Adaption of the Chumbley Score to matching of bullet striation marks

Ganesh Krishnan *
Department of Statistics, Iowa State University
and
Heike Hofmann
Department of Statistics and CSAFE, Iowa State University

March 4, 2018

Abstract

Keywords: 3 to 6 keywords, that do not appear in the title

*The authors gratefully acknowledge ...

Contents

1 Introduction and Background

1.1 Motivation

Same source analyses are a major part of an Forensic Toolmark Examiner's job. In current practice examiners make these comparisons visually under a comparison microscope and come to one of the following four conclusions: identification, inconclusive, elimination or unsuitable for examination (AFTE Glossary 1998). These conclusions are made on the basis of "unique surface contours" of the two toolmarks being in "sufficient agreement" (AFTE Glossary 1998). AFTE describes the term "sufficient agreement" as the possibility of another tool producing the markings under comparison, as practically impossible (AFTE Glossary 1998). This subjectivity in the assessment as well as the lack of error rates are the main points of criticisms in the toolmark examination first raised by the National Research Council in 2009 (National Research Council 2009) and later emphasized further by the President's Council of Advisors on Science and Technology (President's Council of Advisors on Science and Technology 2016).

Technological advances, such as profilometers and confocal microscopy allow to measure 3D surfaces (Vorburger et al. 2016) in a digitized form. This forms the basis of statistical analysis of toolmarks. A statistical approach based on data removes both subjectivity from the assessment and allows a quantification of error rates for both false positive and false negative identifications.

Various toolmarks have been studied in the literature: screwdriver marks digitized using a profilometer have been analyzed in Faden et al. (2007) and Chumbley et al. (2010); Bachrach et al. (2010) have investigated 3D marks from screwdriver, tongue and groove pliers captured using a confocal microscope; digitized marks from slip-joint pliers generated by a surface profilometer have been investigated by Grieve et al. (2014).

Bachrach et al. (2010) define a relative distance metric and use it as similarity measure between two toolmarks. Faden et al. (2007) extract many small segments in the markings of two toolmarks and compare similiarity using a maximum pearson correlation coefficient. The Chumbley scoring method, first introduced by Chumbley et al. (2010), uses a similar but more extensive framework based on a Mann-Whitney U test of the resulting correla-

tion coefficients. This approach was later improved by Hadler and Morris (Hadler & Morris 2017) where they introduce a deterministic approach and address the issue of lack of indepedece between segements of striae. In this paper, we are investigating the applicability of the Chumbley scoring method by Hadler & Morris (2017) to assess striation marks on bullet lands for same-source identification.

Striation marks on bullets are made by impurities in the barrel. As the bullet travels through the barrel, these imperfections leave "scratches" on the bullet surface. Typically, only striation marks in the land engraved areas (LEAs) are considered AFTE Criteria for Identification Committee (1992). Bullet lands are depressed areas between the grooves made by the rifling action of the barrel. Compared to toolmarks made by screwdrivers striation marks on bullets are typically much smaller, both in length and in width. Bullets also have a curved cross-sectional topography sketch?. Figure 1 shows us how the signature from a bullet land lines up with the image of the land from which it was extracted. We can also see in the figure how the depth and relative position of the striation markings seen in the image are interpreted as the signature.

Bullet matching methods depend on the attribute being used for making comparisons. For single feature based matching Chu et al. (2013) used an automatic method for counting consequtive matching striae (CMS). They reported an error rate 52% of the known same source lands comparisons as misidentified and zero false positives for known different source lands. Ma et al. (2004) and Vorburger et al. (2011) discuss about CCF (cross-correlation function) and its discriminating power and application but do not talk about matches and non-matches. Hare et al. (2016) use multiple attributes like CCF, CMS, D (distance measure) etc in a random forrest based method and compare every land against every other land of the Hamby dataset Hamby et al. (2009). They report an out-of-bag error rate of 0.0046.

The Chumbley score provides us with another approach in the same-source assessment of bullet striation marks.Chumbley et al. (2010), in their paper, compare two toolmarks with the intention of determining if it comes from the same tool. The data for their study was obtained from 50 sequentially manufactured screwdriver tips.Chumbley et al. (2010) reported error rates for marking made by the tips at differnt angles. For a 30 degree they

