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 19, 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 by visual inspection under a comparison microscope and come to one of the following four conclusions: identification, inconclusive, elimination or unsuitable for examination˜(**?**). These conclusions are made on the basis of "unique surface contours" of the two toolmarks being in "sufficient agreement" (**?**). AFTE describes the term "sufficient agreement" as the possibility of another tool producing the markings under comparison, as practically impossible (**?**). This subjectivity in the assessment as well as the lack of error rates are the main points of criticisms first raised by the National Research Council in 2009 (**?**) and later emphasized further by the President's Council of Advisors on Science and Technology (**?**).

Technological advances, such as profilometers and confocal microscopy allow to measure 3D surfaces in a high-resolution digitized form. This technology has become more accessible over the last decade, and has made its way into topological images of ballistics evidence, such as bullet lands and breech faces (**????**). Digitized images of 3D surfaces of form the data 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.

In the next page and a half it is easy to lose the red line. It might help to include a table with an overview. The table should include the reference to the paper, the data used, the statistical method and the associated error rates. Various toolmarks have been studied in the literature: **?** and **?** have been analyzing screwdriver marks digitized using a profilometer; **?** have investigated 3D marks from screwdriver, tongue and groove pliers captured using a confocal microscope; **?** have been investigated digitized marks from slip-joint pliers generated by a surface profilometer.

We need an additional sentence here to get from the data to the statistical methods ...

? define a relative distance metric and use it as similarity measure between two toolmarks. **?** 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 **?**, uses a similar but more extensive framework based on a Mann-Whitney U test of the resulting correlation coefficients. This approach is non-deterministic, because segments are chosen randomly. (**?**) make the score deterministic for each pair of toolmarks by choosing segments for comparison systematically. This approach also ensures independence between segements of striae. In this paper, we are investigating the applicability of the Chumbley scoring method by **?** 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 **?**. 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. Figure 1 shows us how the signature from a bullet land (bottom) lines up with the image of the land (top) 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 are usually based on these associated signatures. **?** use an automatic method for counting consecutive matching striae (CMS). The authors report an error rate of 52% of the known same source lands comparisons as misidentified (false negative) and zero false positives for known different source lands. **?** and **?** discuss CCF (cross-correlation function) and its discriminating power and applicability for same-source analyses of bullets, but do not provide any error rates in their discussion. **?** use multiple features like CCF, CMS, D (distance measure) etc in a random forest based method and compare every land against every other land of digitised versions of Hamby 252 and Hamby 44 (**?**) published on the NIST Ballistics Database (**?**). The authors report an out-of-bag overall error rate of 0.46%, comprised of a false positive error rate of 30.05% and a false negative rate of 0.026%.

The Chumbley score provides us with another approach in the same-source assessment of bullet striation marks. **?** compare two toolmarks for same-source. The data for this

