{-2} |
| 1 | 2.58×10^{-1} | 2.64×10^{-1} | 1.11×10^{-1} | 1.2×10^{-1} | 1.26×10^{-1} |
| 2 | 5.55×10^{-1} | 6.23×10^{-1} | 7.80×10^{-1} | 6.4×10^{-1} | 6.65×10^{-1} |
| 3 | 1.42×10^{-1} | $1. \times 10^{-1}$ | 9.55×10^{-2} | $2. \times 10^{-1}$ | 1.94×10^{-1} |

(c) $X_7|X_4$

| | | | | | |
|-----------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| X_7/X_4 | 0 | 1 | 2 | 3 | 4 |
| 0 | 4.26×10^{-1} | 5.28×10^{-1} | 3.86×10^{-1} | 4.95×10^{-1} | 3.52×10^{-1} |
| 1 | 3.98×10^{-1} | 2.87×10^{-1} | 4.94×10^{-1} | 3.27×10^{-1} | 2.96×10^{-1} |
| 2 | 1.18×10^{-1} | 1.36×10^{-1} | 8.30×10^{-2} | 1.10×10^{-1} | 1.48×10^{-1} |
| 3 | 5.81×10^{-2} | 4.96×10^{-2} | 3.73×10^{-2} | 6.78×10^{-2} | 2.04×10^{-1} |

(d) $X_3|X_4$

| | | | | | |
|-----------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| X_3/X_4 | 0 | 1 | 2 | 3 | 4 |
| 0 | 4.70×10^{-2} | 1.55×10^{-1} | 4.46×10^{-1} | 2.65×10^{-2} | 2.08×10^{-1} |
| 1 | 9.44×10^{-1} | 8.23×10^{-1} | 5.46×10^{-1} | 9.54×10^{-1} | 2.26×10^{-1} |
| 2 | 9.41×10^{-3} | 2.19×10^{-2} | 8.33×10^{-3} | 1.91×10^{-2} | 5.66×10^{-1} |

(e) $X_8|X_6$

| | | | | |
|-----------|-----------------------|-----------------------|-----------------------|-----------------------|
| X_8/X_6 | 0 | 1 | 2 | 3 |
| 0 | 7.14×10^{-2} | 2.14×10^{-1} | 1.98×10^{-1} | 1.68×10^{-1} |
| 1 | 1.61×10^{-1} | 2.90×10^{-1} | 2.79×10^{-1} | 1.21×10^{-1} |
| 2 | $5. \times 10^{-1}$ | 3.34×10^{-1} | 3.39×10^{-1} | 2.47×10^{-1} |
| 3 | 2.14×10^{-1} | 1.14×10^{-1} | 1.01×10^{-1} | 3.13×10^{-1} |
| 4 | 5.36×10^{-2} | 4.85×10^{-2} | 8.26×10^{-2} | 1.51×10^{-1} |

(f) $X_1|X_7$

| | | | | |
|-----------|-----------------------|-----------------------|-----------------------|-----------------------|
| X_1/X_7 | 0 | 1 | 2 | 3 |
| 0 | 9.67×10^{-2} | 1.74×10^{-1} | 1.96×10^{-1} | 7.57×10^{-2} |
| 1 | 3.22×10^{-1} | 3.96×10^{-1} | 3.90×10^{-1} | 3.41×10^{-1} |
| 2 | 3.96×10^{-1} | 2.64×10^{-1} | 3.15×10^{-1} | 2.92×10^{-1} |
| 3 | 1.85×10^{-1} | 1.66×10^{-1} | 9.94×10^{-2} | 2.92×10^{-1} |

(g) $X_5|X_3$

| | | | |
|-----------|-----------------------|-----------------------|-----------------------|
| X_5/X_3 | 0 | 1 | 2 |
| 0 | 5.21×10^{-1} | 1.60×10^{-1} | 3.07×10^{-1} |
| 1 | 3.30×10^{-3} | 2.51×10^{-2} | 1.33×10^{-2} |
| 2 | 4.59×10^{-1} | 8.02×10^{-1} | 4.53×10^{-1} |
| 3 | 1.65×10^{-2} | 1.33×10^{-2} | 2.27×10^{-1} |

Table 2.5: Conditional Probability Tables of characteristics of *and* for *cursive writing* needed in the BN shown in Figure 2.13 (a).

(a) $X_1|X_6$

| X_1/X_6 | 0 | 1 | 2 | 3 | 4 |
|-----------|-----------------------|-----------------------|---------------------|-----------------------|-----------------------|
| 0 | 9.56×10^{-1} | 8.15×10^{-1} | $2. \times 10^{-1}$ | 2.13×10^{-2} | 3.45×10^{-2} |
| 1 | 1.32×10^{-2} | 1.07×10^{-1} | $2. \times 10^{-1}$ | 1.06×10^{-1} | 3.45×10^{-2} |
| 2 | 1.10×10^{-3} | 5.95×10^{-3} | $2. \times 10^{-1}$ | 2.13×10^{-2} | 3.45×10^{-2} |
| 3 | 2.31×10^{-2} | 4.17×10^{-2} | $2. \times 10^{-1}$ | 8.09×10^{-1} | 3.45×10^{-2} |
| 4 | 1.10×10^{-3} | 5.95×10^{-3} | $2. \times 10^{-1}$ | 2.13×10^{-2} | 3.45×10^{-2} |

(b) $X_8|X_6$

| X_8/X_6 | 0 | 1 | 2 | 3 | 4 |
|-----------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| 0 | 3.45×10^{-1} | 3.77×10^{-1} | 2.50×10^{-1} | 4.78×10^{-1} | 3.57×10^{-2} |
| 1 | 6.38×10^{-1} | 4.97×10^{-1} | 2.50×10^{-1} | 2.17×10^{-2} | 3.57×10^{-2} |
| 2 | 6.59×10^{-3} | 1.20×10^{-2} | 2.50×10^{-1} | 1.74×10^{-1} | 3.57×10^{-2} |
| 3 | 9.89×10^{-3} | 1.14×10^{-1} | 2.50×10^{-1} | 3.26×10^{-1} | 3.57×10^{-2} |

(c) $X_5|X_1$

| X_5/X_1 | 0 | 1 | 2 | 3 | 4 |
|-----------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| 0 | 1.35×10^{-1} | 5.41×10^{-2} | 3.33×10^{-1} | 3.79×10^{-1} | 7.14×10^{-2} |
| 1 | 7.97×10^{-1} | 7.57×10^{-1} | 3.33×10^{-1} | 2.12×10^{-1} | 7.14×10^{-2} |
| 2 | 6.80×10^{-2} | 1.89×10^{-1} | 3.33×10^{-1} | 4.09×10^{-1} | 7.14×10^{-2} |

(d) $X_4|X_5$

| X_4/X_5 | 0 | 1 | 2 |
|-----------|-----------------------|-----------------------|-----------------------|
| 0 | 4.76×10^{-1} | 2.31×10^{-1} | 2.45×10^{-1} |
| 1 | 2.98×10^{-1} | 6.09×10^{-1} | 3.73×10^{-1} |
| 2 | 2.38×10^{-2} | 9.17×10^{-3} | 9.09×10^{-3} |
| 3 | 1.07×10^{-1} | 4.93×10^{-2} | 1.36×10^{-1} |
| 4 | 9.52×10^{-2} | 1.02×10^{-1} | 2.36×10^{-1} |

Table 2.6: Conditional Probability Tables of characteristics of *and* for *handprint writing* needed in the BN shown in Figure 2.13 (b).

$$P(X_a|X_i \in (X - x_a)) = \frac{P(X)}{\sum_{X_a} P(X)}, \tag{2.6}$$

where, $P(X)$ is the joint probability calculated using the Bayesian Network models.

To sample a value of X_a from $P(X_a|X_i \in (X - x_a))$, we divide the real number line into M parts such that each m^{th} part is proportional to $P(X_a = m|X_i \in (X - x_a))$ as shown in Figure 2.14. Then, a random number between 0 and 1 is generated using the Matlab rand(1) function and its location on the number line is determined. If the random number lies in the m^{th} location on the line, the value of $X_a^{(\tau+1)}$ is set to m . Samples generated for the cursive and hand-print datasets are shown in Table 2.7.

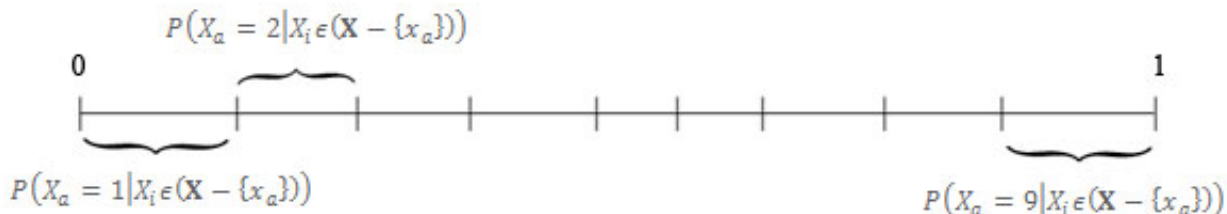
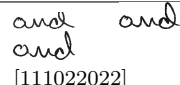
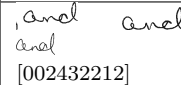
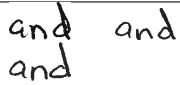
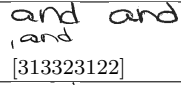
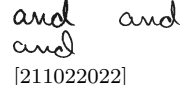
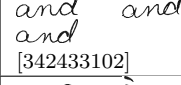
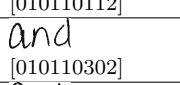
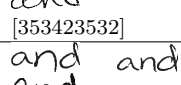
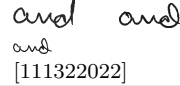
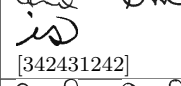
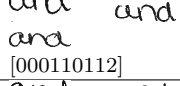
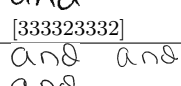
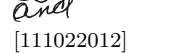
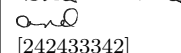
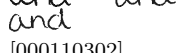
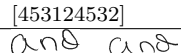
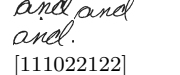
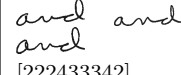
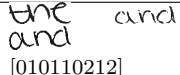
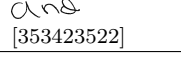


Figure 2.14: Dividing the number line into nine intervals for sampling.

Table 2.7: Examples of samples with highest and lowest joint probabilities: (a) Cursive- Highest (b) Cursive- Lowest (c) Handprint- Highest (d) Handprint- Lowest. The samples were obtained by Gibbs sampling of BNs.

| (a) Cursive-Common | | (b) Cursive-Rare | | (c) Handprint-Common | | (d) Handprint-Rare | |
|--|-----------------------|--|------------------------|---|-----------------------|--|------------------------|
| Samples with Characteristics | Probability | Samples with Characteristics | Probability | Samples with Characteristics | Probability | Samples with Characteristics | Probability |
|  [111022022] | 5.47×10^{-3} |  [002432212] | 2.50×10^{-9} |  [010110112] | 1.67×10^{-2} |  [313323122] | 1.31×10^{-9} |
|  [211022022] | 5.10×10^{-3} |  [342433102] | 1.24×10^{-9} |  [010110302] | 1.64×10^{-2} |  [353423532] | 8.38×10^{-10} |
|  [111322022] | 4.41×10^{-3} |  [342431242] | 3.2×10^{-10} |  [000110112] | 1.29×10^{-2} |  [333323332] | 6.69×10^{-10} |
|  [111022012] | 4.32×10^{-3} |  [242433342] | 1.63×10^{-10} |  [000110302] | 1.27×10^{-2} |  [453124532] | 4.39×10^{-10} |
|  [111022122] | 4.21×10^{-3} |  [222433342] | 1.61×10^{-10} |  [010110212] | 8.43×10^{-3} |  [353423522] | 2.85×10^{-10} |

2.3.3 Markov Networks

Markov networks, or Markov random fields (MRFs), represent probability distributions using undirected graphs. They have been used to represent joint probability distributions in a variety of domains: language processing, image segmentation and restoration, statistical physics, signal processing, computational biology, etc. Although MRFs have been used for several decades, significant open problems still remain in their use. Here we consider the task of discrete MRF structure learning from a given data set. The problem is to identify the MRF structure with a bounded complexity, which most accurately represents a given probability distribution, based on a set of samples from the distribution. By MRF complexity we mean the number of features in the log-linear representation of the MRF. This problem is proved to be *NP*-hard [42].

A lot of attention has been drawn to the problem of structure learning over the past years [3, 64, 47, 55, 91, 68, 24, 9, 30, 27, 65]. Most existing approaches typically focus on specific parametric classes of MRFs. For example, the majority of existing methods allow only for MRFs with pairwise variable dependencies. MRF structure learning algorithms can be roughly divided into two groups by their underlying approach [44].

Constraint-based approach lies in the estimating conditional independences of variables using hypothesis testing on a given data set. Score-based approach defines a score function for any model structure, e.g. log-likelihood with the maximum likelihood parameters. A search algorithm can then be used to obtain the MRF structure with the optimal score. Score-based approach is typically more flexible because there exists a variety of scoring functions and search algorithms, however generally it is more computationally expensive than the constraint-based approach. The latter one often lacks robustness in the presence of noise in the data set and typically requires a large number of samples.

A fast algorithm for MRF structure learning was recently proposed [3]. It is based on evaluating conditional mutual information as inspired by the early works of Chow and Liu [18]. The algorithm introduces a pairwise factor into the MRF structure for each pair of variables with high conditional mutual information value. Another generalization of the Chow-Liu ideas is described in [16], where a greedy approach (step by step construction of the model) is combined with feature selection using KL-divergence gain. Using the constraint that the resultant graphical model should be a bounded-treewidth ('thin') junction tree, they preserve low computational complexity of running inference.

Another recently designed and already popular algorithm for MRF structure learning is described in [64] where the neighborhood of any given variable is estimated by performing logistic regression subject to an L_1 -

constraint. It was originally designed for the Ising model (the pairwise symmetric binary model) selection. The algorithm can be easily generalized to solve discrete pairwise MRF structure learning problems, but not the problem of MRF structure learning with arbitrary size factors. The same approach can be combined with trace norm regularization [30], which was shown to lead to sparse models.

A common deficiency of the above-stated algorithms is that they allow to treat MRFs with pairwise factors only.

A promising algorithm for MRF structure learning that uses a so called 'bottom-up' approach is described in [24]. The algorithm starts by treating each sample in the data set as a feature in the MRF. It considers generalizing each feature to match its k nearest previously unmatched features by dropping variables. If introduction of a generalized feature improves the model's objective function value than it is kept in the model. The algorithm works with features of arbitrary size, but it tends to dense models, which complicates inference. Another way for further improvements is the use of not only positive but also negative variables correlations.

A widely used algorithm for MRF structure learning, called *Grafting* [59], uses a greedy approach. Starting with the MRF with no features in the log-linear representation, i.e., structure without edges, features that offer the greatest improvement to the objective function value are added iteratively. The effect of introducing a particular feature is estimated by computing the gradient of the objective function w.r.t. the features' weights. A successful improvement to this algorithm is known as *Grafting-Light* [91]: it mixes new feature selection and parameter estimation steps that used to be done separately. A similar algorithm for MRF structure learning, also based on the greedy approach, defines the effect of a feature entering the MRF as the increase in the objective function value resulting from the introduction of the feature [47]. The objective function is a sum of the log-likelihood of a given data set and L_1 -regularization on weights.

The main disadvantage of algorithms based on the greedy approach is the need to evaluate a specified expression for *each possible* feature not yet included in the model. This means running inference for the MRF with *each possible* feature at least once at each greedy algorithm iteration, which becomes intractable for problems with even few variables. However, the advantage of the greedy approach is that it is not limited to MRFs with pairwise features only, i.e. features defined on more than two variables can be also considered.

Here we describe an algorithm for structure learning of discrete MRFs with arbitrary size factors that is based on the greedy approach and uses a special heuristics for the search space reduction. The greedy approach allows for features of arbitrary sizes, while heuristics makes the algorithm faster compared to the existing MRF structure learning algorithms relying on the greedy approach. We call it the Fast Greedy Algorithm (FGA) for MRF structure learning since it can be viewed as an improvement and a generalization of the greedy algorithm described in [47].

We conducted a set of experiments on real-world and simulated data sets to assess the performance of the designed algorithm and compare it with the original greedy algorithm [47] and some other state of the art algorithms. The data sets and the FGA source code are available online.

The rest of the section is organized as follows: (i) formal definitions and key concepts, (ii) statement of structure learning problem, (iii) algorithm design, and (iv) experimental results.

Preliminaries

We use a framework of log-linear discrete Markov Random Fields (MRFs) [44]. Let $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ be a set of discrete-valued random variables. The log-linear model is a representation of a joint probability distribution over assignments is \mathbf{X} :

$$P(\mathbf{X} : \theta) = \frac{1}{Z(\theta)} \exp \left(\sum_{i=1}^k \theta_i f_i(D_i) \right) \quad (2.7)$$

where $f_i(D_i)$ is a binary feature function defined over variables $D_i \in \mathbf{X}$, the set of all used feature functions is denoted as $F = \{f_i(D_i)\}_{i=1}^k$, k is a number of features in the model, $\theta = \{\theta_i : f_i \in F\}$ is a set of features' weights, $Z(\theta) = \sum_{\xi} \exp \left(\sum_{i=1}^k \theta_i f_i(\xi) \right)$ is a partition function, $f_i(\xi)$ is a shortened notation for $f_i(\xi \langle D_i \rangle)$ with a given assignment to the set of variables D_i .

For parameter learning of the MRF with a fixed structure we use the maximum likelihood estimation (MLE) approach. The log-likelihood and its partial derivative with respect to θ_i are:

$$l(F, \theta : S) = \sum_{i=1}^k \theta_i (\sum_{s \in S} f_i(\xi[s])) - M \cdot \ln Z(\theta) \tag{2.8}$$

$$\frac{\partial}{\partial \theta_i} \frac{1}{M} l(F, \theta : S) = E_S [f_i] - E_P [f_i]$$

where M is the number of samples in data set S , $E_S [f_i]$ is the empirical probability of f_i given data set S , $E_P [f_i]$ is the expected value of f_i w.r.t. the probability distribution defined by P .

Structure Learning Problem

Assume that data set S consists of M samples. Each sample is an instantiation of the set of variables \mathbf{X} . The structure learning problem is as follows: *With a given algorithm for the MRF parameter estimation find a set of features $F = \{f_i(D_i)\}_{i=1}^k$ of size no greater than a fixed bound C that maximizes the likelihood $l(F, \theta : S)$:*

$$\{f_i(D_i)\}_{i=1}^k = \arg \max_{F'} l(F', \theta_{F'} : S), k \leq C$$

where $\theta_{F'}$ is a set of weights estimated for the MRF with the fixed set of features F' .

A comparatively low number of weights θ is one of the main advantages of using MRFs for representing a joint probability distribution. To control the maximum number of weights, a boundary condition is set on the number of features (without this condition the trivial case of including every possible feature to the MRF would be the optimal solution for the stated problem).

Running inference is a computationally difficult task for MRFs that arises frequently. Since computational complexity of inference is determined by the connectivity in the MRF structure [44], we define the extended problem by adding a bound on feature domain sizes:

$$\max_{f_i \in F} R(f_i) \leq r$$

where $R(f_i)$ is the size of the domain of feature function f_i .

Fast Greedy Algorithm for MRF Structure Learning (FGAM)

Since the MRF structure learning problem NP -hard [42], we use an approximate solution. It is a greedy approach that has no restrictions to any particular class of features. Starting with the MRF without any features (the model where all variables are disjoint), features are introduced one at a time. At each iteration, a feature that brings maximum increase in the objective function value is selected. The key idea of the FGA is that the number of candidate features to enter the MRF can be limited to just two subsets. These subsets contain features whose empirical probability differs most from their expected value with respect to the probability distribution defined by the MRF with the current structure. The methods for estimating the gain in the objective function value and constructing the subsets are described further in this section. The algorithm terminates when no feature can bring an objective function gain that exceeds a pre-set threshold or when the maximum allowed number of features in the MRF is reached. See Algorithm 3 (FGAM) for more details.

