Statistical Analysis of Handwriting: Probabilistic outcomes for closed-set writer identification

Amy Crawford, MSc
Alicia Carriquiry, PhD
Danica Ommen, PhD

AAFS 2020, Anaheim, CA

csafe
Center for Statistics and Applications in Forensic Evidence

Forensic Statistics Research at CSAFE

The **C**enter for **S**tatistics and **A**pplications in **F**orensic **E**vidence

- NIST Center of Excelllence
- Three-part mission: **research**, outreach, training.

Funding Statement

This work was partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement #70NANB15H176 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California Irvine, and University of Virginia.

Introduction

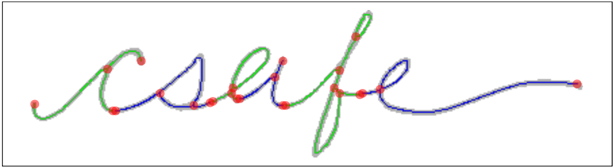
Objective

Use a statistical model to provide probabalistic statements of writership for handwritten documents.

- Without character recognition
- Robust to writing style - cursive, print
- Closed set of writers - search a collection

Data Processing with *handwriter*

The R package *handwriter*[1] takes in a scanned handwritten document. Then,

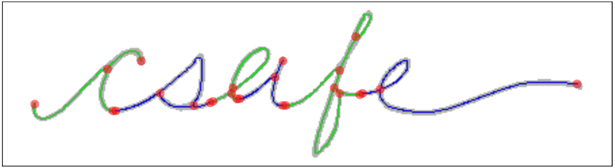


1. Binarize
   - Turn the image to pure black and white.

[1] https://github.com/CSAFE-ISU/handwriter

Data Processing with *handwriter*

The R package *handwriter*[1] takes in a scanned handwritten document. Then,

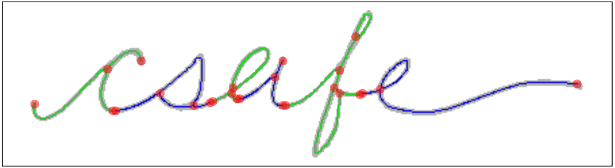


2. Skeletonize
   - Reduce writing to a 1 pixel wide skeleton.

[1] https://github.com/CSAFE-ISU/handwriter

Data Processing with *handwriter*

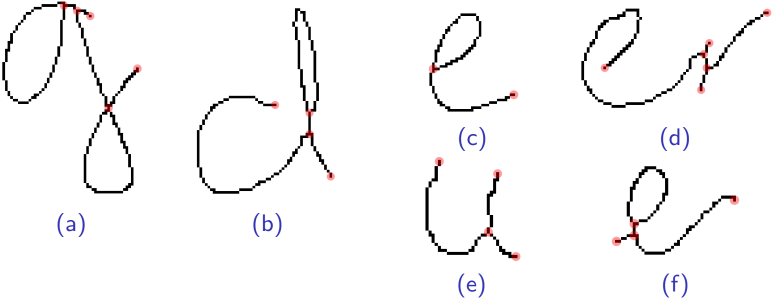
The R package *handwriter*[1] takes in a scanned handwritten document. Then,



3. Break
    - Connected writing is decomposed into small manageable graphical structures.
    - Often, but not always, correspond to Roman letters.

[1] https://github.com/CSAFE-ISU/handwriter

Handwriting as Data



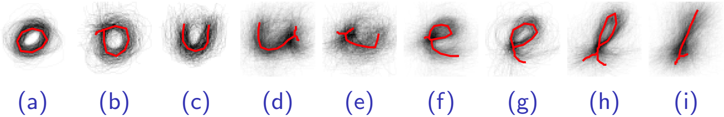
Writing as graphical structures.

- Parziale, et al. (2014), Miller et. al. (2017), others
- For us, attributed graphs with nodes and edge locations

Parziale, Antonio, et al. *An interactive tool for forensic handwriting examination*. 14th International Conference on Frontiers in Handwriting Recognition. IEEE, 2014.

