Studies of Biometric Fusion

Appendix E

Modeling Biometric Score Distributions

Brad Ulery,[1] William Fellner,[1] Peter Hallinan,[2] Austin Hicklin,[1] Craig Watson[3]

[1] Mitretek Systems

[2] Independent consultant to Mitretek Systems

[3] National Institute of Standards and Technology

20 July 2006

Abstract

This report discusses lessons learned in the course of developing models of score distributions from large samples of data using kernel density estimation techniques. The methods detailed here were used in the implementation of the Product of Likelihood Ratios and FAR-based fusion techniques.

Contents

1 Introduction

Accurate modeling of score distributions is the key to optimizing score-level fusion [Neyman-33; Griffin-05; Scott-05]. This paper discusses procedures for modeling score distributions that were explored in the course of this study. It compares the effectiveness of some alternatives and provides examples of how they can fail. It summarizes how the modeling steps were combined to implement several of the techniques evaluated in Part IV, including the Product of Likelihood Ratios that was the basis for quantitative results in Parts VIII and IX, and discusses limitations of this implementation.

The remainder of Section 1 discusses relevant characteristics of the sample data (see also Part III) and the visual tools used to see those characteristics. Section 2 discusses variations on the basic modeling procedures and the sensitivity of the final fused results to specific types of imprecision in the models. Section 3 defines the standard modeling procedure used throughout this series, and discusses known limitations, robustness and validity. Section 4 defines a simple linear fusion technique that is also based on sample score distributions and was evaluated (Part IV) and applied (Parts VIII and IX) in this series.

1.1 Examples of score distributions

When fusing scores from multiple sources, there are two score distributions of interest: the mate distribution (genuines) and non-mate distribution (impostors). These distributions exist in an M-dimensional space, as each comparison of two subjects results in M individual biometric scores, one for each matcher, instance, sample or mode to be fused.

It is common practice to assume that these distributions are parametric (conform to some known, parameterized family of distributions), either to facilitate the use of theoretical results or simply for lack of sufficient data to construct a more precise model. It is clear from large sample distributions that this is rarely a valid assumption. However it is not evident *a priori* how sensitive fusion results are to such simplifying assumptions.

Figure 1 shows histograms of the mate and non-mate scores for one matcher. Notice that both distributions are heavy-tailed in the primary region of score overlap, and that both distributions have a "spike" at Score = 0 (i.e., a relatively high density at this singular score value).

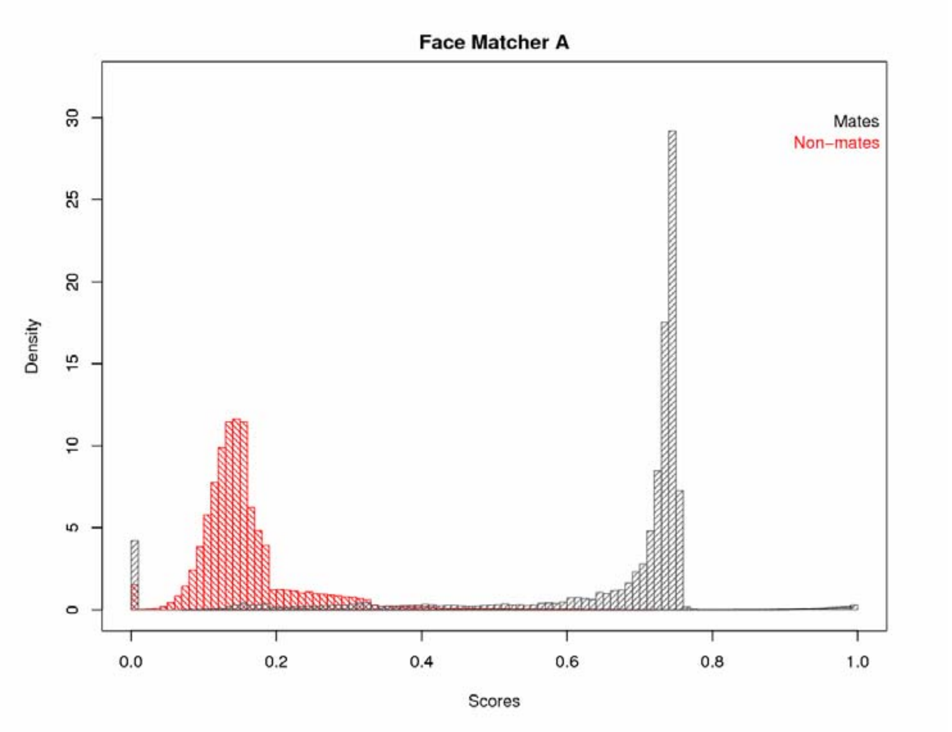

Figure 1: Histograms of non-mate scores (red) and mate scores (black); smoothed density estimates of non-mates (blue) and mates (green); and the ratio of the density estimates (unitless, scaled to chart).