Objective Function

We define the objective function as the average log-likelihood of the data set, see expr. 2.8, with weights θ defined using the MLE approach. The gain in the objective function for feature f_{k+1} , given MRF Q and data set S , is defined as the increase in the objective function value resulting from the introduction of f_{k+1} to MRF Q :

$$G_Q[f_{k+1}] = \frac{1}{M} (l(\{f_i\}_{i=1}^{k+1}, \{\theta''\}_{i=1}^{k+1} : S) - l(\{f_i\}_{i=1}^k, \{\theta'\}_{i=1}^k : S)) \tag{2.9}$$

If there is no upper bound on the number of MRF features in the problem definition then we add

Algorithm 3 FGAM: A fast greedy algorithm for MRF structure learning

```

1: Input: Data set  $S = \{s_i\}_{i=1}^M$ 
2: Output: A set of features in the MRF  $F = \{f_i\}_{i=1}^k$  with their weights  $\theta = \{\theta\}_{i=1}^k$ 
3:  $Set_S \leftarrow FeaturesWithHighEmpiricalProbability(S)$  // see Algorithm 4
4:  $F \leftarrow \emptyset$ ;  $\theta \leftarrow \emptyset$ 
5: repeat
6:    $f_{best} \leftarrow SelectBestFeature(Set_S)$ 
7:   if  $Gain[f_{best}] < threshold$  then break; end if
8:    $F \leftarrow F \cup f_{best}$ ;  $\theta \leftarrow \theta \cup ApproximateValue(\theta_{best})$ 
9:    $Set_Q \leftarrow FeaturesWithHighExpectedValue(F, f_{best})$  // see Algorithm 5
10:  for  $i = 1 \rightarrow numOfIterations$  do
11:     $f_{best} \leftarrow SelectBestFeature(Set_Q)$ 
12:    if  $Gain[f_{best}] < threshold$  then break; end if
13:     $F \leftarrow F \cup f_{best}$ ;  $\theta \leftarrow \theta \cup ApproximateValue(\theta_{best})$ 
14:  end for
15:   $\theta \leftarrow ParametersOptimization(F, \theta)$ 
16: until  $Set_S = \emptyset \vee SizeOf(F) \geq C$ 
    
```

L_1 -regularization on weights to the objective function. In this case $G_Q[f_{k+1}]$ takes the form:

$$\frac{1}{M} (l(\{f_i\}_{i=1}^{k+1}, \{\theta''\}_{i=1}^{k+1} : S) - l(\{f_i\}_{i=1}^k, \{\theta'\}_{i=1}^k : S)) - \frac{1}{\beta} \left(\sum_{i=1}^{k+1} |\theta''_i| - \sum_{i=1}^k |\theta'_i| \right) \quad (2.10)$$

where β is a regularization parameter.

To compute $G_Q[f_{k+1}]$ we have to optimize parameters of the MRF with feature f_{k+1} : $\{\theta''_i\}_{i=1}^{k+1}$. This operation is computationally expensive since it involves running inference and numerical optimization. However, we can use the following observation: introduction of one feature f_{k+1} to the MRF structure only marginally effects the weights of the previously added features: $\theta'_i \approx \theta''_i$, $i = \overline{1, k}$. Hence, the effect of feature entry at each iteration of Algorithm 3 can be assessed under the approximating assumption that all other features' weights hold still: $\theta'_i = \theta''_i$, $i = \overline{1, k}$. In this case, a closed-form solution [62] for the optimal value of the parameter θ_{k+1} is obtained:

$$\theta_{k+1} = \log \frac{(1 - E_Q[f_{k+1}]) \cdot E_S[f_{k+1}]}{E_Q[f_{k+1}] \cdot (1 - E_S[f_{k+1]})} \quad (2.11)$$

for feature f_{k+1} such that: $\exists s_1, s_2 \in S$ $f_{k+1}(\xi[s_1]) = 1$, $f_{k+1}(\xi[s_2]) = 0$.

When L_1 -regularization on weights is used in the objective function, expression 2.11 for the optimal value of the parameter θ_{k+1} takes the following form [47]:

$$\theta_{k+1} = \log \frac{(1 - E_Q[f_{k+1}]) \cdot \left(E_S[f_{k+1}] - \frac{1}{\beta} \text{sign } \theta_{k+1} \right)}{E_Q[f_{k+1}] \cdot \left(1 - E_S[f_{k+1}] + \frac{1}{\beta} \text{sign } \theta_{k+1} \right)} \quad (2.12)$$

After substituting the expression for θ_{k+1} the objective function gain takes the following form:

$$G_Q[f_{k+1}] = E_S[f_{k+1}] \log \frac{E_S[f_{k+1}]}{E_Q[f_{k+1}]} + (1 - E_S[f_{k+1}]) \log \frac{1 - E_S[f_{k+1}]}{1 - E_Q[f_{k+1}]} \quad (2.13)$$

Hence, to calculate the objective function gain $G_Q[f_{k+1}]$ for any feature from a given set we only need to evaluate the empirical probability $E_S[f_{k+1}]$, given data set S , and the expected value $E_Q[f_{k+1}]$ with respect to the probability distribution defined by MRF Q . The empirical probability can be evaluated efficiently after certain transformations of data set S , see section 2.3.3.

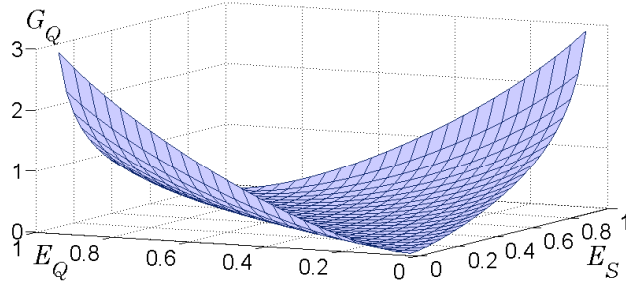


Figure 2.15: Gain G_Q in the objective function, see expr. 2.13, for a given feature f with empirical probability E_S and expected value E_Q w.r.t. the probability distribution defined by MRF Q .

Algorithm 4 Constructing a set of features with high empirical probabilities

- 1: **Input:** Data set $S = \{s_i\}_{i=1}^M$ and the maximum allowed size of features r .
 - 2: **Output:** A set of features with high empirical probability Set_S .
 - 3: $F_I \leftarrow \{f^{ij} : f^{ij}(X) = \mathbf{1}\{X_i = j\} \quad \forall i, \forall j\}$
 - 4: $V^{ij}[m] = \mathbf{1}\{f^{ij}(\xi[s_m]) = 1\}$, $m = \overline{1, M}$, $\forall f^{ij} \in F_I$
 - 5: $F_I^{sel} \leftarrow \{f^{ij} : f^{ij} \in F_I, \frac{1}{M} \sum_{m=1}^M V^{ij}[m] \geq \text{threshold}\}$
 - 6: $Set_S \leftarrow \{f^{1j} : f^{1j} \in F_I^{sel}, \forall j\}$
 - 7: **for** $k = 2 \rightarrow n$ **do**
 - 8: $F^{cur} \leftarrow \{f^{kj} : f^{kj} \in F_I^{sel}, \forall j\} \cup \{f_c : f_c = f_i \wedge f^{kj}, R(f_c) \leq r, \forall f_i \in Set_S, \forall f^{kj} \in F_I^{sel}\}$
 - 9: $V_c[m] = V_i[m] \cdot V^{kj}[m]$, $m = \overline{1, M}$, $\forall f_c = (f_i \wedge f^{kj}) \in F^{cur}$
 - 10: $Set_S \leftarrow Set_S \cup \{f_h : f_h \in F^{cur}, \frac{1}{M} \sum_{m=1}^M V_h[m] \geq \text{threshold}\}$
 - 11: **end for**
 - 12: $Set_S \leftarrow Set_S \setminus F_I^{sel}$
-

Feature Search Space Reduction

The form of expression 2.13 for the objective function gain $G_Q[f]$ for a given feature f (see Figure 2.15) leads to the following key observation: there are two regions in the space of $[E_S, E_Q]$ where G_Q takes high values: **1.** $E_S[f]$ is high and $E_Q[f]$ is low; **2.** $E_Q[f]$ is high and $E_S[f]$ is low. At each iteration of the designed Algorithm 3 a feature f with the maximum value of $G_Q[F]$ is selected, i.e. a feature whose $(E_S[f], E_Q[f])$ belong to one of these two regions. Hence, the feature search space can be reduced to two sets that correspond to the stated regions, ensuring lower computational complexity without loss in accuracy. The problem now is how to efficiently construct these two sets of features. We address this challenge by approximating these regions with larger ones: **1a.** $E_S[f]$ is high; **2a.** $E_Q[f]$ is high. The first method that constructs a set of features corresponding to the region **1a**, see section 2.3.3, can be called before learning the structure of the MRF, see Algorithm 3. The second method is then used to reduce the search space to a set of features with expected values $E_Q[f]$ corresponding to the region **2a**, as described further in section 2.3.3.

Construction of a Set of Features with High Empirical Probabilities

Definition 1. A binary feature function f is called an individual feature if and only if it can be represented in the following form: $f(X) = \mathbf{1}\{X_i = j\}$, $i \in [1, n]$, $j \in \text{Range}(X_i)$, where n is a number of variables in X ; $\mathbf{1}\{X_i = j\} = \begin{cases} 1 & \text{if } X_i = j \\ 0 & \text{otherwise} \end{cases}$. We will denote it as: $f^{ij} = \mathbf{1}\{X_i = j\}$.

For an individual feature f^{ij} and data set S we define a binary vector V^{ij} of size M :

$$V^{ij}[m] = \mathbf{1}\{f^{ij}(\xi[s_m]) = 1\}, \quad m = \overline{1, M}$$

where: s_m is m -th sample in data set S and M is the number of samples in data set S .

Any binary feature f for discrete valued variables \mathbf{X} can be represented as a finite conjunction of a set of

Algorithm 5 Constructing a set of features $\{f\}$ with high expected values $E_Q[f]$ with respect to the probability distribution defined by MRF Q

```

1: Input: A set of features  $F = \{f_i\}$  that are used in MRF  $Q$  with the corresponding weights  $\{\theta\}$  and
   the last added to MRF  $Q$  feature  $f_{last}$ .
2: Output: Set of features  $Set_Q\{f\}$  with high expected values  $E_Q[f]$ .
3:  $F_I \leftarrow \{f^{ij} : \exists f_k \in F, f^{ij} \in f_k, \theta_k > 0\}$ 
4:  $F_I^{last} \leftarrow \{f^{ij} : f^{ij} \in f_{last}, \forall i, \forall j\}$ 
5:  $Set_Q \leftarrow \emptyset$ 
6: for all  $f^{ij} \in F_I^{last}$  do
7:    $F^{cur} \leftarrow \{f^{ij}\}$ 
8:   for all  $f^{kl} \in F_I$  do
9:     for all  $f^c \in F^{cur}$  do
10:      if  $(R(f^c) < r) \wedge (D^{kj} \cap D^c = \emptyset)$  then //where:  $D^c$  - is a domain of the feature  $f^c$ 
11:         $F^{cur} \leftarrow F^{cur} \cup \{f : f = f^{kj} \wedge f^c\}$  end if
12:     end for
13:   end for
14:    $Set_Q \leftarrow Set_Q \cup F^{cur} \setminus F \setminus F_I$ 
15: end for

```

individual features F_I . The corresponding binary vector V for feature f is: $V[m] = \prod_{f^{ij} \in F_I} V^{ij}[m]$, $m = \overline{1, M}$. Empirical probability $E_S[f]$ for feature f estimated using data set S is equal to the average value of the elements in vector V .

Algorithm 4 constructs a set Set_S of features with empirical probabilities exceeding a given threshold. It uses the idea of selecting individual features with empirical probabilities exceeding the threshold and combining them in every possible way. Since the number of features in Set_S grows exponentially with the number of variables in \mathbf{X} , it is managed through choice of threshold value.

Construction of a Set of Features with High Expected Values

To construct Set_Q containing features with high expected values $E_Q[f]$ with respect to the probability distribution defined by MRF Q Algorithm 5 can be used. It relies on the following observation: features $\{f\}$ with high expected values $E_Q[f]$ include individual features that are already used in MRF Q with positive weights. Algorithm 5 constructs all possible conjunctions of such individual features.

To limit the number of features in Set_Q we use the following observation: all of the features in Set_Q must include at least one individual feature from set F_I^{last} of the individual features that were added to the MRF at the last iteration of the Algorithm 3. The reason is that all the features that do not include any of the elements of F_I^{last} were considered in the previous iterations. We assume that the expected value of a feature considered earlier (that does not include any of F_I^{last}) is not increasing with introduction of a new feature that consists of F_I^{last} to the MRF with a positive weight.

Empirical Evaluation

Algorithm 3 (FGAM) for MRF structure learning was evaluated and compared with baseline algorithms with simulated and real-world data sets. The rationale for the choice of baseline algorithms is as follows. The designed algorithm is a generalization of the greedy algorithm described in [47] (which corresponds to using zero values of thresholds and parameter 'numOfIterations' in Algorithms 3, 4 and 5); the latter is based on the ideas of [62] and comparisons can be found with other state of the art methods. Hence, we compare the designed algorithm with the original greedy algorithm for structure learning of MRFs with arbitrary size factors described in [47]. Additionally, on the handwriting data set we also applied the algorithm described in [3] in order to demonstrate that using features of arbitrary size in comparison to pairwise features can possibly lead to better accuracy in real-world applications.

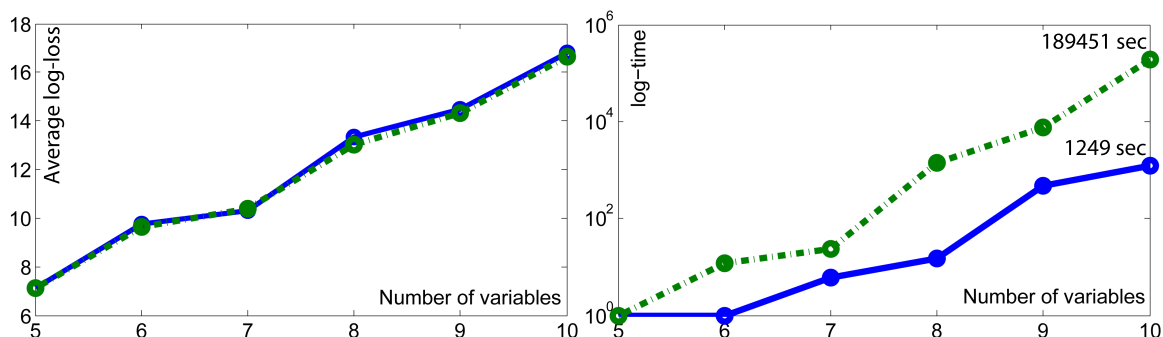


Figure 2.16: Accuracy and construction time comparison of the greedy algorithm with L_1 -regularization on weights [47] (green dot line) and the designed fast greedy algorithm (blue solid line) for data sets with different number of variables.

Test Results with Simulated Data. To assess the performance of the discussed algorithms on problems of varying complexities we created a synthetic data set generator from multivariate normal distribution. The generated values were rounded to the nearest integer from a given interval. The range of each variable $X_i \in \mathbf{X}$ was chosen uniformly on the interval $[1, 20]$ and the distribution parameters were also uniform on variables' ranges. The number of variables in \mathbf{X} in the generated data sets varied from 5 to 10. The size of each data set was equal to 200 times the number of variables in the data set.

Figure 2.16 shows the results of applying the greedy algorithm with L_1 -regularization on weights [47] and the presented FGA algorithm to the generated data sets. We used 3-fold cross validation and 3 data sets for each number of variables. The first graph in Figure 2.16 indicates that the accuracy of both algorithms was almost the same for all the test runs. The second graph shows that the execution times for the FGA algorithm are several orders of magnitude smaller compared to the the original greedy algorithm (1249 sec compared to 189451 sec for the data set with 10 variables on a system with a dual core processor - Intel Core i5-2410M).

Test Results with Handwriting Data. Consider a task of identifying unusual and unique characteristics, i.e., rare letter and word formations. Rarity is the reciprocal of probability. Using the most commonly occurring letter pair th and the characteristics specified in Table 2.2, the highest probability formations and low probability formations in a database were determined. The most crucial task in this problem is learning the probabilistic model structure [72]. The sample of th in Figure 2.17(f) is jointly encoded using the specified characteristics as $\{r^3, t^4, a^1, c^1, b^2, s^3\}$ and the sample in Figure 2.17(g) – as $\{r^0, t^2, a^0, c^3, b^1, s^2\}$. The data set contains 3,125 images authored by 528 individuals, some of whom contributed just one sample while some – upto 7.

Figure 2.17 shows five alternative MRF structures (a)-(e): both the first and the second were constructed independently by domain experts, the third one was created by an algorithm based on thresholding conditional mutual information [3], the fourth resulted from the general greedy algorithm with L_1 -regularization [47], and the fifth MRF is the result of the FGA. Performance of each of the MRF structure learning methods is summarized in Table 2.8. Weights were obtained using maximum likelihood estimation (MLE) with L-BFGS [49] as the optimization algorithm. Table 2.8 shows structure learning time (on a system with a dual core processor Intel Core i5-2410M) and average log-loss (negative log-likelihood) obtained using 3-fold cross-validation. Probabilities of the most common and rare th formations in the data sets are also shown in the table. The designed fast greedy algorithm gives about the same accuracy as the original greedy algorithm [47] but the required construction time is significantly lower.

In terms of the results quality, the instance in Figure 2.17(f) has the highest probability with all the tested MRFs. The instance in Figure 2.17(g) is among the top ten samples with the lowest probability in the data set for all the MRFs, which indicates that handprint writing is more common than cursive in the database.

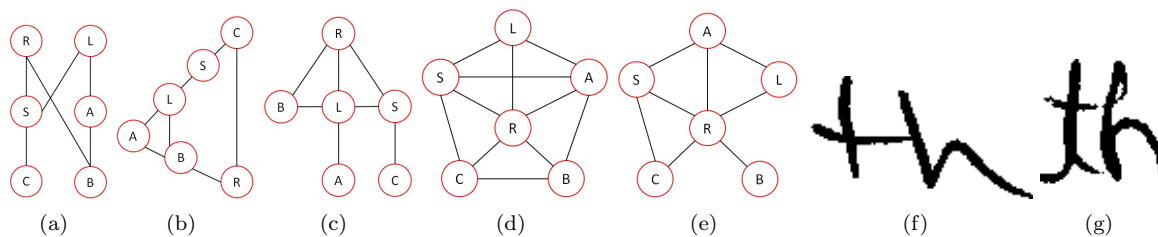


Figure 2.17: Candidate MRFs: first two were manually constructed, the third – by the modified Chow-Liu algorithm, fourth MRF – by the original greedy algorithm and the fifth – by the proposed FGAM; and (f) the highest probability ‘th’ in data set and (g) a low probability ‘th’.

Table 2.8: Results of MRF structure learning with 3-fold cross-validation.

| MRF index | Construction method | Time of structure learning (sec) | Average log-loss | Probability of ‘th’ in Fig. 2.17(f) | Probability of ‘th’ in Fig. 2.17(g) |
|-----------|----------------------|----------------------------------|------------------|-------------------------------------|-------------------------------------|
| MRF_1 | Manual | n/a | 6.428 | $1.59 \cdot 10^{-2}$ | $16 \cdot 10^{-6}$ |
| MRF_2 | Manual | n/a | 6.464 | $1.82 \cdot 10^{-2}$ | $6 \cdot 10^{-6}$ |
| MRF_3 | Mod. Chow-Liu | 2 | 6.426 | $1.7 \cdot 10^{-2}$ | $5 \cdot 10^{-6}$ |
| MRF_4 | Greedy w. L_1 -reg | 53 | 6.326 | $1.94 \cdot 10^{-2}$ | $25 \cdot 10^{-6}$ |
| MRF_5 | FGA | 7 | 6.328 | $1.8 \cdot 10^{-2}$ | $21 \cdot 10^{-6}$ |

2.4 Methods: Statistical Inference

In Sections 2.3.2 and 2.3.3 we have described how to construct directed and undirected probabilistic graphical models from data consisting of characteristics of handwriting samples. Now we consider the problem of given such a model of the probability distribution, how to evaluate the probabilities of interest in forensics.