Miller, J. J. et al. (2017). *A set of handwriting features for use in automated writer identification. Journal of forensic sciences*, 62(3), 722-734.

Group like structures with $K-$Means - I



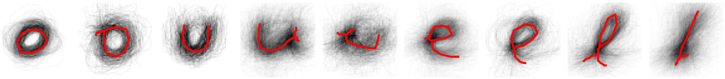
Handwriting elements/measurements into “bins” / “buckets”.

- Bulacu and Schomaker (2007) , Saunders et al. (2011) , others
- For us, flexible and structure based through clustering.

Bulacu, M. and Schomaker, L. (2007). Text-independent writer identification and verification using textural and allographic features. IEEE trans-actions on pattern analysis and machine intelligence, 29(4):701-717.

Saunders, C. P., Davis, L. J., Lamas, A. C., Miller, J. J., and Gantz, D. T. (2011). Construction and evaluation of classifiers for forensic document analysis. The Annals of Applied Statistics, 5(1):381-399.

Group like structures with $K-$Means - II

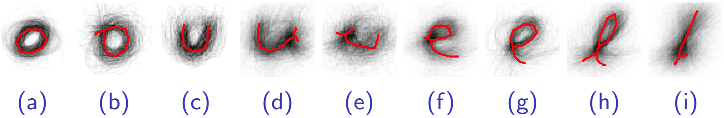


(a) (b) (c) (d) (e) (f) (g) (h) (i)

Joint work with Nick Berry, PhD.

40 clusters → 40 centers that make up the template.

Group like structures with $K-$Means - III



Three data sources for template creation.

1. CSAFE Handwriting Database[2], 25 documents, 1 prompt.
2. CVL Database[3], 25 documents, 6 prompts.
3. IAM Handwriting Database[4], 50 documents, 50 prompts.

[2] A Crawford, A Ray, A Carriquiry, J Kruse, M Peterson (2019). CSAFE Handwriting Database. Iowa State University. Dataset. https://doi.org/10.25380/iastate.10062203.v1

[3] F Kleber, S Fiel, M Diem, R Sablatnig (2013). CVL-DataBase: An Off-Line Database for Writer Retrieval, Writer Identification and Word Spotting in 2013 12th International Conference on Document Analysis and Recognition. pp. 560–564.

[4] UV Marti, H Bunke (2002). The IAM-database: An English sentence database for offline handwriting recognition. Int. J. on Document Analysis Recognit.5, 39–46.

Feature Extraction with Template

All graphs from training and testing documents are filtered throught the template and assigned to the nearest center.

Feature Extraction with Template

All graphs from training and testing documents are filtered throught the template and assigned to the nearest center.

$Y_{doc,writer}$	$Cluster_1$	$Cluster_2$	$Cluster_3$	$Cluster_4$	...	$Cluster_{39}$	$Cluster_{40}$
$Y_{1,1}$	42	21	9	5	...	1	1
$Y_{1,38}$	39	91	23	6	...	0	1
$Y_{1,95}$	38	81	16	14	...	0	0
$\vdots$							

Model #1

$$Y_{doc,writer} \sim \mathbf{f_1}(Y_{doc,writer} | \pi_{writer})$$

[5] A.S. Osborn, 1929. Questioned documents, 2nd edn. New York, NY: Boyd Printing Co.

[6] L. F. Baum, 1900. The Wonderful Wizard of Oz, illustrated by W.W. Denslow. Chicago and New York: G.M. Hill Co.

Model #1

$$Y_{doc,writer} \sim \mathbf{f_1}(Y_{doc,writer} | \pi_{writer})$$

Model data come from 90 writers in the CSAFE Database.

- 3 training documents (most), London Letters[5]

[5] A.S. Osborn, 1929. Questioned documents, 2nd edn. New York, NY: Boyd Printing Co.

[6] L. F. Baum, 1900. The Wonderful Wizard of Oz, illustrated by W.W. Denslow. Chicago and New York: G.M. Hill Co.

Model #1

$$Y_{doc,writer} \sim \mathbf{f_1}(Y_{doc,writer}|\pi_{writer})$$

Model data come from 90 writers in the CSAFE Database.