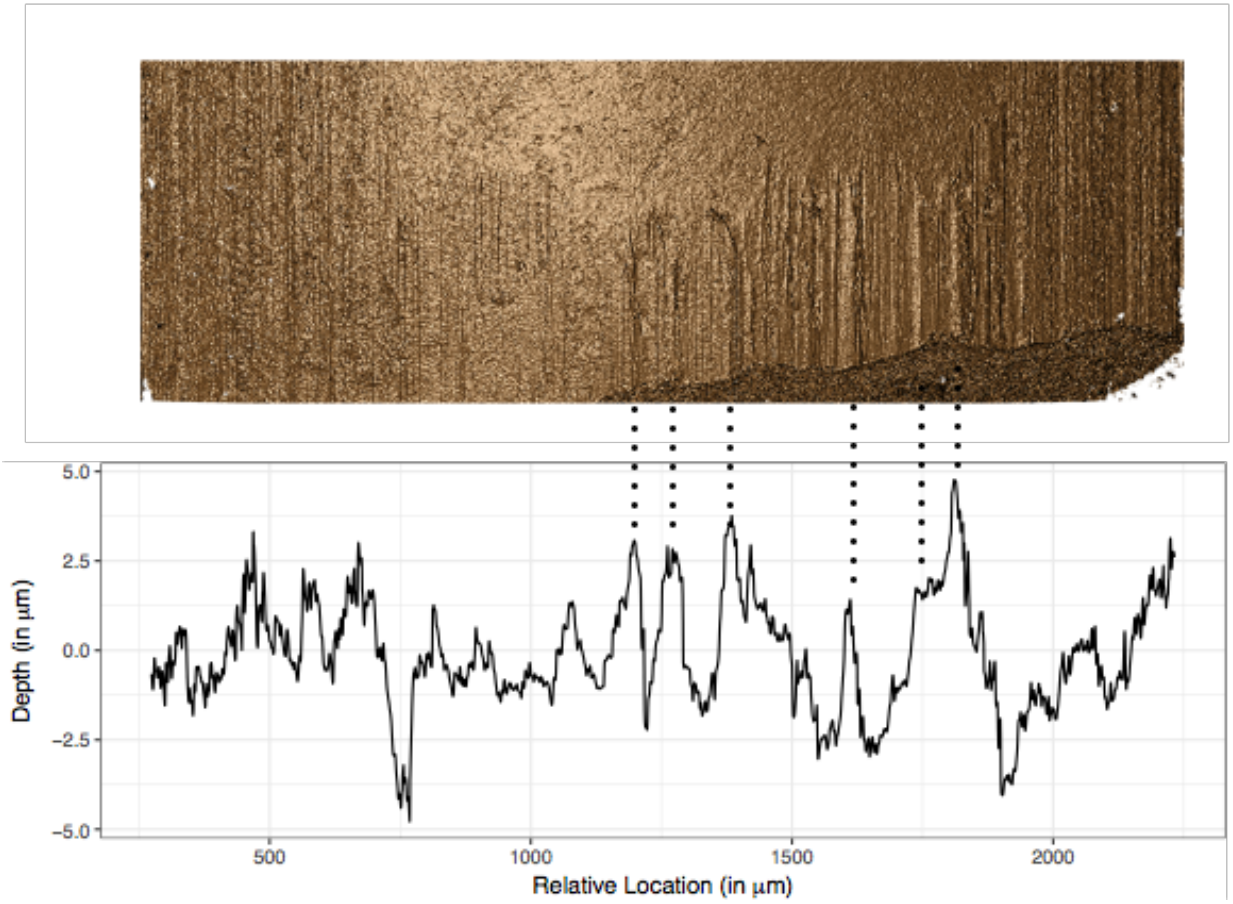

Figure 1: Image of a bullet land from a confocal light microscope at 20 fold magnification (top) and a chart of the corresponding signature of the same land (bottom). The dotted lines connect some peaks visible in both visualizations.

report false negative error rate to be 0.023 and false positive error rate be 0.09. For other angles the error rates remain the same for false negatives and 0.01 for false positives. The improvement proposed by Hadler & Morris (2017) on the other hand, did not consider effect of angles and gave 0.06 false negative error rate and 0 false poisitive error rate for markings made at the same angle.

1.2 Scans for land engraved areas

The land engraved markings across the cross section of the land are termed as bullet profiles (Hare et al. 2016, Ma et al. 2004). These are often used for making land to land comparisons. Bullet signatures on the other hand (Chu et al. 2013, Hare et al. 2016) refer

to a processed version of the raw land engraved markings or profiles. The generation of bullet signatures involves first extraction of a bullet profile by taking the cross-sectional of the surface at a given height. Loess fits are then used to model the structure. The residuals of these fits are called signatures. They are considered to be noise free and a good reflection of the key attributes of the raw marking, and the unique features of a bullet. A more detailed version of the extraction technique of signatures is discussed by Hare et al. (2016), where comprehensive details about the height at which profile is to be selected, removing curvature, smoothing, identifying groove locations are explained.

The majority of Bullet profiles and signatures of the Hamby study (Hamby et al. 2009) used in this paper are extracted by procedures mentioned by Hare et al. (2016). The scans for the Hamby study come from two sources namely, NIST scans (Zheng 2016) and scans produced at Iowa State University. Both scans were generated using confocal microscopes but at different resolutions. The NIST database contains scans of the bullet sets Hamby 44 and 252 with the resolution 1.5625 μm per pixel. The Iowa State scans are available for only the Hamby 44 bullet set and at a resoultion of 0.065 μm per pixel. The data used by Chumbley et al. (2010), Hadler & Morris (2017) on the other hand, was produced by a profilometer at a resolution of about 0.73 μm per pixel. The different resolutions indicate that the length of the digitized markings would be significantly different for each setting.

1.3 Potential Challenges in Chumbley Score Adaptation

Markings made by Screw driver tips (Chumbley et al. 2010) are longer and pronounced and about 7 mm in length (Faden et al. 2007), while the groove to groove length of the bullet lands from the Hamby study (Hamby et al. 2009) are about 2mm in length. The Ruger pistol barrels used in the Hamby study have a 9mm caliber. This means that bullets used in these pistols have land markings of almost $(1/4)^{th}$ the size of screwdriver toolmarks used by Chumbley et al. (2010), Hadler & Morris (2017), therefore raising a problem of distinctiveness. One way to investigate this is to evaluate bullet striations at different resolutions. This gives us an incentive to use both the NIST and Iowa State scans and investigate the effect of length of the digitized mark on the working of the algorithm. Apart from this, the algorithm developed by Hadler & Morris (2017) has only been tested

for flatter and wider surfaces which have negligible curvature, whereas striations on bullets are made on their curved surfaces. Therefore, using methods proposed for toolmarks need adaptation in order to give tangible results for bullets. Another challenge that arises from these topographical differences, is the understanding of that effect of different levels of smoothing. Especially investigating its effect on the algorithm performance when signatures are being compared and when raw markings or profiles are being compared.

Some other challenges associated with the adaptation process call for a brief understanding of the methodology involved. The Chumbley method and algorithm works on the premise of comparison windows that are supposed to represent small segments of pixels of equal sizes in the two markings. These windows are to be compared in a predetermined fashion and are termed as windows of optimization and validation. The optimization window is selected in the first step called the optimization step where areas/windows of best agreement between the the two markings are found.

The window of optimization for bullet striations are shorter as bullet signatures in terms of absolute length (in mm), are smaller compared to toolmarks. Even in terms of digitized pixels, there is no straightforward way of establishing what window size is to be chosen. One important way to address this is to keep the number of windows of optimization sufficiently large or not too small. This means, if the number of windows are too small we inadvertently use bigger window sizes as we want to compare all segements of one mark with the segments of the other mark. This leads to not giving enough emphasis to smaller segements of the two marking that might be in high agreement. In terms of absolute length we end up comparing smaller segments between the two markings while dealing with bullet lands, than those being compared for toolmarks.

For too large window sizes, the weight of small features that would otherwise uniquely classify a signature and hence identify the region of agreement is vastly reduced. On the other hand, this introduces another problem as, if we go too small in the window of optimization, the unique features of the trace segments are lost and all seem similar. The problem of distinctiveness in the structure of bullet striations may also lead to potential failure of the algorithm to identify correct windows of agreements or not be able to identify these windows at all. This may lead to incorrect classification of bullets as coming from

the same-source (or matches) or from different source (or non-matches).

Thus, the comaparison windows have a direct influence on the false positives and false negatives, making identification of the optimum sizes of these windows very important. This raises important questions about how and what parameter settings need to be chosen for bullet land comparison. Something similar is done in Grieve et al. (2014) for toolmark comparisons of slip-joint pliers where optimum window sizes are determined. There is a need to figure out the best parameter settings which minimizes the errors for unsmoothed markings or profiles and pre-processed signatures. An analysis of these error rates and comparison with other methods will help us understand the adaptibility of the chumbley score to bullets.