Two types of probabilistic queries are of potential interest in QD examination: (i) those relating to a given handwritten item, e.g., a single handwritten item which is either known or questioned, and (ii) those relating to the comparison of two handwritten items, e.g., both questioned and known. We refer to the first as the *probability of evidence*: a high probability of characteristics implies that the characteristics relate to a class, and a low probability implies rarity or individualizing characteristics. Another probability relating to a single handwritten item is the probability of random correspondence (PRC) which can be calculated from the entire distribution of the characteristics. A third probability relating to a given handwritten item is that of type, e.g., probability of cursive/hand-print.

The second type of probabilistic query relates to both the evidence and known, called as the *probability of identification*. This probability is useful for forming a forensic conclusion. While computing the probability of evidence is computationally complex, probability of identification is even more so.

2.4.1 Probability of Evidence

Type of Writing

One of the first decisions the QD examiner has to make is whether a given handwritten item is cursive or hand-printed. This can be formulated probabilistically as discussed in Appendix 5. For the propose of statistical analysis of characteristics, the handprint/cursive decision was made manually by QD examiners.

Rarity of Characteristics

An important decision for the QD examiner is to determine individualizing characteristics, as opposed to class characteristics which are common to a group. This means that we are interested in characteristics that are *unusual* or *rare* or *have a low probability*. Thus the simplest probabilistic query of interest to the QD examiner is the probability of the characteristics found in the handwritten item. This is evaluated in the case of a Bayesian network by using Eq. 2.1 and in the case of a Markov network by Eq. 2.7.

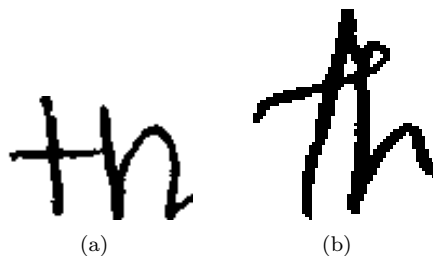


Figure 2.18: Examples of rarity evaluation: (a) the highest probability *th* in data set, and (b) a low probability *th*.

Since rarity is the reciprocal of probability its definition follows from that of probability. We can formally define rarity in discrete and continuous spaces as follows.

Def. 1 (discrete)

Given a probability space (Ω, \mathcal{F}, P) , the *rarity* of a random event $\xi \in \mathcal{F}$ is defined by

$$R(\xi) = \frac{1}{P(\xi)}, \quad (2.14)$$

where Ω is the sample space, $\mathcal{F} \subseteq 2^\Omega$ is the set of events, P is the probability measure, and $P(\xi) (\neq 0)$ is the probability of the event ξ .

Def. 2 (continuous)

Let $\mathbf{x} = (x_1, \dots, x_n)^T$ be a *continuous* n -dimensional random vector with the p.d.f. $p(\mathbf{x})$ defined on a domain S . Suppose for every assignment of $\mathbf{x} \in S$, there is a *confidence interval* $(\mathbf{x} - \epsilon/2, \mathbf{x} + \epsilon/2)$ associated with \mathbf{x} at a given confidence level $1 - \alpha$, where $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$, and α is a small positive value less than 1. This interval $(\mathbf{x} - \epsilon/2, \mathbf{x} + \epsilon/2)$ is a n -dimensional *region* whose volume is $\prod_{i=1}^n \epsilon_i$. Then the rarity of the event that \mathbf{x} takes value \mathbf{x}_0 is given by

$$R(\mathbf{x} = \mathbf{x}_0) = \frac{1}{\int_{\mathbf{x}_0 - \epsilon/2}^{\mathbf{x}_0 + \epsilon/2} p(\mathbf{x}) d\mathbf{x}}. \quad (2.15)$$

Since the magnitude of ϵ is usually small, the density in the region $(\mathbf{x}_0 - \epsilon/2, \mathbf{x}_0 + \epsilon/2)$ can be considered as constant, therefore (2.15) can be approximated by

$$R(\mathbf{x} = \mathbf{x}_0) = \frac{1}{p(\mathbf{x}_0) \prod_{i=1}^n \epsilon_i}, \forall \mathbf{x}_0 \in S. \quad (2.16)$$

Examples of Rarity Evaluation:

Probabilities assigned by BN_{th} to each element in the database was evaluated using Eq. 2.1. The highest probability assigned by the model is to the feature value $\{r^1, l^0, a^0, c^0, b^2, s^1\}$ with probability 0.0304. It corresponds exactly to the features assigned to writer 100 in the database whose writing is shown in Figure 2.18(a). The lowest probability assigned is to $\{r^2, l^3, a^2, c^2, b^0, s^4\}$ with value 7.2×10^{-8} which does not have a corresponding element in the database. A low probability *th* is shown in Figure 2.18(b) $\{r^3, l^1, a^0, c^2, b^0, s^1\}$.

Examples of highest and lowest probability *and* are shown in the examples of Table 2.7 where common and rare forms of handwritten *and* are given for cursive and hand-print cases together with probabilities. The rare cases are those with low probabilities assigned by the model, i.e., they are in the tail of the distribution. The corresponding characteristics can be considered to be individualizing. The common ones, i.e., those with high probabilities, are more representative of class characteristics.

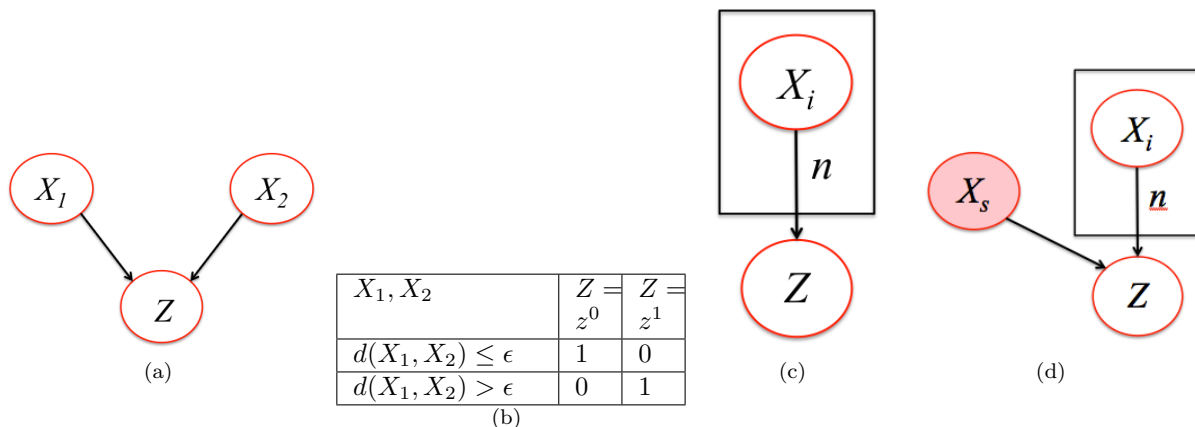


Figure 2.19: Graphical models for determining random correspondence: (a) PRC, the probability of two samples having the same value within ϵ , where the *indicator* variable Z : $P(Z|X_1, X_2)$ has the distribution shown in (b), (c) the probability of some pair of samples among n having the same value, n PRC, and (d) conditional n PRC, the probability of finding X_s among n samples and.

Random Correspondence

For a distribution $P(X)$, the Probability of Random Correspondence (PRC), can be defined as the probability that a random pair of samples have the same value. Since PRC is a function of the distribution, it is mathematically a *functional* just like entropy. It is a measure of the discriminatory power of X .

We can extend PRC to define the probability that at least one pair among n have the same value, called n PRC. We can also define the conditional PRC as the probability that a known value X_s is found among n such samples[80]. These definitions are formalized below.

PRC

Probability that two independent, identically distributed samples X_1 and X_2 , each with distribution $P(X)$, have similar values is given by the graphical model (Bayesian network) in Figure 2.19(a). It is evaluated as follows:

$$\rho = P(Z = z^0) = \sum_{X_1} \sum_{X_2} P(z^0|X_1, X_2)P(X_1)P(X_2) \quad (2.17)$$

where Z is a binary indicator variable which takes values $\{z^0, z^1\}$ and has the *deterministic* CPD [44] shown in Figure 2.19 (b), also given by

$$P(z^0|X_1, X_2) = \begin{cases} 1 & \text{if } d(X_1, X_2) \leq \epsilon \\ 0, & \text{otherwise,} \end{cases} \quad (2.18)$$

d measures the difference between its arguments, the quantity ϵ represents as to how different two samples can be while they are considered to correspond (be the same), and $P(z^1) = 1 - P(z^0)$. The quantity ϵ is a tolerance that takes value 0 when the two variables X_1 and X_2 are identical. If d is the number of characteristics that are the same then $\epsilon = 1$ would lead to X_1 and X_2 being considered to be the same if they do not differ in more than one variable.

The PRC of *th* with $\epsilon = 0$, i.e., exact match, and determined using the BN was evaluated to be 2.62×10^{-13} . Using the BN models for *and* constructed as in Figure 2.12, the PRC was evaluated as: 7.90×10^{-4} for cursive and 6.85×10^{-3} for hand-print. This value of PRC can be used to compare the discriminative power of *and* to those of other letters and combinations. For instance, cursive writing is more individualistic than hand-print.

It should be noted that the PRC can be computationally intensive. The double summation in Eq. 2.17 indicates the root of the problem. When $\epsilon = 0$ the evaluation can proceed without having an inordinately large number of joint probabilities. However, if we were to allow correspondence with one or more character-

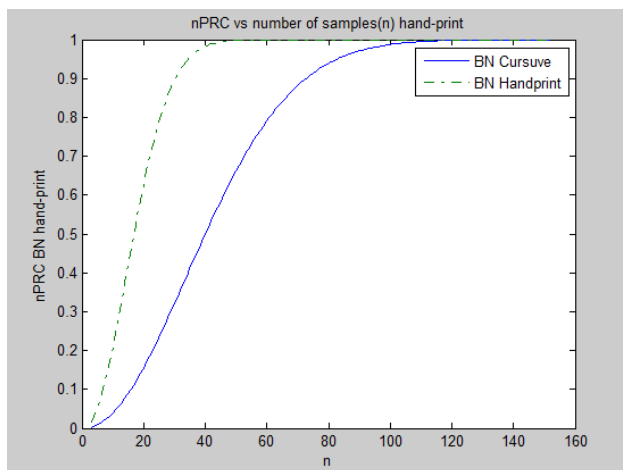


Figure 2.20: Probability of finding a random match for *and* among n writers for cursive and handprint. These plots of n PRC show the discriminative power of the characteristics, with cursive writing of *and* being more individualistic than hand-printing.

istics not matching, the number of terms in the summation is exponential, e.g., if we have nine characteristics with each having four possible values, the number of possible values for X_1 and X_2 can go unto 4^{18} which is greater than billion. Thus approximate inference algorithms will be needed.

n PRC

The probability that among a set of $n \geq 2$ independent, identically distributed samples $\mathbf{X} = \{X_1, \dots, X_n\}$, some pair have the same value within specified tolerance is given by the graphical model in Figure 2.19(b). The n PRC, can be written in terms of the PRC as

$$\rho[n] = 1 - (1 - \rho)^{\frac{n(n-1)}{2}}. \quad (2.19)$$

Note that when $n = 2$, $\text{PRC} = n\text{PRC}$. Since there are $\binom{n}{2}$ pairs involved this probability can be much higher than PRC. For instance, in the famous birthday paradox, while the probability of a birthday (PRC) is $1/365$, the value of n PRC for $n = 24$ is 0.5.

Values of n PRC for *and* obtained for different values of n are plotted in Figure 2.20. While n PRC for both cursive and handprint gradually increases until it reaches 1, it increases faster for hand-print.

Conditional n PRC

The probability that given a specific value it coincides, within tolerance, one in a set of n samples drawn from the same distribution is given by the graphical model in Figure 2.19(c). Since we are trying to match a specific value it depends on the probability of the conditioning value. It is smaller than n PRC and can be lower than the PRC. The exact relationship with respect to PRC depends on the distribution. The conditional n PRC is given by the marginal probability

$$p(Z = 1|X_s) = \sum_{\mathbf{X}} p(Z = 1|X_s, \mathbf{X})p(\mathbf{X}). \quad (2.20)$$

In the case of identical match this can be shown to be equivalent to

$$1 - (1 - P(X_s))^n \quad (2.21)$$

Conditional n PRC for the two writers in Figure 2.18 were evaluated using Eq. 2.21. Plots as a function of n for $d(X, Y) = 0$ and $d(X, Y) = 1$ are shown for the two writers in Figure 2.21 considering exact match

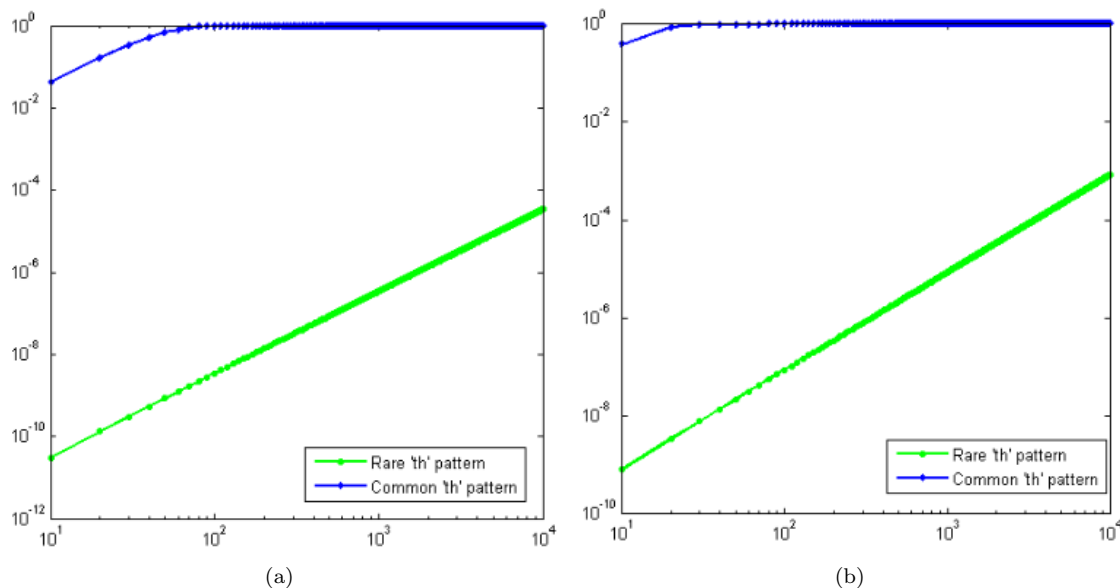


Figure 2.21: Probability of finding a matching entry for th in a database of size n , called conditional n PRC: (a) exact match, and (b) match with one feature mismatch allowed. The two graphs in each figure correspond to X_s being the most common th and a rare th whose forms are shown in Figure 2.18.

Table 2.9: Probability of finding an identical match for a given and among 7000 writers.

| | Sample (X_s) | Conditional n PRC, with $n = 7,000$ |
|-----------------|---------------------|--|
| Cursive and | [1,1,1,0,1,2,0,0,2] | 2.64×10^{-6} |
| | [0,1,0,2,2,3,0,3,2] | 1.13×10^{-5} |
| Handprint and | [0,5,0,1,1,0,1,1,2] | 7.28×10^{-4} |
| | [0,0,0,0,2,0,0,1,2] | 1.09×10^{-4} |

and with a tolerance of mismatch in one feature. With $n = 10$ the probabilities of exact match for the two writers were 0.041 and 3.1×10^{-11} respectively, and probability allowing one mismatch were 0.387 and 7.69×10^{-10} .

For the two samples each for cursive and handprint datasets obtained from our model using Gibbs' sampling in Section 2.3.2, the probability of finding an identical match with $n = 7000$ is shown in Table 2.9.

2.4.2 Probability of Identification

We now consider how a probabilistic model for rarity can be combined with a probabilistic model of similarity so that the probability of identification can be determined. Such a probability can then be discretized into an opinion scale.

Let $S = \{s_i\}$ be a set of sources. They correspond to, say, a set of individual writers. Let \mathbf{o} be a random variable representing an object drawn from a source s_i , e.g., a handwriting specimen of a known writer. Let \mathbf{e} be a random variable representing evidence drawn from a source s_j , e.g., a questioned document. The task is to determine the odds of whether \mathbf{o} and \mathbf{e} came from the same or different source.

We can state two opposing hypotheses:

h^0 : \mathbf{o} and \mathbf{e} are from the *same* source ($i = j$); and

h^1 : \mathbf{o} and \mathbf{e} are from *different* sources ($i \neq j$), which are the *identification* and *exclusion* hypotheses of forensics; some forensic statistics literature also refers to them as *prosecution* and *defense* hypotheses[1].

We can define two joint probability distributions $P(\mathbf{o}, \mathbf{e}|h^0)$ and $P(\mathbf{o}, \mathbf{e}|h^1)$ which specify as to how often each instance of the object and evidence occur together when they belong to the same source or to different

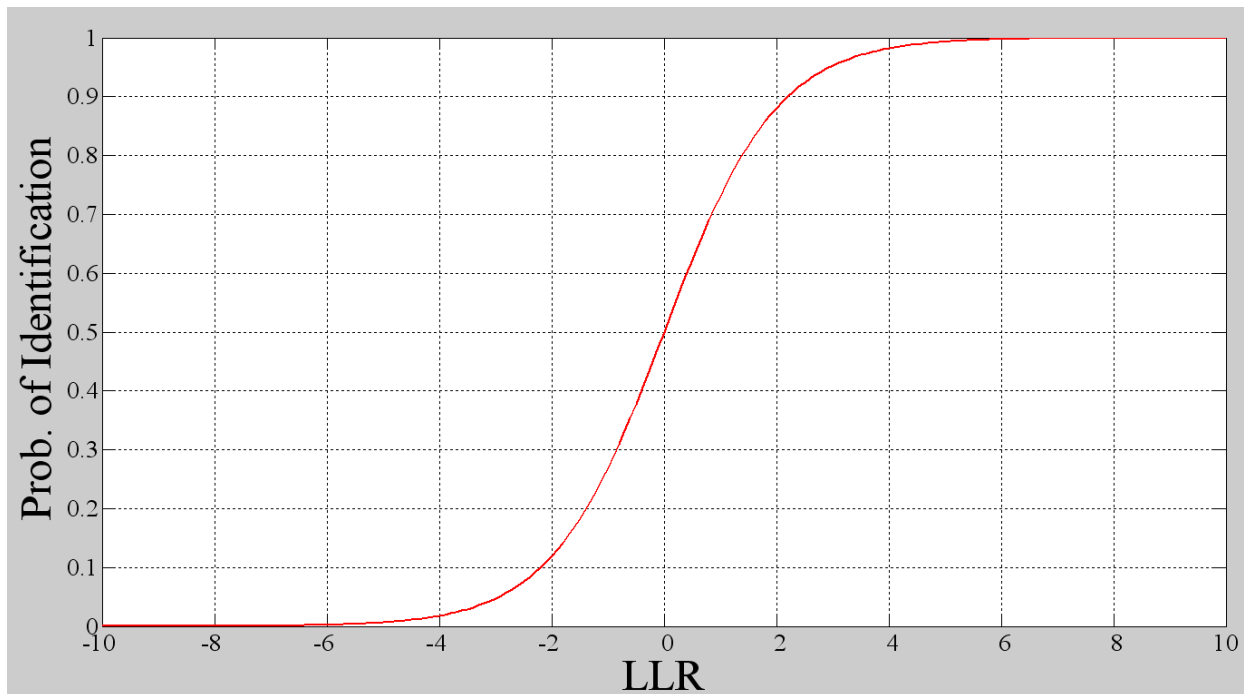


Figure 2.22: Probability of identification is a sigmoid function of the log-likelihood ratio.

sources. The relative strengths of evidence supporting the two hypotheses is quantified by the likelihood ratio

$$LR_J = LR(\mathbf{o}, \mathbf{e}) = \frac{P(\mathbf{o}, \mathbf{e}|h^0)}{P(\mathbf{o}, \mathbf{e}|h^1)}. \quad (2.22)$$

The corresponding log-likelihood ratio, $LLR(\mathbf{o}, \mathbf{e}) = \ln P(\mathbf{o}, \mathbf{e}|h^0) - \ln P(\mathbf{o}, \mathbf{e}|h^1)$, has representational advantages: its sign is indicative of same or different source, it has a smaller range than LR, and additivity of contributions of independent features².