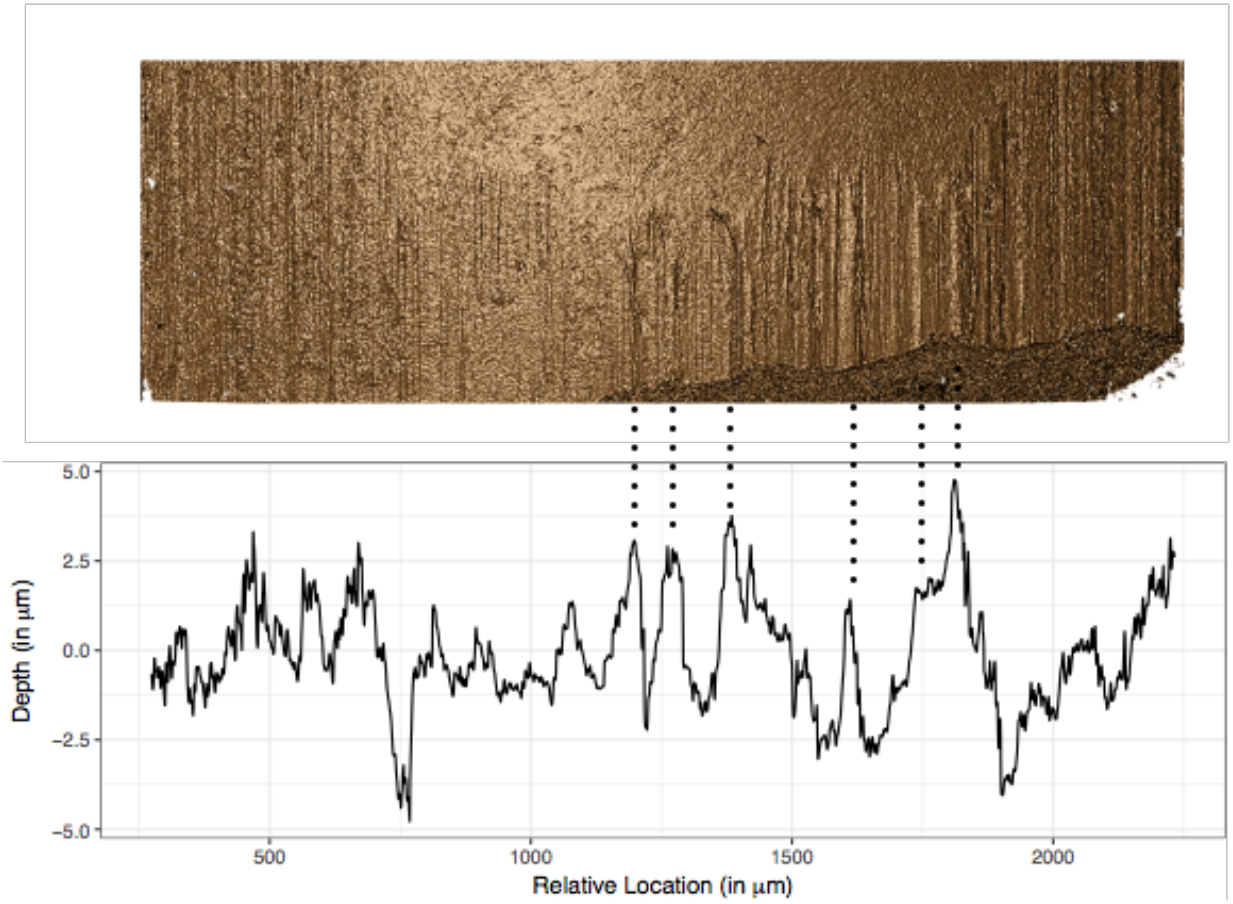

Figure 1: Image of a bullet land from a confocal light microscope at 20 fold magnification (top) and a chart of the corresponding signature of the same land (bottom). The dotted lines connect some peaks visible in both visualizations.

study was obtained from 50 sequentially manufactured screwdriver tips. **?** report error rates for markings made by the tips at different angles. For markings made under a 30 degree the authors report an average false negative error rate of 0.023 and an average false positive error rate of 0.09. For other angles the error rates for false negatives stay the same while the rate of false positives decreases to 0.01. are the angles steeper? The paper by **?** is based on the same data but the authors focus on markings made under the same angle. The error rates associated with the deterministic version of the score are 0.06 for false negatives and a false poisitive error rate of 0.

1.2 Scans for land engraved areas

Comparisons of striae from bullets are usually based on comparisons of striae in land engraved areas, which are extracted in form of cross sections, called *profiles* (**??**). From profiles bullet *signatures* (**??**) are extracted as residuals of a loess fit or Gaussian filter. Signatures 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. Do you have a reference for the previous sentenced? A detailed discussion of the extraction technique for signatures is given in **?**.

don't split the discussion on the size. between the next paragraph and There are two sources of scans for sets from the Hamby study available to us: scans of Hamby 44 and Hamby 252 are available from the NIST database (**?**). Hamby 44 has also been made available to us and has been scanned locally for CSAFE at the Roy J. Carver High Resolution Microscopy Facility using a Sensofar confocal light microscope. Scans in the NIST database are made with a NanoFocus at 20x magnification. The resolutions of the two instruments are different: the NIST scans are taken at a resolution of 1.5625 μm per pixel, while the CSAFE scans are available at a resolution of 0.645 μm per pixel. The length of an average bullet land from Hamby (9 mm Ruger P85) is about 2 millimeter, resulting in signatures of about 1200 pixels for NIST scans, and about 3000 pixels for CSAFE scans.

In comparison, scans from the profilometer used by **??** were taken at a resolution of about 0.73 μm per pixel. The screw driver toolmarks are about 7 mm in length (**?**), for a total of over 9000 pixels for the width of these scans.

This severe limitation in the amount of available data poses the main challenge in adapting the Chumbley score to matching bullet lands, because of the resulting loss in power.

1.3 Potential Challenges in Chumbley Score Adaptation

The Chumbley score allows a separation of a toolmark into raw and normalized digitized versions. Originally, this mechanism is intended to separate between class characteristics and individual markings, however, in the setting of matching bullet striae, we could also use it to separate bullet curvature in profiles from signatures before matching signatures.