1.4 The Chumbley Score Test

The Chumbley score algorithm takes input as two vectorized processes. Here the processes are of the form $z(t)$ which is a spatial process for some location indexed with t. This means that while $z(t)$ represents any striae, $z(t_1)$ is a realization of the spatial process. t here denotes a vector of equally spaced pixel locations for the striation marks under consideration. The word 'pixel' refers to the resolution of the confocal light microscope. In the case for bullet signatures extracted from NIST ballistics database (Zheng 2016) of the Hamby study, a pixel corresponds to approximately 1.56 microns. The vectorized processes representing two sets of striaes are therefore shown as $x(t_1)$, $t_1 = 1, 2, ...T_1$ and $y(t_2)$, $t_2 = 1, 2...T_2$. Here x and y denote two instances of the process $z(*)$ which means two striation marks, while the indexing t or pixel range is shown as t_1 and t_2 for x and y respectively. This representation of t as t_1 and t_2 lets us identify them as either of same or different lengths. The striation marks under consideration are potentially from two different bullets whose source needs to be identified as being same or different. T_1 and T_2, as represented above, are the final pixel indexes of each marking and therefore give the respective lengths of the markings. The similarity in the two markings is then judged by the algorithm on the basis of cross-correlation of a fixed and constant number consecutive pixels (say k) taken from the two markings. This can for example be k taken from the two indexed marking $x(t_1)$ and $y(t_2)$ such that in theory k remains smaller than length of

the two striae or marks. Depending on what stage of the algorithm we are in, matching of different pixel lengths and locations is done. This in the end, effectively compares all possible windows that would guarantee in quantifying the two marks or striae as coming from the same source or not.

The algorithm works in two phases, namely, an optimization step and a validation step, at the end of which a Mann Whitney U statistic is calculated as a measure to differentiate between matches and non-matches. A pre-processing step to the algorithm is to choose a coarseness value which is used as a parameter to the lowess smoothing function. The coarseness essentially gives the proportion of points which influence the smooth at each value, which means larger values lead to more smoothness. The lowess smoothing is applied to each of two sets of vectorized striae or marks $x(t_1)$ and $y(t_2)$, before proceeding to the algorithm. The algorithm starts with the optimization step where the area of best agreement in the two markings being compared is identified. Comparison window size is predefined. Each window of the first marking is compared with all windows of the second marking. A maximum correlation statistic is used to identify the region of best agreement, with the maximum usually seen being near 1 for both cases which is what is intuitively expected for matches, and not expected intuitively for non matches. Hadler & Morris (2017) in their paper proposed an improvement to this algorithm by trying to remove mutual dependence of parameters. The mututal dependence was due to the serial correlation in surface depth values of a marking. Also a random sampling sub step in the validation phase makes a group of pixels to be chosen more than once and hence introduces lack of independence in certain steps. The reason for removing the mutual dependence was the Mann-Whitney U statistic works under the assumption of independence of parameters. The validation step builds on this with two sub-steps namely, same shift and different shift. As the aim is to come up with a non-parameteric U statistic, Hadler & Morris (2017) proposed a normalization procedure in the validation step which goes to some extent to address the issue of mutual dependence. A series of windows are chosen in both the same and different shift sub-steps for the purpose of comparison. Both substeps were modified by Hadler & Morris (2017) who introduced a deterministic rule for sampling of windows in the same-shift and different shift as opposed to the originally proposed random sampling

by Chumbley et al. (2010). In th same shift substep the chosen series of windows are at a common distance (rigid-shifts) from the window earlier identified as the region of best agreement in the optimization step. The correlation of these windows generally turn out to be lower than the maximum correlation window. The significance is that these same shift windows still have large enough correlation values for the two markings being compared that are in reality a match. Any choice of maximum correlation windows (where correlation values are often near 1) for non-matches in the optimization step are also validated by this sub-step. This is because the same-shift correlations would not be anywhere as large for same-shift windows when all windows in the sub-step are compared. The different shift substep on the other hand gives perspective to the correlation values of the same shift window correlation values. There are no rigid-shifts but different shifts where distance from the maximum correlation window are chosen randomly by Chumbley et al. (2010) and deterministically as per Hadler & Morris (2017). This means that there is an equal possibilty of comparing a trace segment from one marking to any trace segment or window in the second marking.

Neither of above sets of correlation in the two sub-steps are allowed to include the maximum correlation window as identified earlier. Therefore the assumption is that if two markings match each other, the same-shift correlations would be larger than the different-shift windows. And if they are not a match the correlations in the two sets will be very similar. The U-statistic tests for the null hypothesis that the two markings are not a match and are therefore not made by the same source. As given by Hadler & Morris (2017), it is computed from the joint rank of all correlations of both the same and different shift samples.

endedit end of intro: remaining paper is structured as follows: introduce to the data we get from confocal microscopy, introduce to profiles and signatures. Introduce to chumbley score method, apply chumbley and discuss results ....

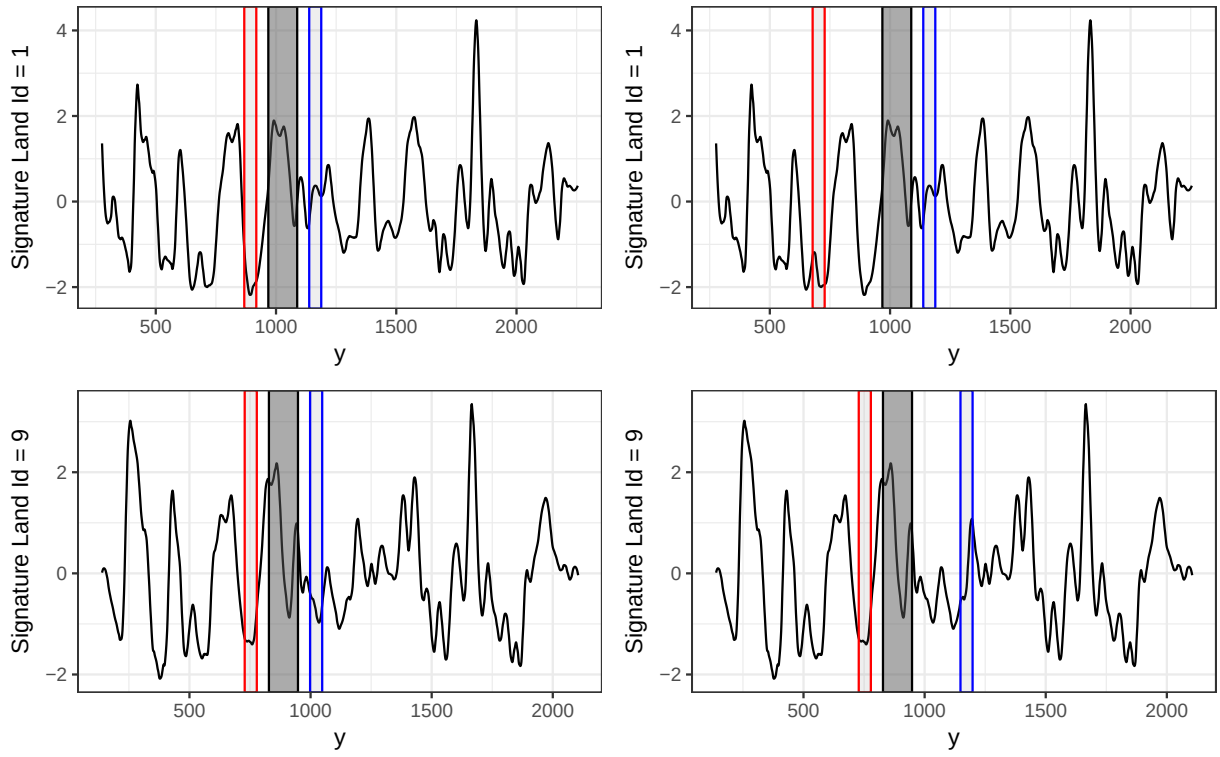

Figure 2: The two plots on the left show how the same shift behaves in case of a matching pair and the two plots on the right show how the different shift behaves in case of a matching pair.