It is useful to convert the LR into a probability of identification (and exclusion) using a Bayesian formulation. Let the *prior* probabilities of the hypotheses be $P(h^0)$ and $P(h^1)$ with $P(h^0) + P(h^1) = 1$. Defining the prior odds as $O_{prior} = \frac{P(h^0)}{P(h^1)}$, we can express the prior probability of the same source as $P(h^0) = O_{prior}/(1 + O_{prior})$. The prior odds can be converted into posterior odds as $O_{posterior} = \frac{P(h^0|\mathbf{o}, \mathbf{e})}{P(h^1|\mathbf{o}, \mathbf{e})} = O_{prior} \times LR(\mathbf{o}, \mathbf{e})$. Thus we can write the posterior probability of the same source as $P(h^0|\mathbf{o}, \mathbf{e}) = O_{posterior}/(1 + O_{posterior})$. The particular case of equal priors is of interest in forensics, as opinion without prior bias. In this case we get a simple form for the probability of identification as

$$P(h^0|\mathbf{o}, \mathbf{e}) = \frac{LR(\mathbf{o}, \mathbf{e})}{1 + LR(\mathbf{o}, \mathbf{e})} = \frac{\exp(LLR(\mathbf{o}, \mathbf{e}))}{1 + \exp(LLR(\mathbf{o}, \mathbf{e}))} = \frac{1}{1 + e^{-LLR(\mathbf{o}, \mathbf{e})}} = \sigma[LLR(\mathbf{o}, \mathbf{e})]. \quad (2.23)$$

where σ is the sigmoid function $\sigma(a) = \frac{1}{1+e^{-a}}$. The probability of exclusion is $P(h^1|\mathbf{o}, \mathbf{e}) = 1 - P(h^0|\mathbf{o}, \mathbf{e}) = 1/[1 + LR(\mathbf{o}, \mathbf{e})] = 1/[1 + e^{LLR(\mathbf{o}, \mathbf{e})}]$.

The probability of identification with respect to the LLR follows a sigmoid function as shown in Figure 2.22. This function reaches either 1 or zero very quickly with the value of LLR. Thus the LLR has to be computed quite precisely for the probability of identification to provide meaningful discrimination.

²In performing LLR additions, since LR values in the interval $(1, \infty)$ convert to positive LLRs and LR values in the interval $(0, 1)$ convert to negative LLRs, the precisions of LR values < 1 must be high, otherwise the ranges of positive and negative LLRs will not be symmetric.

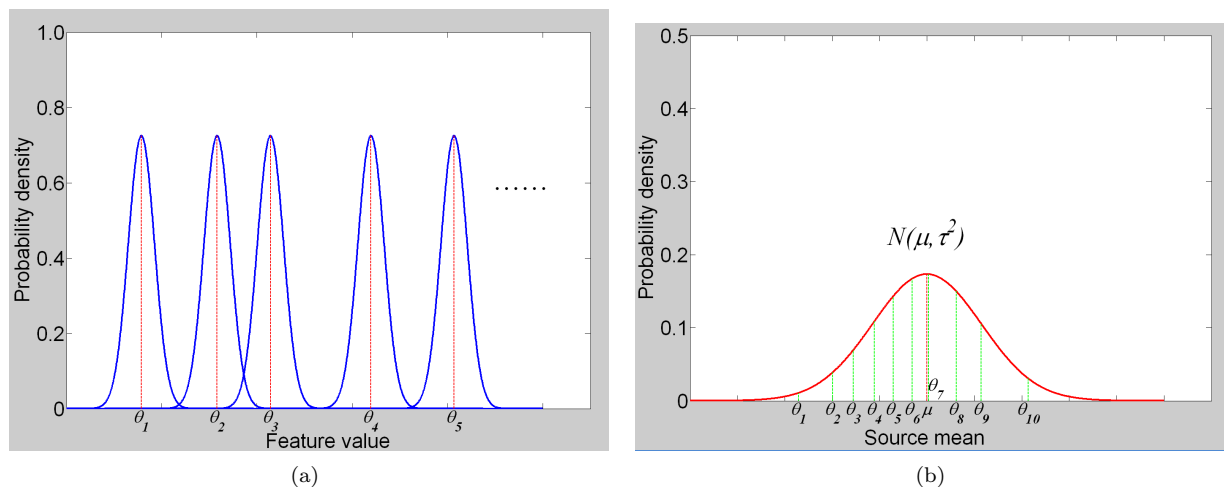


Figure 2.23: Distributions of sources for object and evidence: (a) Normally distributed sources, where each source s_i ($i = 1, 2, \dots$) is $N(\theta_i, \sigma^2)$. Samples may come from a common source s_i or different sources s_i and s_j ($s_i \neq s_j$). (b) Source means $\{\theta_1, \theta_2, \dots\}$, are assumed to have distribution $N(\mu, \tau^2)$, with $\tau \gg \sigma$. Samples coming from sources θ_1, θ_{10} are rarer (less frequent) than samples from θ_6 and θ_7 , suggesting that information about the distribution of source means is useful to assess strength of evidence.

The key to determining the probability of identification is to determine LR defined by Eq. 2.22, which in turn requires the distributions $P(\mathbf{o}, \mathbf{e}|h^i)$ ($i = 0, 1$), defined over all possible values of objects and their evidential forms. If \mathbf{o} and \mathbf{e} are n -dimensional binary vectors with each feature taking K possible values, then $2K^{2n}$ parameters are needed to specify the joint distribution. Determining these distributions is computationally and statistically infeasible. Computationally, kernel density estimation [2] and finite mixture models [53] have been proposed, but they have limitations as well. More important is the statistical limitation of having a sufficient number of samples for so many parameters. Today, objects and evidence can be represented by ever finer features due to higher camera resolution and automatic feature extraction methods and their possible evidential forms is infinite.

One method of simplification is to use a (dis)similarity function between object and evidence. The approach is to define $d(\mathbf{o}, \mathbf{e})$ as a scalar *distance* between object and evidence and define another likelihood ratio as follows

$$LR_D = \frac{P(d(\mathbf{o}, \mathbf{e})|h^0)}{P(d(\mathbf{o}, \mathbf{e})|h^1)}. \quad (2.24)$$

The number of parameters needed to evaluate LR_D is constant, or $O(1)$, and is independent of the number of features n . Due to its simplicity, this method has been widely used in fingerprint identification [56], handwriting analysis [74], pharmaceutical tablet comparison [17], etc.

For certain features spaces and distance measures, e.g., continuous features with Euclidean distance, this approach is equivalent to a *kernel* method [69]. The scalar distance d is just the magnitude of the vector difference \mathbf{d} . However, because it maps two distributions of $2n$ variables each into two scalar distributions there is severe loss of information (many pairs of \mathbf{o} and \mathbf{e} can have the same distance). A natural extension is to use vector difference \mathbf{d} , which quantifies the distribution of both the magnitude and the orientation of the difference between \mathbf{o} and \mathbf{e} , giving a much fine-grained characterization of the difference between \mathbf{o} and \mathbf{e} . While this likelihood ratio LR_{VD} , provides the simplification of mapping two distributions of $2n$ variables each into two distributions of n variables each, there is still a loss of information in the many to one mappings.

Lindley's Result

Evaluating LR+ was considered by Lindley [48] when the object and evidence arise from univariate Gaussian distributions. This is illustrated in Fig. 2.23. There was a single characteristic, the refractive index of glass.

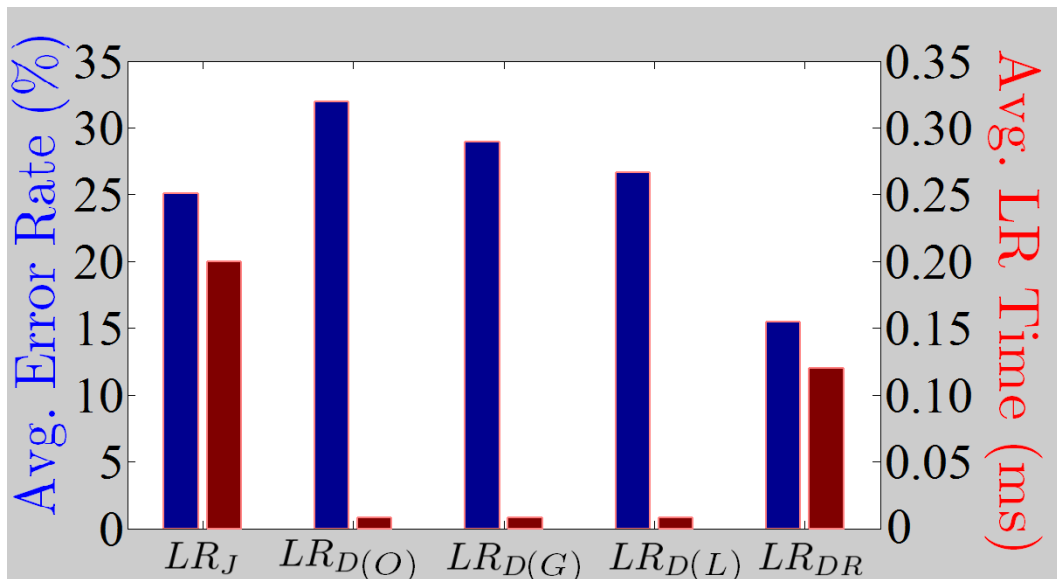


Figure 2.24: Comparison of five methods of computing the likelihood ratio. For each method the average error rates are on the left (blue) and time per sample on the right (red).

In this case the object is a fragment sample on a window and evidence is a sample on a suspect's clothing.

It is assumed that the object and evidence are continuous scalar random variables drawn from the same or different sources. Samples are normally distributed about its source mean with a known constant variance σ^2 ; the source mean is normally distributed with mean μ and variance τ^2 with $\tau \gg \sigma$, and there are p object samples with mean o , and q evidence samples with mean e with $p = q$, then the likelihood ratio can be approximated by

$$LR(o, e) = \frac{\tau}{\sigma\sqrt{2/p}} \exp\left\{-\frac{(o-e)^2}{4\sigma^2/p}\right\} \exp\left\{\frac{(m-\mu)^2}{2\tau^2}\right\}, \quad (2.25)$$

where $m = (o + e)/2$ is the mean of o and e .

Generalization of Lindley's Result

Lindley's result only pertains to a single variable. In the generalization of Eq. 2.25 to several variables [84], LR is approximated as the product of two factors, one based on *difference* and the other on *rarity*:

$$LR_{DR} = P(\mathbf{d}(\mathbf{o}, \mathbf{e})|h^0) * \frac{1}{P(\mathbf{m}(\mathbf{o}, \mathbf{e}))}, \quad (2.26)$$

where $\mathbf{d}(\mathbf{o}, \mathbf{e})$ is the *difference* between \mathbf{o} and \mathbf{e} , and $\mathbf{m}(\mathbf{o}, \mathbf{e})$ is the *mean* of \mathbf{o} and \mathbf{e} .

Average error rates, obtained by thresholding the LR into correct/wrong decisions, using the different methods of computing LR discussed earlier, are shown in Fig. 2.24. The distance and rarity method is seen to perform much better than both the simple distance methods and a joint distribution method that assumes independent characteristics.

2.4.3 Opinion Scale

The task of QD examination of handwritten items is typically to make a decision of whether or not two items have the same origin. Providing a strength of opinion or evidence for any such decision is an integral part. SWGDOC recommends a nine- point scale: [1-Identified as same, 2-Highly Probable same, 3-Probably same, 4-Indicating same, 5-No conclusion, 6-Indicating different, 7-Probably different, 8-Highly probably different and 9-Identified as different].

In the context of an automatic system we have previously introduced a statistical model which, for the sake of completeness, is summarized in Appendix 6. It is based on the philosophy that the strength of evidence should incorporate: (i) the amount of information compared in each of the two items (line/half page/full page etc.), (ii) the nature of content present in the document (same/different content), (iii) characteristics used for comparison and (iv) the error rate of the model used for making the decision.

Adequacy of Evidence

Adequacy of evidence can be naturally included in the likelihood ratio approach. In likelihood ratio based decision methods, often a variable number of input evidences is used. A decision based on many such inputs can result in nearly the same likelihood ratio as one based on few inputs. We consider methods for distinguishing between such situations. One of these is to provide confidence intervals together with the decisions and another is to combine the inputs using weights. We suggest a new method that generalizes Bayesian approach and uses an explicitly defined discount function. Empirical evaluation shows greater flexibility of the designed method.

Consider statistical decision methods based on likelihood ratio. Assume that: h_0 – a hypothesis of interest, h_1 – the alternative hypothesis to h_0 , $F = \{f_i\}_{i=1}^n$ – a set of n observed features. Then:

$$P(h_0|F) + P(h_1|F) = 1 \quad (2.27)$$

Applying Bayes' rule and assuming that all features are independent we get:

$$P(h_0|F) = \frac{P(h_0) \cdot \prod_i LR(f_i|h_0, h_1)}{P(h_1) + P(h_0) \cdot \prod_i LR(f_i|h_0, h_1)} \quad (2.28)$$

where: $LR(f_i|h_0, h_1) = \frac{P(f_i|h_0)}{P(f_i|h_1)}$ – is a likelihood ratio of f_i for h_0 to h_1 .

A variable number of features can be used. Nearly the same likelihood ratio value $LR(F|h_0, h_1)$ (and the same overall result) can be based on many features and on few of them.

For example, assume the following two cases:

1. A discrete uniform prior ($P(h_0) = P(h_1) = 0.5$) and one feature $F = f_1$ with $LR(f_1|h_0, h_1) = 96$.
2. A discrete uniform prior ($P(h_0) = P(h_1) = 0.5$) and nine features $F = \{f_i\}_{i=1}^9$ with likelihood ratios $\{3, 4, 2, \frac{1}{4}, 2, 2, \frac{1}{3}, 6, 2\}$ accordingly.

For the first case we get: $P_1(h_0|F) = \frac{96}{97}$, which is the same as $P_2(h_0|F) = \frac{96}{97}$ – a posterior probability of h_0 in the second case. Consider the QD application where: h_0 – two compared documents were written by the same person, h_1 – written by different people. In the first case the result is based on one feature (a discriminating characteristic in QDE) and could be just a chance, whereas in the second case the support of h_0 tends to be more trustworthy since it is based on nine features. We consider methods for distinguishing between such cases. This problem is especially important in case when probabilities of features are estimated approximately.

One of these methods is to use not a single numerical value for posterior probability of h_0 but a credible interval [19]. When there are no parameters of h_0 then we can assume that there is a distribution of likelihood ratios and estimate a credible interval for it. A crucial part of this approach is to define a proper prior probability for likelihood ratios.

Another way of dealing with the problem at hand is to use statistical hypothesis testing [23]: to define likelihood ratio confidence interval with a given level of confidence. If we assume a log-normal distribution with unknown mean and variance as a distribution of likelihood ratios than we can calculate confidence interval for log-likelihood ratio mean and use it in equation 2.28.

A completely different approach to the problem is to use a score instead of posterior probability of h_0 . This approach allows us not to make any assumptions about likelihood ratio distribution. We have designed such scoring method based on Bayesian approach with weighted likelihoods and a discount function that diminishes the score when features in a comparison have a small sum of weights.

Assume two hypotheses: h_0 – Q and K have the same writer; h_1 – different writers, with prior probabilities: $P(h_0)$, $P(h_1)$. Set $F = \{f_i\}_{i=1}^n$ of n extracted features (e.g. F is a set of discriminating characteristics in compared documents). The problem is to get a score for hypothesis h_0 that reflects how likely h_0 is true in comparison with its alternative and that explicitly includes how trustworthy a hypotheses comparison is.

Here we assume that all features X are independent from each other. A common way to solve the problem is to use Bayesian approach (see equation 2.28). To adapt its ideas in order to explicitly include a hypotheses comparison trustworthiness we discuss several existing methods and suggest a new one.

Methods for Relative Scoring of Hypotheses

Assume that likelihood ratios of features are themselves i.i.d. samples of some distribution. For example, we can assume that it is a log-normal distribution. Then it is possible to identify a credible interval [4]: an $\alpha \cdot 100\%$ credible interval is a set C such that $P(C) = \alpha$. There are different ways to evaluate credible intervals approximately or exactly when it is possible [4]. Values of equation 2.28 determined on credible interval boundaries together with level α define a result for the stated problem.

An analogy in some way of a credible interval in statistical hypothesis testing is a confidence interval [23]. It is an interval $(L_\alpha(F), R_\alpha(F))$ evaluated from a given set of features F , that frequently includes the parameter of interest, if the experiment is repeated. The frequency here is determined by confidence level α . For example, the parameter of interest could be mean of log-likelihood ratios if they comply with the log-normal distribution. Values of equation 2.28 determined on confidence interval boundaries along with confidence level α define a solution for the stated problem.

A useful property of using credible or confidence intervals is that in most cases with increase in the number of features the distance between interval boundaries is decreasing. A problem of using credible or confidence intervals is a necessity of combining interval boundaries and level α into one score if only one single value is required by the application (as a score for hypothesis h_0).

To drop any assumptions on likelihood ratio distributions, we have designed a new method to solve the problem. To include credibility of the hypotheses comparison into the score we suggest to use the following expression instead of posterior probabilities ratio:

$$S\left(\frac{h_0|F}{h_1|F}\right) = \frac{P(h_0)}{P(h_1)} \cdot \left(\prod_{i=1}^n \left(\frac{P(f_i|h_0)}{P(f_i|h_1)}\right)^{w_i}\right)^{d(\sum_i w_i)} \quad (2.29)$$

where: $S\left(\frac{h_0|F}{h_1|F}\right)$ – is the relative score of hypothesis h_0 against h_1 . The following term:

$$\prod_{i=1}^n \left(\frac{P(f_i|h_0)}{P(f_i|h_1)}\right)^{w_i} \quad (2.30)$$

is a weighted likelihood ratio [87], where weights $w_i \in \mathbb{R}^+$. A trivial set of weights is:

$$w_i = 1, \quad i = \overline{1, n} \quad (2.31)$$

which defines that all of the used features are equal in their contribution to the overall value. A more discriminating approach is to use higher weights for more accurate and influential features.

Function $d(\sum_i w_i) : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is a discount function which reflects a credibility of the comparison. It ought to give a support to a comparison with big sum of features weights and discount the score for a comparison with the small sum. A trivial discount function is:

$$d_0\left(\sum_i w_i\right) = \frac{n}{\sum_i w_i} \quad (2.32)$$

which does not make any discount but normalizes weighting scheme. An example of discount function which

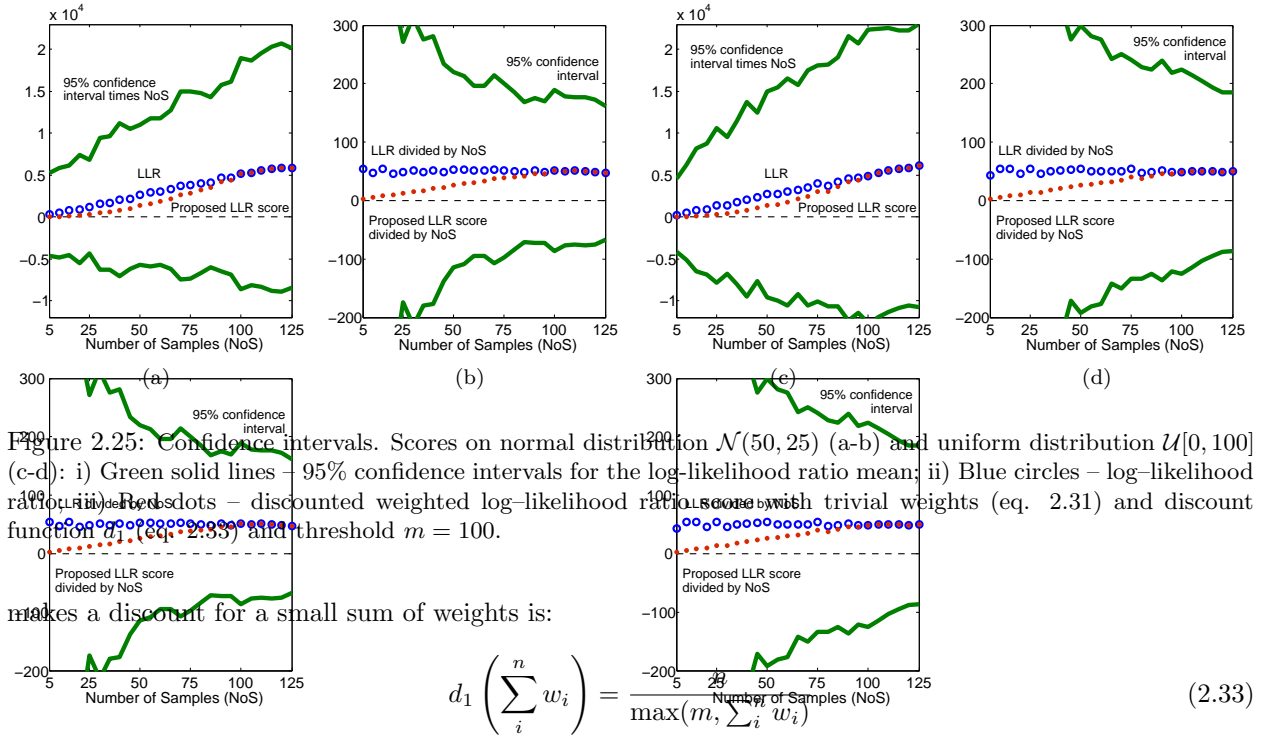


Figure 2.25: Confidence intervals. Scores on normal distribution $\mathcal{N}(50, 25)$ (a-b) and uniform distribution $\mathcal{U}[0, 100]$ (c-d): i) Green solid lines – 95% confidence intervals for the log-likelihood ratio mean; ii) Blue circles – log-likelihood ratio; iii) Red dots – discounted weighted log-likelihood ratio score with trivial weights (eq. 2.31) and discount function d_1 (eq. 2.33) and threshold $m = 100$.

makes a discount for a small sum of weights is:

$$d_1 \left(\sum_i^n w_i \right) = \frac{\sum_i^n w_i}{\max(m, \sum_i^n w_i)} \quad (2.33)$$

where: m is a constant threshold. When trivial weights are used and a number of features is more than a threshold then score S is equal to the ratio of posterior probabilities defined in equation 2.28.