1.4 The Chumbley Score Test

The Chumbley score algorithm takes input in form of two digitized toolmarks. The toolmark is in form of $z(t)$ which is a spatial process for location indexed by t. t here denotes equally spaced pixel locations for the striation marks under consideration, $t = 1, ..., T$. Let further $z(s, t)$ denote the vector of markings between locations s and t.

Let $x(t_1)$, $t_1 = 1, 2, ...T_1$ and $y(t_2)$, $t_2 = 1, 2...T_2$ be two digitized toolmarks (where T_1 and T_2 are not necessarily equal). The toolmarks under consideration are potentially from two different sources or the same source. 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.

In a pre-processing step the two markings are smoothed using a lowess (**?**) with coarseness parameter c. The purpose of this smoothing is to remove drift and (sub)class characteristics.

The Chumbley scores is calculated in two phases, namely, an optimization step and a validation step. In the optimization step, the two markings are aligned horizontally such that within a pre-defined window of length w_o the correlation between $x(t_1)$ and $y(t_2)$ is maximized:

$$(t_1^o, t_2^o) = \underset{1 \leq t_1 \leq T_1, 1 \leq t_2 \leq T_2}{\arg\max} \operatorname{cor}\left(x(t_1, t_1 + w_o), y(t_2, t_2 + w_o)\right)$$

This results in an optimal vertical (in-phase) shift of $t_1^o - t_2^o$ for aligning the two markings.

In the validation step, two sets of windows of size w_v are chosen from both markings (see Figure 2). In the first set, pairs of windows are extracted from the two markings using the optimal vertical shift as determined in the first step, whereas for the second set the windows are extracted using a different (out-of-phase) shift.

For both samples the correlations between the pairs of markings is then calculated. The intuition here is that for two markings from the same source the correlation for the in-phase sample should be high, while the correlations of the out-of-phase sample provide a measure for the base-level correlation for non-matching marks of a given length w_v. The Chumbley score is then computed as a Mann Whitney U statistic to compare between in-phase sample and out-of-phase sample. In the original method proposed in **?** both in-phase and out-of-phase sample are extracted randomly, whereas **?** proposed deterministic rules for both samples to make the resulting score deterministic while simultaneously avoiding

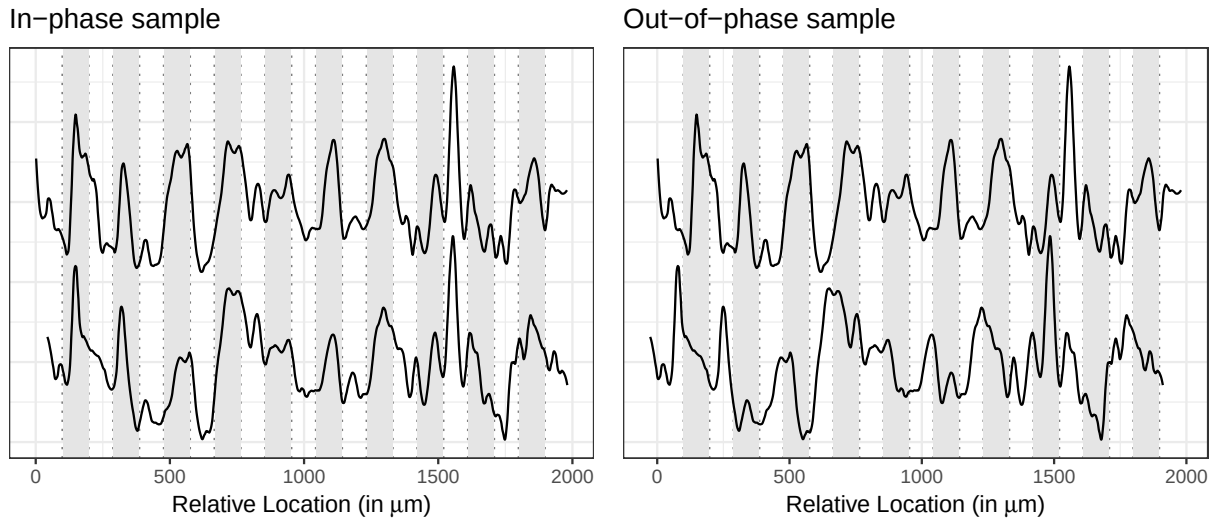

Figure 2: Two markings made by the same source. For convenience, the markings are moved into phase on the left and out-of phase on the right. In-phase (left) and out-of-phase (right) samples are shown by the light grey background. The Chumbley-score is based on a Mann-Whitney U test of the correlations derived from these two sets of samples.

overlaps within selected marks to ensure independence.