1.5 Scans for land engraved areas

- scans available: NIST database (citation), Hamby 44 and Hamby 252 (Hamby citation)
- move figure up
- discuss cross section, profiles and signatures

2 Simulation setup

Following on similar lines to the setup of toolmarks, the first step here is to first identify what difference does different window sizes of optimization and the validation step have, when adapting the toolmark method to bullets.

The marking made on bullets are smaller than toolmarks and is also less wider. The idea is to find out possible areas of error while adapting the score based method proposed for toolmarks, using cross-validation setup to identify appropriate parameter settings for (a) signatures and (b) profiles directly

2.0.1 Signatures

Signatures of lands for all Hamby-44 and Hamby-252 scans made available through the NIST ballistics database (Zheng 2016) were considered. Both of these sets of scans are part of the larger Hamby study (Hamby et al. 2009) and each consist of twenty known bullets (two each from ten consecutively rifled Ruger P85 barrels) and fifteen questioned bullets (each matching one of the ten barrels). Ground truth for both of these Hamby sets is known and was used to assess correctness of the tests results.

Bullet signatures being compared at this time are therefore from the Hamby 44 and Hamby 252 data. The database setup and pre-processing system used for choosing the Bullet signatures are as described by Hare et al. (2016). In order to choose the bullet signatures we first filter out Land_id for Profiles from the Hamby 44 and Hamby 252 data and remove all NA values. Then run_id = 3 is chosen as the signatures generated from this run_id give the closest match. Different run_id's have some different settings for generating the signatures.(The level of smoothing does not seem to be one of them)

The bullet signatures when generated by this process already includes a loess smoothing. Therefore, the coarseness factor is set to 1 while running the chumbley non random algorithm for comparing different optimization windows.The algorithm generates the same_shift, different_shift, U-Stat and P_value parameters which are then used to calculate the errors associated with different sets of window sizes.

2.0.2 Profiles

The profiles are cross-sectional values of the the bullet striation mark which are chosen at an optimum height (x as used by Hare et al. (2016)). This x or height is not a randomly chosen level. The rationale behind the choice has been explained by Hare et al. (2016). A region is first chosen where the cross-correlation seems to change very less and in this

region an optimum height is chosen. The profiles generally resemble a curve which is more or less similar to a quadratic curve (a quadratic fit to the raw data values of the profile is not an exact fit but it does show a similar trend). Profiles are the set of raw values representing the striation marks, and signatures are generated from these by removal of the inherent curvature and applying some smoothing (the signatures generated by Hare et al. (2016) use a loess function for smoothing).

Similar to signatures the run_id = 3 was used when applying the chumbley algorithm using the database setup given by Hare et al. (2016) of Hamby-44 and Hamby-252 datasets, on the profiles. The run_id not only defines the level of smoothing but also signifies the chosen height at which the profiles were selected initially. Another important aspect is the range of horizontal values (which is referred to as the y values in Hare et al. (2016)) in the signatures. These have already been pre-processed in the database to not include any grooves.

Therefore for the sake of comparison the run_id = 3 is still chosen so as to ensure that the horizontal values remain the same as that of the signatures. This also gives us profiles with the grooves removed.

The idea therefore is to first use these raw values of the profile directly in the chumbley algorithm, and see how the algorithm performs for different coarseness values (smoothing parameter as referred in the function lowess used in the chumbley algorithm).

3 Results

We used the adjusted Chumbley method as proposed in Hadler & Morris (2017) and implemented in the R package `toolmaRk` (Hadler 2017) on all pairwise land-to-land comparisons of the Hamby scans (a total of 85,491 comparisons) with the pairwise sets for the comparisons given in the table 1.

[!h]

Table 1: Overview of parameter settings used for optimization and validation windows for bullet land signatures.

wo	50	50	60	60	80	80	90	90	100	100	110	110	120	120	120	120	120	120
wv	30	50	30	50	30	50	30	50	30	50	30	50	10	20	30	40	50	60
wo	130	130	140	140	150	150	160	160	200	200	200	240	240	280	280			
wv	30	50	30	50	30	50	30	50	20	30	50	30	50	30	50			

3.1 Signatures

Figure 3 gives an overview of type II error rates observed when varying the window size in the optimization step. Two levels of validation window size 30 and 50 were chosen as to compare the error rates for different nominal type I errors. We notice that the trends for these nominal type I errors are similar and in most cases a validation window of 50 has higher type II error than for 30. A change in this trend is seen for a 0.05 α level, although the difference between the two windows is very small for this case. We can also notice an obvious trend of increase in the Type II error as the window of optimization increases and see a minimum around the optimization window size of 120 pixels. Hence we are inclined to choose a smaller validation window size and optimization window as 120.

Table 2 shows the confusion tables with the classification of type I and type II errors and how the numbers change with a change in the optimization window. The windows represent areas to the left of the window with minimum type II, near the minimum type II window and to the far right of the minimum type II error