Another example of discount function applies different discounts for different values of sum of weights:

$$d_2 \left(\sum_i^n w_i \right) = \beta \cdot \log \left(\sum_i^n w_i \right) \quad (2.34)$$

where: β – is a magnitude parameter.

Accuracy of using score S instead of posterior probability ratio can be at least the same since the ratio is a special case of S with trivial weights and the trivial discount function. Another advantage of using score S is absence of additional assumptions in comparison with using credible or confidence intervals. The disadvantage however is that in general it results in a score for the hypotheses of interest in comparison with its alternative, but not in the posterior probability.

Experimental Results

To reveal pros and cons of the described methods we have conducted a set of experiments. Synthetically created data sets were used since it allows to compare methods on different (initially specified) sample distributions and with a wide range of data sets sizes. Two types of data sets were used:

- I) Data set consists of values derived from normal distribution $\mathcal{N}(50, 25)$ with mean $\mu = 50$ and variance $\sigma^2 = 25$.
- II) Data set consists of values derived from uniform distribution $\mathcal{U}[0, 100]$ defined on interval $[0, 100]$.

Data sets of each type consisted of 5 through 125 samples with step size of 5. Every sample in each data set represents a log-likelihood ratio value for one feature.

Throughout testing we use a discrete uniform prior:

$$P(h_0) = P(h_1) = 0.5$$

Hence, we focus on a score for log-likelihood ratio (LLR) rather than on a posterior score for the hypotheses h_0 . The following methods were used in the comparison:

- I) Original Bayesian approach where score is defined as posterior probabilities ratio of the hypotheses.
- II) An approach based on estimation of confidence interval for LLR distribution mean. We assume that likelihood ratios have a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with unknown mean and variance.
- III) Discounted weighted likelihood ratio with trivial weights (see equation 2.31) and discount function d_1 (see equation 2.33) and threshold $m = 100$.

We did not include an approach based on credible intervals since if we assume that log-likelihood has a normal distribution with unknown mean and precision and use a normal-gamma distribution as a conjugate prior then (with certain parameters) credible and confidence intervals coincide[52].

Results on data sets generated from normal distribution $\mathcal{N}(50, 25)$ are shown in figures 2.25(a) and 2.25(b). Results on data sets generated from uniform distribution $\mathcal{U}[0, 100]$ are shown in figures 2.25(c) and 2.25(d). Blue circles show log-likelihood ratio (method I); green solid lines and blue circles on figures 2.25(b) and 2.25(d) show 95% confidence intervals and log-likelihood ratio average accordingly (method II); red dots show discounted weighted log-likelihood ratio score (method III).

In both sets of tests method III converges to log-likelihood ratio for large numbers of samples and gives a discounted value for small number of them. In particular, the score average is close to zero for less than 10 samples while LLR average is the same as for large number of samples (see figures 2.25(b) and 2.25(d)).

Let us refer to the example from the introduction: comparison of two cases (in the first case one feature is given and in the second - nine features). Applying Bayesian approach we get:

$$P_1(h_0|F) = P_2(h_0|F) = 0.99$$

where: $P_1(h_0|F)$ – posterior probability of h_0 in the first case, which is the same as posterior probability $P_2(h_0|F)$ of h_0 in the second case. Using discounted weighted likelihood ratio with same parameters as earlier (trivial weights and discount function d_1 but with threshold $m = 20$) we get the following posterior probability scores for the first and the second cases:

$$P_1^s(h_0|F) = 0.89, P_2^s(h_0|F) = 0.56$$

The use of the proposed score enables distinguishing between such cases.

Experimental results show that there are the following disadvantages of using confidence intervals: i) a necessity to make an assumption about log-likelihood ratios distribution; ii) confidence interval boundaries are wide (even for 95% confidence level) and especially in case when real data distribution is different from the assumed one (see figure 2.25(c-d)); iii) there is no common way to combine confidence interval boundaries and confidence level into one score (see figures 2.25(a), 2.25(c): if we use confidence interval boundaries times the number of samples then the result scores become even more distant).

The advantages of using a discounted weighted likelihood are: i) it generalizes Bayesian approach and allows to explicitly set a discount for hypotheses comparison based on a few features; ii) no need to make additional assumptions on likelihood ratios distribution; iii) flexibility of choosing weights and discount function allows tuning the approach for every particular application. However, there are disadvantages also: i) the likelihood score results not in a posterior probability but in a relative score; ii) the flexibility of choosing weights and a discount function makes it non-trivial to find the best settings for a particular application.

Summary of Evidence Combination

The problem of estimating a posterior score of a hypothesis given prior beliefs and a set of evidences (observed features) has been considered. A common way of using posterior probability as the score has a drawback. Nearly the same result can be based on many features and on few of them, which can be important to distinguish. We consider methods for distinguishing between such situations. One of these is to provide confidence intervals for the obtained result and another is to combine features using weights and a discount function. The usage of intervals has the following problem: there is no common way to combine interval boundaries and the corresponding level into one score.

The designed method has the following advantages: i) flexibility (it can be tuned for every particular application); ii) accuracy of the designed approach at least the same as of Bayesian approach; iii) there are no assumptions on likelihood distribution (in comparison to approaches based on credible or confidence

intervals). Although the scoring method does not in general provide probabilities as output (which is a disadvantage), it distinguishes the relative strengths of having different numbers of features.

Automated methods for selecting a weighting scheme and choosing a discount function are topics for future research.

2.5 Methods: QD Work-flow

We have described methods for modeling the distribution of characteristics and performing inference with the models. Next we indicate as to how the methods can be incorporated into the work-flow of the QD examiner. We begin with the procedure laid down in the ASTM document *Standard Guide for Examination of Handwritten Items* [7] that lists steps that must be followed. It can be regarded as representing the knowledge engineering necessary for an expert system. For the validation purpose, the standard procedure has been vetted and accepted by the FDE community.

Following the standard procedure, the examiner often needs to make several decisions, since every case has special needs, e.g., *ransom notes* could be written by multiple writers thus requiring comparison of document sub-parts, with *historical manuscripts* different writers may be more similar to each other than with contemporary writers thus requiring recalibration of individualizing characteristics [13].

2.5.1 Standard Work-Flow

The standard work-flow for the examination of handwritten items [7] was given in Section 2.1.1. It involves making several decisions and item comparisons, which need not be sequential. We annotate steps in the standard work-flow by indicating as to where existing and future computational tools can be useful.

2.5.2 Use of Probabilistic Methods

Methods such as those developed in this research can be used in several steps of the QD process. They are indicated in the pseudocode given in Algorithm 6.

Algorithm 6 Comparison of handwritten items with statistical tools

```

1: Determine Comparison Type:
2:    $Q \vee Q$  (no suspect or determine no. of writers)
3:    $K \vee K$  (to determine variation range)
4:    $K \vee Q$  (to determine/repudiate writership)
5: for each  $Q$  or  $K$  do
6:   Quality: determine visually or by automatic detection of noise.
7:   Distortion: detect manually or by use distortion measures.
8:   Type determination: manually or by automatic classification.
9:   Internal consistency: within document, e.g., multiple writers.
10:  Determine range of variation: compare subgroups.
11:  Identify individualizing characteristics: those with low probability.
12: end for
13: for each Comparison do
14:  Comparability: Both of same Type (Step 8).
15:  Comparison: Determine likelihood ratio (LR) based on characteristics and adequacy.
16:  Form Opinion: by quantizing LR or probability of identification.
17: end for

```

Note that statistical models of characteristics such as those discussed in Section 2.3 are used in Step 11 for choosing individualizing characteristics. They are also used in Step 15 in computing the likelihood ratio as discussed in Section 2.4.2. Statistical models of type (handprint or cursive) discussed in Appendix ?? can be used in Step 8. Statistical models involving quantity of comparisons, called adequacy, can be used in Step 16. In fact statistical models can be usefully developed for all the remaining steps as well.

2.6 Conclusions

The principal conclusions from this research are:

- I) Statistical characterization of handwriting characteristics can be useful to assist the QD examiner in the examination of handwritten items. Particularly in determining individualizing characteristics and in expressing a quantitative opinion.
- II) Since probability distributions of handwriting characteristics involve too many parameters, even for few letter combinations such as *th* and *and*, the complexity can be handled using probabilistic graphical models (PGMs) which are either directed (Bayesian networks) or undirected (Markov random fields). The PGMs can be learnt from a database of handwriting characteristics using new algorithms proposed.
- III) The PGMs can be used to
 - Determine the rarity (inverse of probability) of given characteristics. The probability of random correspondence of an input print in a database of size n can be determined. The measure can be used to determine as to what extent a given set of characteristics are useful in individualizing.
 - Rarity can be combined with a distribution of similarity to determine a likelihood ratio for identification. The likelihood ratio can be mapped to a nine-point scale.
- IV) Software interfaces for creating databases of handwriting characteristics for different commonly occurring letter combinations have been developed.
- V) A dataset of handwritten *and* written by over 1,000 writers, together with their characteristics and probability, has been made publicly available at <http://www.cedar.buffalo.edu/srihari/cursive-and> and at <http://www.cedar.buffalo.edu/srihari/handprint-and>
- VI) An automatic method for determining handwriting type was introduced. It classifies a full page with 92% accuracy. It can be further improved for use at the word level.
- VII) Statistical methods can be used in the work-flow of the FDE. These are the steps where the FDE isolates individualizing characteristics and when he/she expresses an opinion.

2.7 Implications for policy and practice

We have proposed methods for constructing probabilistic models for several steps in the QD examination process for handwritten items. Nearly every step in the process requires human judgement. Tools such as those indicated in this research will benefit the QD examiner in associating quantitative measures with each step. They will make it possible to associate probabilities with conclusions analogous to other forensic domains such as DNA. There will be lesser criticism of examiner bias when quantitative methods are used. Since many of these probabilities are likely to be very large or very small they will not take away from the QD examiner expressing strong opinions about individualization or exclusion.

Methods developed will make it feasible to repeat the compilation as new sample sets become available; which is important since handwriting characteristics of the general population can be expected to change in time, e.g., less time is spent in the schools on teaching penmanship in favor of keyboarding skills. Characteristics specified by QD examiners for other languages and scripts can also benefit from the methods. We have only considered a handwriting of *th* and *and* in this research. Such data will have to be collected for more letter combinations so that the results become more applicable.

Methods developed in this research have applications beyond QD examination. Statistical formulations such as those for determining rarity and opinion can be useful in other forensic domains, particularly in other areas of impression evidence such as latent prints, footwear marks, etc. Algorithms for Bayesian and Markov structure learning have wide applications beyond the forensic sciences.

2.8 Implications for further research

Advances are needed in all areas described: data preparation, model construction, efficient algorithms for inference and software tools to integrate results into the QD work-flow.

- I) Data sets of more letter combinations needs to be extracted from handwriting samples and models constructed.
- II) Since the probability specification involves the evaluation of a large number of parameters, we have described how probabilistic graphical models can be useful. Their automatic learning is important and this area of machine learning needs further exploration.
- III) Inference algorithms become quickly intractable. Approximate inference methods need to be developed.
- IV) The algorithms need to become software tools for the QD examiner.

2.9 Dissemination

2.9.1 Publications

The following papers were published:

- I) G. R. Ball and S. N. Srihari, “Statistical Characterization of Handwriting Characteristics Using Automated Tools” in *Proceedings Document Recognition and Retrieval (DRR XVIII)*, SPIE Conference, San Francisco, CA, January 27, 2011. The paper discusses the need for statistical characterization.
- II) G. R. Ball, D. Pu and S. N. Srihari, “Determining Cursive or Printed Nature of Handwriting Samples,” in *Proceedings International Graphonomics Society (IGS) Conference*, Cancun, Mexico, June 2011. Evaluation of a preprocessing step already incorporated into CEDAR-FOX.
- III) S. N. Srihari, “Evaluation of Rarity of Handwriting Formations,” in *Proceedings International Conference on Document Analysis and Recognition (ICDAR)*, Beijing, China, September 2011.
- IV) Y. Tang, S. N. Srihari and H. Srinivasan, “Handwriting Individualization Using Distance and Rarity,” in *Proceedings Document Recognition and Retrieval XIX*, San Jose, CA, January 2012.

- V) K. Das, S. N. Srihari and H. Srinivasan, “Questioned Document Workflow for Handwriting with Automated Tools,” in *Proceedings Document Recognition and Retrieval XIX*, San Francisco, CA, January 2012.
- VI) S. N. Srihari and K. Singer, “Role of Automation in the Examination of Handwritten Items,” in *Proceedings International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Bari, Italy, September 2012. This paper has been invited to appear in the *Pattern Recognition* journal.
- VII) D. Kovalenko and S. N. Srihari, “On Methods for Incorporating Evidences into Posterior Scoring of Hypotheses”, in *Proc. Int. Conf. Pattern Recognition*, Tsukuba, Japan, Nov. 2012.
- VIII) Y. Tang and S. N. Srihari. “An Efficient Algorithm for Learning Bayesian Networks for Multinomials,” in *Proc. Int. Conf. Pattern Recognition*, Tsukuba, Japan, Nov. 2012.
- IX) Y. Tang and S. N. Srihari, “Learning Bayesian Networks for Likelihood Ratio Computation,” in *Proc. Int. Workshop on Computational Forensics*, Tsukuba, Japan, Nov, 2012.
- X) S. N. Srihari, D. Kovalenko, Y. Tang and G. R. Ball, “Combining Evidence using Likelihood Ratios in Writer Verification,” in *Proceedings Document Recognition and Retrieval XIX*, San Francisco, CA, February 2013.

2.9.2 Presentations

The following presentations were made:

- I) The PI presented part of this work in the ICDAR award talk in Beijing, China in September 2011.
- II) The PI presented part of this work as an invited talk at the Evaluating the Probability of Identification in the forensic sciences.
- III) Two papers were presented at the *SPIE Document Recognition and Retrieval* conference in San Francisco in January 2012. The first of these is on work-flow in handwriting comparison and the role of automated tools within the work-flow. The second is on using distance and rarity in handwriting comparison.
- IV) The PI presented part of this work as an invited talk “Evaluating the probability of Identification in the forensic sciences” at the *International Conference on Handwriting Recognition* in Bari, Italy in September 2012. This paper has been invited to appear in the *Pattern Recognition* journal.
- V) A paper coauthored by Sargur Srihari and Kirsten Singer was presented as a submitted paper at the *International Conference on Handwriting Recognition* in Bari, Italy in September 2012.
- VI) Two papers will be presented at the *AAFS annual meeting* in Washington DC in February 2013.

2.9.3 Students

- I) Yi Tang completed his doctoral dissertation titled *Likelihood Ratio Methods in Forensics* in July 2012.
- II) Dmitry Kovalenko did a one-year Fulbright under this research effort.

2.10 Acknowledgement

The project benefitted from the advise of several QD examiners: Kirsten Singer of the Department of Veteran's Administration, Traci Moran of the Financial Management Service, and Lisa Hanson of the Minnesota Criminal Apprehension Laboratory. Nonetheless, the views expressed here are of the author alone and do not reflect the opinions of the QD examiners nor of the Department of Justice.

Kirsten Singer and Traci Moran spent nearly a hundred hours ground-truthing the *and* data. They were also instrumental in suggesting the statistical characterization work.

Graduate students Yi Tang and Mukta Puri worked on Bayesian network learning. Chang Su and Yu Liu contributed to the inference methods. Dmitry Kovalenko of Moscow State University spent his Fulbright scholar year with me working on MRF learning.

Bibliography

- [1] AITKEN, C., AND TARONI, F. *Statistics and the Evaluation of Evidence for Forensic Scientists*. Wiley, 2004.
- [2] AITKEN, C. G. G., AND LUCY, D. Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society* 53, 1 (2004), 109–122.
- [3] ANANDKUMAR, A., TAN, V., AND WILLSKY, A. High-dimensional Gaussian graphical model selection: Tractable graph families. *CoRR* (2011).
- [4] ANTELMAN, G., MADANSKY, A., AND MCCULLOCH, R. *Elementary Bayesian Statistics*. Cheltenham, UK, 1997.
- [5] ARIVAZHAGAN, M., SRINIVASAN, H., AND SRIHARI, S. A statistical approach to handwritten line segmentation. In *Document Recognition and Retrieval XIV* (2007), SPIE.
- [6] ARMISTEAD, T. Issues in the identification of handprinting: A case study in anonymous death threats. *Journal of Police Science and Administration* 12(1) (1984), 81–98.
- [7] ASTM. Standard Guide for Examination of Handwritten Items, 2007. Designation: E2290-03.
- [8] ASTM. Standard Terminology for Expressing Conclusion of Forensic Document Examiners, 2007. Designation: E1658-04.
- [9] BACH, F., AND JORDAN, M. Thin junction trees. In *Advances in Neural Information Processing Systems 14* (2001), MIT Press, pp. 569–576.
- [10] BALDING, D. J., AND DONNELLY, P. Inferring identify from DNA profile evidence. *Proceedings of the National Academy of Sciences, USA* 92, 25 (1995), 1174111745.
- [11] BALL, G. R., PU, D., AND SRIHARI, S. N. Determining cursive or printed nature of handwritten samples. In *Proceedings International Graphonomics Society Conference, Cancun, Mexico* (2011), pp. 189–192.
- [12] BALL, G. R., AND SRIHARI, S. N. Statistical characterization of handwriting characteristics using automated tools. In *Proc Document Recognition and Retrieval* (2011), SPIE.
- [13] BALL, G. R., SRIHARI, S. N., AND STRITMATTER, R. Writer identification of historical documents among cohort writers. In *Proc. Int. Conf. Frontiers Handwriting Recognition Kolkata, India* (Nov. 2010), IEEE Comp. Soc. Press.
- [14] BHARADWAJ, A., SINGH, A., SRINIVASAN, H., AND SRIHARI, S. N. On the use of lexeme features for writer verification. In *Proc of the International Conference on Document Analysis and Recognition* (2007).
- [15] BISHOP, C. *Pattern Recognition and Machine Learning*. Springer, 2006.

