Addressing the National Academy of Sciences’ Challenge:
A Method for Statistical Pattern Comparison of
Striated Tool Marks


Nicholas D. K. Petraco,1,2 Ph.D.; Peter Shenkin,3 Ph.D.; Jacqueline Speir,4 M.S.; Peter Diaczuk,1,5 DABC, B.S.; Peter A. Pizzola,6 DABC, Ph.D.; Carol Gambino,7 M.S. and Nicholas Petraco,1,8,9 DABC, M.S.

1 Department of Sciences, John Jay College of Criminal Justice, City University of New York, 899 10th Avenue, New York, NY, 10019.

2 Faculty of Chemistry, Graduate Center, City University of New York, 365 5th Avenue, New York, NY, 10016.

3 Department of Mathematics and Computer Science, John Jay College of Criminal Justice, City University of New York, 899 10th Avenue, New York, NY, 10019.

4 Department of Chemical and Physical Sciences, Forensic Science Program, Cedar Crest College, 100 College Drive, Allentown, PA, 18104.

5 PEDICO Research, RR2 Box 62, Waymart, PA, 18472.

6 New York City Office of Chief Medical Examiner, Special Investigations Unit, 421 East 26th Street, New York, NY, 10016.

7 Department of Sciences, Borough of Manhattan Community College, City University of New York, 199 Chambers Street, New York, NY, 10007.

8 Petraco Forensic Consulting, 73 Ireland Place, Amityville, NY, 11701.

9 New York City Police Department Crime Laboratory, 150-14 Jamaica Avenue, Jamaica, NY, 11432.
ABSTRACT: In February 2009 the National Academy of Sciences published a report entitled “Strengthening Forensic Science in the United States: A Path Forward”. The report notes research studies must be performed to “…understand the reliability and repeatability…” of comparison methods commonly used in forensic science. Numerical classification methods have the ability to assign objective quantitative measures to these words. In this study, reproducible sets of ideal striation patterns were made with nine slotted screwdrivers, encoded into high dimensional feature vectors and subjected multiple statistical pattern recognition methods. The specific methods employed were chosen because of their long peer reviewed track record, widespread successful use for both industry and academic applications, rely on few assumptions on the data’s underlying distribution, can be accompanied by standard confidence levels and are falsifiable. For PLS-DA, correct classification rates of 97% or higher were achieved by retaining only eight dimensions of data. PCA-SVM required even fewer dimensions, 4D, for the same level of performance. Finally, for the first time in forensic science, it is shown how to use conformal prediction theory to compute identifications of striation patterns at a given level of confidence.

KEYWORDS: forensic science, Daubert, Frye, National Academy of Sciences, tool marks, multivariate, pattern recognition, error rates
Over the last two decades, DNA profiling has become one of the most widely applied techniques for the identification of biological samples in forensic science. This is largely due to the volumes of allele frequency data and the clear applicability of simple statistical methods. The unparalleled success of DNA profiling is likely responsible for the recent National Academy of Sciences’ (NAS) report on the raising of standards for scientific examination of all forms of physical evidence (e.g. tool marks, soils, dust, questioned documents, shoe prints, fire debris, fingerprints, gunshot residue, tire tracks, hairs, fibers, etc.) (1). Tool mark impression evidence, for example, has been successfully used in courts for decades, but its examination has lacked scientific, statistical proof that would independently corroborate conclusions based on morphology characteristics (2-7). In our study, we will apply methods of statistical pattern recognition (i.e. machine learning) to the analysis of tool mark impressions.

The process of associating, tool mark evidence to a specific tool involves classifying the object into groups of similar objects. To conclude that two striation patterns have a high likelihood of being associated, a tool mark examiner usually states that the “quantity and quality” of the markings were such as to allow them to properly draw their conclusion.

The Association of Firearms and Toolmark Examiners (AFTE), the de facto group that set the standards for firearm and tool mark examination, states the theory of identification as it relates to tool marks as (2, 8):

“A. The theory of identification as it pertains to the comparison of tool marks enables opinions of common origin to be made when unique surface contours of two tool marks are in sufficient agreement”.

“B. This sufficient agreement is related to significant duplication of random toolmarks by correspondence of pattern or combination of patterns of surface contours.”

“a. Significance is determined by comparative examination of two or more sets of surface contour patterns comprised of individual peaks, ridges and furrows.”

Impressions and striations made by tools and firearms can be viewed as mathematical
patterns composed of peaks, ridges and furrows which we will refer to as features. Numerical classification methods (9-11) are of particular interest because they have the potential of assigning objective quantitative measures to the words “sufficient agreement” and “comparative examination”. Unfortunately only a few numerically based studies for tool mark and firearm impression/striation comparisons (mostly firearms) have appeared in the literature (12-41). Nichols’s reviews give excellent background and insightful overviews for most of these studies (42-44).

In 1959, Biasotti recognized the need for empirically based statistical studies on tool mark patterns (20). His study recorded the number of matching “striations” between bullets fired from the same gun and bullets fired from different guns. He parameterized his model (now known as consecutively matching striae or the CMS model) in terms of groups of matching “lines”. Biasotti defined striation as “… engravings or striations appearing on the bullet as a result of being engraved by the individual irregularities on characteristics of the barrel plus any foreign material present in the barrel capable of engraving the bullet.” He identified groups of matching lines between bullets in phase; matching occurred “…when the lines appear to be similar in contour and of common origin.” Building on Biasotti’s work, Neel and Wells have recently published a large scale statistical study where they examined 4188 striated tool mark comparisons for various sized runs of consecutive matching striations. They found very low empirical frequencies for CMS runs larger than 4X if two different tools were known to have generated the striation patterns (19).