2 Testing setup

2.1 The Data

More on NIST database, some on CSAFE

Introduce profiles and signatures, as shown in figure 3.

Lands for all Hamby-44 and Hamby-252 scans are made available through the NIST ballistics database (**?**) and were considered, here. Both of these sets of scans are part of the larger Hamby study (**?**) 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.

Profiles and signatures were extracted from the Hamby 44 and Hamby 252 data as described in **?**.

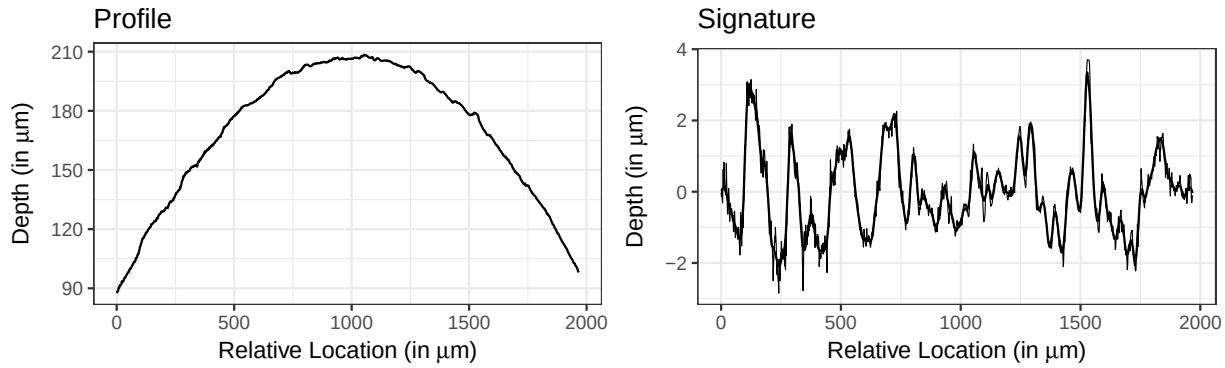

Figure 3: Bullet land profile (left) and the corresponding signature (right) for one of the lands of Hamby-44.

2.2 Setup

using (a) signatures and (b) profiles, run chumbley score across scans from NIST and CSAFE for various settings of w_o and w_v (and coarseness c for profiles).

We used the adjusted Chumbley method as proposed in **?** and implemented in the R package `toolmaRk` (**?**) on all pairwise land-to-land comparisons of the Hamby scans provided by NIST (a total of 85,491 comparisons). The settings for optimizing and validating window sizes, w_o and w_v, ranged from $w_o \in [50, 280]$ and $w_v \in \{30, 50\}$, see also figure 4.

2.3 Results

For signatures from NIST scans we see three problems:

1. type-2 error rate is at best 30% for a type-1 error rate of 5%, which is well above the error rates we see for tool marks from screw drivers,

2. the observed type-1 error, which generally close to the nominal type-1 error rate, depends on the size of the optimization window: as the window size increases, the observed type-1 error decreases,

3. the Chumbley-score fails to provide a result for up to 3% of the cases. This drop-out rate is considerably higher when the two lands are from different sources than when the lands are from the same source.