- 3 training documents (most), London Letters[5]

$$Y_{1,writer} \sim \mathbf{f_1}(Y_{1,writer}|\pi_{writer})$$
$$Y_{2,writer} \sim \mathbf{f_1}(Y_{2,writer}|\pi_{writer})$$
$$Y_{3,writer} \sim \mathbf{f_1}(Y_{3,writer}|\pi_{writer})$$

[5] A.S. Osborn, 1929. Questioned documents, 2nd edn. New York, NY: Boyd Printing Co.

[6] L. F. Baum, 1900. The Wonderful Wizard of Oz, illustrated by W.W. Denslow. Chicago and New York: G.M. Hill Co.

Model #1

$$Y_{doc,writer} \sim \mathbf{f_1}(Y_{doc,writer} | \pi_{writer})$$

Model data come from 90 writers in the CSAFE Database.

- 3 training documents (most), London Letters[5]

$$\begin{aligned} Y_{1,writer} &\sim \mathbf{f_1}(Y_{1,writer} | \pi_{writer}) \\ Y_{2,writer} &\sim \mathbf{f_1}(Y_{2,writer} | \pi_{writer}) \\ Y_{3,writer} &\sim \mathbf{f_1}(Y_{3,writer} | \pi_{writer}) \end{aligned}$$

- 1 testing document, Wizard of Oz[6] Excerpt

[5] A.S. Osborn, 1929. Questioned documents, 2nd edn. New York, NY: Boyd Printing Co.

[6] L. F. Baum, 1900. The Wonderful Wizard of Oz, illustrated by W.W. Denslow. Chicago and New York: G.M. Hill Co.