Figure 2 shows a selection of 2-score (bivariate) scatterplots from the NBDF06 dataset to illustrate the wide range of joint score distributions with which a fusion algorithm must cope. Generally, any simplifying assumptions made by an algorithm about the nature of the score distributions will lead to suboptimal fusion performance on at least some datasets.

Notice, for instance, that the C-I scores are inherently separated well by a linear boundary, but those of A-C are not so well separated by that technique. These data reveal numerous patterns that violate common simplifying assumptions: dissimilarity of distributions across multiple matchers (e.g., B-C), discontinuous distributions (e.g., Q nonmates – low scores), censored distributions (e.g., Q mates – high scores), and spikes (e.g., A mates – low scores). Another important characteristic of this data (indicated, but not fully evident in these charts) is the great differences in matcher strength (e.g., C-H).

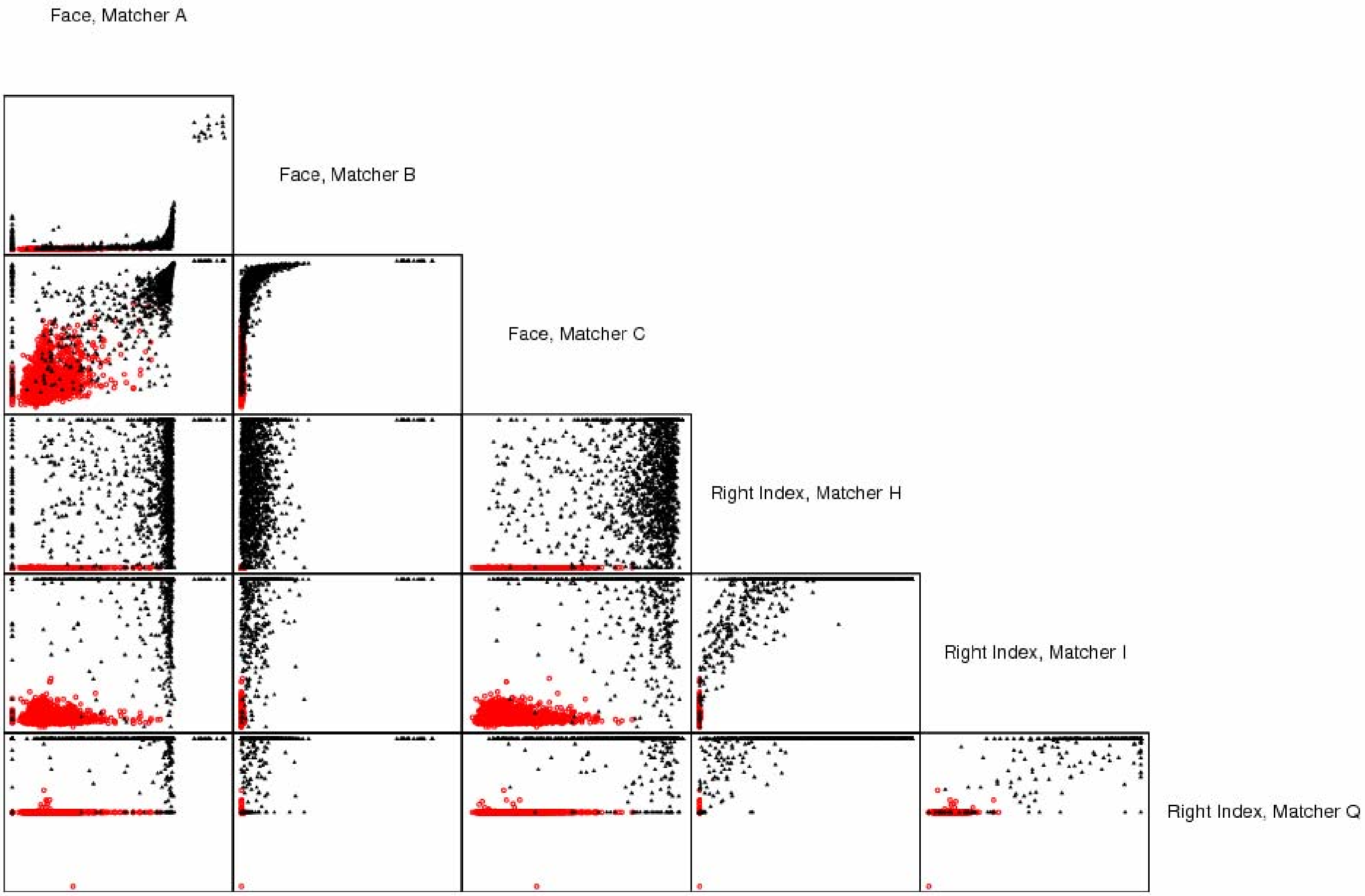

Figure 2: Joint score distributions for each possible pair of six raw score sets. Non-mates are shown in red; mates in black. Each axis covers the full score range, which differs widely by matcher; scales were omitted to focus attention on the distributions. Each plot is constructed from the same set of 3000 randomly selected, mated subjects and 3000 non-mated subjects.

As discussed in Part III, the NBDF06 dataset includes face and fingerprint scores for 64,867 pairs of mated subjects (genuines) and 122,000 non-mates (impostors). The large size of this dataset makes good density estimates possible.

2 Density Estimation

This section addresses the problem of estimating probability densities from random samples of biometric score data. As we have seen, these score distributions are not well described by parametric models (i.e., those having a fixed form, such as Gaussian or exponential).

The histograms shown in Figure 1 belong to the class of non-parametric solutions. Histograms reveal the shape of the distributions effectively, but the result is not smooth, and depends on the width and end points of the bins.

Kernel density estimation is a general method for obtaining smooth, non-parametric estimators from sample data. In this method, a kernel function is centered at each data point, and a probability density estimate for the entire sample is obtained by summing over these functions. Selecting a smooth kernel function produces an estimator that is also smooth. The Parzen window method, which uses kernel of fixed width, has been proposed for modeling biometric score distributions (e.g., [Jain-00], [Dass-05], [Jain-05]). Parzen windows address several of the modeling problems identified above.

