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

February 5, 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

Compairing pairs of toolmarks with the intention of matching it to a tool has been studied many times in the past. Extensive examples in literature for tools and toolmark research ranging from screwdrivers to groove pliers to slip-joint pliers can be found in the work of Faden et al. (2007), Bachrach et al. (2010), Miller (1998), Grieve et al. (2014) and many more. In comparison to this, same source matching of bullets to firearms has not been examined as prominently as that of toolmarks, with even less information available on validity of methods and error rates associated with firearms examination. The National Academy of Sciences in its report in 2009 (National Research Council 2009) discussed the need for determining error rates in methods proposed for firearms examination. In the context of same source matching and error rates determination, as seen in the case of most forensic applications, the first step involves identification of unique features that are characteristic of the object at hand. For the case of bullets and firearms, striation marks on the surface of the bullet are considered to be such markings that can be used in methods for same source matching. These marks are often a product of rifling and impurities and defects due to manufacturing in the barrel of the gun, which leads to engravings on the bullet surface (AFTE Criteria for Identification Committee 1992). In current practice, firearm examiners invariably make visual comparisons of bullet striae and use visual assesment tools to dignify bullets as being matches and non-matches. One way of accomplishing anykind of comparison between bullets is to do a comparison between surface marking of two or more bullet lands. Bullet Lands are considered to be areas between grooves made by the rifling action of the barrel. These marking are considered to be unique. The land engraved markings or sometimes termed as Bullet profiles (Hare et al. 2016) are striation marks made on Bullet lands and often used for these land to land comparisons. Bullet Signatures is another word used in literature as seen in the work of Chu et al. (2013) and Hare et al. (2016). In our context bullet signatures 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 and then using loess fits

to model the structure. The residuals of this fit are called signatures, which are considered to be noise free and a good reflection of the class charecteristics and 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.

In the study conducted by Hare et al. (2016) a machine learning based algorithm was developed for same source matching of bullets and error rates were discussed using the database from the Hamby Study (Hamby et al. 2009). In this paper, we first try to adapt a deterministic algorithm and method developed for toolmarks by Chumbley et al. (2010) and improved by Hadler & Morris (2017), to bullets. Then we consequently discuss about the efforts in doing so along with the associated error rates. The data used in this paper also belongs to the Hamby Study (Hamby et al. 2009). This gives us a common platform for comparing the performance of the chumbley method on bullets with an already existing method proposed by Hare et al. (2016) for bullets. The proposed algorithm and method of Chumbley et al. (2010), in their paper, compares two toolmarks with the intention of determining if it comes from the same source (same tool). The method also provides a means to determine error rates and claims to reduce subject bias. As mentioned earlier, subject bias and error rate determination have been a long standing issue in firearm examination (National Research Council 2009). This remains one of the motivations to explore the adaptibility of the Chumbley score methodology to bullets. Chumbley et al. (2010) used an empirical based setup to validate their proposed algorithm and quantitative method which calculates a U-statisic for the purpose of classification of toolmarks as matching or non-matching. The data for their study was obtained from 50 sequentially manufactured screwdriver tips, and preselected comparison window sizes were given as inputs to the algorithm. The algorithm then compares the two toolmarks and comes up with a U-statistic and an associated p-value to designate them as matches or non-matches. The performance for every 100 comparisons, of the algorithm proposed by Chumbley et al. (2010) and the improvement proposed by Hadler & Morris (2017) are listed in the table below.

Table 1: Chumbley et al. 2010

Classification	Match	Non-Match
Match	41	9
Non-Match	2	48

Table 2: Hadler et. al. (2017)

Classification	Match	Non-Match
Match	47	3
Non-Match	0	50

1.2 Potential limitations of the Chumbley Score adaptation to bullets

Bullets are much smaller in length, width, are not flat and curved in the cross-sectional topography as opposed to tools like screw driver tips which produced longer and pronounced markings. This means the makings made by barrels in comparison to toolmarks may have a problem in distinctiveness. The majority of Bullet profiles and signatures extracted by procedures mentioned by Hare et al. (2016) are almost 1/4 th the size of toolmarks as used by Chumbley et al. (2010) or even smaller. Striations on Bullets are made on their curved surfaces, whereas the algorithm developed by Chumbley et al. (2010) and Hadler & Morris (2017) has only been tested for flatter and wider surfaces which have negligible curvature. Therefore, using methods proposed for toolmarks may need adaptation in order to give tangible results for bullets. Moreover, in order to to get flat bullet signatures and remove the curvatures some kind of smoothing needs to be applied as a pre-step which needs further investigation as to whether the level of smoothing does effect the working of the algorithm on Bullets.

Also when in the optimization step, the Window of optimization for bullets will be shorter as the signatures are smaller. The idea is to keep the number of windows of optimization sufficiently large, which means shorter Trace segments (partition of signature or toolmark with length = size of window of optimization) that lets us compare smaller segments of one signature to another. This introduces a problem as, if we go too small in the window of optimization, the unique features of the trace segments are lost and seem similar, while too large sizes vastly reduces the weight of small features that would otherwise uniquely classify a signature and hence identify the region of agreement.