One of the first efforts for the computational comparison of striated firearm data was by Gardner in the late 1970s (13). His system attempted to use the same striation data a ballistics examiner would use, and echoed much of the same ideas that were implemented two decades later in the IBIS system marketed by Forensic Technology Inc. Using an SEM, Gardner obtained multiple scans of land-and-groove areas from .38 caliber bullets fired by four different revolvers. Striations were quantified by comparing derivative peak heights/depths to the derivative mean line. Gardner then employed a heuristic probability based formulation to generate similarity scores between all possible comparisons of engraved areas (land-to-land and grove-to-grove). Reasonably correct classification rates were obtained on the small test set used—two bullets fired from each of three different guns.
Databases for tools and tool marks mostly came into existence in the early 1990s (32, 45-47). The IBIS system is probably the most widely used database system for semi-automated firearm identification and functions like a computer aided comparison microscope (47). The version in widest use employs 2D gray scale images of various striation and impression patterns on bullets and casings, and converts them into a “signature”. This signature is compared against a large database of signatures collected by local, state and national law enforcement agencies. The image of a pattern from questioned firearms evidence is “scored” by a system “correlation” server for “similarity” against entries in the database. The reason the words “signature”, “scored”, “correlation” and “similarity” appear in quotes is because Forensic Technology Inc. does not scientifically define these very important technical terms and because the comparison algorithms are considered trade secrets of a private company. In 2005, a committee of the National Research Council (NRC) was assembled to assess the feasibility, accuracy and technical capability of a national ballistics database. The committee noted that IBIS is a tool for search, not verification (48). If a tool mark comparison system is to evolve to meet the needs of forensic science and be able to stand up to the Daubert challenge, the inner workings absolutely must be public knowledge.

With the advent of confocal microscopy and laser scanners, the acquisition of the entire 3D surface of a tool mark can be obtained. Very recently an excellent study by Bachrach et al. appeared where confocal microscopy was used to digitally record the surfaces of striated tool marks made by screwdrivers and tongue and groove pliers (52). The surfaces were then filtered and averaged to form a surface “signature”. Similarity scores for all possible pairs of signatures were generated based on the cross-correlation function and used to produce matching and non matching distributions (histograms). Algorithm generated identifications were found to be highly reliable so long as the screwdrivers’ angles of attack were consistent (angle of attack was obviously not an issue for tongue and groove pliers). A NIST study on an automated bullet signature identification system has also recently been published (53). The NIST scheme used the same principles as the previous study to establish likely bullet-gun associations. This new system is reported to have a 10% increase in accuracy over current commercially available systems which is a marked increase.
Howitt, Tulleners et al. recently published a model for the computation of correspondence probabilities (i.e. “matching”) between striation patterns imparted on bullets fired by the same gun vs. bullets fired from different guns (60). Again, the authors use the same functional definition of a “line” in a striation pattern as was given by Biasotti. This definition makes their theory very general and equally applicable to striation patterns imparted by actual tools and striation patterns found on firearms evidence. Their theory can also take into account arbitrary magnification levels (which would be a parameter of a comparison microscope) and the number of lines found in a striation pattern. Output of the approach is a probability that various CMS runs on striation patterns generated by different sources would match purely by chance. The entire model rests on the assumption that the possible patterns which the lines can form, are probabilistically independent of each other and are identically distributed. Their study shows some evidence for this assumption. The Howitt-Tulleners computed probability of random correspondence between 2X CMS runs on bullets (called doublets in the study) from different sources is between 0.1-0.16 at 20 mm resolution (i.e. at 10%-16% chance) and 0.14-0.24 at 30 mm resolution. Biasotti’s empirically derived probabilities for the same situation are 0.2-0.46 depending on whether or not the bullet is jacketed. Results however are in less agreement if 3X CMS runs (called triplets in the study) are considered (20, 60). The Howitt-Tulleners computed probability of random correspondence for this situation is 0.003-0.005 at 20 mm resolution and 0.007-0.01 at 30 mm resolution. Biasotti’s empirically derived probabilities however are 0.01-0.1 depending on jacketing (20, 60).

Despite these studies, there are no standard methods for the application of probability and statistics to the analysis of tool mark evidence. Our work is intended to help in the establishment of standard protocols. We believe that when applying statistical pattern recognition methods to legal issues, they should rely on as few underlying assumptions about the data (i.e. the evidence) as possible. For these reasons, we wanted to choose methods which: 1) make few assumptions about the form of underlying statistical distributions of the data, 2) function adequately with small or limited data sets and 3) have been shown to work well on real world problems (e.g. in industry and medicine). Unlike past studies, we take an entirely multivariate approach to the statistical discrimination of striated tool marks. This approach has been successfully exploited numerous times in the forensic science literature (61-68). We utilize two of the most successful multivariate pattern recognition methods: partial least squares discriminant analysis (PLS-DA)
and support vector machines (SVM). Both make no assumptions about underlying probability densities, are designed to work well with small sample sizes, and enjoy an extensive record of performance in the literature (69, 70). Standard confidence intervals (at the 95% level) for the classification algorithms’ output (tool mark I.D.s) are computed from conformal prediction theory and reported as well. This study focuses on striated tool marks made by screwdrivers, which can be applied in practice quickly. It is expected that the methodology developed in this study will translate to any kind of striated tool mark or firearm evidence.