In this study, kernel density estimation was used as a basis for developing estimators, but several further refinements were made. This paper discusses those refinements in detail.

The following steps outline the general modeling procedure. They are explained in detail in the following subsections. Each univariate score distribution is first modeled separately, then combined as a product to form the joint estimate:

1. Manually identify "spikes" in the data
2. Kernel fit the remaining data (excluding spikes)
3. Adjust the initial kernel fit:
    a. Adjust limits of bounded distributions using flat linear extension (Gaussian kernels do not fit well)
    b. Log-linearly extend non-mate right tail to maximum score
4. Compute FAR = non-mate right tail integral (probability estimator)
5. Model the joint density of each distribution (mates and nonmates) as a product of independent distributions

2.1 Density Estimation: Variable Bandwidth Kernels

In kernel density estimation, bandwidth selection is critical. If the kernel is too large, oversmoothing and bias result. If it is too small, the model preserves patterns of sample noise. By varying the bandwidth as a function of the matcher score, it is possible to model regions of high sample data density precisely with a narrow bandwidth, and regions of low sample density smoothly with a high bandwidth.

High performance at very low FAR requires modeling the right tail of the non-mate distribution accurately, even though training data in the tail is sparse. The standard Parzen method of kernel density estimation – using fixed bandwidth kernels - does not work well in this case; in particular, standard Parzen fits, using default bandwidth parameters, produce far too much variance in the tails. For typical heavy-tailed, non-mate distributions, it was found that increasing the bandwidth produced unacceptable bias (flattening of the curve) before smoothing the tails.

Variable bandwidth kernels provide a means of addressing this problem. The following example (using data from FpVTE MST) shows results obtained using the KernSec function of R (available in the GenKern library).[1]

Figure 3 shows a histogram of the score distributions. The red curve is a standard Parzen fit using a default bandwidth (approx 16.5). Note that some variance (lack of smoothness) is evident for scores near 1000. The green curve represents an attempt at greater smoothing by increasing the bandwidth to 50. Notice how this results in increased bias, especially in the left tail. The black curve was obtained using a variable bandwidth (described below). Notice how this results in a tight fit (similar to the red line), but with less variance.

[1] A spike at score=0 (12.5% of the data) was excluded from this histogram, per step 2.