Thus the Window of Optimization has a direct influence on whether we are Falsely

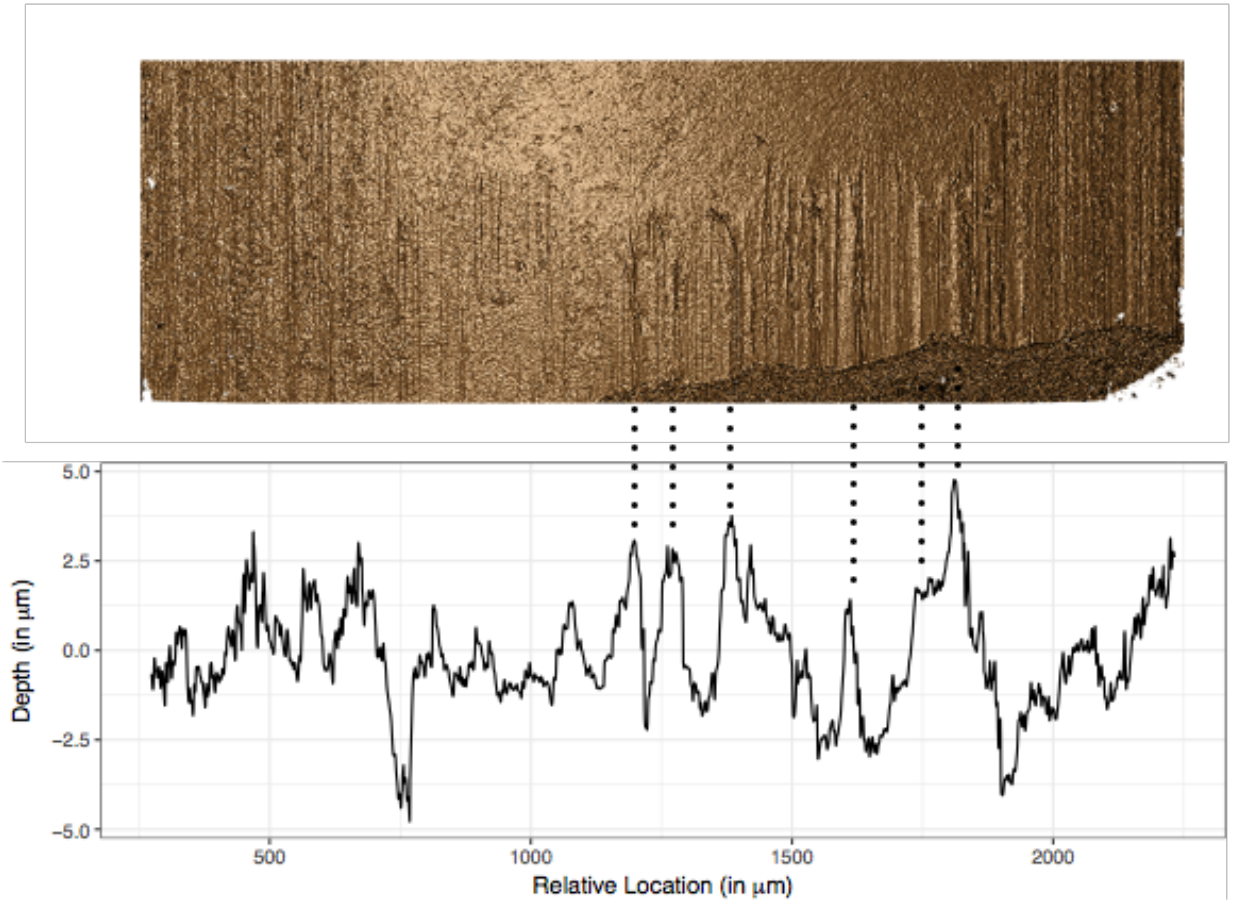

Figure 1: Image of a bullet land from a confocal light microscope at 20 fold magnification (top) and a chart of the corresponding signature of the same land (bottom). The dotted lines connect some peaks visible in both visualizations.

rejecting the null when it is true (Type I) or Falsely accepting the null when it is false (Type II) making identification of the optimum size of window of optimization very important. This raises important questions about 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 and pre-processed signatures as determined by Hare et al. (2016). An analysis of these error rates and comparison with other methods will help us understand the adaptibility of the chumbley score to bullets.