**Methodology**

Nine identical high quality Craftsman® quarter inch slotted screwdrivers were purchased at a local hardware store. The screwdrivers, shown in Figure 1, were brand new and came in packages of three. The screwdriver shafts were manufactured by drop forging blanks followed by grinding off of resulting flashing. Subclass characteristics on the tool’s working surface are a possibility with some manufacturing methods, but grinding is not one of them. In general, statistical pattern comparison algorithms would detect both similarities (possible subclass characteristics) and dissimilarities (individual characteristics) between sets of features for many different tool marks. However, even though some of the features making up a tool’s working surface may be subclass characteristics, it is unlikely that the majority of the multitudes of features used by a statistical pattern recognition scheme are subclass in nature. If subclass characteristics were a significant issue for a set of tools, the problem would manifest itself in a study of this design by producing relatively high error rates in the testing phase of the algorithm. We, in fact, did not encounter this problem (see below). However, note that any statistical pattern recognition scheme should be thoroughly tested before being put into practice. With proper validation it is unlikely that subclass characteristics would cause major problems with an algorithmic discrimination process and go unnoticed.

The screwdrivers were used to generate multiple reproducible striation pattern standards. For this pilot study the striation patterns were made by hand while holding the tip of the screwdriver parallel to the striation medium surface. This was done to introduce some extra “case work” realism into patterns to test how well the classification algorithms hold up to the variability expected from the pattern-reproduction process. A jig to hold the screwdriver could also be used to make even more uniform patterns if desired.
Number 2 Roma Plastina modeling clay ([www.SculptureHouse.com](http://www.SculptureHouse.com), last accessed April 2010) was used as impression medium. Figure 2 illustrates the process. Further details of striation pattern generation are available in reference (71). The striation patterns made by each of the nine screwdrivers were digitally photographed with a Nikon DSFi1 digital camera under a Nikon SMZ1000 stereo-microscope at 20x magnification. The positions and widths of a small number of striation lines/grooves were measured with a stage micrometer. The measurements were later used to calibrate the digital image processing program ImageJ (72). Striation patterns were generated for only one side of each screwdriver’s working surface. A total of 75 actual patterns were recorded. Using these real striation patterns 732 more striation patterns were simulated for testing purposes. See the statistical methods section for simulation details.

*Striation Pattern Quantification*

The surface topography of a striated tool mark obviously contains a tremendous amount of physical information. Fortunately the human mind, in particular the professional tool mark examiner’s mind, can efficiently handle such a voluminous amount of information. Unfortunately computers and mathematics are not as efficient as the human brain and they require “models” of physical reality in order to execute their orders in a reasonable amount of time.

For the striation patterns examined in this study, the positions of striation lines/groves measured with the stage micrometer were used to calibrate the image processing program ImageJ (72). Distances of each line or groove from the left edge of each striation pattern were measured to the nearest 0.05 mm. Each striation pattern is no more than 7mm (~0.25in) wide. For each pattern a list of 140 pieces of information (7 mm / 0.05 mm slots) is created. Each piece of information is a 1 or 0, *i.e.* a “bit”. The list thus consists of “slots” for information and is superimposed over each striation pattern. A 1 is recorded in a slot of the list if a line or groove is present or spans the slot. A 0 is recorded otherwise. The procedure yields a 140-dimentional binary feature vector for each pattern, which is reminiscent of a “bar-code” (cf. Figure 3). In this study it was found that of the 140 components in the feature vector, 19 always had value 0 across all recorded striation patterns. These non varying components were excluded and thus feature vectors of 121D (140D-19D =121D) were used in the statistical analyses described below.
Note from the stack of striation patterns in Figure 2 (rightmost picture in the figure), some of the patterns required alignment (registration) because a striation pattern was not entirely complete. Figure 2 (rightmost picture in the figure) depicts 9 registered (aligned) striation impression patterns collected using a single screwdriver. Registration was performed by aligning the left edge and/or an obvious groove shared between patterns generated by a single tool. As can be seen from Figure 2, exemplar impressions generated by a single tool contain inherent variability. Although this can be controlled to some degree during exemplar pattern reproduction in a controlled laboratory setting, it must be considered an unavoidable consequence of patterns generated during the commission of a crime, and therefore should be represented in an exemplar database. All data measurements made with ImageJ were preprocessed and registered using a program written in Mathematica 7 (73). Our Mathematica notebooks are available upon request. A more automated registration algorithm is in development in our lab.

Statistical Methods

The binary feature vectors generated from the striation patterns were arranged into an $n \times p$ data matrix ($X$) for analysis:

$$X = \begin{bmatrix}
X_{11} & \ldots & X_{1j} & \ldots & X_{1p} \\
\vdots & \vdots & \vdots \\
X_{ij} & \ldots & X_{ij} & \ldots & X_{ip} \\
\vdots & \vdots & \vdots \\
X_{nj} & \ldots & X_{nj} & \ldots & X_{np}
\end{bmatrix}$$

where $n = 75$ is the number of striation patterns and $p = 121$ is the number of components in each feature vector. For each $X_{ij}$

$$X_{i,j} = \begin{cases} 
1 & \text{if a line/groove falls in slot } j \text{ of striation pattern } i \\
0 & \text{if a line/groove does not fall in slot } j \text{ of striation pattern } i
\end{cases}$$

The symbol $X_i$ designates row $i$ of $X$ and is a vector of data representing striation pattern $i$. Multivariate statistical methods were used to transform the data set ($X$) into a new data set ($Z$). These transformed variables may be used in classification algorithms (supervised or
unsupervised) to discriminate between different screwdrivers. In this study, \( k = 9 \) different screwdrivers of the same brand were used. The multivariate transformation analyses of data set \((X)\) undertaken in this study were partial least squares discriminant analysis (PLS-DA) and principal component analysis (PCA). The multivariate statistical discrimination methods used in this study were PLS-DA (i.e. it is both a transformation and discrimination method), one-vs.-one multiclass support vector machines (SVM) and conformal prediction theory utilizing PCA-SVM and 3-nearest neighbor classifiers (3-NN). Further details on the properties of PCA and PLS derived variables as well as the above stated discrimination methods are available in references (10,70,76, 79). We note however that none of the methods used in this study depend on the data being Gaussian.