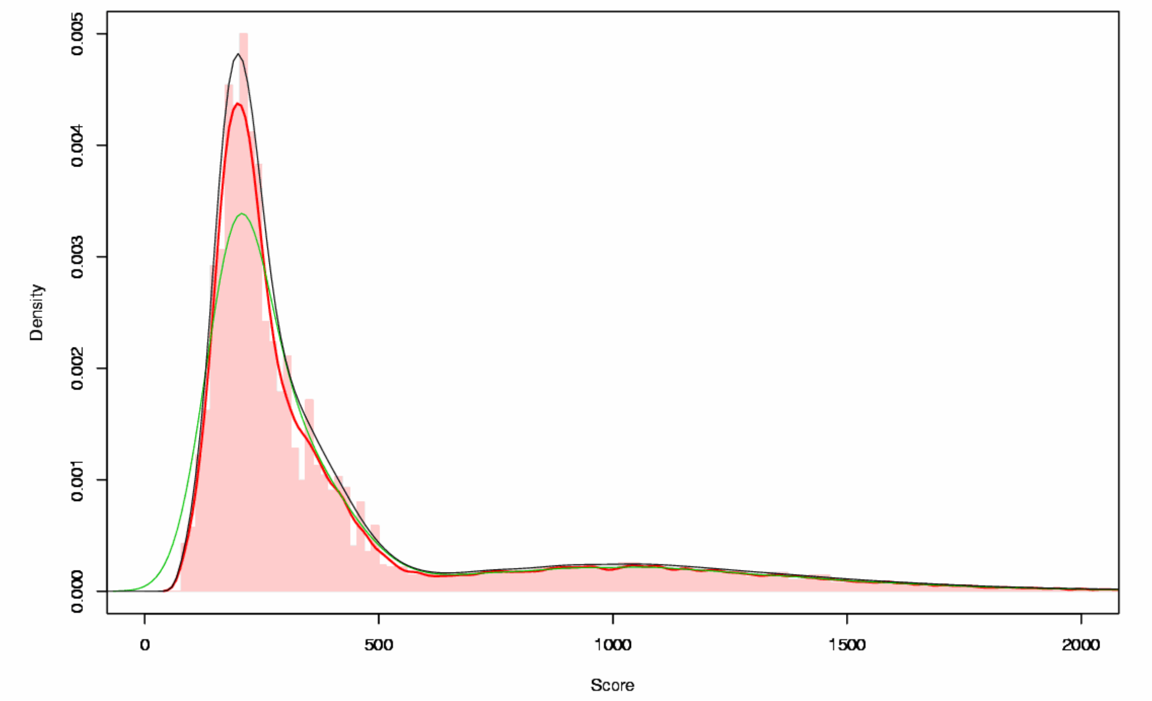

Figure 3. Histogram and several kernel density estimates

Figure 4 shows a detail of Figure 3, revealing unacceptable variance in the right tail of both the default (red) and smoother (green) Parzen fits.

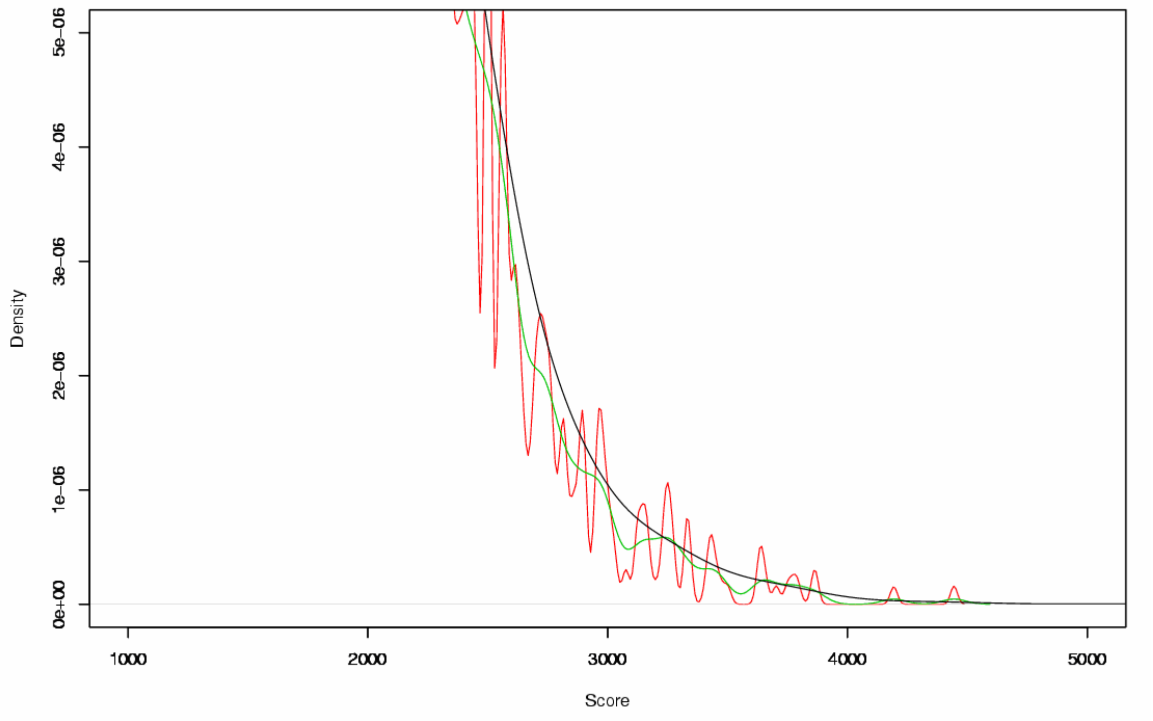

Figure 4. Kernel density estimates (Detail: y-scale)