figure 1 needs to be mentioned somewhere in the write-up

2 The Chumbley Score Test

The Chumbley score algorithm takes input as two vectorized processes $x(t_1)$, $t_1 = 1, 2, ...T_1$ and $y(t_2)$, $t_2 = 1, 2...T_2$ which denote two sets of marks or striae. Here the processes are of the form $z(s)$ which is a spatial process for some t, while $z(s_1)$ is the realization of the process in s_1. The t_1 and t_2 can be better understood as equally spaced pixel locations for the two marks under consideration, where a 'pixel' is the resolution of the confocal light microscope. In the case for bullet signatures a pixel corresponds to 0.645 microns. These marks or striae are potentially from two different bullets or two different toolmarks whose source needs to be identified as being same or different. The marks or striae are indexed by the pixel location where t_1 is for the first striae referred to as x and t_2 is for the second striae which is referred as y. T_1 and T_2 being the respective lengths of the markings, need not be the same but are usually of similar lenghtsThe similarity 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 indexed marks $x(t_1)$ and $y(t_2)$ such that in theory k remains smaller that length of the two striae or marks which is the same as t_1 and t_2. Depending on the what stage of the algorithm we are in, matching of different pixel lengths and locations is done, which at 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, which also lets us assess the error rates by checking for a large number of cases.

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. 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.

Hadler & Morris (2017) in their paper proposed an improvement to this algorithm by trying to remove mutual dependence of parameters (due to serial correlation in surface depth values of a toolmark and because of a random sampling sub step in the validation phase which makes a group of pixels to be chosen more than once and hence introduces

lack of independence) in certain steps, especially because the Mann-Whitney U statistic that is later calculated in the algorithm and used as a measure to differentiate between matches and non-matches works under the assumption of independence of parameters.

Since we are only interested in a non-parameteric U statistic, Hadler et al. proposed a normalization procedure in the Validation step that goes to some extent to address the issue of mutual dependence. Also the same shift and different shift substep was modified to use a deterministic rule for sampling sam-shhift and dofferent shift-samples as opposed to the originally proposed random samples.

The data that is used by Chumbley et al. (2010) is generated by a surface profilometer that gives the height in terms of distance along a linear trace. This is taken perpendicular to the striations present in the toolmark. Two such trace are then compared using the algorithm.

2.1 Detailed algorithm

2.1.1 Optimization step

The idea behind this step is to first identify the area of best agreement in the two toolmark data. Comparison window size is predefined by the user, which in case of screwdriver toolmarks was chosen by Chumbley et al and Hadler et al as around 10 percent of the length of the toolmarks, which was around 500. This window is henceforth referred to as Window of optimization.

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 we intuitively expect for matches, and something which we do not intuitively expect for non matches

First, there are a very large number of cross-correlations calculated for the two series of striae or marks $x(t_1)$ and $y(t_2)$, for eg if the window of optimization was defined as 200 and toolmark pixel length is around 1000 then we have 1200-(200-1) = 1001.

This is the number of windows we have for one toolmark and each window is compared with all windows of the second toolmark (the number of windows is again similar), and the window with the maximum correlation is identified to be the region where the toolmarks are in maximum agreement with each other