Model #1

Fit/train with

$$Y_{doc,writer} \sim \mathbf{f_1}(Y_{doc,writer}|\pi_{writer})$$

Model #1 Results

Data for testing document, $Y_{????}$

Model #1

Fit/train with

$$Y_{doc,writer} \sim \mathbf{f_1}(Y_{doc,writer} | \pi_{writer})$$

Model #1 Results

Data for testing document, $Y_{????}$

$$\mathbf{f_1}(Y_{????} | \pi_{writer})$$

Model #1

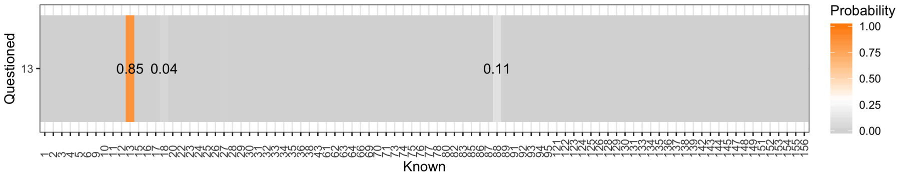
Fit/train with

$$Y_{doc,writer} \sim \mathbf{f_1}(Y_{doc,writer} | \pi_{writer})$$

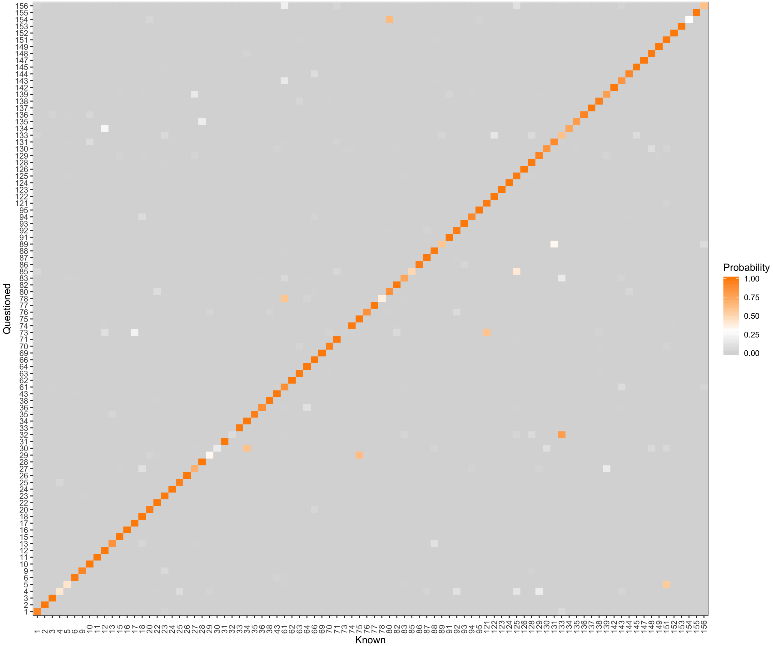
Model #1 Results

Data for testing document, $Y_{????}$

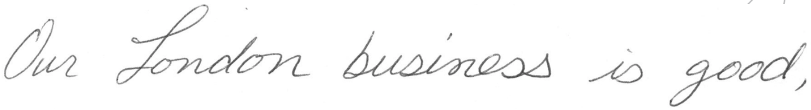
$$\mathbf{f_1}(Y_{????} | \pi_{writer})$$



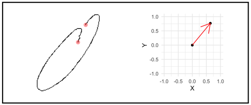
88.05% probability is on-diagonal.



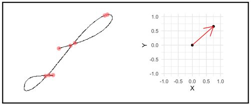
Rotation Angles



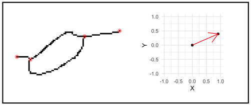
Our London business is good,



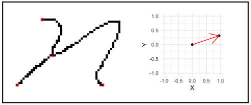
(a)



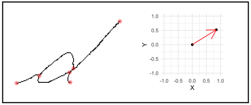
(b)



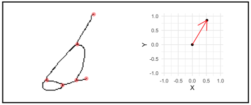
(c)



(d)

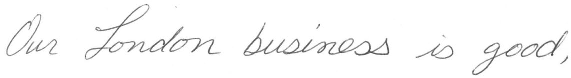


(e)



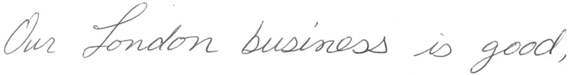
(f)