The raw data was mean centered and variance scaled before all transformation/discrimination computations. All statistical computations were performed with the statistical software R (74).

**Principal Component Analysis**

Principal component analysis (PCA) is a multivariate procedure that is used to reduce the dimensionality of a data set \((X)\) to a new data set \((Z_{PC})\) of “derived variables” which account for successively decreasing amounts of variance (76-78). The variance order of the variables in \(Z\) provides guidance for the reduction of the data’s dimensionality while retaining an adequate representation. See references (76) and (78) for a fuller account of the details of PCA. The native R PCA program prcomp was used in this study (74).

**Support Vector Machines**

Small sample sizes are inevitable for many statistical studies of tool marks. Statistical learning theory and its practical application, the support vector machine (SVM) was developed in response to the need for reliable statistical discriminations within small sample studies (79). SVMs seek to determine efficient decision rules in the absence of any knowledge of probability densities for the data by determining maximum margins of separation (cf. Figure 4) (69, 79). This procedure produces an algorithm which determines linear decision rules with (typically) large margins for error. In this study a linear kernel was used with the SVM algorithm. Also, the penalty parameter, \( C \) was found to be 10 using a standard line search (80).
The SVM methodology was originally designed to separate two groups but it does have several incarnations which can handle multi-group (multicategory) problems. The most popular, because it works well in practice, is to consider all possible pairs of groups of striation patterns and determine a two-group (binary) SVM for each pair. Thus if there are \( k \) samples of striation patterns generated from \( k \) screwdrivers, \( k(k-1)/2 \) binary SVMs are computed and group identity of a pattern is determined by voting of the decision rules (81). This multicategory approach to SVMs is called the one-vs.-one method and is what we use in this study. The R package kernlab was used to perform all SVM computations (82).

**Partial Least Squares Discriminant Analysis**

Multivariate linear regression is a method to find linear relationships between dependent “response” variables, and a data matrix of “predictor” variables \( \mathbf{X} \) by exploiting the covariance between these entities. If the response variables are coded as class labels, the method can be used for supervised classification (i.e. discriminant analysis) (83). Partial least squares (PLS) is a method to determine the relationship between predictor and response variables when there are many more predictors than samples and/or when many of the predictors are correlated. The method determines the PLS “latent vectors” (LVs) analogous to PCs from PCA. Also, just as in PCA, the first two or three LVs can be use to make an approximate graphical representation of the data. In this study we used the “softmax” function to map output of the PLS procedure to class assignments (i.e. striation pattern-screwdriver identifications) (84). The R packages pls and caret were used to perform PLS-DA computations (80, 84). For more of the computational details of PLS see references (70, 84).

**Error Rate Analysis—Hold-One-Out Cross Validation**

An error for the computational learning methods used in this study is defined as a misidentification of a striation pattern. There are many methods that can be used to assess error rates for pattern comparisons. All methods produce estimates of error rates based on samples. The estimation is of “global” error rate, or how often an algorithm would make a misidentification on a population of striation patterns it was not trained on. The simplest method to empirically estimate the error rate is resubstitution (85). This is the application of the computed classification rules to the set of data used to derive them. The percentage of
misclassifications via the resubstitution method is called the apparent error rate and is simply the empirical risk, $R_{\text{emp}}$. This is a biased estimate and tends to be overly optimistic and should be corrected. One option to improve the estimate of error rate is to use hold-one-out cross validation (HOO-CV) (77). This method computes the decision rules using all but one of the tool mark patterns in the data set. The hold-one-out procedure is repeated for each tool mark pattern in the data set and the results are averaged to compute an estimated error rate (86).

**Error Rate Analysis—Bootstrapping**

Another improved error rate estimate is the refined bootstrap (86, 87). First $B$ sets of bootstrap data, $X^*$ are generated by randomly selecting (with replacement) $n$ striation pattern feature vectors from the original data set $X$. Note that each bootstrap data set contains the same number of elements (striation pattern feature vectors) as the original data set, thus some patterns may be repeated. The decision rules, $g^*$ are recomputed for each bootstrap sample and an average error rate is computed using them on the original data as well as the bootstrapped data used to compute them (86). An alternative to the refined bootstrap is the .632 bootstrap error rate estimate (87). The prediction error is more likely to be larger for test patterns not contained in a given bootstrap sample. Thus in order to give a more conservative error rate estimate the .632 bootstrap focuses on these larger errors. All cross validation and bootstrap error rate estimates were computed with either native functions in the kernlab and pls packages or with the ported R package, bootstrap (87).