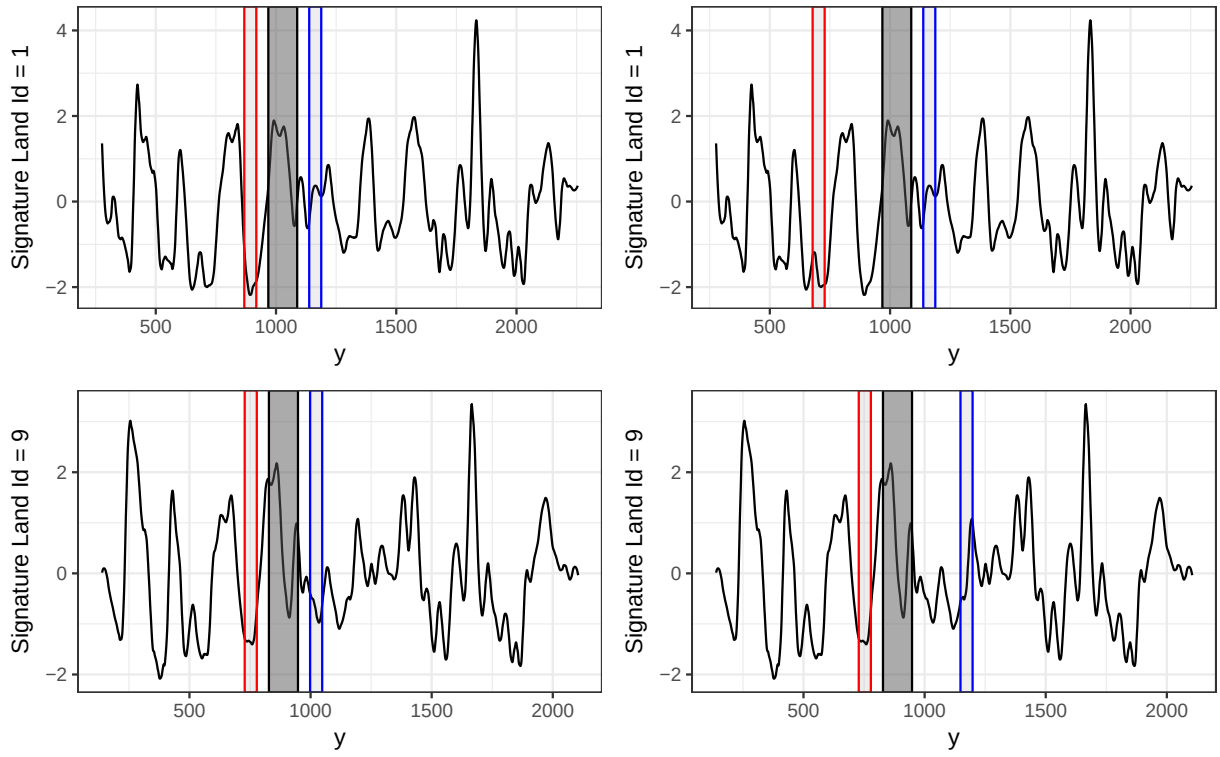

Figure 2: The two plots on the left show how the same shift behaves in case of a matching pair and the two plots on the right show how the different shift behaves in case of a matching pair.

2.1.2 Validation Step

Same-shift: In this step a series of windows are chosen at random (originally by chumbley et al) and deterministically (by hadler et al), but at a common distance (rigid-shift) from the window identified as the region of best agreement in the vaildation step. The correlation of these windows would be as intuitively assumed i.e. lower than the maximum correlation window (Optimization step), but the significance is that these same shift windows will still have large enough correlation values for two toolmarks or signatures that are in reality a match.

This also validates that if in the optimization step the maximum correlation window chosen (with correlation value near 1) was by accident (like in case of signatures that in reality are a non-match), all these same- shift correlations (A fixed number of these trace segments are identified) would not be anywhere large enough for all same-shift windows.

Different Shift: Primary reason for this substep is to give perspective to the correlation values of the same shift window correlation values.

This time there are no rigid-shifts but different shifts (distance from Window of Opt with max correlation) chosen randomly by Chumbley et al. (2010) and deterministically by Hadler et al. such that there is an equal possibilty of comparing a trace segment from one signature or toolmark to any one in the second signature or toolamrk.

Neither of above sets of correlation are allowed to include the maximum correlation window as identified earlier.

Therefore the assumption is that if two toolmarks or signatures 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.

2.1.3 U Statistic:

This is computed from the joint rank of all correlations of both the same and different shift samples. As given by Hadler & Morris (2017)

Null Hypothesis: If the toolmarks were not match i.e not made by the same tool.

Let ns and nd be the number of same shift and different shift windows

$$N = n_s + n_d$$

The mann whitney U statistic is given by

$$U = \sum_{i=1}^{ns} R_s\left(i\right)$$

with the standardized version which includes provision for rank ties

$$\overline{U} = \frac{U - M}{\sqrt{V}}$$

where prior to normalization the U-statistic has the mean as

$$M = n_s\left(\frac{N+1}{2}\right)$$

and variance

$$V = \frac{n_s n_d}{N\left(N-1\right)}\left[\Sigma^{n_s} R_s\left(i\right)^2 + \Sigma^{n_d} R_d\left(j\right)^2\right] - \frac{n_s n_d\left(N+1\right)^2}{4\left(N-1\right)}$$

2.2 Testing 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

2.2.1 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.

2.2.2 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 3.

[!h]

Table 3: 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 4 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