- [16] BLEI, D. M., AND JORDAN, M. I. Variational inference for Dirichlet process mixtures. *Bayesian Analysis* 1, 1 (2004), 121–144.
- [17] BOLCK, A., WEYERMANN, C., DUJOURDY, L., ESSEIVA, P., AND BERG, J. Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons. *Forensic Science International* 191, 1 (2009), 42–51.
- [18] CHOW, C. K., AND LIU, C. N. Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Information Theory* 14, 3 (1968), 462–467.
- [19] CONGDON, P. *Bayesian Statistical Modelling*. Wiley, USA, 2007.
- [20] CONWAY, J. The identification of handprinting. *The Journal of Criminal Law, Criminology, and Police Science* 45(5) (1955), 605–612.
- [21] CONWAY, J. V. P. *Evidential Documents*. C. C. Thomas, 1959.
- [22] COOPER, G., AND HERSKOVITS, E. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 9, 4 (1992), 309–347.
- [23] COX, D., AND HINKLEY, D. *Theoretical Statistics*. Chapman and Hall/CRC, USA, 1979.
- [24] DAVIS, J., AND DOMINGOS, P. Bottom-up learning of markov network structure. In *Proceedings of the ICML'10* (2010).
- [25] DE CAMPOS, C. P., AND JI, Q. Efficient structure learning of Bayesian networks using constraints. *Journal of Machine Learning Research* 12 (2011), 663–689.
- [26] DE CAMPOS, C. P., ZENG, Z., AND JI, Q. Structure learning of Bayesian networks using constraints. In *Proc. of ICML 2009* (2009), pp. 113–120.
- [27] DING, S. Learning undirected graphical models with structure penalty. *CoRR abs/1104.5256* (2011).
- [28] EVETT, I. W., LAMBERT, J. A., AND BUCKLETON, J. S. A Bayesian approach to interpreting footwear marks in forensic casework. *Science and Justice* 38, 4 (1998), 241 – 247.
- [29] GILBERT, A. N., AND WYSOCKI, C. J. Hand preference and age in the United States. *Neuropsychologia* 30 (1992), 601–608.
- [30] GRAVE, E., OBOZINSKI, G., AND BACH, F. Trace lasso: a trace norm regularization for correlated designs. *CoRR abs/1109.1990* (2011).
- [31] HAND, D., MANNILLA, H., AND SMYTH., P. *Principles of Data Mining*. MIT Press, 2001.
- [32] HARRISON, D., BURKES, T. M., AND SIEGER, D. P. Handwriting examination: Meeting the challenges of science and the law. *Forensic Science Communications* 11, 4 (2009).
- [33] HARRISON, W. R. *Suspect Documents*. Sweet & Maxwell, 1966.
- [34] HECKER, M. Forensic information system for handwriting (FISH), 1993. Technical Document from the Kriminaltechnisches Institut, Bundeskriminalamt.
- [35] HILTON, O. *Scientific Examination of Questioned Documents*. Elsevier, 1982.
- [36] HILTON, O. *Scientific Examination of Questioned Documents, Revised Edition*. CRC Press, 1993.
- [37] HUANG, C., AND SRIHARI, S. Mapping transcripts to handwritten text. In *Proc. International Workshop on Frontiers in Handwriting Recognition (IWFHR-10), La Baule, France* (2006), pp. 15–20.
- [38] HUANG, C., AND SRIHARI, S. Word segmentation of off-line handwritten documents. In *Proc. Document Recognition and Retrieval (DRR) XV* (2008), vol. 6815, SPIE.

- [39] HUBER, R. A., AND HEADRICK, A. M. *Handwriting Identification: Facts and Fundamentals*. CRC Press, 1999.
- [40] KABRA, S., SRINIVASAN, H., HUANG, C., AND SRIHARI, S. N. On computing the strength of evidence for writer verification. In *Proc. International Conference on Document Analysis and Recognition (ICDAR-2007), Curitiba, Brazil (2007)*, pp. 844–848.
- [41] KAM, M., AND LIN, E. Writer identification using hand-printed and non-hand-printed questioned documents. *Journal of forensic sciences* 48 (2003).
- [42] KARGER, D., AND SREBRO, N. Learning markov networks: Maximum bounded tree-width graphs. In *Proceedings of ACM-SIAM, Washington DC (2001)*, pp. 392–401.
- [43] KIRSHNER, S., SMYTH, P., AND ROBERTSON, A. W. Conditional Chow-Liu tree structures for modeling discrete-valued vector time series. In *Proc. of the 20th conference on UAI (2004)*, pp. 317–324.
- [44] KOLLER, D., AND FRIEDMAN, N. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.
- [45] KUZHINJEDATHU, K., SRINIVASAN, H., AND SRIHARI, S. Robust line segmentation for handwritten documents. In *Proc. Document Recognition and Retrieval (DRR) XV (2008)*, vol. 6815, SPIE.
- [46] LEE, S., CHA, S., AND SRIHARI, S. N. Combining macro and micro features for writer identification. *SPIE, Document Recognition and Retrieval IX, San Jose, CA (2002)*, 155–166.
- [47] LEE, S., GANAPATHI, V., AND KOLLER, D. Efficient structure learning of markov networks using l_1 -regularization. *Proceedings of Neural Information Processing Systems (2006)*. Vancouver, Canada.
- [48] LINDLEY, D. V. A problem in forensic science. *Biometrika* 64, 2 (1977), 207–213.
- [49] LIU, D., AND NOCEDAL, J. On the limited memory bfgs method for large scale optimization. *Mathematical Programming* 45 (1989), 503–528.
- [50] LOHR, S. *Sampling: design and analysis*. Duxbury Press, 1999.
- [51] MUEHLBERGER, R. J., NEWMAN, K. W., REGENT, J., AND WICHMANN, J. G. A statistical examination of selected handwriting characteristics. *Journal of Forensic Sciences (1977)*, 206–210.
- [52] MURPHY, K. *Conjugate Bayesian analysis of the Gaussian distribution*. UBC, 2007.
- [53] NANDAKUMAR, K., CHEN, Y., DASS, S. C., AND JAIN, A. K. Likelihood ratio-based biometric score fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 2 (2008), 342–347.
- [54] NAS. *Strengthening the Forensic Sciences in the United States: A Path Forward*. National Academy of Sciences Press, 2009.
- [55] NETRAPALLI, P., BANERJEE, S., SANGHAVI, S., AND SHAKKOTTAI, S. Greedy learning of markov network structure. In *Proceedings of Allerton Conference on Communication, Control and Computing (2010)*.
- [56] NEUMANN, C., CHAMPOD, C., PUCH-SOLIS, R., EGLI, N., ANTHONIOZ, A., AND BROMAGE-GRIFFITHS, A. Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *Journal of Forensic Sciences* 52, 1 (2007), 54–64.
- [57] NEUMANN, C., CHAMPOD, C., PUCH-SOLIS, R., MEUWLY, D., EGLI, N., AND ANTHONIOZ, A. Computation of likelihood ratios in fingerprint identification for configurations of three minutiae. *Journal of Forensic Sciences* 51 (2006), 1255.
- [58] OSBORN, A. *Questioned Documents*. Nelson Hall Pub, 1929.

- [59] PERKINS, S., LACKER, K., AND THEILER, J. Grafting: fast, incremental feature selection by gradient descent in function space. *J. Mach. Learn. Res.* 3 (2003), 1333–1356.
- [60] PERVOUCHINE, V. *Ph.D. Thesis: Discriminative Power of Features Used by Forensic Document Examiners in the Analysis of Handwriting*. Nanyang University, 2006.
- [61] PETERS, J., JANZING, D., AND SCHOLKOPF, B. Causal inference on discrete data using additive noise models. *IEEE TPAMI* 33, 12 (Dec. 2011), 2436–2450.
- [62] PIETRA, S., PIETRA, V., AND LAFFERTY, J. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, 4 (1997), 380–393.
- [63] PLAMONDON, R., AND SRIHARI, S. N. On-line and off-line handwriting recognition: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 1 (2000), 63–84.
- [64] RAVIKUMAR, P., WAINWRIGHT, M., AND LAFFERTY, J. High-dimensional ising model selection using l_1 -regularized logistic regression. *Annals of statistics* 38 (2010), 1287.
- [65] ROY, S., LANE, T., AND WERNER-WASHBURNE, M. Learning structurally consistent undirected probabilistic graphical models. In *Proceedings of ICML'09* (NY, USA, 2009), pp. 905–912.
- [66] SAUNDERS, C. P., DAVIS, L. J., AND BUSCAGLIA, J. Using automated comparisons to quantify handwriting individuality. *J. Forensic Sciences* 56, 3 (2011), 683–689.
- [67] SAUNDERS, P., DAVIS, L. J., LAMAS, A. C., MILLER, J. J., AND GANTZ, D. T. *Construction and evaluation of classifiers for forensic document analysis*. George Mason University, 2009.
- [68] SCHMIDT, M., MURPHY, K., FUNG, G., AND ROSALES, R. Structure learning in random fields for heart motion abnormality detection. In *Proceedings of Computer Vision and Pattern Recognition Conference, Alaska, USA* (2010).
- [69] SCHOLKOPF, B. The kernel trick for distances. *Advances in Neural Information Processing Systems* 13 (2001).
- [70] SRIHARI, S. N. Computational methods for handwritten questioned document examination. *National Criminal Justice Research Report* (2010).
- [71] SRIHARI, S. N. Computing the scene of a crime. *IEEE Spectrum* 47 (2010), 38–43.
- [72] SRIHARI, S. N. Evaluating the rarity of handwriting formations. In *Proc. Int. Conf. Doc. Anal. Recog. (ICDAR), Beijing, China* (Sept. 2011), IEEE Comp. Soc. Press, pp. 618–622.
- [73] SRIHARI, S. N., CHA, S., ARORA, H., AND LEE, S. Individuality of handwriting. *Journal of Forensic Sciences* 44(4) (2002), 856–872.
- [74] SRIHARI, S. N., HUANG, C., AND SRINIVASAN, H. On the discriminability of the handwriting of twins. *Journal of Forensic Sciences* 53(2) (2008), 430–446.
- [75] SRIHARI, S. N., AND LEEDHAM, G. A survey of computer methods in forensic handwritten document examination. In *Proc Int Graphonomics Soc Con Scot. AZ* (2003), pp. 278–281.
- [76] SRIHARI, S. N., AND SRINIVASAN, H. Comparison of ROC and likelihood decision methods in automatic fingerprint verification. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)* 22, 3 (2008), 535–553.
- [77] SRIHARI, S. N., SRINIVASAN, H., AND DESAI, K. Questioned document examination using CEDAR-FOX. *J. Forensic Document Examination* 18 (2007), 1–20.

- [78] SRIHARI, S. N., ZHANG, B., TOMAI, C., LEE, S., SHI, Z., AND SHIN, Y.-C. A system for handwriting matching and recognition. In *Proceedings Symposium on Document Image Understanding Technology* (2003), pp. 67–75.
- [79] SRIKANTAN, G., LAM, S., AND SRIHARI, S. Gradient based contour encoding for character recognition. *Pattern Recognition* 7 (1996), 1147–1160.
- [80] SU, C., AND SRIHARI, S. Probability of random correspondence of fingerprints. In *Proceedings International Workshop on Computational Forensics* (2009), Springer, pp. 55–66.
- [81] SU, C., AND SRIHARI, S. Evaluation of rarity of fingerprints in forensics. In *Advances in Neural Information Processing Systems*, vol. 23. NIPS, 2010, pp. 1207–1215.
- [82] SUZUKI, J. A generalization of the Chow-Liu algorithm and its application to statistical learning. In *Proc. of Intl. Conf. on Artificial Intelligence* (2010), pp. 478–486.
- [83] TANG, Y., AND SRIHARI, S. N. Efficient and accurate learning of Bayesian networks using chi-squared independence tests. In *Proc. of International Conference on Pattern Recognition, Tsukuba, Japan* (2012).
- [84] TANG, Y., AND SRIHARI, S. N. Handwriting individualization using distance and rarity. In *Proc. of Document Retrieval and Recognition, San Francisco, CA* (2012), SPIE.
- [85] TARONI, F., AITKEN, C. G. G., GARBOLINO, P., AND BIEDERMANN, A. *Bayesian networks and probabilistic inference in forensic science*. Wiley, 2006.
- [86] TRIGGS, C., HARBISON, S., AND BUCKLETON, J. The calculation of DNA match probabilities in mixed race populations. *Science & Justice* 40, 1 (2000), 33 – 38.
- [87] WANG, X. Asymptotic properties of maximum weighted likelihood estimators. *Journal of Statistical Planning and Inference* 119, 1 (2004), 37–54.
- [88] WASSERMAN, L. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2003.
- [89] WEIR, B. S. Matching and partially-matching DNA profiles. *Journal of Forensic Sciences* 49, 5 (2004), 1009 – 1014.
- [90] WING, J. Computational Thinking. *Comm. ACM* 49, 3 (2006), 33–35.
- [91] ZHU, J., LAO, N., AND XING, E. Grafting-light: Fast, incremental feature selection and structure learning of markov random fields. In *Proceedings of the 16th ACM SIGKDD, Washington DC* (2010).

2.11 Appendix 1: Handwriting sample source

The handwriting sample set was collected by us in 1999 [73]. This dataset consists of 4701 handwritten documents where 1567 writers wrote the CEDAR letter exactly three times each. The CEDAR letter has content designed to contain 156 words which include all characters (letters and numerals), punctuation, and distinctive letter and numeral combinations (ff, tt, oo, 00). In particular, this document set out to contain each letter of the alphabet in capital form at the initial position of a word and in lowercase form in the initial, middle, and terminal locations in a word (a minimum of 104 forms of each letter). The vocabulary size is 124 (that is, 32 of the 156 words are duplicate words, mostly stop words such as the, she, etc.). The letters were written three times in each writers most natural handwriting using plain unlined sheets with a medium black ball-point pen. The repetition was to determine, for each writer, the variation of handwriting from one writing occasion to the next. The samples are scanned at 300 dpi and 8-bit grayscale.

In preparing the data set, our objective was to obtain a set of handwriting samples that would capture variations in handwriting between and within writers. This meant that we would need handwriting samples from multiple writers, as well as multiple samples from each writer. The handwriting samples of the sample

population should have the following properties (loosely based on [39]): (i) they are sufficient in number to exhibit normal writing habits and to portray the consistency with which particular habits are executed, and (ii) for comparison purposes, they should have similarity in texts, in writing circumstances and in writing purposes.

Several factors may influence handwriting style, e.g., gender, age, ethnicity, handedness, the system of handwriting learned, subject matter (content), writing protocol (written from memory, dictated, or copied out), writing instrument (pen and paper), changes in the handwriting of an individual over time, etc. For instance, we decided that document content would be such that it would capture as many features as possible. Only some of these factors were considered in the experimental design. The other factors will have to be part of a different study. However, the same experimental methodology can be used to determine the influence factors not considered.

There were two design aspects to the collection of handwriting samples: content of the handwriting sample and determining the writer population.

2.11.1 Source Document

A source document in English, copied by each writer, is shown in Figure 2.26(a). It is concise (156 words) and complete in that it captures all characters (alphabets and numerals) and certain character combinations of interest. In the source document, each alphabet occurs in the beginning of a word as a capital and a small letter and as a small letter in the middle and end of a word (a total of 104 combinations). The number of occurrences in each position of interest in the source text is shown in Table 2.10(a). In addition, the source document also contains punctuation, all ten numerals, distinctive letter and numeral combinations (ff, tt, oo, 00), and a general document structure that allows extracting macro-document attributes such as word and line spacing, line skew, etc. Forensic literature refers to many such documents, including the London Letter and the Dear Sam Letter [58]. We set out to capture each letter of the alphabet as capital letters and as small letters in the initial, middle, and terminal positions of a word. This creates a total of 104 possibilities (cells) for each of the 26 letters in the alphabet. A measure of how "complete" the source text is is given by the expression: $\frac{104 - \text{No. of empty cells}}{104}$. While our source text scores 99% on this measure, the London Letter scores only 76%. Each participant (writer) was required to copy-out the source document three times in his/her most natural handwriting, using plain, unruled sheets, and a medium black ballpoint pen provided by us. The repetition was to determine, for each writer, the variation of handwriting from one writing occasion to the next.

Table 2.10: CEDAR Letter Data Attributes: (a) positional frequency of letters in text, and (b) demographics of writers.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z |
|-------------|----|---|---|----|----|---|---|----|----|---|---|----|---|----|----|---|---|----|----|----|----|---|---|---|---|---|
| Init | 4 | 2 | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 3 | 1 | 1 | 1 |
| | a | b | c | d | e | f | g | h | i | j | k | l | m | n | o | p | q | r | s | t | u | v | w | x | y | z |
| Init | 17 | 4 | 1 | 1 | 6 | 1 | 2 | 9 | 4 | 2 | 1 | 2 | 2 | 1 | 6 | 2 | 1 | 5 | 8 | 14 | 1 | 1 | 8 | 1 | 3 | 1 |
| Mid | 33 | 2 | 8 | 6 | 59 | 4 | 5 | 20 | 32 | 1 | 3 | 14 | 3 | 35 | 36 | 4 | 1 | 30 | 19 | 25 | 18 | 7 | 5 | 2 | 2 | 2 |
| Term | 5 | 2 | 1 | 21 | 20 | 3 | 3 | 5 | 1 | 0 | 3 | 5 | 2 | 7 | 5 | 1 | 1 | 12 | 15 | 17 | 2 | 1 | 2 | 1 | 8 | 1 |

| Ethnicity/ Gender | White Female | White Male | Black Female | Black Male | API Female | API Male | AIEA Female | AIEA Male | Hispan Female | Hispan Male |
|----------------------|-----------------|---------------|-----------------|---------------|---------------|-------------|----------------|--------------|------------------|----------------|
| Age/Total | 872/371 | 333/359 | 103/64 | 36/56 | 38/16 | 31/14 | 19/5 | 4/5 | 91/54 | 40/56 |
| 12-14 | 49/17 | 25/16 | 2/4 | 2/4 | 1/1 | 2/1 | 0/0 | 0/0 | 22/4 | 16/4 |
| 15-24 | 158/66 | 111/64 | 25/15 | 13/13 | 16/4 | 18/2 | 4/1 | 1/2 | 22/13 | 10/14 |
| 25-44 | 252/140 | 76/136 | 31/25 | 8/22 | 12/6 | 7/6 | 11/3 | 2/1 | 34/24 | 11/24 |
| 45-64 | 267/87 | 69/85 | 24/13 | 10/11 | 6/4 | 2/3 | 3/1 | 1/1 | 7/10 | 1/10 |
| 65-84 | 139/56 | 50/55 | 20/6 | 3/5 | 3/1 | 2/1 | 1/0 | 0/0 | 6/3 | 2/4 |
| 85 ~ | 7/5 | 2/5 | 1/1 | 0/1 | 0/0 | 0/1 | 0/0 | 0/1 | 0/0 | 0/0 |

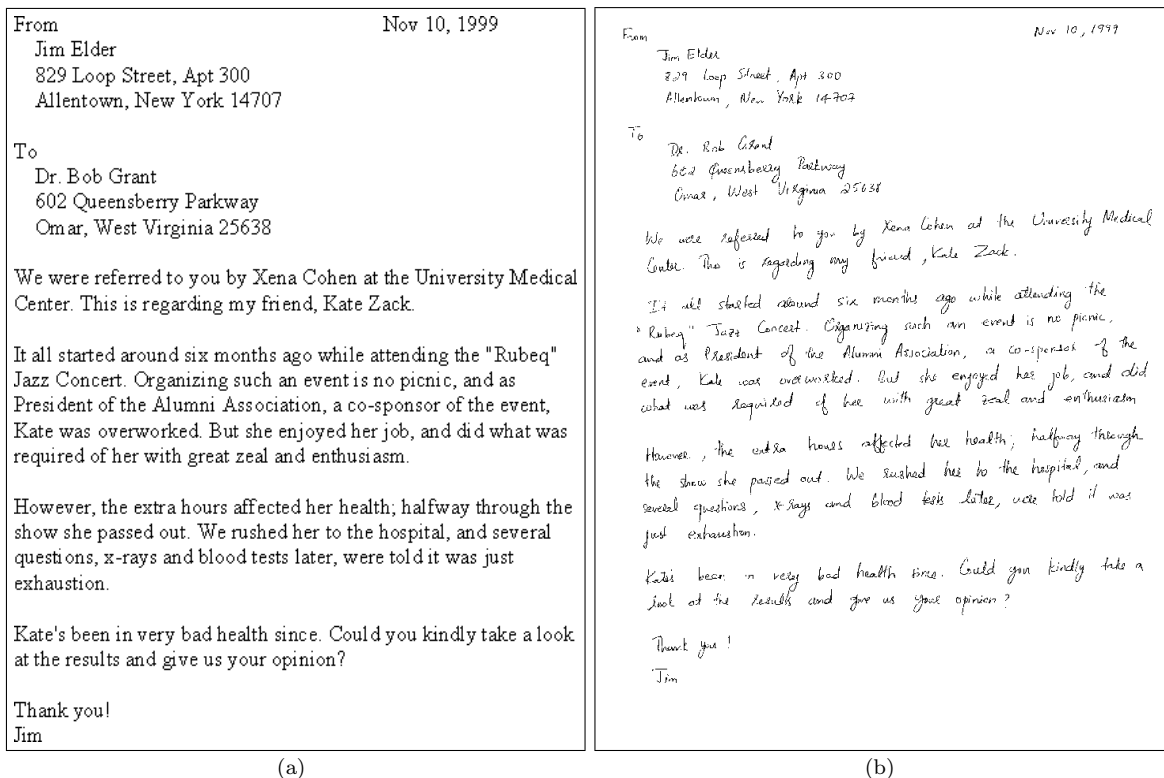


Figure 2.26: Handwriting Source: (a) document copied by writers includes all alphabets, and (b) a digitally scanned handwritten sample provided by a writer.