Figure 5 shows a better way to view the variance: density on a logarithmic scale. Note on this chart that the variable bandwidth kernel method (black) is very well-behaved, with minimal bias relative to the default (red) and very smooth (low variance) over the entire range. Note that starting at 4800, where the variable kernel method begins to suffer from excessive variance, the fit is extended to the maximum score using a log-linear taper (see section 2.2).

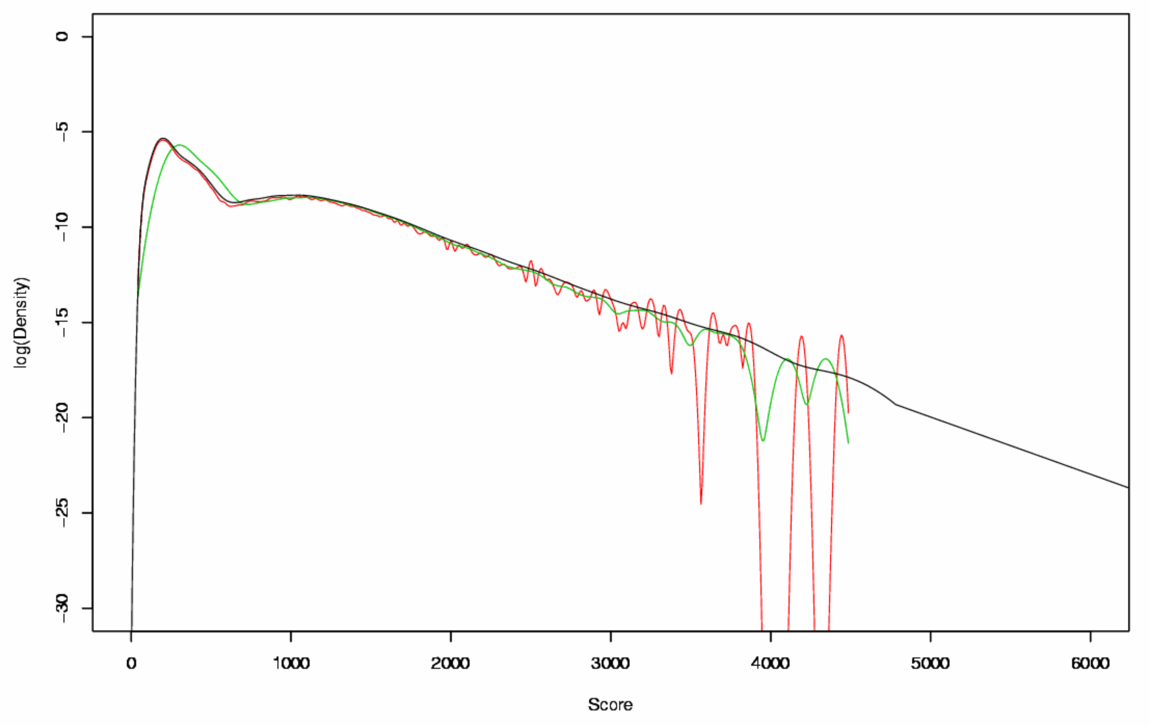

Figure 5. Kernel density estimates (logarithmic y-scale). Black is variable bandwidth (up to score = 4800, after which the fit is extended to the maximum score using a log-linear extension). Red and green are fixed bandwidth.

Figure 6 contrasts the performance resulting from variable bandwidth and fixed bandwidth kernels. In summary, although the fixed bandwidth Parzen technique is much simpler to use, it performs too poorly in general for low FAR biometric systems. Variable bandwidth kernels provide better modeling control, and facilitated tapering (see next section) to improve the tail fit.