**Error Rate Analysis—Conformal Prediction and Confidence Regions**

Solomonoff’s and Kolmogorov’s algorithmic theory of randomness is a mathematically sophisticated way to gauge the amount of true information in a string of symbols (88). Recently, a method which gives confidence levels to identification of unknown patterns and control over error rates has arisen from the study of algorithmic randomness (89). This method, called conformal prediction can be applied to any statistical pattern comparison algorithm and holds a great deal of potential if applied to tool mark analysis. Prediction regions (confidence intervals) produced by conformal prediction can give a judge or jury an easy to understand measure of reliability for tool mark pattern identification because the method yields confidence on a scale of 0%-100%.
The way the method works is actually very simple (89, 90). Given a training set of striation patterns with known identities (called a bag) and at least one striation pattern of unknown identity, an estimate of randomness is computed for the bags containing the unknown striation pattern with all possible labels for its identity. The only assumption is that the striation patterns of the training set are drawn independently from the same, but unknown probability distribution.

Randomness of the bag is tested in a way analogous to what is done in traditional hypothesis testing (89, 91, 92). The null hypothesis is that unknown striation pattern \( x \) with assigned identity label \( y \) [i.e. the pair \( (x, y) \)] belongs in the bag and does not significantly decrease the bag’s randomness. The alternative hypothesis is that the pair \( (x, y) \) does not belong in the bag and thus \( y \) must be a different label than the one assigned. \( P \)-values are computed for randomness estimates. Thus conformal prediction regions for tool mark pattern identities can be thought of as generalizations of confidence intervals known from textbook hypothesis testing. Traditionally confidence intervals are computed for population parameters (e.g. a sample average) to give an indication of the regions where their true values may fall. Technically, the Neyman-Pearson interpretation of \( (1-\alpha) \times 100\% \) confidence interval for an estimated population parameter (here, striation pattern identities) constructed from a random sample of a given sample size, will contain the true population parameter \( (1-\alpha) \times 100\% \) of the time (90-92, 98). The value \( \alpha \) is called the level of significance and is the probability that any given confidence interval constructed from a random sample will not contain the true population parameter.

Note that the null hypothesis can be accepted for multiple label prediction regions of the striation pattern’s identity. In such cases the identity assignment (i.e. the prediction region) at the \( (1-\alpha) \times 100\% \) confidence level is ambiguous. While multi-label output is not wholly uninformative, hopefully the prediction region will contain only one label with a \( p \)-value less than or equal to 0.05. This means that the conformal prediction algorithm has produced a prediction region with only one label and a confidence level of at least 95%.

\( P \)-values for striation pattern test identities (I.D.s) are found by computing nonconformity scores. The nonconformity score, \( \alpha_i \) for the \( i \)th striation pattern using one-vs.-one multiclass SVMs was computed as
\[ \alpha_i = \frac{1}{k-1} \sum_{j=1}^{k(k-1)/2} \lambda_{i,j} \]

where \( \lambda_{i,j} \) is a matrix element of an \( n \) row by \( k(k-1)/2 \) column matrix of Lagrange multipliers (89). The formula just sums all the columns in this matrix and weights the resulting \( n \)-dimensional vector by \( 1/(k-1) \). The Lagrange multipliers were all computed by the binary SVM function of kernlab (82). Since a suitable nonconformity measure has not yet been developed for PLS, we did not use this method with CPT. Instead 3-nearest neighbor classification (3-NN), which has been shown to perform well in a number of machine learning tasks was used (10, 89, 93, 94). A 3-NN nonconformity score for striation pattern \( i \) is computed by first finding the distance matrix between all striation patterns and selecting three closest to \( i \) with the same I.D. and the three closest to \( i \) with a different I.D. The actual score is then computed as ratio

\[ \alpha_i = \frac{\sum_{j=1}^{3} \text{dist}(i,j) \text{ same I.D.}}{\sum_{j=1}^{3} \text{dist}(i,j) \text{ different I.D.}} \]

A \( p \)-value for each possible labeling \( \text{tlab}_i \in \{1,2,...,k\} \) of a test striation pattern is computed as

\[ P_{\text{tlab}_i} = \frac{\# \left\{ j \in \{1,2,...,n\} : \alpha_j^{\text{tlab}_i} \geq \alpha_{\text{test-pattern}}^{\text{tlab}_i} \right\}}{n} \]

where \( \alpha_j^{\text{tlab}_i} \) is the nonconformity score of the \( j \)th pattern when the test pattern is labeled as screwdriver \( \text{tlab}_i \), \( \alpha_{\text{test-pattern}}^{\text{tlab}_i} \) is nonconformity score of the test pattern labeled as screwdriver \( \text{tlab}_i \), and \( P_{\text{tlab}_i} \) means the \( p \)-value of the test pattern labeled as screwdriver \( \text{tlab}_i \) (89). A set of \( k \) \( p \)-values is computed for each test pattern, one for each possible screwdriver identity. For a chosen significance level \( \epsilon \) (i.e. level of confidence \( 1-\epsilon \) ) a \( 1-\epsilon \) confidence region of labels is determined by selecting those labels of the test pattern with \( p \)-values \( \geq \epsilon \). Ideally the output confidence region contains only one label. If a multi-label confidence region is output, it counts as a correct I.D. if it contains the true label of the striation pattern, though obviously it is less informative.
Empty regions can also be output if the CPT algorithm cannot confidently identify the striation pattern. Empties automatically count as errors (89).