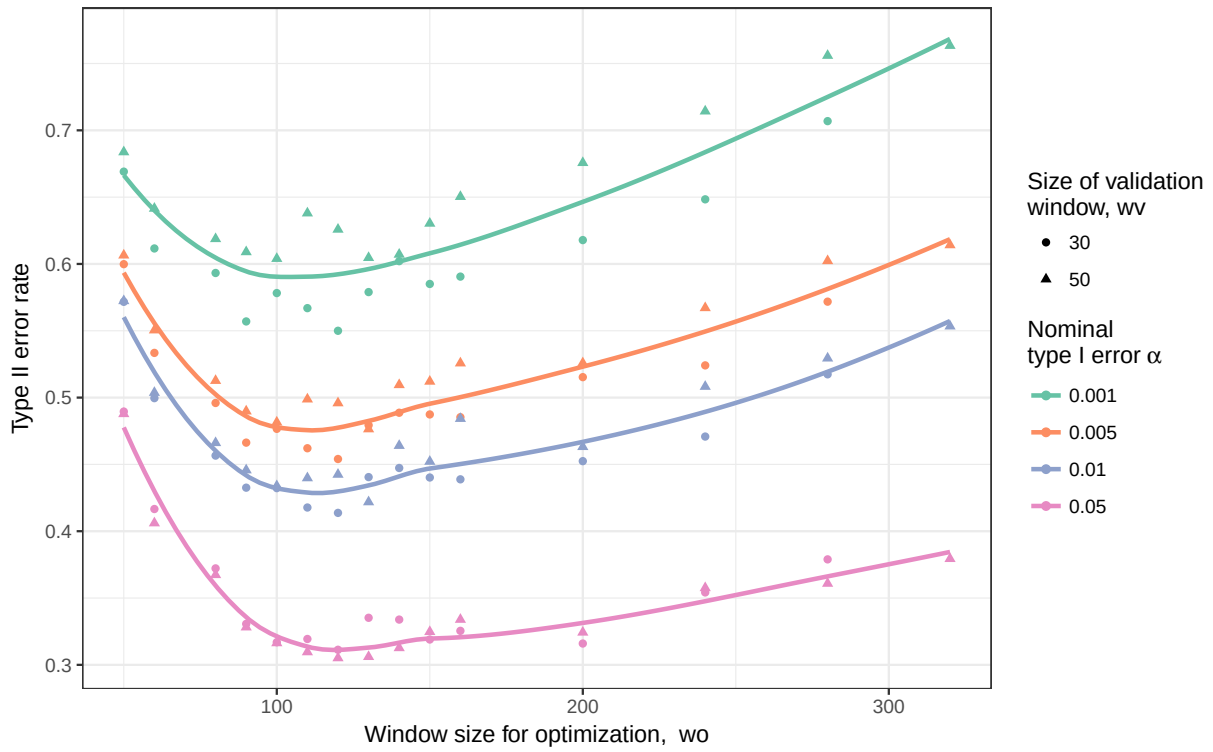

Figure 3: Type II error rates observed across a range of window sizes for optimization wo. For a window size of $wo = 120$ we see a drop in type II error rate across all type I rates considered. Smaller validation sizes wv are typically associated with a smaller type II error.

Table 2: Confusion Table for different optimization window sizes with validation window size as 30.

signif	match	Freq	Type
Size of Optimization Window = 280			
FALSE	FALSE	78909	True Negative
TRUE	FALSE	4192	False Positive (Type I)
FALSE	TRUE	446	False Negative (Type II)
TRUE	TRUE	731	True Positive
signif	match	Freq	Type
Size of Optimization Window = 120			
FALSE	FALSE	79249	True Negative
TRUE	FALSE	4523	False Positive (Type I)
FALSE	TRUE	386	False Negative (Type II)
TRUE	TRUE	854	True Positive
signif	match	Freq	Type
Size of Optimization Window = 80			
FALSE	FALSE	79477	True Negative
TRUE	FALSE	4503	False Positive (Type I)
FALSE	TRUE	463	False Negative (Type II)
TRUE	TRUE	781	True Positive

Figure 4 compares nominal (fixed) type I error and actually observed type I errors for the parameter settings in table 1. With an increasing size of the window used in the optimization step the observed type I error rate decreases (slighty). This means as the optimization window increase the observed type I error rate gets smaller. A smaller validation window on the other hand, tends to be associated with a higher type I error rate. This can be better imagined for a given window of optimization, where the actual Type I error is comparable to the nominal level for only a select few validation window sizes. For

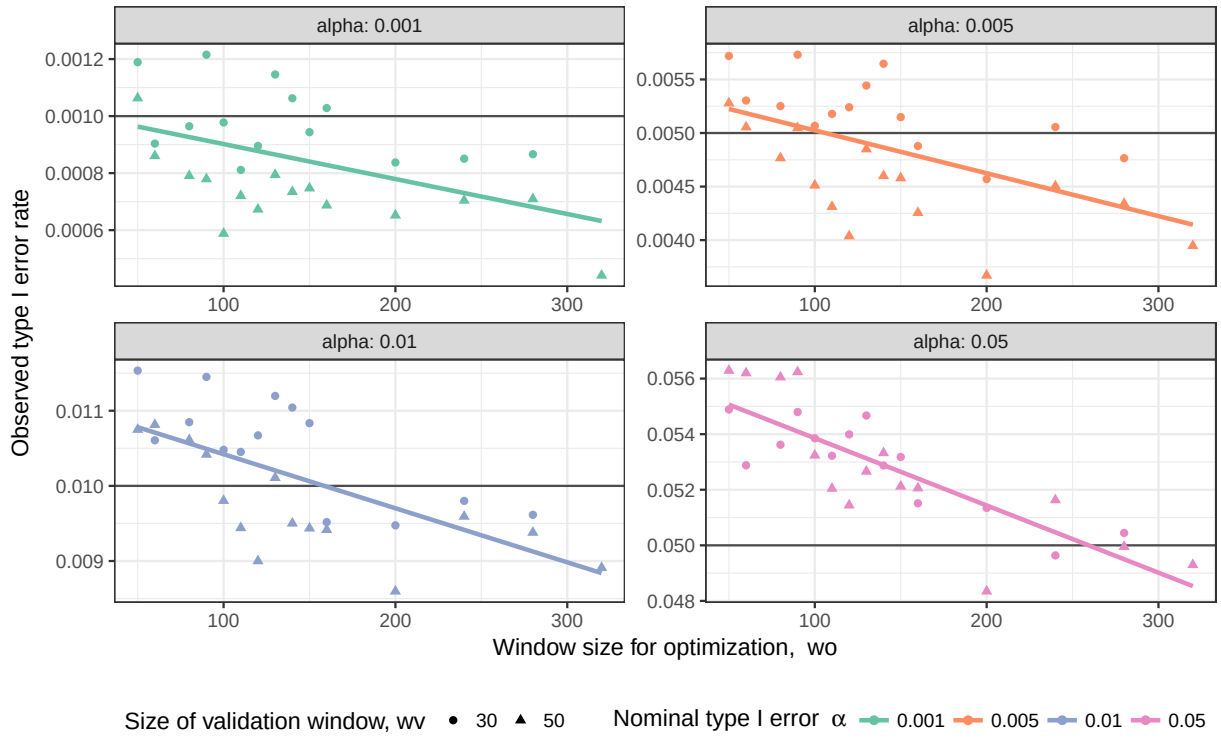

Figure 4: Comparison of observed and nominal type I error rates across a range of window sizes for optimization *wo*. The horizontal line in each facet indicates the nominal type I error rate.

these comparable validation window sizes of 30 and 50 as done here, the actual type I error increases very slightly and can be seen in Figure 4. This increase is not as much when compared to the variation seen with the optimization window sizes. This effect might be related to the increasing number of tests that fail for larger optimization window sizes, in particular for non-matching striae (see fig 6).