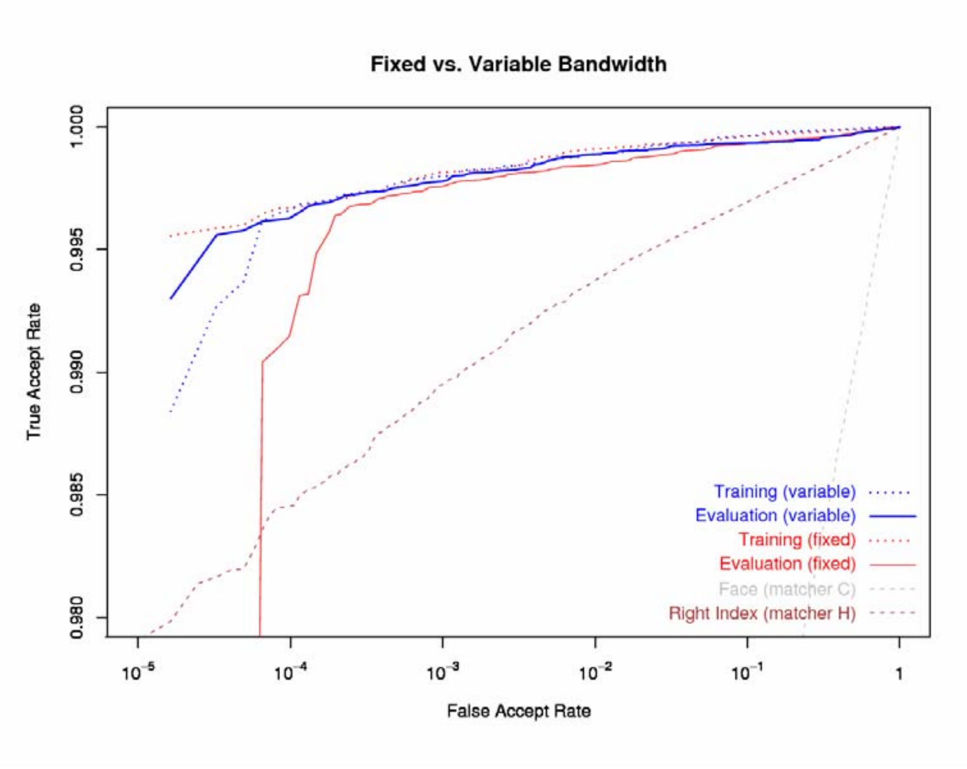

Figure 6: ROC curves for training and evaluation datasets for variable bandwidth and fixed bandwidth density estimation techniques. Fixed bandwidth kernel estimation overfits at low FAR.

2.2 Density Estimation: Tapered Tails

Properly modeling the right tail of the non-mate density is critical when maximizing TAR at very low FAR. We tried two methods: linear and log-linear tapering. The log-linear taper was implemented as a log-linear descent from FAR=5*10^{-5} to max(mates). The linear taper is a linear descent over the same interval. Figure 5 shows an example of a log-linear taper beginning at Score = 4800.

Figure 7 and Figure 8 below show the differences in results between the two methods. Log-linear tapering is clearly more effective. In fact, with linear tapering, we found that the Product of Likelihood Ratios deteriorated below FAR=10^{-3} and that Product of FARs deteriorated below FAR=10^{-2}. We also observed that a rough approximation to log-linear works quite well, e.g., a small number of points with linear interpolation.