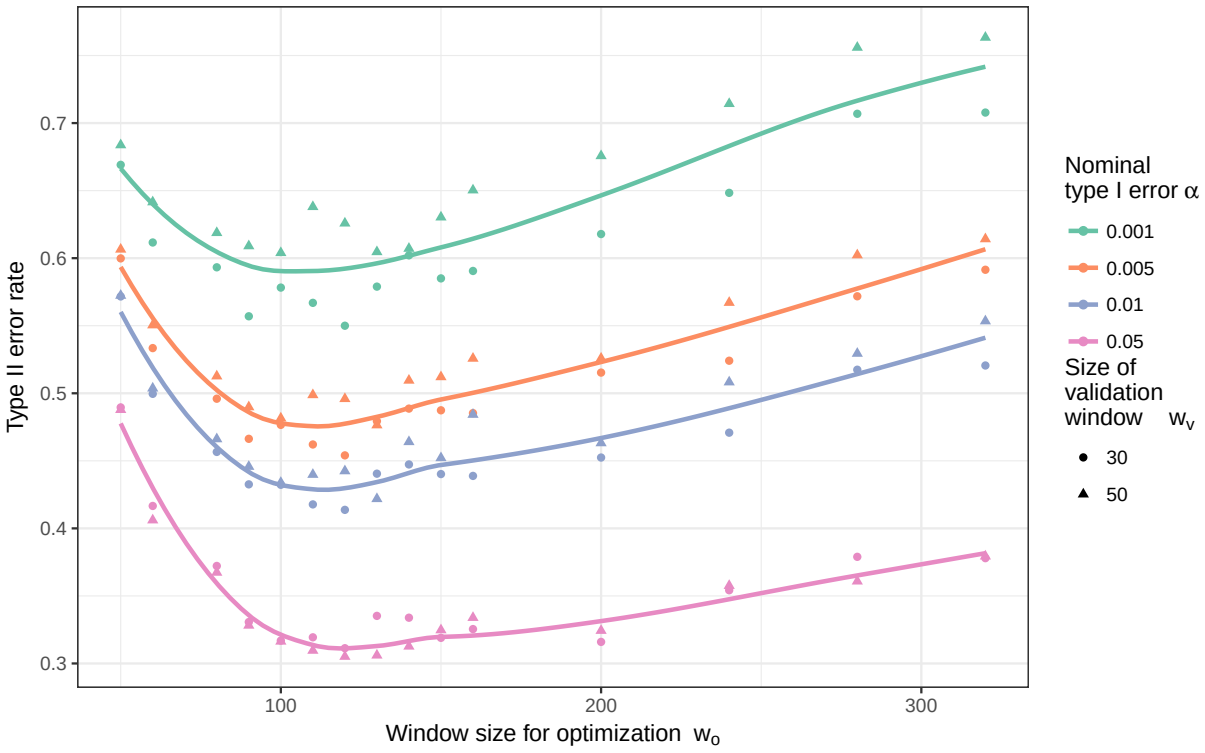

Figure 4: Type II error rates observed across a range of window sizes for optimization w_o. For a window size of $w_o = 120$ we see a drop in type II error rate across all type I rates considered. Smaller validation sizes w_v are typically associated with a smaller type II error.

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

2.3.1 Signatures

Signatures of lands for all Hamby-44 and Hamby-252 scans made available through the NIST ballistics database (**?**) were considered. Both of these sets of scans are part of the larger Hamby study (**?**) 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 **?**. 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.

2.3.2 Profiles

The profiles are cross-sectional values of the the bullet striation mark which are chosen at an optimum height (x as used by **?**). This x or height is not a randomly chosen level. The rationale behind the choice has been explained by **?**. 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 **?** 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 **?** 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 **?**) 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 **?** and implemented in the R package `toolmaRk` (**?**) 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 4 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

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 5 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