The actual type I error and type II error for signatures were also compared for different validation window sizes. Figure 5 shows the actual rates for different nominal α levels. We can see that the type II error rises with higher validation windows for the smaller nominal α levels while for the nominal $\alpha = 0.05$ its almost constant.

Figure 6 gives an overview of the number of failed tests, i.e. tests in which a particular parameter setting did not return a valid result. This happens, when the shift to align two signatures is so large, that the remaining overlap is too small to accommodate windows for

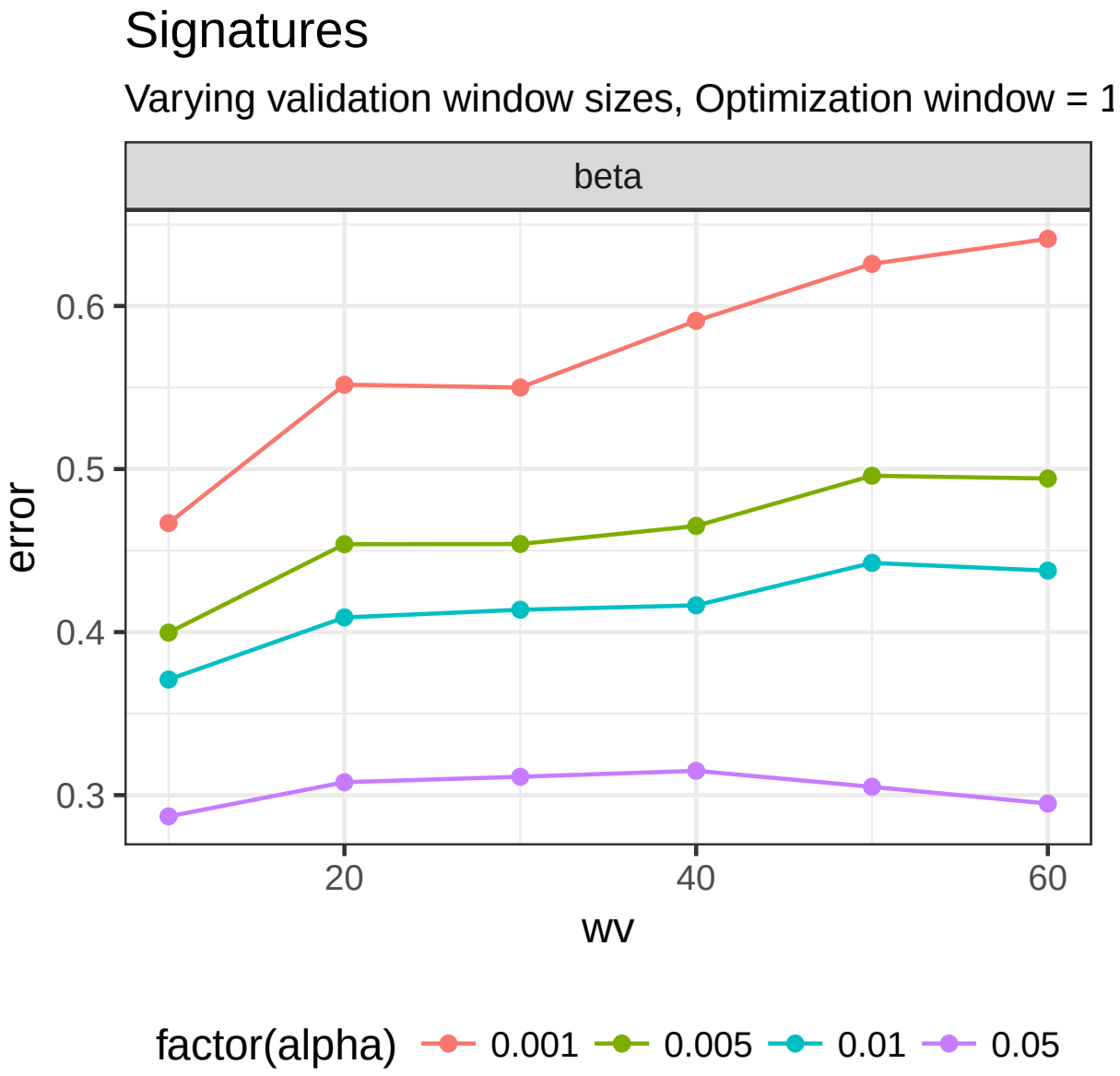

Figure 5: The figure on the left shows the actual Type I error while the figure on the right shows the Type II error for different validation window sizes and different chosen nominal alpha levels when the size of the optimization window = 120

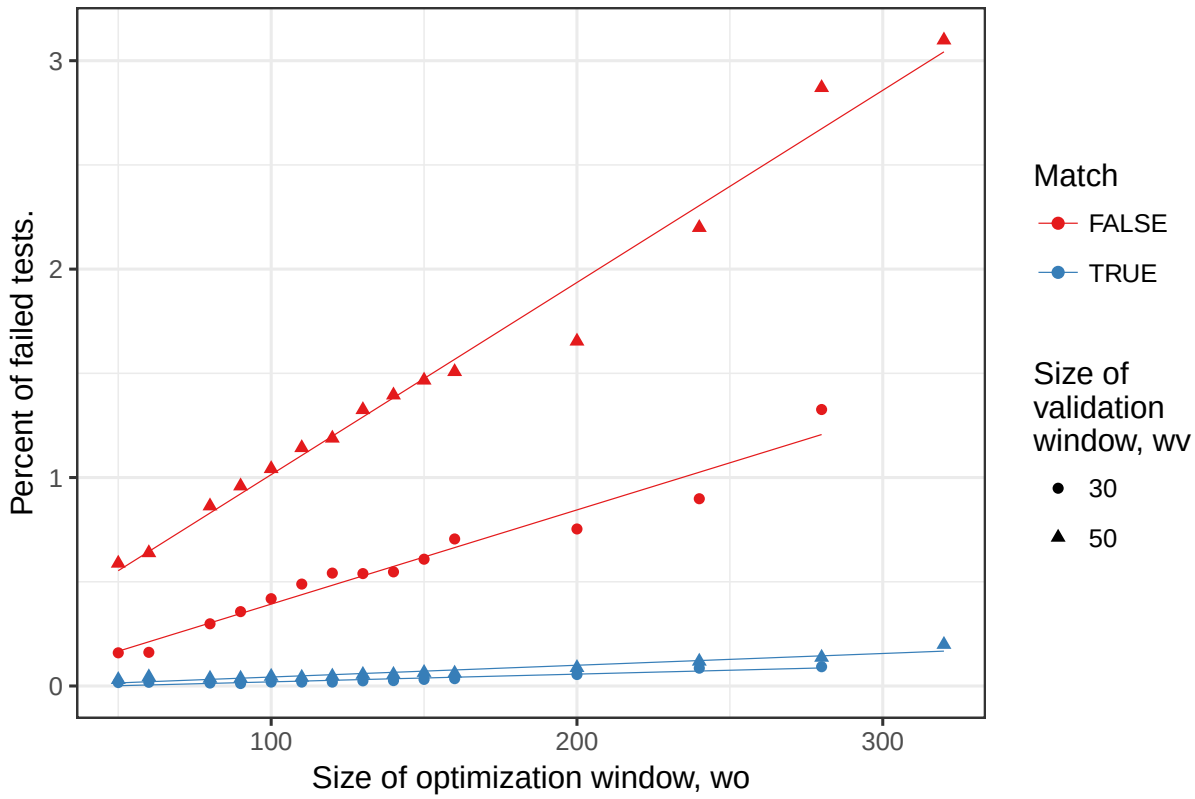

Figure 6: Number of failed tests by the window optimization size, wo, and ground truth.