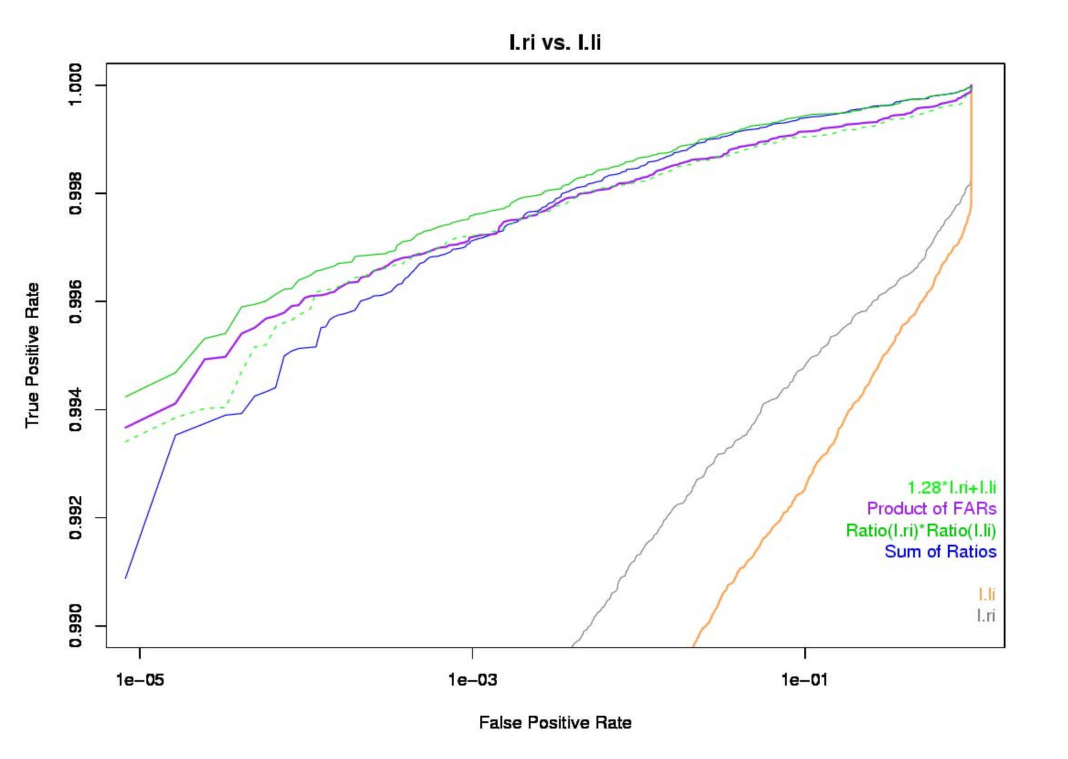

Figure 7: Performance using log-linear tapering.

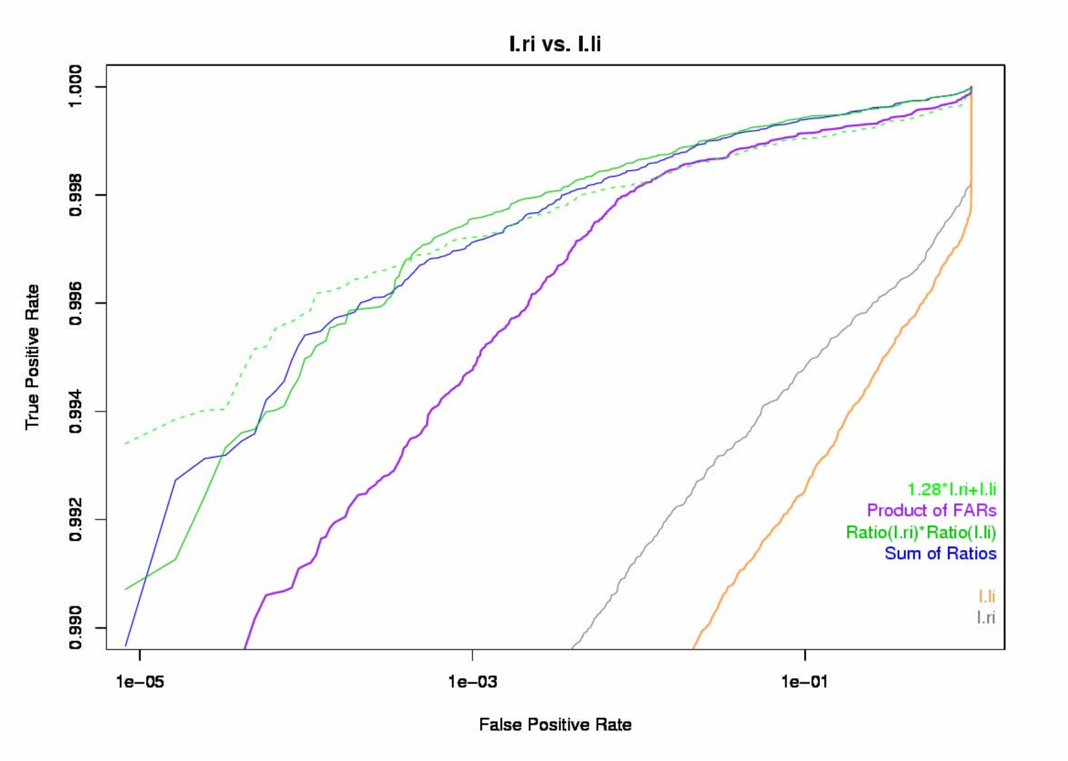

Figure 8: Performance using linear tapering.

2.3 Density Estimation: Visualizing the Fit

A standard set of four charts (see Figure 9) was used to visualize the density models during the fitting process. Such charts were very helpful in developing an understanding of the score distributions, and of bias and variance in the estimators.