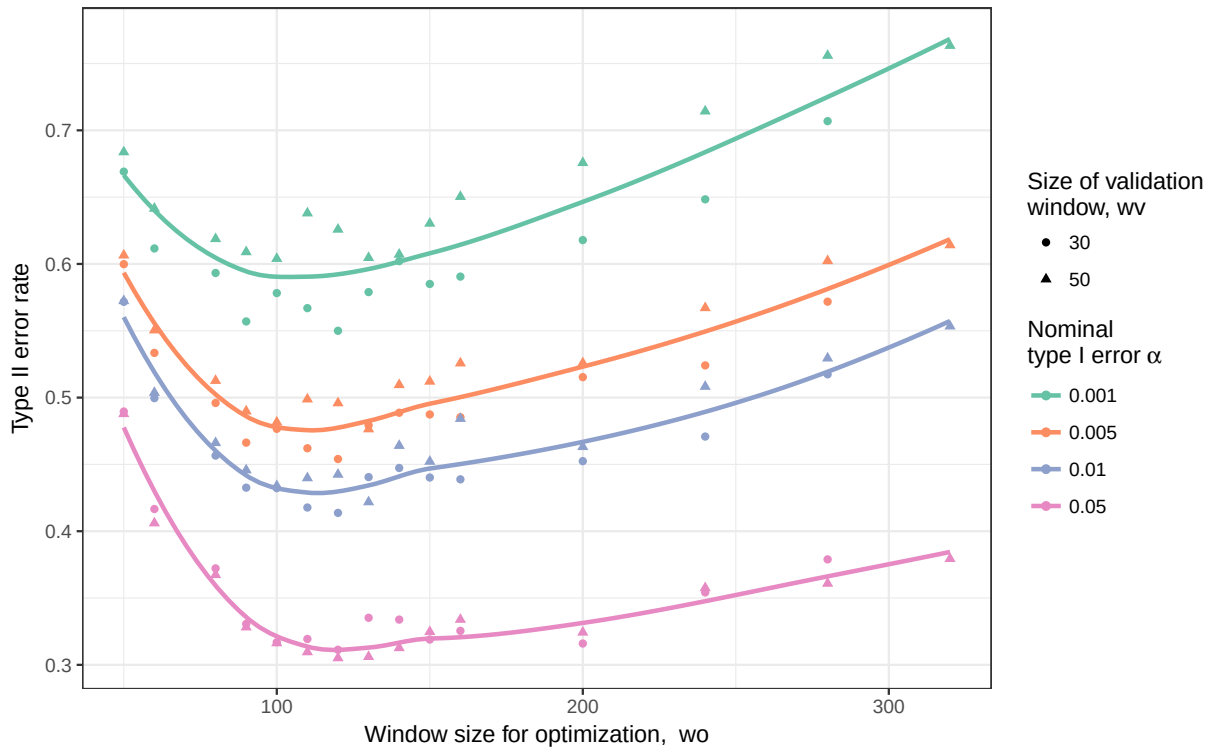

Figure 3: Type II error rates observed across a range of window sizes for optimization wo. For a window size of $wo = 120$ we see a drop in type II error rate across all type I rates considered. Smaller validation sizes wv are typically associated with a smaller type II error.

Table 4: 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 3. 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

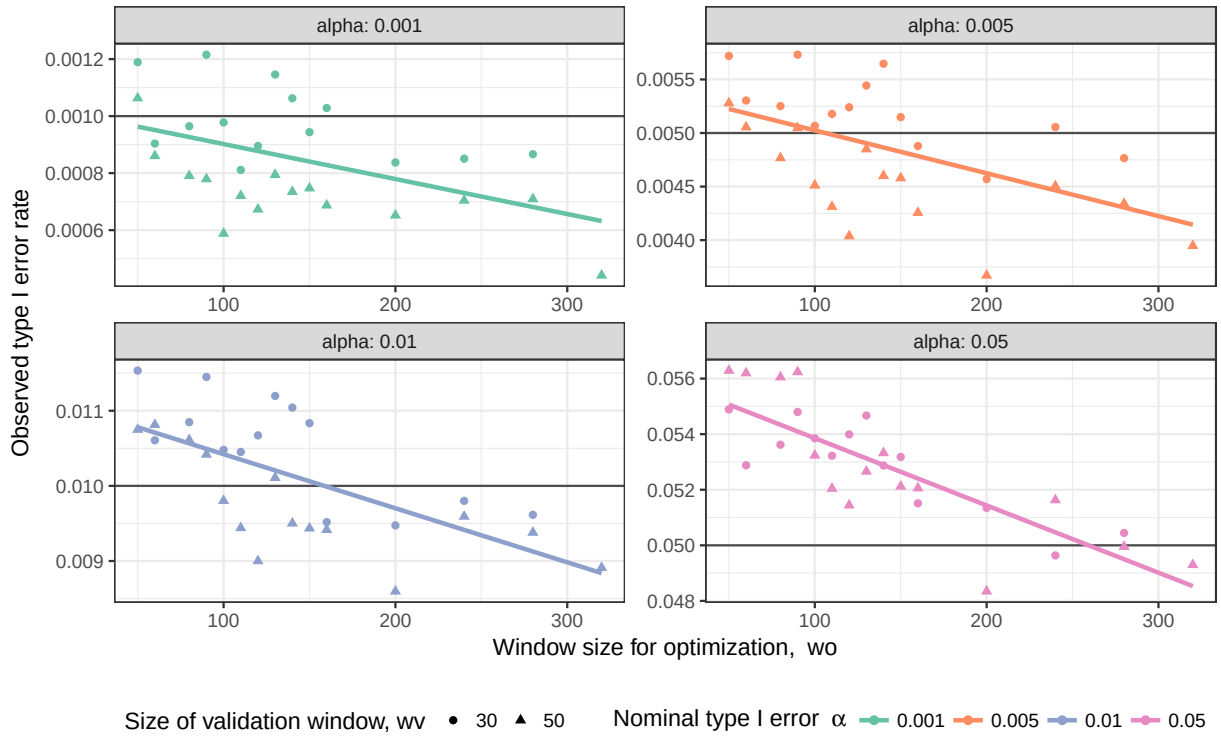

Figure 4: Comparison of observed and nominal type I error rates across a range of window sizes for optimization *wo*. The horizontal line in each facet indicates the nominal type I error rate.