Rotation Angles by Cluster, Writer 95

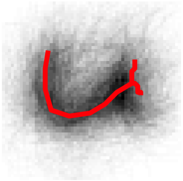


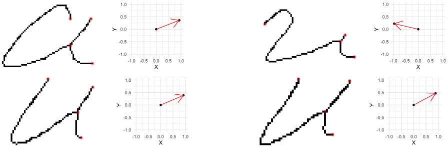
Our London business is good,

Rotation Angles by Cluster, Writer 95

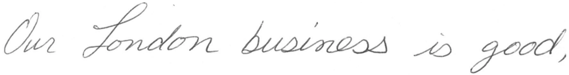


Our London business is good,

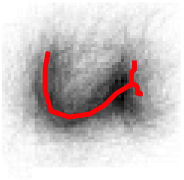


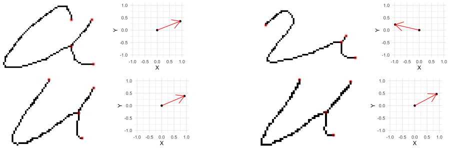


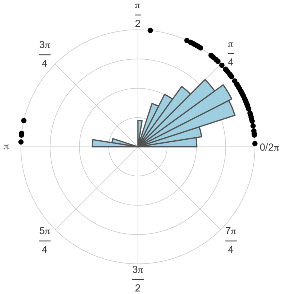
Rotation Angles by Cluster, Writer 95



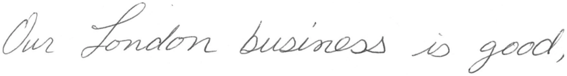
Our London business is good,



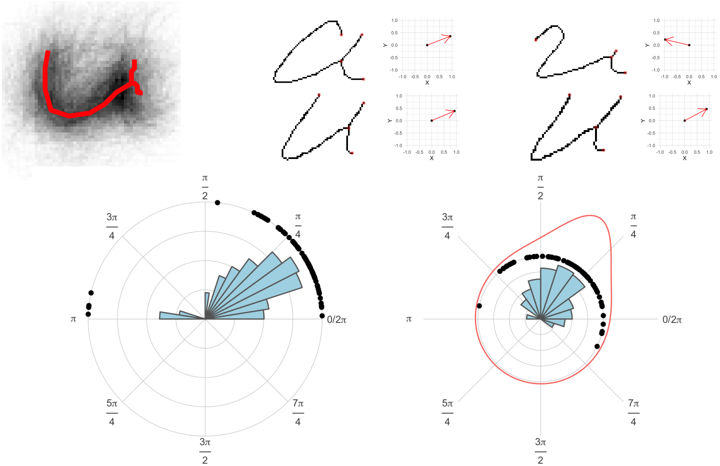




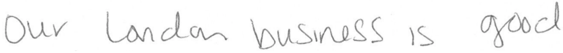
Rotation Angles by Cluster, Writer 95



Our London business is good,

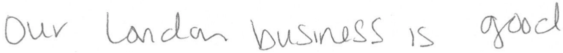


Rotation Angles by Cluster, Writer 1

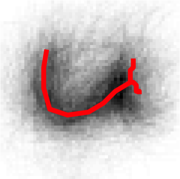


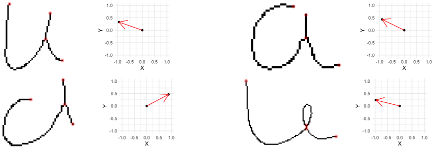
Our London business is good

Rotation Angles by Cluster, Writer 1

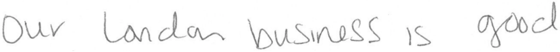


Our London business is good

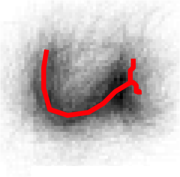


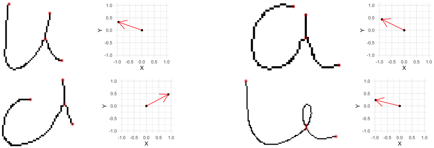


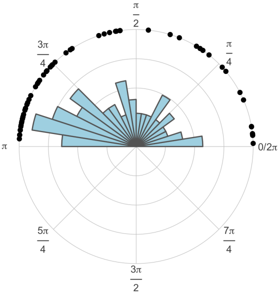
Rotation Angles by Cluster, Writer 1



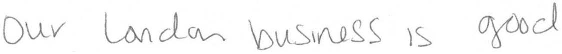
Our London business is good



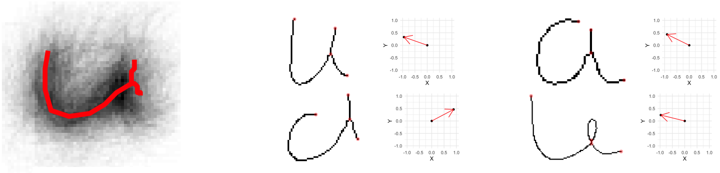


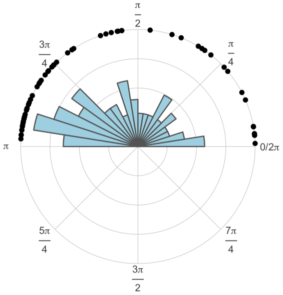


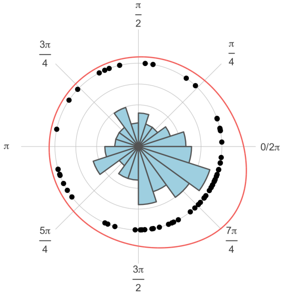
Rotation Angles by Cluster, Writer 1



Our London business is good







Model #2

$$Y_{doc,writer}, RA_{cluster,writer} \sim \mathbf{f_2}(Y_{doc,writer}, RA_{cluster,writer} | \pi_{writer}, \alpha_{cluster,writer})$$

Model #2

$$Y_{doc,writer}, RA_{cluster,writer} \sim \mathbf{f_2}(Y_{doc,writer}, RA_{cluster,writer} | \pi_{writer}, \alpha_{cluster,writer})$$

Model #2 Results

Data for testing document, $Y_{????}$ & $RA_{cluster,????}$ for all 40 clusters.