**Results and Discussion**

*Partial Least Squares Discriminant Analysis*

Figure 5 shows all 75 striation patterns in the basis of the first two latent vectors (LVs). The advantage of a 2D or 3D plot of the data is to give some insight into the overall data structure, which can be instructive for judges and juries. One must also bear in mind that a good deal of information may have to be discarded to make such plots. This is the case for Figure 5. The first two LVs only account for 33% of the data’s overall variance. Note however that a good deal of discrimination (large inter-group separation, small intra-group separation) is already apparent between the groups of striation patterns.

Examining Figure 5 further, one can see that two screwdriver #3 patterns (at close to coordinate [-4,-4.5]) are clearly separated from the other screwdriver #3 patterns. They appear as outliers in this 2D-PLS space and it is reasonable to assume that almost any statistical discrimination algorithm will mistakenly identify them as screwdriver #7 patterns. Their departure from the other screwdriver #3 patterns can be explained by examining the actual physical striation patterns which are labeled patterns 6 and 8 in Figure 6. The rightmost groove imparted by screwdriver #3 is highlighted in the figure. It spans from where the right edge of the screwdriver made contact with the clay, and across to the left about 1.2 mm (cf. Figure 6, pattern 3). Replicate patterns 6 and 8 are clearly aligned with replicate pattern number 7 in Figure 6 (note all the common lines vertically down the patterns). Note however how the remaining 0.7mm of the groove in question is missing in patterns 6 and 8, likely caused by striking the clay surface at a slight angle. This groove is apparent on all of the striation patterns generated by screwdriver #3 with the exception of pattern 6 and 8 (photographs available on request) and its absence in these patterns is the cause of their distance from the other screwdriver #3 patterns in Figure 5.
Such a phenomenon may be of concern to the reader because obviously a perpetrator of a crime will not take the time to make sure all the striation patterns they leave are made at the same angle! This however was not to be a problem for the discrimination methods used in this study. With the addition of a few more dimensions to the data, these patterns are easily classified as stemming from screwdriver #3 (i.e. the classification holds up to cross-validation and bootstrapping.). Figure 7 shows that by retaining 8 dimensions in PLS-space (72% variance retained), the HOO-CV error rate can be dramatically reduced. Thus 8D PLS-space has more than enough discrimination power to identify most of the striation patterns in this data set. Table 1 shows the HOO-CV, refined bootstrap and .632 bootstrap methods all estimate the global correct identification rate at 97% to 98%. A 100% correct classification rate was achieved on a randomly generated test set of consisting of 25% of the data. Scaling up to 38D PLS-space (95% variance retained) the estimated global correct identification rates are all nearly perfect. Note that this is still a relatively low dimensional model considering that the raw data set is 121D.

Principal Component Analysis with Support Vector Machines

Figure 8 shows the data in the basis of the first two principal components (PCs). There is a strong resemblance to the 2D-PLS plot (Figure 5), which is typical. The advantage of PCA however, is that it reduces the dimension of the data set in a completely unsupervised way. Thus, if natural clustering between groups of data exist in 2D or 3D PCA space there is strong evidence that the differences are real. PCA however requires that a separate classification method be used to numerically identify to which group a data point belongs. While PLS is a totally supervised method, it is its own classification method. The classification method we chose for PCA in this study is support vector machines (SVMs). An SVM is a supervised identification method, however the procedure assume very little about the distribution of the data. This is quite unlike most other classification methods in machine learning, the dropping of a distributional assumption is in the authors’ minds, a major advantage in forensic applications.

Figure 9 shows a plot of HOO-CV classification error rate with increasing dimension of PC-space. To make this figure, a test pattern is first held out, PCA followed by SVM is performed on the remaining data and then the held out pattern is projected and classified. This retains some noise in the test pattern to stress the method and thus yields a higher HOO-CV error rate estimate. It is also computationally more expensive.
First note that the error rates are lower than for PLS-DA. An HOO-CV error rate of ~3% is achieved with 4 dimensions for PCA (52% variance retained) as opposed to 8 dimensions for PLS. Table 2 shows the estimates of apparent and global correct identification rates using PCA-SVM. Note that for all of the global correct classification rate procedures used to construct this table, PCA is performed first on the entire data set and then hold-out or bootstrapping is applied (This is how it would be performed in practice once an identification model is fit). Thus some variance in the test observations is initially removed, which is likely why the 4D HOO-CV error is slightly lower in Table 2 compared to that seen in Figure 9.

For reference, Figure 10 shows the data in the basis of the first 3 PCs. Even though only 33% of the variance is retained with 3 PCs, the estimated correct classification rates are very good. By 4D, they are nearly perfect. For reference the 31D (95% variance retained) estimated global correct classification rates are shown. All dimensions shown in Table 2 performed well on a randomly chosen test set of 25% of the data.

Conformal Prediction with PCA-SVM

Table 3 gives the results for conformal prediction theory (CPT) utilizing PCA-SVM for classification in both the on-line and off-line modes at the 95% level of confidence. Columns two and three show CPT results when applied to the 75 real striation patterns when reduced to 4D (indicated as adequate for good discrimination by HOO-CV, cf. discussion above) and 121D (not reduced in dimension at all, i.e. the full data set). In the on-line mode, immediately after I.D. prediction regions are generated by the CPT algorithm, the true I.D. of the striation pattern being tested is fed into the computations and an I.D. prediction region is generated for the next striation pattern to be tested. This “on-line” process of continually updating the data set used to make I.D. predictions guarantees that the CPT algorithm yields erroneous prediction regions at a level of less than or equal to $\alpha \times 100\%$, up to small statistical fluctuation (89).