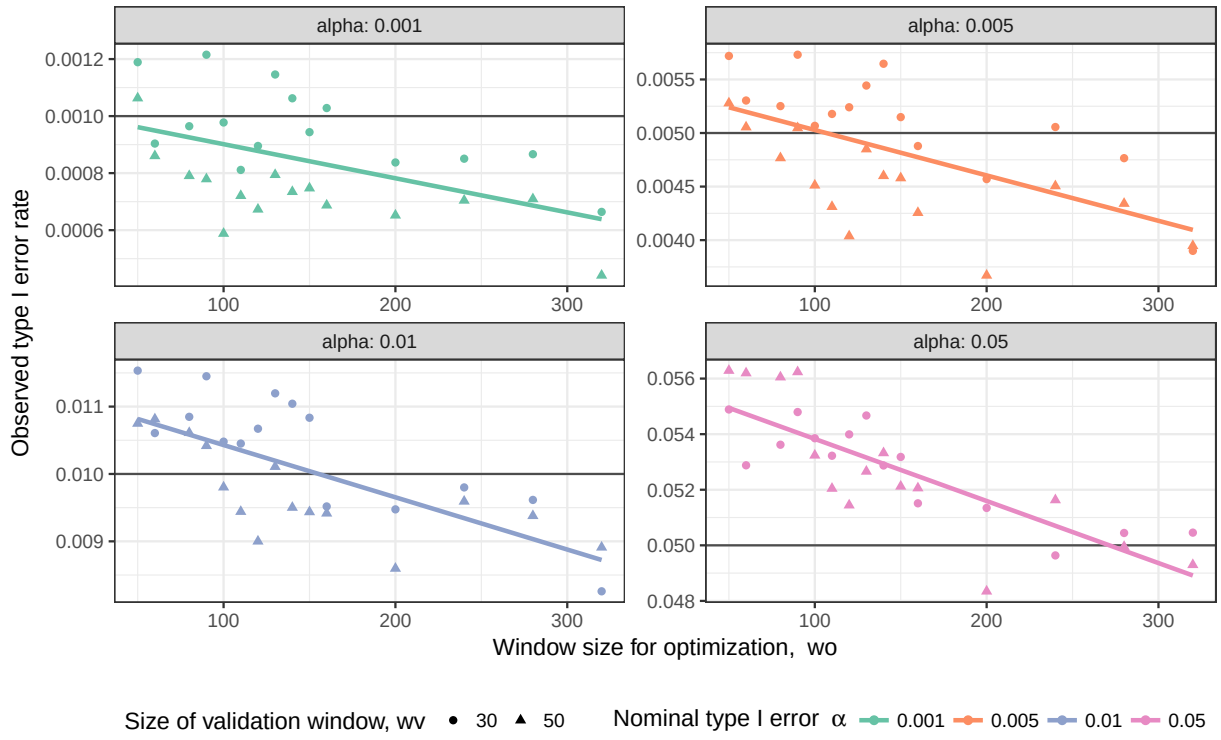

Figure 5: Comparison of observed and nominal type I error rates across a range of window sizes for optimization *wo*. The horizontal line in each facet indicates the nominal type I error rate.

error is comparable to the nominal level for only a select few validation window sizes. For 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 5. 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 7).

The actual type I error and type II error for signatures were also compared for different validation window sizes. Figure 6 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 7 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