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

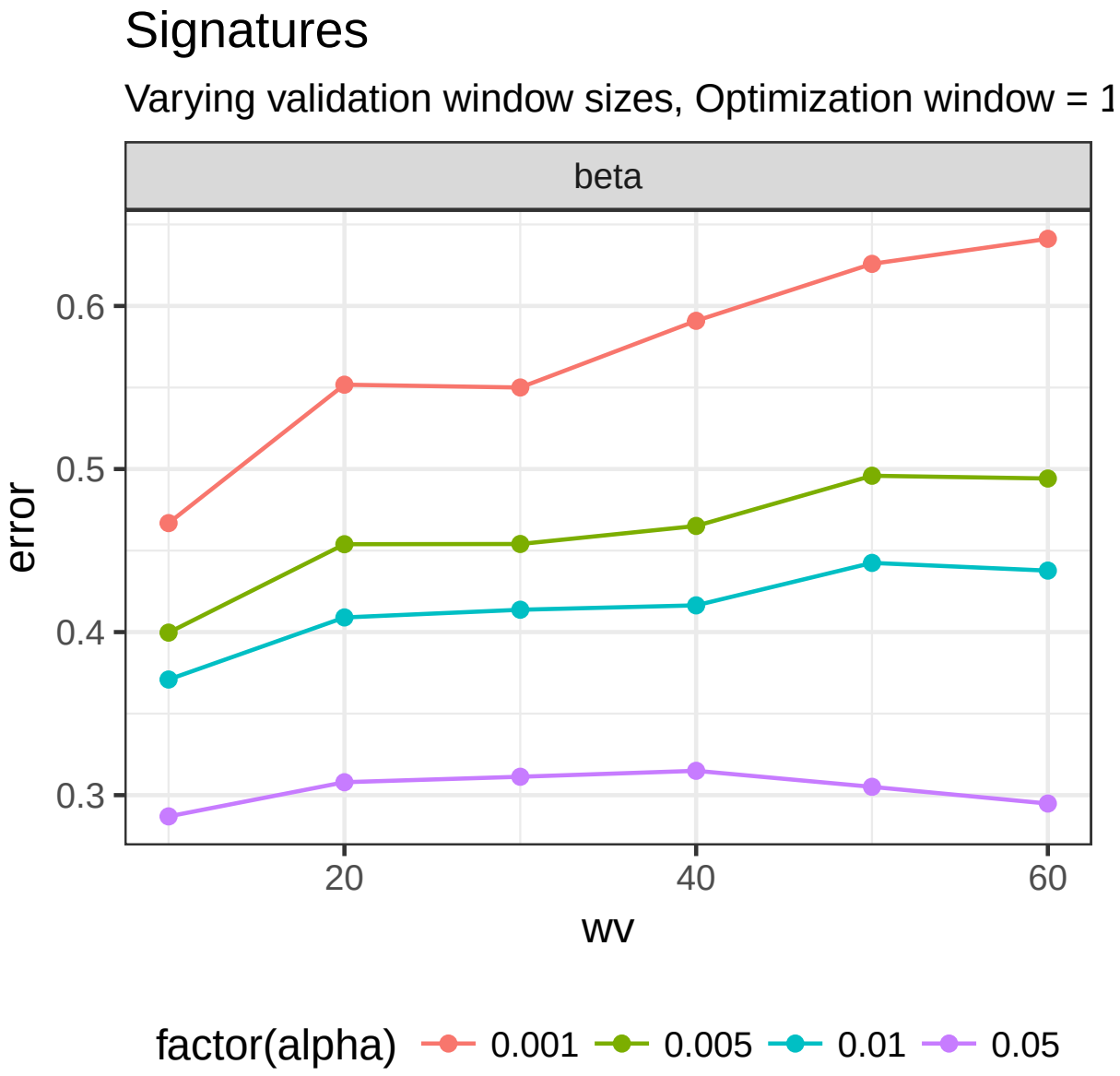

Figure 5: The figure on the left shows the actual Type I error while the figure on the right shows the Type II error for different validation window sizes and different chosen nominal alpha levels when the size of the optimization window = 120