As is apparent from columns two and three of Table 3, application of CPT to the real striation patterns produced confidence regions containing the correct identity of the test pattern to within a reasonable statistical fluctuation of the theoretical error rate. An error rate of 6% was found for both 4D and 121D which is within statistical fluctuation of the expected 5% error rate.
In other words, incorrect 95% confidence intervals (empty or not containing the true I.D.) were produced only about 5% of the time (89, 94).

Note that prediction regions produced by CPT can be multi-label. Such multi-label results are not counted as incorrect if they contain the true I.D. of the striation pattern being tested. Obviously though, while not totally useless, correct but multi-label prediction regions are not maximally informative. Thus an important measure of the CPT algorithm’s performance is also the rate at which it produces regions with one unique I.D., and that I.D. is correct. Table 3 shows that on-line CPT produced unique and correct I.D.s of the striation patterns well above 90% of the time at the 95% level of confidence. Such results are very rigorous numerical identifications of striation patterns and are accompanied by a definitive level of confidence. They can also be stated easily in a courtroom situation.

The original developers of conformal prediction theory acknowledge that strict on-line classification is not always realistic in practice (89, 95). That is, one cannot expect that the true label for each test striation pattern can be produced immediately after classification occurs. In order to circumvent this problem there are three solutions. The first is to repeat the CPT identification sequence on each test pattern, building up to it by first classifying a fairly large sequence of randomly selected striation patterns with known labels. Essentially this is on-line prediction using many independent random samples from a large data set augmented with one unknown striation pattern as the last pattern to be tested. In this way, a valid confidence level is maintained for the label set produced for the unknown pattern. A drawback to this approach is that it is very computationally intensive. A second possibility is to use CPT in “slow teacher” mode where the true identities of unknown test patterns are eventually presented to the algorithm. This also produces a confidence level “as advertised” but larger statistical fluctuations in the error rate can occur for small data sets (89, 95).

The last possibility is what we pursue in this study; use the CPT algorithm in the “off-line” mode where the same sequence of striation patterns with known identities is used to predict identities of a set of unknown patterns. CPT in the off-line mode can no longer theoretically guarantee that the (1-\(\alpha\))\times100% confidence regions produced are erroneous \(\alpha\times100\%\) of the time. However, despite this fact it has been shown empirically that off-line mode CPT confidence region error rates actually do remain close to \(\alpha\times100\%\) (96). Table 3 shows the off-line
performance of CPT with PCA-SVM classification. Results are indeed commensurate with CPT run in the on-line mode, \textit{i.e.} only about 5\% of the striation patterns were incorrectly identified. Also correct, unique label prediction regions were produced approximately 90\% of the time or higher. Computations however were much faster in off-line mode.

\textit{Conformal Prediction with k-Nearest Neighbors}

While a nonconformity measure for PLS-DA which can be used with CPT is imminent, none has yet been released in the literature. Thus we employed an alternative classification method which has been extensively shown in the literature to perform well when combined with CPT, k-nearest neighbors (89, 94, 95). Table 4 shows the 95\% confidence CPT results using three-nearest neighbors (3-NN) for classification in both the on-line and off-line modes. Again, the error rates are all on or close to 5\%. Note that a 0\% error rate for 4D 3-NN CPT classification on the real striation patterns is a bit optimistic (cf. Table 4 column two). The uniqueness and efficiency of the prediction regions produced by the algorithm run in both modes are consistently \textasciitilde90\% or more as was the case for PCA-SVM. Again, this means that most of the time CPT produced a 95\% confidence interval with only one I.D. for the test pattern and that I.D. was correct.

\textbf{Conclusion}

The intention of this study was to show that it is possible to associate relatively flat ideal striation patterns with a “scraping” type tool using objective numerical measures of similarity, error rate and confidence. The criteria of selection for the statistical methods used in this study were that they have a peer-reviewed track record, a high rate of success in the fields in which they have been applied and that they are relatively free of many of the underlying assumptions that typically underlie comparison methods in statistics (\textit{e.g.} assumed parametric distributions, dichotomous decisions, etc.). Partial Least Squares Discriminant Analysis was able to differentiate striation patterns made by screwdrivers at or higher than a 97\% correct classification rate (\textless3\% error rate) with 8D feature vectors. Principal component analysis with support vector machines showed comparable high performance with only 4D feature vectors. Conformal prediction theory, which has grown out of Solomonoff’s and Kolmogorov’s algorithmic theory of randomness, can in fact be used to control identification error rates. For the
first time in forensic science we used CPT to produce tool mark identifications (conformal prediction regions) at the 95% level of confidence. As is advertised by the theory, error rates are always 5% to within small statistical fluctuation. Uniquely labeled and correct conformal prediction regions were produced at or greater than 90% of the time using CPT in both on-line and off-line modes.

Our sample size for this pilot study was necessarily small, however the results strongly indicate the feasibility of using machine learning techniques to identify tool marks. Studies are currently underway to drastically increase the sample size, vary the angle at which the tool strikes the impression media, use incomplete striation patterns (more akin to what a tool mark examiner encounters in practice) and most importantly use 3D metrological instrumentation and software to carry out the same set of tasks. Also the techniques presented in this study could be extended to striation patterns found on firearms evidence. Lastly, the authors would be happy to share the data set and the Mathematica and R software written by them for this study, upon request.

Acknowledgements
We are grateful to Professor Patrick Buzzini for enlightening discussions of subjectively based Bayesian methods. We also thank Ms. Loretta Kuo, Officer Patrick Mac Laughlin, Ms. Helen Chan and Mr. Manny Chapparo, for many valuable discussions and for reading our manuscript.