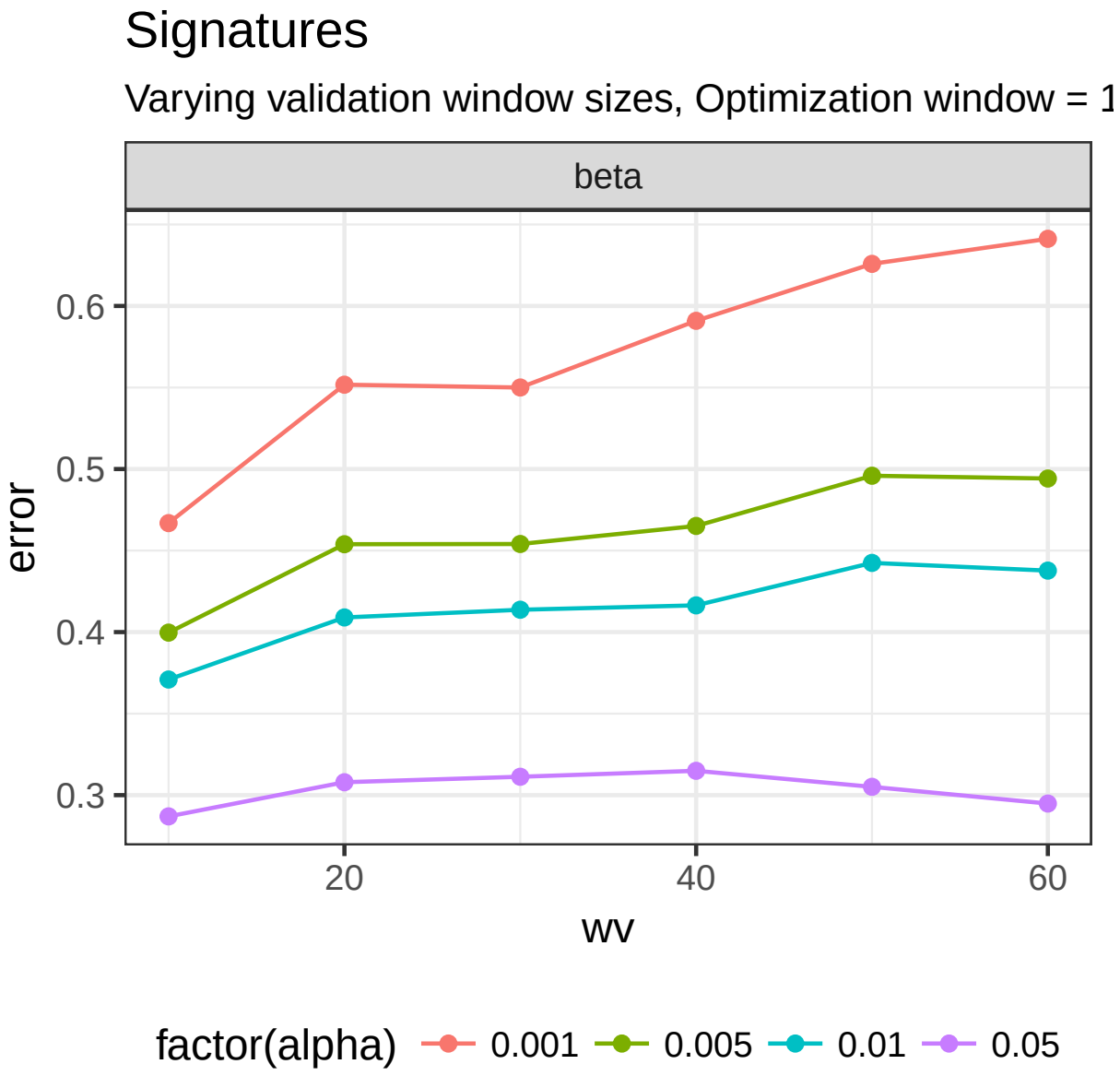

Figure 6: The figure on the left shows the actual Type I error while the figure on the right shows the Type II error for different validation window sizes and different chosen nominal alpha levels when the size of the optimization window = 120

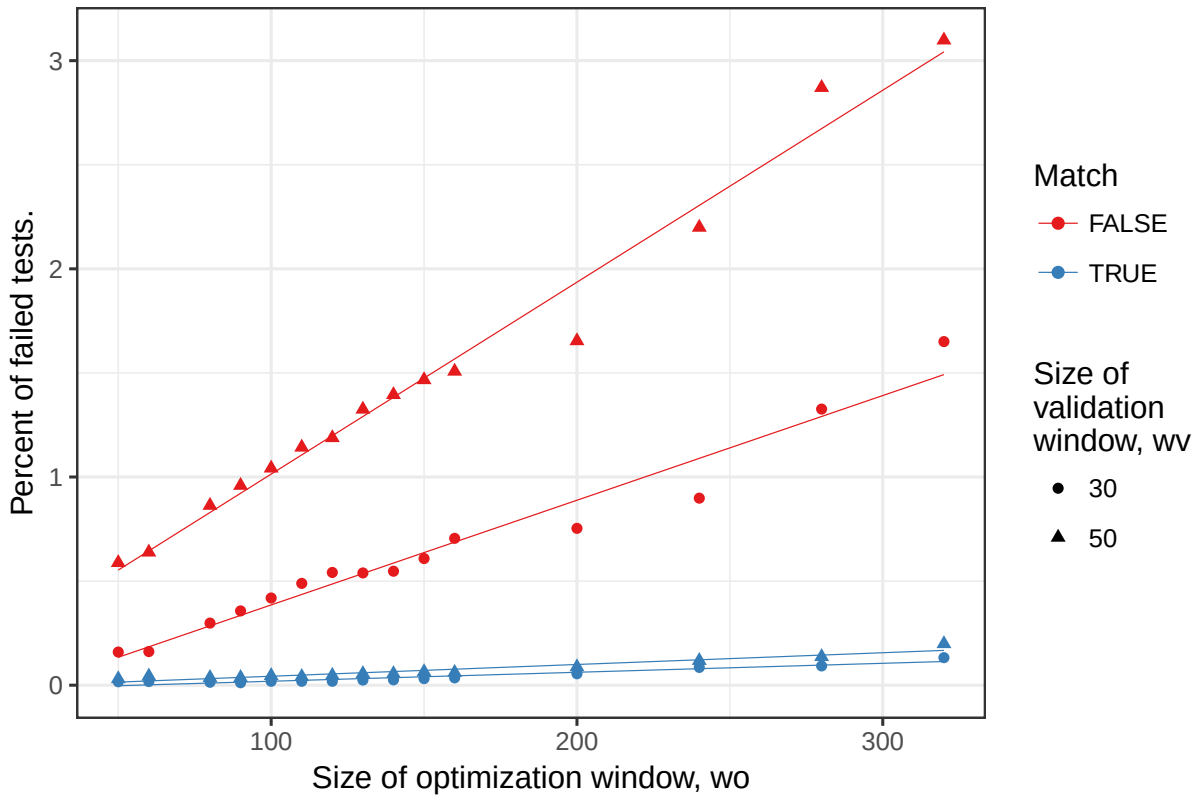

Figure 7: Number of failed tests by the window optimization size, wo, and ground truth.

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	0.503	0.028
FALSE	50	0.922	0.035
TRUE	30	0.043	0.004
TRUE	50	0.056	0.005

signatures is so large, that the remaining overlap is too small to accommodate windows for validation. The problem is therefore exacerbated by a larger validation window. Figure 7 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 8 (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 8 this coarseness value is used.