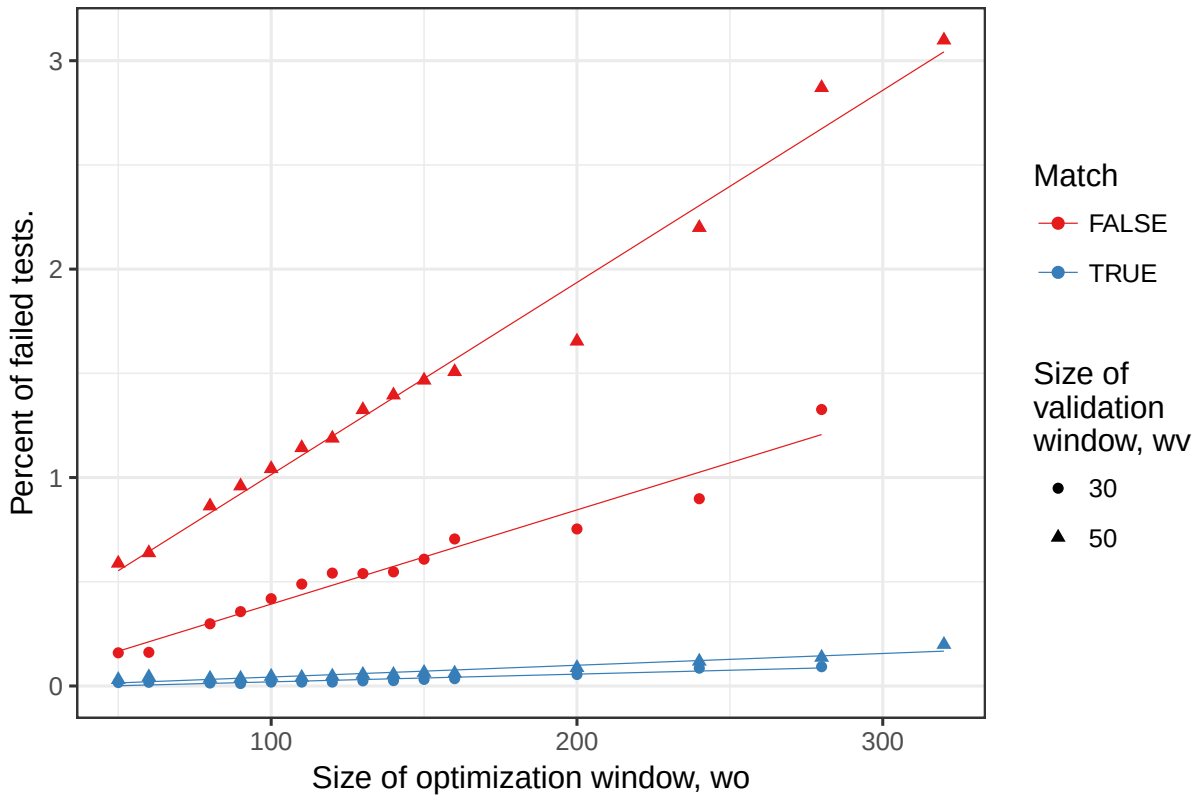

Figure 6: Number of failed tests by the window optimization size, wo, and ground truth.

Table 5: 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.