Table 3: Estimates of the increase in percent of failed tests corresponding to a 100 point increase in the optimization window.

match	wv	estimate	std.error
FALSE	30	0.452	0.027
FALSE	50	0.922	0.035
TRUE	30	0.037	0.004
TRUE	50	0.056	0.005

validation. The problem is therefore exacerbated by a larger validation window. Figure 6 also shows that the number of failed tests is approximately linear in the size of the optimization window. Test results from different sources have a much higher chance to fail, raising the question, whether failed tests should be treated as rejections of the null hypothesis of same source. For known non-matches there is a higher possibility that in the optimization pair of windows where cross-correlations are maximum are too far apart, and same shifts of this order hit the end of the signature.

3.2 Profiles

Figure 7 (a) shows the type II error rates for profiles for the optimization window 120 and validation window 30 with varying level of coarseness. We can see that the type II error for all the nominal α levels is lowest in the range of 0.20 to 0.35. Therefore, a value of 0.25 can be used keeping in mind it keeps the type II error lowest while runnning simulations. Thus for comparisons of different window sizes etc as seen in the differnt parts of Figure 7 this coarseness value is used.

On the other hand Figure 7 (b) shows if the coarseness level set in the chumbley agorithm has any effect on the signatures, which are pre-processed and already smoothed to a certain extent. From Figure 7 (b) we can notice that for different nominal α levels, the type II error fluctuates slightly but does not change much, thereby helping us conclude that the coarseness levels set in the LOWESS smoothing in the chumbley alggorithm does effect the type II error much for signatures.

3.2.1 Comparison of profiles and signatures

Another reason for failed tests can be incorrect identification of maximum correlation windows in the optimization step as seen in figure 7(d) because of the level of smoothing, as too much smoothing would subdue intricate features that might otherwise help in the correlation calculations and correct identification of maximum correlation windows irrespective of the size. This would again cause a simiar effect as explained for figure 6 with validation windows, irrespective of size, during the shifts end up at the ends of the markings resulting in an invalid calculation and failed comparison attempt.

In figure 7(d) and (f), we compare profiles and signatures on the basis of number of failed tests. The profiles chosen for figure 7(f) have a constant coarseness of 0.25 and window of optimization as 120. The signatures in this case are not smoothed using the chumbley algorithm step of LOWESS smoothing. Instead signatures are used as calculated by Hare et al. (2016). The smoothing in these signatures were determined and fixed on the basis of their performance in the random forest based algorithm proposed by Hare et al. (2016). The comparison of profiles and signatures with variation of validation window size therefore is made on even footing. The trends are similar to figure 6 in the sense that for known non-matches the number of failed tests are more for both signatures and profiles and increasing linearly with the validation window size. The problem is however, worse for profiles which has higher number of failed tests than signatures for all validation windows.

The total error for different validation window sizes for signatures and profiles can be seen in figure 7 (e).The optimization window size is 120 and profiles are calculated at a default 0.25 coarseness level while signatures as before are not smoothed again in the modified chumbley algorithm. We can see that the total error is always higher for profiles as compared to signatures for all sizes of validation window.