On the other hand Figure 8 (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 8 (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 8(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 7 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 8(d) and (f), we compare profiles and signatures on the basis of number of failed tests. The profiles chosen for figure 8(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 **?**. The smoothing in these signatures were determined and fixed on the basis of their performance in the random forest based algorithm proposed by **?**. The comparison of profiles and signatures with variation of validation window size therefore is made on even footing. The trends are similar to figure 7 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 8 (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.

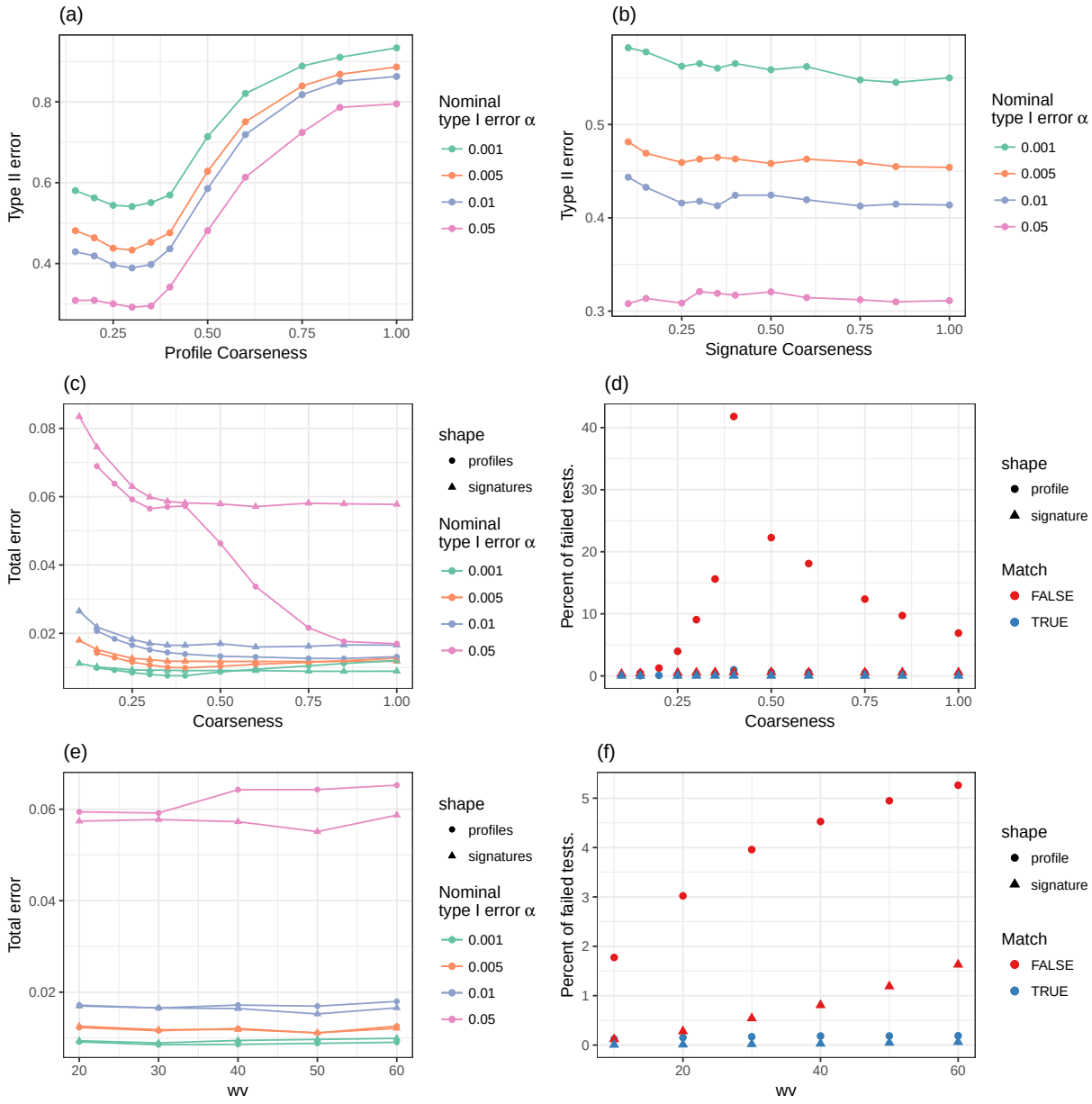

Figure 8: Row 3: Total error and Number of failed tests by the window validation size, wv, and ground truth, Row 2: Total error and Number of failed tests with Coarseness for both profiles and signatures, Row 1: Type II error for different coarseness levels as used in the modified chumbley algorithm 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 **?**, and some level of modification to the deterministic algorithm needs to be identified and tested that would reduce the number of incorrect classifications.