2.11.2 Writer Population

The writer population was representative of the U.S. population. Statistical issues in determining the writer population were: the number of samples needed to make statistically valid conclusions and the population distribution needed to make conclusions that apply to the US population, which are issues in the design of experiments [50].

Randomness

If the samples are random, then every individual in the US should have an equal chance of participating in the study. The samples were obtained by contacting participants in person, by mail, by advertising the study with the use of flyers and internet newsgroups, and by manning a university booth. For geographic diversity, we obtained samples by contacting schools in three states (Alaska, Arizona, and New York) and communities in three states (Florida, New York, and Texas) through churches and other organizations.

Sample Size

The sample population should be large enough to enable drawing inferences about the entire population through the observed sample population. The issue of large enough is related to sampling error, the error that results from taking one sample instead of examining the whole population, i.e., how close is an estimate of a quantity based on the sample population to the true value for the entire population?

Public opinion polls that use simple random sampling specify using a sample size of about 1100, which allows for a 95% confidence interval, with a margin of error of 0.03 [73]. Higher precision levels would entail a larger number of samples. Our database has a sample size of about 1500, and our results are therefore subject to such a margin of error.

The sample population should be representative of the US population. For instance, since the US population consists of an (approximately) equal number of males and females, it would be unwise to perform the study on a sample population and expect the conclusions of the study to apply to the entire US population consisting of males and females (especially in the absence of any scientific evidence that proves or disproves the association between handwriting and gender). The sample was made representative by means of a stratified sample with proportional allocation [50].

We divided the population into a pre-determined number of sub-populations, or strata. The strata do not overlap, and they constitute the whole population so that each sampling unit belongs to exactly one stratum. We drew independent probability samples from each stratum, and we then pooled the information to obtain overall population estimates. The stratification was based on US census information (1996 projections).

Proportional allocation was used when taking a stratified sample to ensure that the sample reflects the population with respect to the stratification variable and is a miniature version of the population. In proportional allocation, so called because the number of sampled units in each stratum is proportional to the size of the stratum, the probability of selection is the same for all strata. Thus, the probability that an individual will be selected to be in the sample is the same as in a simple random sample without stratification, but many of the bad samples that could occur otherwise cannot be selected in a stratified sample with proportional allocation. The sample size again turns out to be about 1000 for a 95% confidence interval, with a margin of error of 0.03.

A survey designed as above would allow drawing conclusions only about the general US population and not any subgroup in particular. In order to draw any conclusions about the subgroups, we would need to use allocation for specified precision within data. This would entail having 1,000 in each cell of the cross-classification.

From the census data, we obtained population distributions pertaining to gender, age, ethnicity, level of education, and country of origin; we also obtained a distribution for handedness from [29]. Based on this information, a proportional allocation was performed for a sample population of 1000 across these strata. Among these variables, only gender, age, and ethnicity can be considered as strata (by definition). Due to the limited amount of census data on other combinations, we were unable to stratify across handedness and level of education.

Each writer was asked to provide the following writer data, enabling us to study the various relationships: gender (male, female), age (under 15 years, 15 through 24 years, 25 through 44 years, 45 through 64 years, 65 through 84 years, 85 years and older), handedness (left, right), highest level of education (high school graduate, bachelors degree and higher), country of primary education (if US, which state), ethnicity (Hispanic, white, black, Asian/Pacific Islander, American Indian/Eskimo/Aleut), and country of birth (US, foreign).

The details (actual/target) of the distribution for a sample size of 1568 writers are given in Table 2.10(b). The strata are sometimes under-represented (actual < target) or over-represented (actual > target). Parameters considered in addition to strata shown in Table 2.10(b) are handedness and country of origin - Male: handedness (right, left): 382/429, 61/61, and country of origin (US, foreign): 373/451, 71/39; Female: handedness (right, left): 1028/461, 95/49, and country of origin (US, foreign): 1026/469, 98/41.

There may be other relevant strata that could have been considered, such as the system of writing learned (e.g., the Palmer method), country in which writing was learned, etc. We were constrained by the limited information we have on these distributions. Moreover, a perfect sample (a scaled-down version of the population that mirrors every characteristic of the whole population) cannot exist for complicated populations. Even if it did exist, we would not know it was a perfect sample without measuring the whole population.

2.12 Appendix 2: Tool for extracting image snippets

The transcript-mapping function of CEDAR-FOX [37] is useful to locate image snippets of interest. Screenshots of this function are given in Figure 2.27. Results can be filtered to get desired letter combination image snippets.

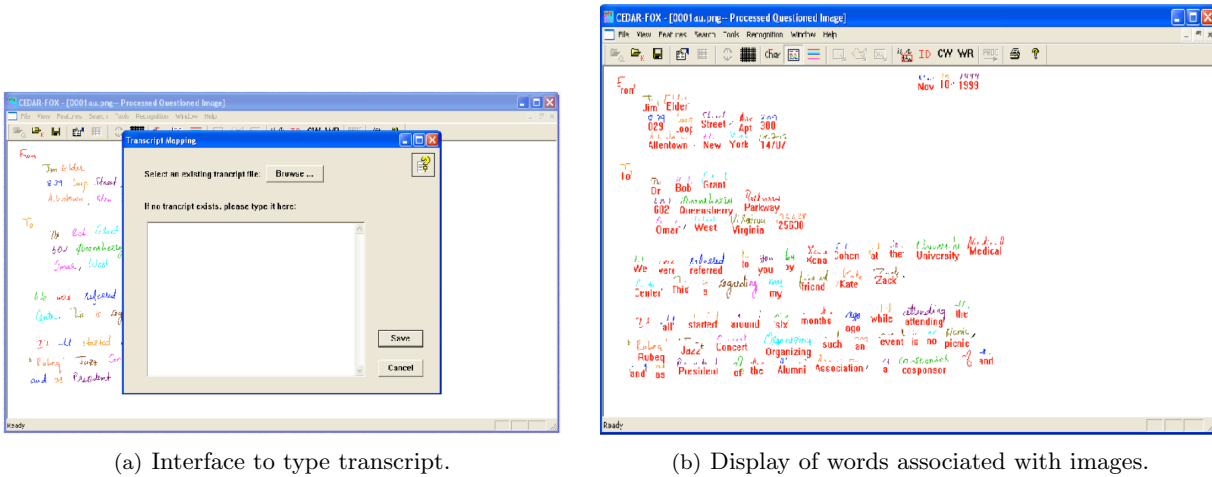


Figure 2.27: Transcript mapping function of CEDAR-FOX for extracting letter combinations from handwritten pages: (a) window or an input text file, and (b) truth super-imposed on handwriting image.

2.13 Appendix 3: Comparison of *th* marginals with previous work.

We provide here a comparison of the marginal distributions of the six characteristics of *th* given in Figure 2.9(b) with those give in [51]. Both sets of marginal probabilities are given in Table 2.11.

Table 2.11: Comparison of two marginal distributions

| (a) Marginal probabilities of BN_{th} in Figure 2.9(a) | | | | | | | (b) Marginal probabilities from [51] | | | | | | |
|--|------|-------|------|-------|------|------|--------------------------------------|-------|-------|------|-------|-------|-------|
| Val. | R | L | A | C | B | S | Val. | R | L | A | C | B | S |
| 0 | 0.23 | 0.69 | 0.41 | 0.53 | 0.11 | 0.09 | 0 | 0.78 | 0.275 | 0.18 | 0.715 | 0.375 | 0.015 |
| 1 | 0.37 | 0.05 | 0.44 | 0.28 | 0.1 | 0.61 | 1 | 0.015 | 0.32 | 0.66 | 0.105 | 0.11 | 0.32 |
| 2 | 0.16 | 0.006 | 0.16 | 0.008 | 0.49 | 0.02 | 2 | 0.055 | 0.025 | 0.16 | 0.01 | 0.105 | 0.14 |
| 3 | 0.24 | 0.08 | - | 0.18 | 0.29 | 0.05 | 3 | 0.15 | 0.17 | - | 0.17 | 0.41 | 0.315 |
| 4 | - | 0.17 | - | - | - | 0.22 | 4 | - | 0.21 | - | - | - | 0.21 |

2.13.1 Chi-squared Test Short Description

Chi-square test is a statistical test commonly used in comparing observed data with data we would expect to obtain, according to a certain specific hypothesis. The formula for Chi-square test is:

$$\chi^2 = \sum (O_i - E_i)^2 / E_i$$

where O_i the observed value and E_i is expected data. So chi-square is the sum of the squared difference between observed and expected data in all possible categories.

By given significance level α and the degree of freedom $d = n - 1$, we have rejection region: $W = [0, \chi_{1-\alpha/2}^2(n-1)] \cup [\chi_{\alpha/2}^2(n-1), +\infty]$. If the calculated value of χ^2 lies in this region, we reject the hypothesis, else we accept the hypothesis.

2.13.2 Results of χ^2 Tests

For R

Hypothesis H_0 :The R's distribution between two sources agree.

The degree of freedom $d=n-1=3$, $\chi^2 = \frac{(0.23-0.78)^2}{0.78} + \frac{(0.37-0.015)^2}{0.015} + \frac{(0.16-0.055)^2}{0.055} + \frac{(0.24-0.15)^2}{0.15} = 9.04182$ and as we see $\chi_{0.05}^2(3) = 7.815$, so this value lies in W, we reject the hypothesis, they don't fit each other;

For L

Hypothesis H_0 :The L's distribution between two sources agree.

The degree of freedom $d=n-1=4$, $\chi^2 = \frac{(0.69-0.275)^2}{0.275} + \frac{(0.05-0.32)^2}{0.32} + \frac{(0.006-0.025)^2}{0.025} + \frac{(0.08-0.17)^2}{0.17} + \frac{(0.17-0.21)^2}{0.21} = 0.925976$ and as we see $\chi_{0.05}^2(4) = 9.488$, $\chi_{0.95}^2(4) = 0.711$, so this value lies outside W, we accept the hypothesis, they fit each other;

For A

Hypothesis H_0 :The A's distribution between two sources agree.

The degree of freedom $d=n-1=2$, $\chi^2 = \frac{(0.41-0.18)^2}{0.18} + \frac{(0.44-0.66)^2}{0.66} + \frac{(0.16-0.16)^2}{0.16} = 0.367$ and as we see $\chi_{0.05}^2(2) = 5.991$, $\chi_{0.95}^2(2) = 0.103$, so this value lies outside W, we accept the hypothesis, they fit each other;;

For C

Hypothesis H_0 :The C's distribution between two sources agree.

2.13. APPENDIX 3: COMPARISON OF TH MARGINALS WITH PREVIOUS WORKBIBLIOGRAPHY

The degree of freedom $d=n-1=3$, $\chi^2 = \frac{(0.53-0.715)^2}{0.715} + \frac{(0.28-0.105)^2}{0.105} + \frac{(0.008-0.01)^2}{0.01} + \frac{(0.18-0.17)^2}{0.17} = 0.340855$ and as we see $\chi_{0.05}^2(3) = 7.815$, $\chi_{0.95}^2(3) = 0.352$, so this value lies in W, we reject the hypothesis, they don't fit each other;

For B

Hypothesis H_0 :The B's distribution between two sources agree.

The degree of freedom $d=n-1=3$, $\chi^2 = \frac{(0.11-0.375)^2}{0.375} + \frac{(0.1-0.11)^2}{0.11} + \frac{(0.49-0.105)^2}{0.105} + \frac{(0.29-0.41)^2}{0.41} = 1.63458$ and as we see $\chi_{0.05}^2(3) = 7.815$, $\chi_{0.95}^2(3) = 0.352$, so this value lies outside W, we accept the hypothesis, they fit each other;

The degree of freedom $d=n-1=3$;

For S

Hypothesis H_0 :The S's distribution between two sources agree.

The degree of freedom $d=n-1=4$, $\chi^2 = \frac{(0.09-0.015)^2}{0.015} + \frac{(0.61-0.32)^2}{0.32} + \frac{(0.02-0.14)^2}{0.14} + \frac{(0.05-0.315)^2}{0.315} + \frac{(0.22-0.42)^2}{0.42} = 1.0588$ and as we see $\chi_{0.05}^2(4) = 9.488$, $\chi_{0.95}^2(4) = 0.711$, so this value lies outside W, we accept the hypothesis, they fit each other;;

Choice of Sig-value

By selecting the significance level's value of $\alpha = 0.1$, we have $1 - \alpha = 90\%$ possibility that our estimation is correct.

2.13.3 Discussion of Results

Thus we have that four of the six distributions agree while two disagree. The correlation is not as strong as we can expect due to several reasons:

- No distinction was made between cursive and hand-print in the *th* data. Their proportions in the two data sets may be quite different.
- In [51] no attempt was made for the writers to be representative of the population. Our samples are more representative.
- The characteristics of handwriting in the population may have changed over a period of 23 years.
- There were only 200 writers considered in [51] while 500 writers were considered here.
- Marginal probabilities may be less important than the joint probabilities in FDE. In [51] only the marginal probabilities and a few joint probabilities of pairs of variables are given. In our method marginal or any desired joint probability can be calculated.

2.14 Appendix 4: *and* examples

The images of *and* were extracted from the images of the CEDAR letter described in Appendix 1. There are one- to five-samples of the word in one page of writing, with three pages per writer. In the following we give some samples for cursive and handprint where for each image the characteristics entered by the document examiner are given. The probability of the characteristic encoding is given in two ways: as determined by a Bayesian network and as determined by assuming that the characteristics are independent. The highest and lowest probability samples, ordered by the BN determined joint probability are given.

2.14.1 Cursive

High probability cursive samples

Following are some high probability samples in the cursive data set. They are ordered by the joint probability computed by the Bayesian network described in Section 2.3.2.

| # | Samples | Writer ID | Characteristics | BN Joint Probability | Joint Probability assuming independence |
|----|--------------------------------|-----------|-----------------|----------------------|---|
| 1 | <i>and and and and and</i> | 0584a | 111022022 | 5.47e-003 | 5.06e-003 |
| 2 | <i>and and and</i> | 0584b | 111022022 | 5.47e-003 | 5.06e-003 |
| 3 | <i>and and and and</i> | 0584c | 111022022 | 5.47e-003 | 5.06e-003 |
| 4 | <i>and</i> | 1127a | 111022022 | 5.47e-003 | 5.06e-003 |
| 5 | <i>and and and.</i> | 1127b | 111022022 | 5.47e-003 | 5.06e-003 |
| 6 | <i>and and and</i> | 1127c | 111022022 | 5.47e-003 | 5.06e-003 |
| 7 | <i>and my and</i> | 1274a | 111022022 | 5.47e-003 | 5.06e-003 |
| 8 | <i>and and</i> | 1284a | 111022022 | 5.47e-003 | 5.06e-003 |
| 9 | <i>and and</i> | 1360b | 111022022 | 5.47e-003 | 5.06e-003 |
| 10 | <i>and and</i> | 1364b | 111022022 | 5.47e-003 | 5.06e-003 |
| 11 | <i>and and</i> | 1490c | 111022022 | 5.47e-003 | 5.06e-003 |
| 12 | <i>and and and</i> | 1491a | 111022022 | 5.47e-003 | 5.06e-003 |
| 13 | <i>and and and and and</i> | 1491b | 111022022 | 5.47e-003 | 5.06e-003 |
| 14 | <i>and and and and</i> | 0715b | 211022022 | 5.10e-003 | 6.23e-003 |
| 15 | <i>and and and and</i> | 0815c | 211022022 | 5.10e-003 | 6.23e-003 |
| 16 | <i>and and and</i> | 0896a | 211022022 | 5.10e-003 | 6.23e-003 |
| 17 | <i>and and and matters and</i> | 0896c | 211022022 | 5.10e-003 | 6.23e-003 |
| 18 | <i>and matters</i> | 0359a | 211022022 | 5.10e-003 | 6.23e-003 |
| 19 | <i>such and and</i> | 0577a | 211022022 | 5.10e-003 | 6.23e-003 |

Low probability cursive samples

Following are some low probability cursive samples.

| # | Samples | Writer ID | Characteristics | BN Joint Probability | Joint Probability assuming independence |
|------|----------------------------|-----------|-----------------|----------------------|---|
| 3055 | <i>anel in anel anel</i> | 0967b | 112432112 | 9.23e-008 | 1.42e-005 |
| 3056 | <i>and and and and and</i> | 1154c | 342422322 | 7.90e-008 | 1.43e-005 |
| 3057 | <i>and and and</i> | 1306c | 312432022 | 7.41e-008 | 9.58e-006 |
| 3058 | <i>and and</i> | 1042a | 042423112 | 5.32e-008 | 1.29e-006 |
| 3059 | <i>Int</i> | 0131c | 130101322 | 4.85e-008 | 1.42e-007 |
| 3060 | <i>and, and an</i> | 1098b | 300203302 | 4.80e-008 | 2.65e-007 |
| 3061 | <i>and and and</i> | 0293a | 020133132 | 3.51e-008 | 7.13e-008 |
| 3062 | <i>and and and</i> | 1529c | 132332022 | 3.27e-008 | 3.88e-007 |
| 3063 | <i>and and</i> | 1530a | 132332022 | 3.27e-008 | 3.88e-007 |
| 3064 | <i>and and and</i> | 1530b | 132332022 | 3.27e-008 | 3.88e-007 |
| 3065 | <i>and and and</i> | 1530c | 132332022 | 3.27e-008 | 3.88e-007 |
| 3066 | <i>and and</i> | 0058c | 142432122 | 2.35e-008 | 1.42e-005 |
| 3067 | <i>and and and</i> | 1200a | 302402322 | 2.28e-008 | 1.04e-005 |
| 3068 | <i>and and and and and</i> | 1200b | 302402322 | 2.28e-008 | 1.04e-005 |
| 3069 | <i>and and and</i> | 1200c | 302402322 | 2.28e-008 | 1.04e-005 |
| 3070 | <i>and and and</i> | 0042c | 142403222 | 1.53e-008 | 2.00e-006 |
| 3071 | <i>and and and and</i> | 0636b | 002432212 | 2.50e-009 | 3.10e-006 |
| 3072 | <i>and and</i> | 0603a | 342433102 | 1.24e-009 | 8.59e-007 |
| 3073 | <i>and and</i> | 0129a | 342431242 | 3.26e-010 | 4.81e-008 |
| 3074 | <i>and and and</i> | 1205b | 242433342 | 1.63e-010 | 9.30e-007 |
| 3075 | <i>and and</i> | 1205c | 222433342 | 1.61e-010 | 1.31e-007 |

2.14.2 Hand-print

High probability hand-print samples

Following are some high probability samples in the hand-print data set. They are ordered by the joint probability computed by the Bayesian network described in Section 2.3.2.