Model #2

$$Y_{doc,writer}, RA_{cluster,writer} \sim \mathbf{f_2}(Y_{doc,writer}, RA_{cluster,writer} | \pi_{writer}, \alpha_{cluster,writer})$$

Model #2 Results

Data for testing document, $Y_{????}$ & $RA_{cluster,????}$ for all 40 clusters.

$$\mathbf{f_2}\left(Y_{????}, RA_{cluster,????} | \pi_{writer}, \alpha_{cluster,writer}\right)$$

Model #2

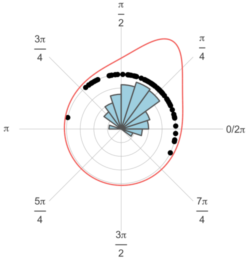
$$Y_{doc,writer}, RA_{cluster,writer} \sim \mathbf{f_2}(Y_{doc,writer}, RA_{cluster,writer} | \pi_{writer}, \alpha_{cluster,writer})$$

Model #2 Results

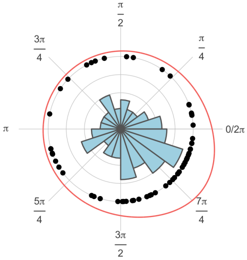
Data for testing document, $Y_{????}$ & $RA_{cluster,????}$ for all 40 clusters.

$$\mathbf{f_2}\left(Y_{????}, RA_{cluster,????} | \pi_{writer}, \alpha_{cluster,writer}\right)$$

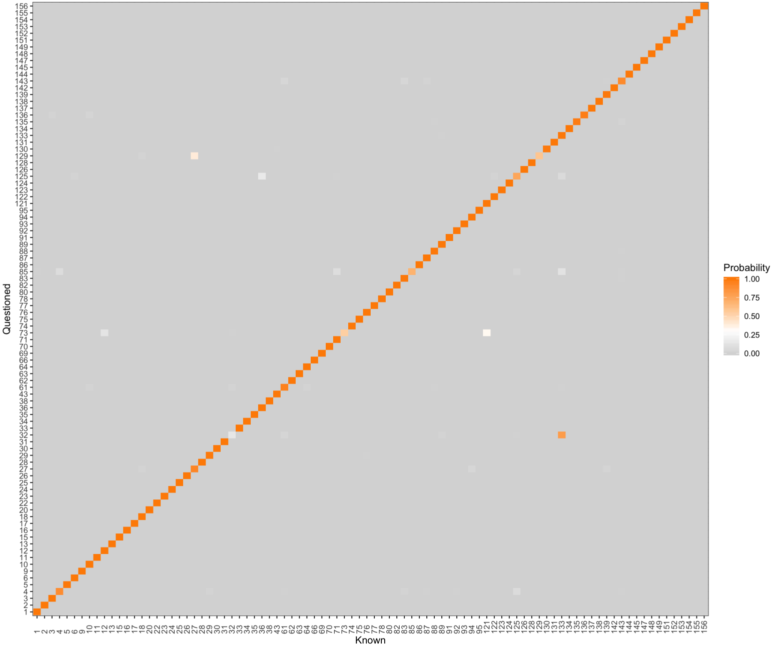
Writer 95, Cluster 29



Writer 1, Cluster 29



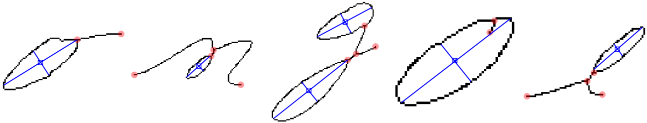
96.99% probability is on-diagonal.



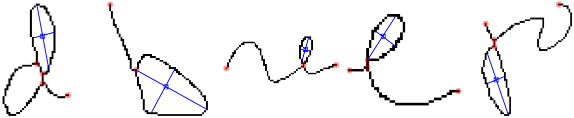
More measurements...

Loops, for example.

Writer 95:



Writer 1:



Thank you!