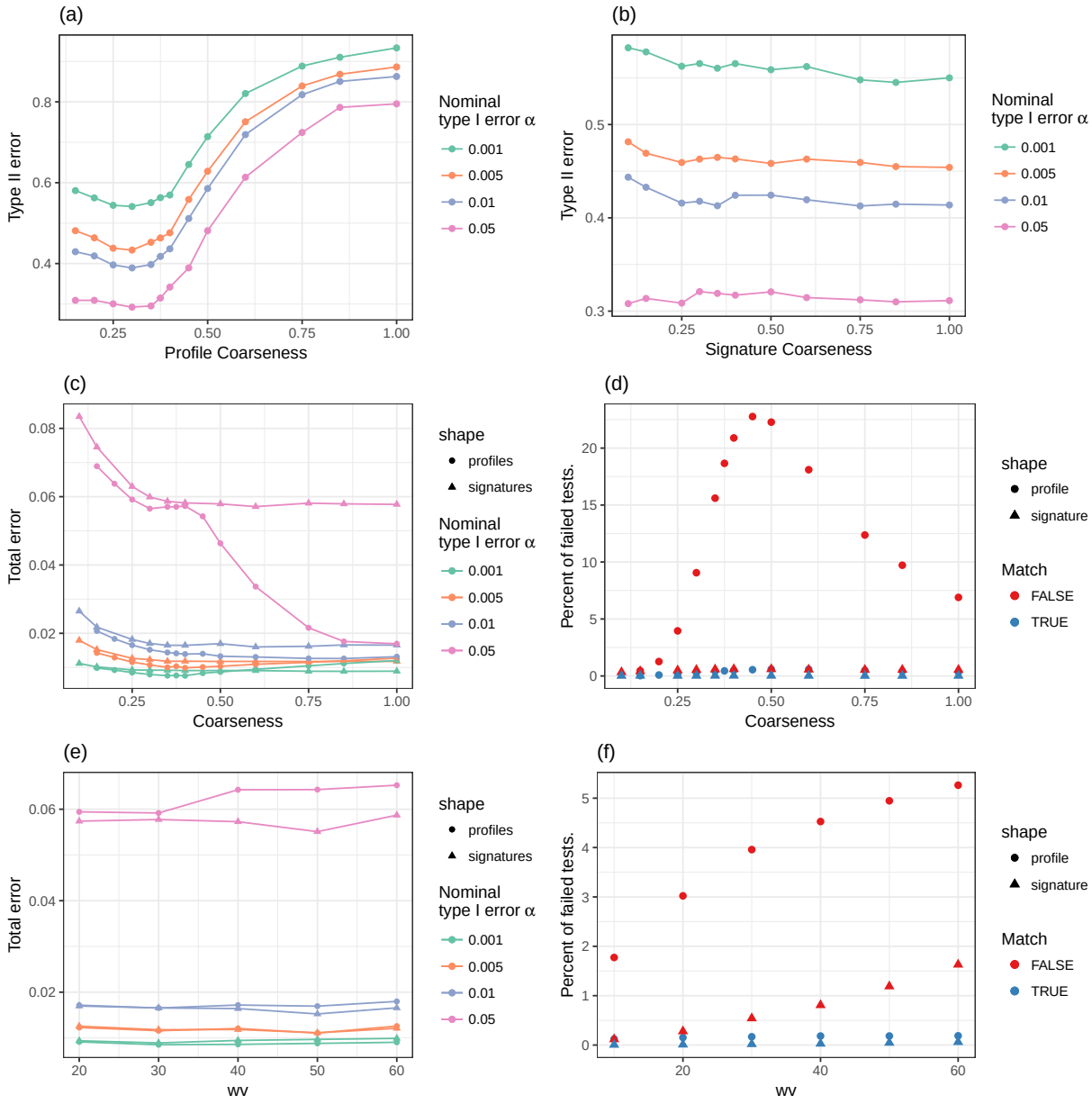

Figure 7: Row 3: Total error and Number of failed tests by the window validation size, wv, and ground truth, Row 2: Total error and Number of failed tests with Coarseness for both profiles and signatures, Row 1: Type II error for different coarseness levels as used in the modified chumbley algorithm for profiles and signatures

3.3 Conclusion

The results suggest that the Nominal type I error α value shows dependence on the size of the window of optimization. For a given window of optimization the actual Type I error is comparable to the nominal level for only a select few validation window sizes and for comparable validation window sizes of 30 and 50 as done here, the actual type I error does not seem to vary as much as it varies with the optimization window sizes . A Test Fail, i.e. tests in which a particular parameter setting did not return a valid result, happens, when the shift to align two signatures is so large, that the remaining overlap is too small to accommodate windows for validation, depends on whether known-match or known non-matches has predictive value, with test results from different sources having a much higher chance to fail. On conducting an analysis of all known bullet lands using the adjusted chumbley algorithm, Type II error was identified to be least bad for window of validation 30 and window of optimization 120. In case of unsmoothed raw marks (profiles), Type II error increases with the amount of smoothing and least for LOWESS smoothing coarseness value about 0.25 or 0.3. In an effort to identify the level of adaptiveness of the algorithm, comparisons were made between signatures and profiles. Their comparison with respect to validation window size for a fixed optimization window size suggested that, profiles have a total error (i.e all incorrect classification of known-matches and known non-matches) greater than or equal to the total error of signatures for all sizes of validation window. Profiles also fail more number of times than signatures in a test fail (for different coarseness keeping windows fixed and also for different validation windows keeping coarseness fixed) which lets us conclude that the behaviour of the algorithm for the profiles instead of pre-processed signatures is not better. Finally it should be noted that the current version of the adjusted chumbley algorithm seems to fall short when compared to other machine-learning based methods Hare et al. (2016), and some level of modification to the deterministic algorithm needs to be identified and tested that would reduce the number of incorrect classifications.

References

AFTE Criteria for Identification Committee (1992), ‘Theory of identification, range striae

comparison reports and modified glossary definitions', *AFTE Journal* **24**, 336–340.

AFTE Glossary (1998), 'Theory of identification as it relates to toolmarks', *AFTE Journal* **30**, 86–88.

Bachrach, B., Jain, A., Jung, S. & Koons, R. (2010), 'A statistical validation of the individuality and repeatability of striated tool marks: Screwdrivers and tongue and groove pliers', *Journal of Forensic Sciences* **55**(2), 348–357.

Chu, W., Thompson, R. M., Song, J. & Vorburger, T. V. (2013), 'Automatic identification of bullet signatures based on consecutive matching striae (cms) criteria.', *Forensic Science International* **231**, 137–141.

Chumbley, L. S., Morris, M. D., Kreiser, M. J., Fisher, C., Craft, J., Genalo, L. J., Davis, S., Faden, D. & Kidd, J. (2010), 'Validation of tool mark comparisons obtained using a quantitative, comparative, statistical algorithm', *Journal of Forensic Sciences* **55**(4), 953–961.
URL: *http://dx.doi.org/10.1111/j.1556-4029.2010.01424.x*

Faden, D., Kidd, J., Craft, J., Chumbley, L. S., Morris, M. D., Genalo, Lawrence J.and Kreiser, M. J. & Davis, S. (2007), 'Statistical confirmation of empirical observations concerning toolmark striae', *AFTE Journal* **39**(2), 205–214.

Grieve, T., Chumbley, L. S., Kreiser, J., Ekstrand, L., Morris, M. & Zhang, S. (2014), 'Objective comparison of toolmarks from the cutting surfaces of slip-joint pliers', *AFTE Journal* **46**(2), 176–185.

Hadler, J. (2017), 'toolmaRk: Tests for Same-Source of Toolmarks'. R package version 0.0.1.
URL: *https://github.com/heike/toolmaRk*

Hadler, J. R. & Morris, M. D. (2017), 'An improved version of a tool mark comparison algorithm', *Journal of Forensic Sciences* pp. n/a–n/a.
URL: *http://dx.doi.org/10.1111/1556-4029.13640*

Hamby, J. E., Brundage, D. J. & Thorpe, J. W. (2009), 'The Identification of Bullets Fired from 10 Consecutively Rifled 9mm Ruger Pistol Barrels: A Research Project Involving 507 Participants from 20 Countries', *AFTE Journal* **41**(2), 99–110.

Hare, E., Hofmann, H. & Carriquiry, A. (2016), 'Automatic Matching of Bullet Lands', *Annals of Applied Statistics* .

Ma, L., Song, J., Whitenton, E., Zheng, A., Vorburger, T. V. & Zhou, J. (2004), 'Nist bullet signature measurement system for rm (reference material) 8240 standard bullets', *Journal of Forensic Sciences* **49**, 649–659.

National Research Council (2009), *Strengthening Forensic Science in the United States: A Path Forward*, The National Academies Press, Washington, DC.
URL: *http://www.nap.edu/catalog/12589/strengthening-forensic-science-in-the-united-states-a-path-forward*

President's Council of Advisors on Science and Technology (2016), 'Report on forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods'.
URL: *https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf*

Vorburger, T., Song, J., Chu, W., Ma, L., Bui, S., Zheng, A. & Renegar, T. (2011), 'Applications of cross-correlation functions', *Wear* **271(3-4)**.
URL: *https://doi.org/10.1016/j.wear.2010.03.030.*

Vorburger, T., Song, J. & Petraco, N. (2016), 'Topography measurements and applications in ballistics and tool mark identifications. surface topography: metrology and properties', *Surface topography: metrology and properties* **4(1)**.

Zheng, X. A. (2016), 'NIST Ballistics Toolmark Research Database (NBTRB)'.
URL: *https://tsapps.nist.gov/NRBTD*