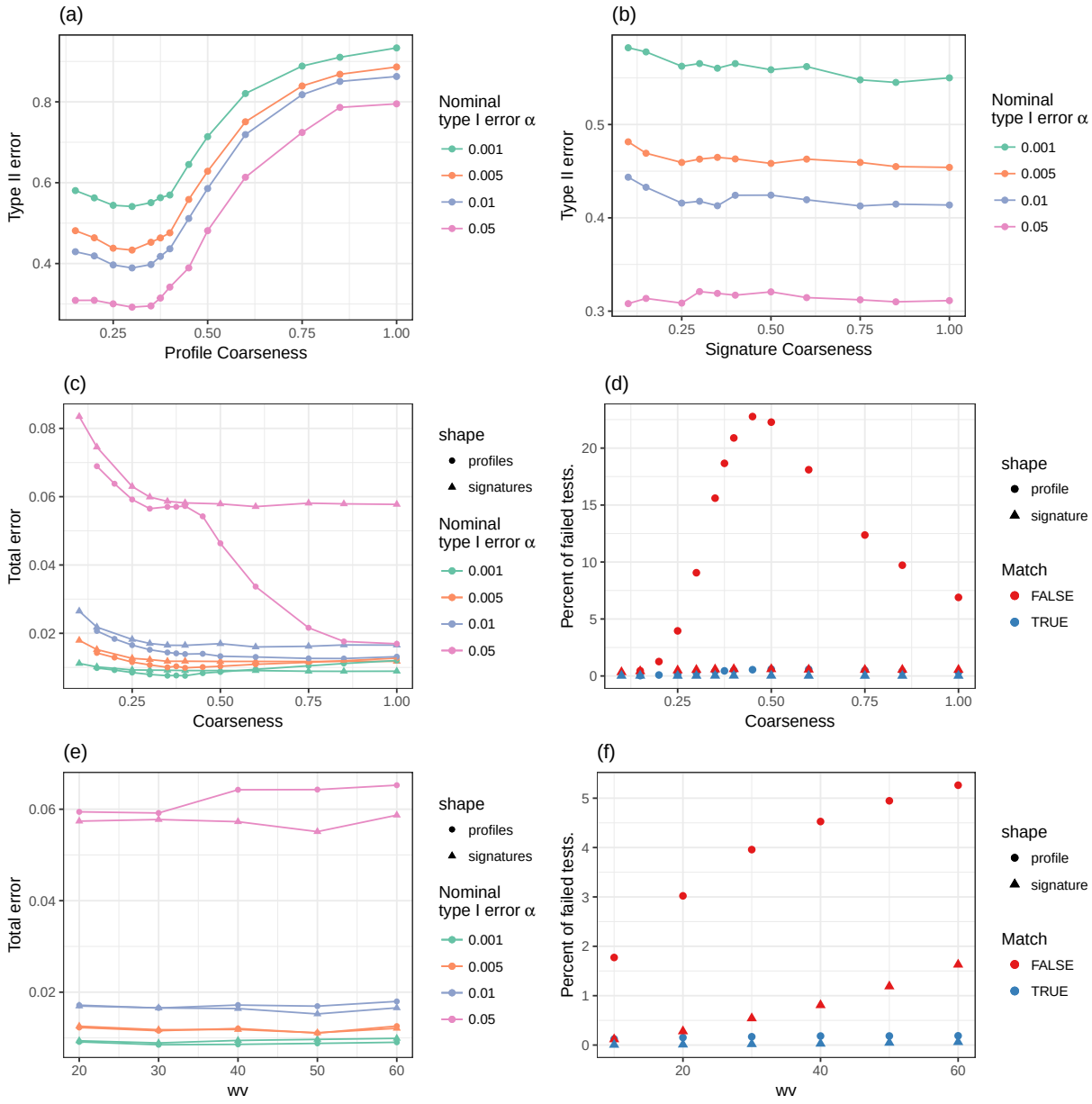

Figure 7: Row 3: Total error and Number of failed tests by the window validation size, wv, and ground truth, Row 2: Total error and Number of failed tests with Coarseness for both profiles and signatures, Row 1: Type II error for different coarseness levels as used in the modified chumbley algorithm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

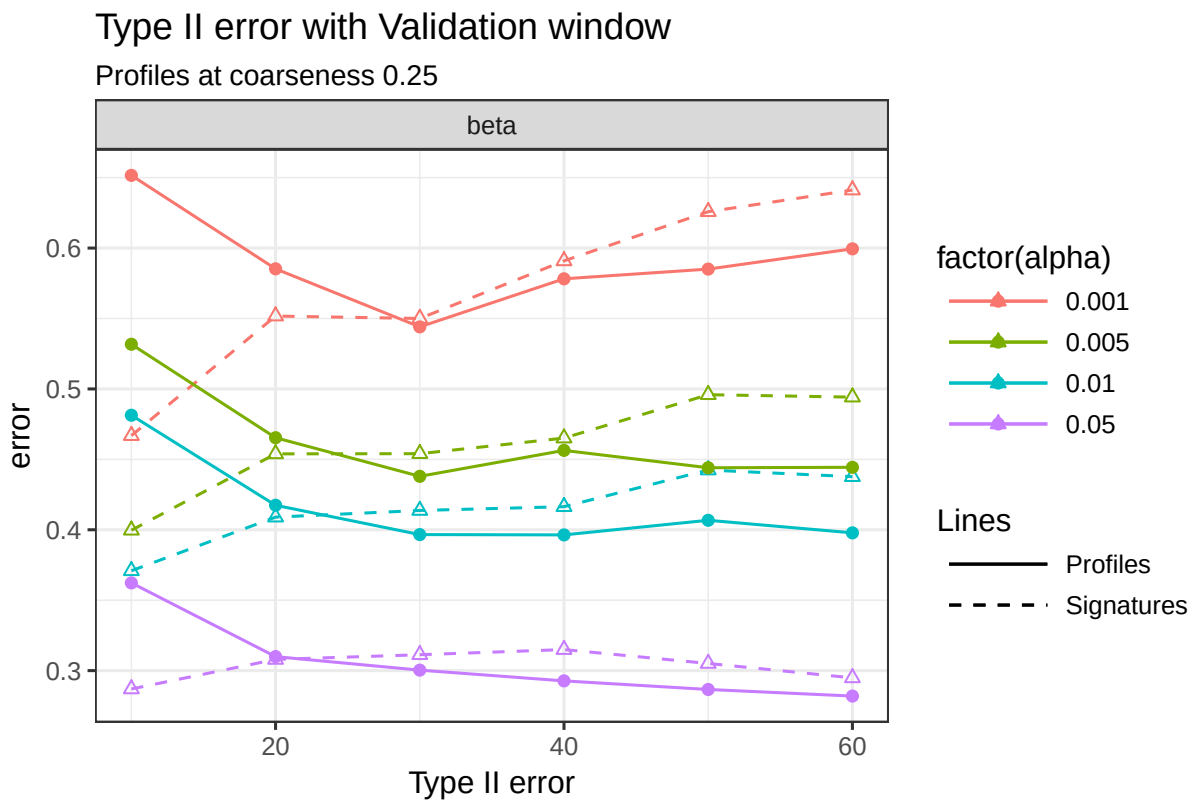

Figure 8: The figure shows the Type II error for different validation window sizes and different chosen nominal alpha levels when the size of the optimization window = 120

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

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* .

Miller, J. (1998), 'Criteria for identification of toolmarks', *AFTE Journal* **30**, 15–61.

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*

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