Additional information, supplementary material and reprint requests:
Nicholas D. K. Petraco, Ph.D.
Department of Science
John Jay College of Criminal Justice
899 10th Ave.
New York, NY, 10019

References


Figure Legends

FIG. 1–Nine quarter inch standard slotted screwdrivers used in this study.

FIG. 2–Process to generate striation pattern standards. As one can see from the figure, screwdriver number 1 was used to make nine striation pattern standards.
FIG. 3–Graphical representations of feature vectors for two striation patterns. They are replicate pattern 3 for screwdriver #1 and replicate pattern 8 for screwdriver #3.

FIG. 4–Graphical representation of support vector machine output for the separation of two data sets.

FIG. 5–All striation patterns projected into the space of the first two PLS latent vectors (33% variance retained). Each point represents a striation pattern. The boldfaced numbers to the left of the points tell which screwdriver generated the pattern.

FIG. 6–Three screwdriver #3 patterns with the rightmost groove highlighted. The first two, pattern 6 and pattern 8, appear as outliers in this 2D-PLS space.

FIG. 7–Hold-one-out cross validation (HOO-CV) error rates vs. PLS dimension (i.e. number of LVs retained). As expected classification error generally decreases as dimension of the space is increased.

FIG. 8–All striation patterns projected into the space of the first two principal components (33% variance retained). Each point represents a striation pattern. The boldfaced numbers to the left of the points tell which screwdriver generated the pattern.

FIG. 9–Hold-one-out cross validation (HOO-CV) error rates vs. PCA dimension (i.e. number of PCs retained). As expected classification error generally decreases as dimension of the space is increased. However the decrease is faster than for PLS.

FIG. 10–All striation patterns projected into the space of the first three principal components (44% variance retained). Each point represents a striation pattern. The boldfaced numbers to the left of the points tell which screwdriver generated the pattern. Cf. Table 2 for correct identification rates of striation patterns in this space.
TABLE 1-PLS correct classification rate estimates.

<table>
<thead>
<tr>
<th>Method of Estimation</th>
<th>Data Dimension(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2D</td>
</tr>
<tr>
<td>Apparent</td>
<td>48%</td>
</tr>
<tr>
<td>Hold-One-Out C.V.</td>
<td>36%</td>
</tr>
<tr>
<td>Refined Bootstrap</td>
<td>78%</td>
</tr>
<tr>
<td>.632 Bootstrap</td>
<td>37%</td>
</tr>
<tr>
<td>Random Test Set(^c)</td>
<td>47%</td>
</tr>
</tbody>
</table>

\(^a\)Data dimension is the number of latent vectors retained to represent the data.

\(^b\)The method’s estimated error rate = (100-correct classification rate estimate)%.

\(^c\)Preprocessing transformation computed using corresponding random training set.
TABLE 2-PCA-SVM correct classification rate estimates.

<table>
<thead>
<tr>
<th>Method of Estimation</th>
<th>Data Dimension&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2D</td>
</tr>
<tr>
<td>Apparent</td>
<td>87%</td>
</tr>
<tr>
<td>Hold-One-Out C.V.</td>
<td>81%</td>
</tr>
<tr>
<td>Refined Bootstrap</td>
<td>88%</td>
</tr>
<tr>
<td>.632 Bootstrap</td>
<td>81%</td>
</tr>
<tr>
<td>Random Test Set&lt;sup&gt;c&lt;/sup&gt;</td>
<td>76%</td>
</tr>
</tbody>
</table>

<sup>a</sup>Data dimension is the number of principal components retained to represent the data.

<sup>b</sup>The method’s estimated error rate = (100-correct classification rate estimate)%.

<sup>c</sup>Preprocessing transformation computed using corresponding random training set.
TABLE 3-95% Confidence prediction results using PCA-SVM classification.

<table>
<thead>
<tr>
<th></th>
<th>On-Line Mode</th>
<th>Off-Line Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Dimension</td>
<td>4D&lt;sup&gt;b,c&lt;/sup&gt;</td>
<td>121D&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>% Error</td>
<td>6%</td>
<td>0%</td>
</tr>
<tr>
<td>% Unique and Correct I.D. Produced</td>
<td>94%</td>
<td>100%</td>
</tr>
<tr>
<td>% Efficiency</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>% Empty Intervals</td>
<td>6%</td>
<td>0%</td>
</tr>
</tbody>
</table>

<sup>a</sup>Data dimension is the number of PCs retained to represent the data. 121D is the full dimensionality of the data set.

<sup>b</sup>Estimated minimal dimension to adequately represent the striation pattern data.

<sup>c</sup>Results based on a random ~75%:25% randomized split of the real striation pattern data into training:test sets (58 training patterns:17 test patterns).
TABLE 4-95% Confidence prediction results using 3-nearest-neighbor classification.

<table>
<thead>
<tr>
<th>Data Dimension</th>
<th>On-Line Mode</th>
<th>Off-Line Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4D&lt;sup&gt;a,c&lt;/sup&gt;</td>
<td>121D&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>% Error</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>% Unique and Correct I.D. Produced</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>% Efficiency</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>% Empty Intervals</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

<sup>a</sup>Data dimension is the number of PCs retained to represent the data. 121D is the full dimensionality of the data set.

<sup>b</sup>Estimated minimal dimension to adequately represent the striation pattern data.

<sup>c</sup>Results based on a random ~75%:25% randomized split of the real striation pattern data into training/test sets (58 training patterns:17 test patterns).