| # | Samples | Writer ID | Characteristics | BN Joint Probability | Joint Probability assuming independence |
|----|---------------------|-----------|-----------------|----------------------|---|
| 1 | andandandandand | 0008a | 010110112 | 1.67e-002 | 2.22e-002 |
| 2 | and | 0120a | 010110112 | 1.67e-002 | 2.22e-002 |
| 3 | and and and and and | 0183a | 010110112 | 1.67e-002 | 2.22e-002 |
| 4 | and and and and | 0183b | 010110112 | 1.67e-002 | 2.22e-002 |
| 5 | the and and my and | 0183c | 010110112 | 1.67e-002 | 2.22e-002 |
| 6 | and and and and and | 0196a | 010110112 | 1.67e-002 | 2.22e-002 |
| 7 | and and and and and | 0203b | 010110112 | 1.67e-002 | 2.22e-002 |
| 8 | and | 0205c | 010110112 | 1.67e-002 | 2.22e-002 |
| 9 | and | 0209a | 010110112 | 1.67e-002 | 2.22e-002 |
| 10 | and | 0209b | 010110112 | 1.67e-002 | 2.22e-002 |
| 11 | and no and | 0209c | 010110112 | 1.67e-002 | 2.22e-002 |
| 12 | and and no and and | 0214a | 010110112 | 1.67e-002 | 2.22e-002 |
| 13 | and and and and and | 0214b | 010110112 | 1.67e-002 | 2.22e-002 |
| 14 | and and and | 0240b | 010110112 | 1.67e-002 | 2.22e-002 |
| 15 | and | 0240c | 010110112 | 1.67e-002 | 2.22e-002 |
| 16 | and and and and and | 0258a | 010110112 | 1.67e-002 | 2.22e-002 |
| 17 | and and | 0641a | 010110112 | 1.67e-002 | 2.22e-002 |
| 18 | and and and | 0693a | 010110112 | 1.67e-002 | 2.22e-002 |
| 19 | and and | 0693b | 010110112 | 1.67e-002 | 2.22e-002 |

Low probability hand-print samples

Following are some low probability hand-print samples.

| # | Samples | Writer ID | Characteristics | BN Joint Probability | Joint Probability assuming independence |
|------|-----------------------|-----------|-----------------|----------------------|---|
| 1115 | andand | 0158a | 450211300 | 9.23e-009 | 4.03e-010 |
| 1116 | andandandand | 0158b | 450211300 | 9.23e-009 | 4.03e-010 |
| 1117 | andandandandand | 0158c | 140211300 | 5.35e-009 | 1.72e-008 |
| 1118 | andandandandand | 0698c | 140211300 | 5.35e-009 | 1.72e-008 |
| 1119 | ANDANDAND | 0548a | 130124532 | 5.24e-009 | 5.52e-009 |
| 1120 | ANDANDANDAND | 0598a | 130323332 | 5.08e-009 | 4.80e-007 |
| 1121 | ANDANDANDANDAND | 0598b | 130323332 | 5.08e-009 | 4.80e-007 |
| 1122 | ANDANDANDAND | 0598c | 130323332 | 5.08e-009 | 4.80e-007 |
| 1123 | andANDANDANDAND | 1110b | 453123332 | 3.56e-009 | 2.15e-008 |
| 1124 | ANDAND | 0275b | 313423122 | 2.28e-009 | 3.13e-006 |
| 1125 | ANDANDANDANDAND | 0801b | 353423002 | 2.11e-009 | 3.37e-007 |
| 1126 | ANDAND | 0103a | 353423322 | 1.34e-009 | 1.84e-006 |
| 1127 | AND | 0103b | 353423322 | 1.34e-009 | 1.84e-006 |
| 1128 | AND, AND AND AND AND | 0103c | 353423322 | 1.34e-009 | 1.84e-006 |
| 1129 | AND | 0275a | 313323122 | 1.31e-009 | 1.80e-006 |
| 1130 | ANDANDandAND | 0107a | 353423532 | 8.38e-010 | 7.32e-007 |
| 1131 | AND ^{en} | 0107b | 353423532 | 8.38e-010 | 7.32e-007 |
| 1132 | AND ^{en} AND | 0107c | 353423532 | 8.38e-010 | 7.32e-007 |
| 1133 | ANDAND | 1233b | 333323332 | 6.69e-010 | 5.88e-007 |
| 1134 | andandandandand | 1110c | 453124532 | 4.39e-010 | 4.53e-010 |
| 1135 | ANDANDANDANDAND | 0629a | 353423522 | 2.85e-010 | 3.90e-007 |

2.15 Appendix 5: Type Determination

Comparability is a requirement in QD examination. If it can be performed automatically, the data collection process can be speeded-up. We describe here such a method which also assigns a probability to whether a handwritten item is cursively written or hand-printed. This work described in [11] has been incorporated into the CEDAR-FOX system [78, 74]. Formal analysis of its performance was made possible with ground truth over entire documents provided by QD examiners.

2.15.1 Characteristics of Type

Character connectivity is useful to differentiate type. If most characters within each word are connected, the item is cursive. If most are disconnected, it is hand printed. Mixed type is characterized by the term *running hand print*, but often grouped with cursive since connected characters are useful in analysis. To capture the idea of whether or not most characters are connected, we identified several characteristics that could be automatically determined.

- Discreteness f_1 : The presence of disconnected characters in a handwritten item is given by the ratio of *Isolated Character Count (ICC)* and *Word Count (WC)*. ICC is the cardinality of the set of groups of connected components recognized as individual characters. WC is the cardinality of the set of words in a given document. We use WC simply as a way to normalize ICC; thus, the feature we found to best capture the intuitive document examiner feature is the ratio $f_1 = \frac{ICC}{WC}$.
- Loopiness f_2 : As cursive writing tends to have more loops present within each connected component, a second feature considered is loopiness. We capture this with the ratio of interior to exterior contours. Due to the complexity, this feature was not explored in depth.

To illustrate these characteristics, consider Figure 2.29 which shows two instances of the word *several*, one cursive and the other hand-printed. In the cursive instance, there are two large connected components; one containing *sev* and one containing *eral*: $ICC = 0$ and $WC = 1$, yielding $f_1 = 0$. The hand-print instance, however, contains six large connected components. Five are recognized as individual characters (*s, e, v, e* and *r*), giving a ratio of 5. Thus, our hypothesis was that cursive documents would have overall very low ratios and hand printed documents would have relatively high ratios. Even ideal cursive documents are likely to have some isolated characters (due to the words *a* and *I*). To complicate matters, however, some spurious breaks often occur, even in predominantly cursive documents. An example is the instance of the word *enthusiasm* which occurred in a predominantly cursive document shown in Figure 2.29(c). In this case, there are six connected components (*en, t, hus, i, as, m*). Since three of these are individual characters in a single word, $f_1 = 3/1$.

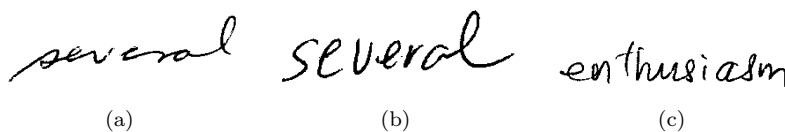


Figure 2.28: Examples of word type: (a) cursive: $f_1 = 0, f_2 = 2.5$, (b) hand-printed: $f_1 = 5, f_2 = 0.5$, and (c) predominantly cursive: $f_1 = 3, f_2 = 1.33$.

Approximating WC

In order to process arbitrary documents which lack ground truth, we first approximate the WC. The document is first segmented into lines and then into a set of numbered words. To segment lines, a stroke tracing method is used to intelligently segment the overlapping components. Slope and curvature information of the stroke is used to disambiguate the path of the stroke at cross points. Once the overlapping components are segmented into strokes, a statistical method is used to associate the strokes with the appropriate lines. This method is capable of disentangling lines which collide by splitting and associating the correct character strokes to the appropriate lines [45]. To segment lines into words, a gap metric approach is used [38]; it is designed to separate a line of unconstrained (written in a natural manner) handwritten text into words. Both local and global features are used to mimic human cognition. Local features are distance between a pair of components, distance between neighboring components, width of left and right components, and height of the left and right components. Global features include ratio of the number of exterior contours and the number of interior contours, the average height of grouped components, average width of grouped components, and average distance between components. Two distance metrics are computed and averaged

depending on the circumstances the first metric is either the bounding box or minimum run-length distance and the second is the convex hull distance. The actual classification based on these features is done using a three-layer neural network.

From Jim Elder
819 Long Street, Apt 300
Allentown, New York 14707

Nov 10, 1999

To Dr. Bob Givand
602 Queensberry Parkway
Oman, West Virginia 25638

We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Eck.

It all started around 6:30 months ago while attending the "Pitney" Jazz Concert. Organizing such an event is intricate, and as President of the Alumni Association, a co-sponsor of the event, Kate was overwhelmed. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm.

However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, X-rays and blood tests later, were told it was just exhaustion.

Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?

Thank you!

Jim.

(a)

From Jim Elder
819 Long Street, Apt 300
Allentown, New York 14707

Nov 10, 1999

To Dr. Bob Givand
602 Queensberry Parkway
Oman, West Virginia 25638

We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Eck.

It all started around 6:30 months ago while attending the "Pitney" Jazz Concert. Organizing such an event is intricate, and as President of the Alumni Association, a co-sponsor of the event, Kate was overwhelmed. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm.

However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, X-rays and blood tests later, were told it was just exhaustion.

Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?

Thank you!

Jim.

(b)

Micro Features

Auto Truthed Cropped

Number of Chars : 59

Enter Character ID

Number of Words : 131

Number of Lines : 21

Document Type :
 Print Cursive

(c)

Micro Features

Auto Truthed Cropped

Number of Chars : 326

Enter Character ID

Number of Words : 159

Number of Lines : 22

Document Type :
 Print Cursive

(d)

Figure 2.29: Determination of cursive (a) and hand print (b) within the CEDAR-FOX system. Screenshots in (c) and (d) show result on a continuous scale as predominantly cursive and predominantly hand-printed.

Determining ICC

Utilizing a character model recognizer, the full preprocessed page image is separated into its connected components. Each connected component is then passed to a character recognition algorithm which determines if the connected component is likely to consist of a single character. The features used for this determination are the Gradient, Structural, Concavity (GSC) features, a 512-dimensional feature vector, which are used in automatic character recognition for interpreting handwritten postal addresses. We extract local contour features based on the pixel

gradients present in an image. A gradient map, i.e., gradient magnitude and direction at every pixel, is computed; this map is thresholded to avoid responses to noise and spurious artifacts, the map is then partitioned coarsely. Character contours are encoded by quantizing gradient directions into a few representative direction bins. This method does not require that images be normalized to a fixed grid-size before feature extraction. The features were used to train a 2-layer neural network classifier [79].

2.15.2 Dataset

The dataset used for our experiments is the entire CEDAR letter dataset described in Appendix 1. The documents in this data set were examined and tagged by two QD examiners with a Boolean flag as either hand printed or not hand printed. The not hand printed group consists of mainly cursive documents with a few questionable running hand print documents. This feature was created by the document examiners as it captured the sets information necessary for their work. An example of a predominantly cursive and a hand print document is shown in Figure 2.29. The QD examiners tagged 621 documents as hand printed with the remaining 4080 being cursive with a few running hand print.

2.15.3 Type Distribution

We began by processing the CEDAR letter dataset in its entirety, generating our measurement on all documents to determine whether or not it had good predictive value. That is, we wanted to see how well the IsolatedCharacterCount/WordCount feature correlated with the Boolean feature value provided by the QD examiners.

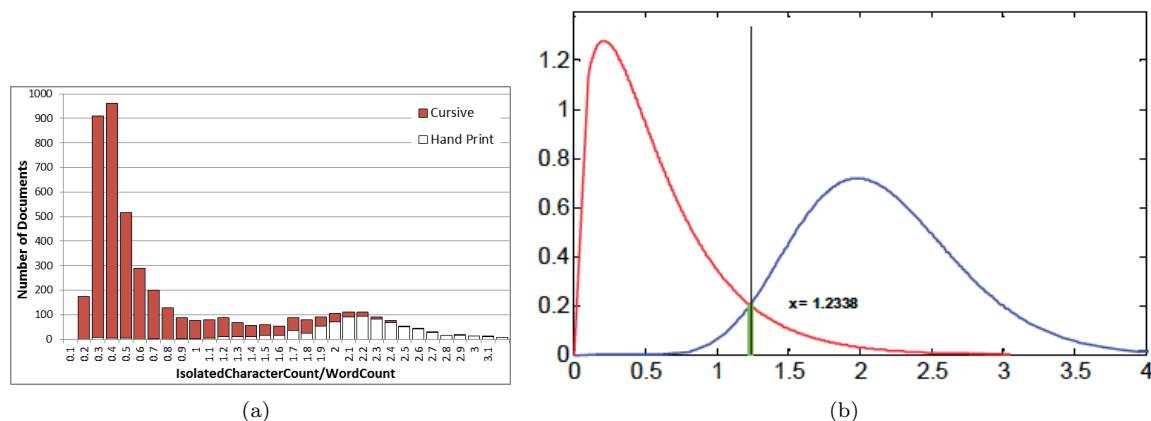


Figure 2.30: Histogram of binned feature f_1 (left) and representative Gamma distributions (right) with their thresholds.

The ratio had a range of 0.10 to 6.86 over all 4701 documents. Of the hand printed documents, the range was 0.20 to 6.36 with a mean of 2.14. Cursive documents have values ranging from 0.10 to 3.05 with a mean of 0.57. Figure 2.30 is a stacked histogram of (99% of the data is included; the few points with values > 3.2 are omitted for space) as well as the data modeled with Gamma distributions. We determined a threshold by using the midpoint of the means of the two subsets (cursive and hand print) in the training set. The mean values of the cursive and hand print training sets were found to be 0.57 and 2.08 respectively. The threshold was set at 1.33. We performed a second experiment by modeling the two subsets with two Gamma distributions and choosing the class with higher probability; for the experiment, we approximated the distributions using Gaussian distributions.

2.15.4 Results

We performed experiments both based on a threshold and Gaussian model. The mean values of the cursive and hand print training sets were found to be 0.57 and 2.08 respectively. The threshold was set at 1.33. 1775 documents in the cursive validation set were found to have values below the threshold, yielding the correct classification of 90.8%. 375 of the hand printed validation set were found to have values above the threshold, yielding correct classification of 92.0% of the documents. Overall, this led to the correct classification of an average of 94.5%. The Gaussian

experiment yielded very similar results with performance improving on the hand printed set to 95.0% and decreasing slightly to 90.5% on the cursive set.

2.15.5 Conclusions

In conclusion, feature f_1 identifies writing type correctly roughly 92.8% of the time. The method needs to be further evaluated at the word level where a higher level of discrimination will be needed. Incorporation of the f_2 feature will likely improve classification performance. Such a tool can eventually be incorporated into the ground-truthing tool described in Section 2.2.5 so that the menu of characteristics can be automatically displayed.

2.16 Appendix 6: Mapping Likelihood Ratio to Opinion Scale

We describe below a method for converting a likelihood ratio between the identification and exclusion hypotheses into a discrete opinion on a nine-point scale. It has been implemented previously in the CEDAR-FOX system using automatically determined characteristics [40]. The distance space distribution is modeled parametrically using gamma and Gaussian densities obtained by comparing ensembles of pairs of documents. Experiments and results show that with increase in information content from just a single word to a full page of document, the accuracy of the model increases.

The formulation of strength of evidence is parameterized based on two factors: (i) the amount of information compared (\mathcal{I}) and (ii) whether the documents being compared have the same or different content (\mathcal{C}). Values that \mathcal{I} can take is discretized as one of *word, line, multiple lines, half page, full page*. Words of different length (short, medium and long) and words made of purely numbers were all considered as belonging to the type word. The two different values \mathcal{C} can take on are *Same Content, Different Content*.

The value of \mathcal{I} can be automatically found during line and word segmentation. The automatic line segmentation method is discussed in [5] and for word segmentation an artificial neural network decides whether a gap between two connected components is a word gap or not. Using the number of lines L and the number of words in each line W_i $i \in \{1 \dots L\}$, the value of \mathcal{I} is decided as: $[L = 1, W_1 = 1 \Rightarrow \mathcal{I} = \text{word}]$, $[L = 1, W_1 > 1 \Rightarrow \mathcal{I} = \text{line}]$, $[L > 1 \ \& \ L < 4 \Rightarrow \mathcal{I} = \text{multiple-lines}]$, $[L \geq 4 \ \& \ L < 8 \Rightarrow \mathcal{I} = \text{half-page}]$, $[L \geq 8 \Rightarrow \mathcal{I} = \text{full-page}]$. Similarly for the value of \mathcal{C} , if the number of common words between the two documents is greater than 80% of the number of words in the smaller document (in terms of number of lines), then the value of \mathcal{C} is “Same Content”, or else it is different content.

Mathematical formulation

For each possible pairs of settings for \mathcal{I} and \mathcal{C} , the distribution of the LLR as observed on a validation set of ensemble of pairs is obtained. This ensemble of pairs consists of both, pairs from same as well as from different writers. The number of such pairs from the same and different writers are kept the same to avoid a bias in the distribution of LLR. Let D_{ic} represent the distribution of LLR for $\mathcal{I} = i$ and $\mathcal{C} = c$. Further, let D_{ic}^S represent the subset of D_{ic} where the samples truly belonged to the same writer and let D_{ic}^D represent the subset of D_{ic} where the samples truly belonged to different writers. It is clear that $D_{ic} = D_{ic}^S \cup D_{ic}^D$. Here it is important to note that the distribution D_{ic}^S and D_{ic}^D will be different for different sets of features used. For eg., the distribution can be further parameterized by a third variable that measures which feature set was used (macro only or macro+micro). We leave the discussion of inclusion of this third parameter to the experiments and results section. Using the distributions D_{ic}^S and D_{ic}^D , and for any given value of LLR \mathcal{L} , two percentages can now be obtained (i) P_{ic}^S : Percentage of samples in D_{ic}^S that had LLR values $> \mathcal{L}$ and (ii) P_{ic}^D : Percentage of samples in D_{ic}^D that had LLR values $> \mathcal{L}$. To be verbose, P_{ic}^S represents the percentage of same writer cases in the validation set that had LLR values even larger than the one for this. This implies that P_{ic}^S represents the percentage of same writer cases in the validation set that were *stronger* than the current case. Similarly, P_{ic}^D represents percentage of different writer cases that were *weaker* than the current case. Mathematically, they are defined as in Eq. 2.35.

$$P_{ic}^S = \frac{|D_{ic}^S > \mathcal{L}|}{|D_{ic}^S|} \times 100$$

$$P_{ic}^D = \frac{|D_{ic}^D > \mathcal{L}|}{|D_{ic}^D|} \times 100 \quad (2.35)$$

where $|\cdot|$ represents cardinality. It is clear that $P_{ic}^{S'} = 100 - P_{ic}^S$ will represent that percentage of samples in D_{ic}^S that had LLR values $\leq \mathcal{L}$ and similarly we define $P_{ic}^{D'} = 100 - P_{ic}^D$. $P_{ic}^{S'}$ and $P_{ic}^{D'}$ represent the complement of P_{ic}^S and P_{ic}^D respectively.

The sign(+ve,-ve) of the LLR \mathcal{L} between a pair of documents makes a decision of same or different writer. The strength of evidence is based on this decision. The scale 1...9 for a particular pair of document can be obtained using either P_{ic}^S (if $\mathcal{L} > 0$) or $P_{ic}^{D'}$ (if $\mathcal{L} < 0$). (In both cases, we are evaluating the percentage of samples that were stronger than the current case.) These two values can be calculated using Equation 2.35 provided i and c are known. The beginning of this section described the method to calculate these i and c . If the LLR \mathcal{L} is +ve, then the opinion scale is in the range 1-5 and in the range 5-9 if it is -ve. Note that, in either cases, the scale “5-No conclusion” can be obtained. Table 2.12 summarizes the rules for obtaining the nine-point scale for +ve and -ve LLR values.

Table 2.12: Rules for obtaining an opinion on the 9 point scale

(a) LLR \mathcal{L} is +ve(b) LLR \mathcal{L} is -ve

| Scale | Opinions for same | P_{ic}^S | Scale | Opinions for different | $P_{ic}^{D'}$ |
|-------|----------------------|----------------|-------|---------------------------|----------------|
| 1 | Identified as same | 0.00 ~ 22.21 | 5 | No conclusion | 88.88 ~ 100.00 |
| 2 | Highly probably same | 22.22 ~ 44.43 | 6 | Indicating different | 66.66 ~ 88.87 |
| 3 | Probably same | 44.44 ~ 66.65 | 7 | Probably different | 44.44 ~ 66.65 |
| 4 | Indicating same | 66.66 ~ 88.87 | 8 | Highly probable different | 22.22 ~ 44.43 |
| 5 | No conclusion | 88.88 ~ 100.00 | 9 | Identified as different | 0.00 ~ 22.21 |