Each chart presents the same basic set of information with varying range and scale[2]. The top two charts present the data on a linear y-scale; the bottom charts use a logarithmic y-scale. The two charts on the left limit the x-axis to the region of non-mate scores (99.9% of the non-mate sample data); the charts on the right show the entire range of scores.[3]

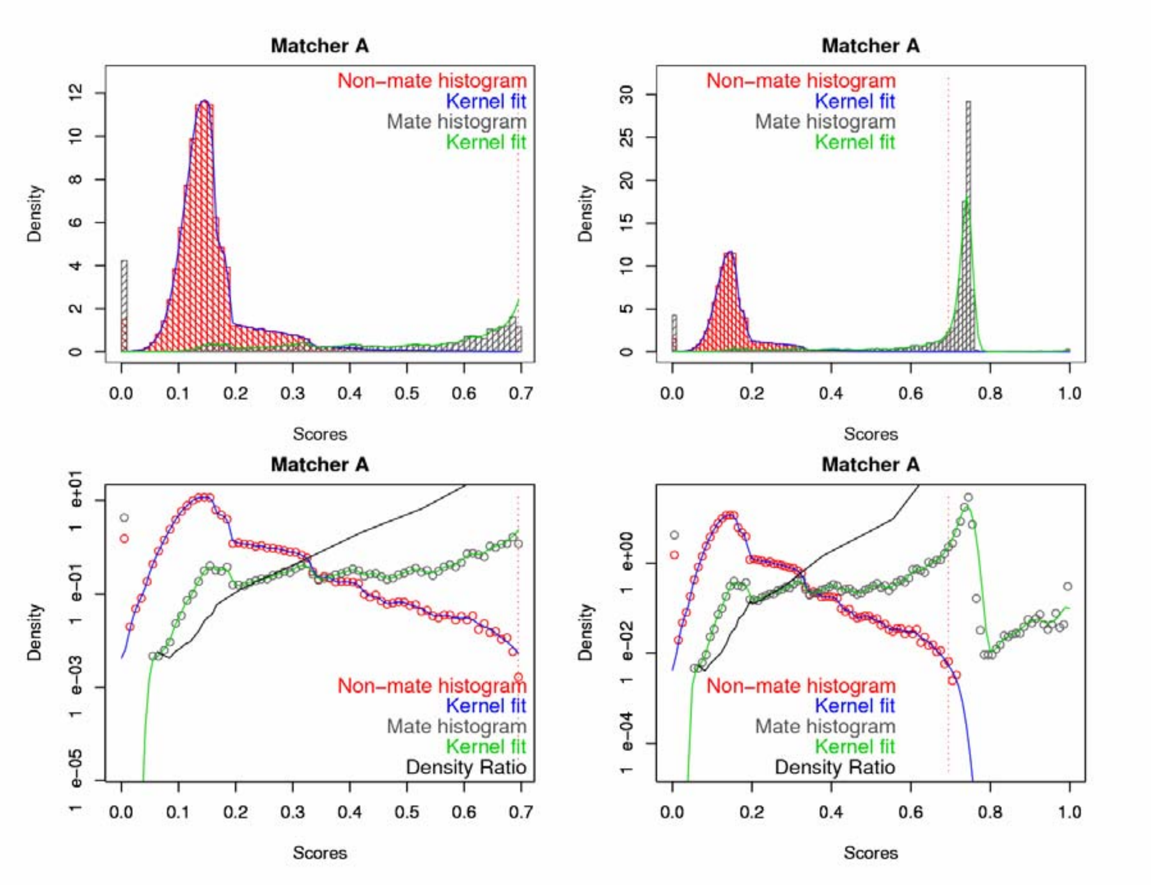

Figure 9: A "4-chart" showing mate and non-mate score distributions for face matcher A, showing density estimates and ratio of density estimates. Goodness of fit is reviewed visually using the histograms as a reference.

[2] The red and black circles in the lower charts represent the height of histogram columns. This stylistic difference is the unfortunate result of technical difficulties with the plotting software.

[3] Some details of the actual models are not fully represented in these charts, e.g., fitting to the spikes and special handling of the extremes of the distributions.

2.4 Density Estimation: Gaussians

This section demonstrates the effects of modeling distributions as normal curves. Although clearly the distributions are not normal, it is not obvious *a priori* how much effect this simplification has on the ROC curves. Several researchers have assumed normal distributions, if only because they lacked sufficient data for a more accurate model.

An example is shown based on fusing Matcher H right index finger scores with Matcher A face scores. Figure 10 and Figure 11 show the normal fits to the score distributions. Figure 12 and Figure 13 compare the results of using variable bandwidth density estimation and normal densities respectively. Product of Likelihood Ratios and Product of FARs both perform better at FAR = 10^{-4} using accurate density estimates. Note that in both methods of density estimation, spikes were handled discretely.

Comparison of a representative set of 12 scatterplots and ROCs for Kernel vs. Normal fits of two biometrics revealed that normal-based ROCs tend to show slightly lower slopes than Kernel ROCs, i.e., the TAR reduces over the entire range of FAR. Occasionally, the normal assumption causes a severe drop in TAR at very low FAR.

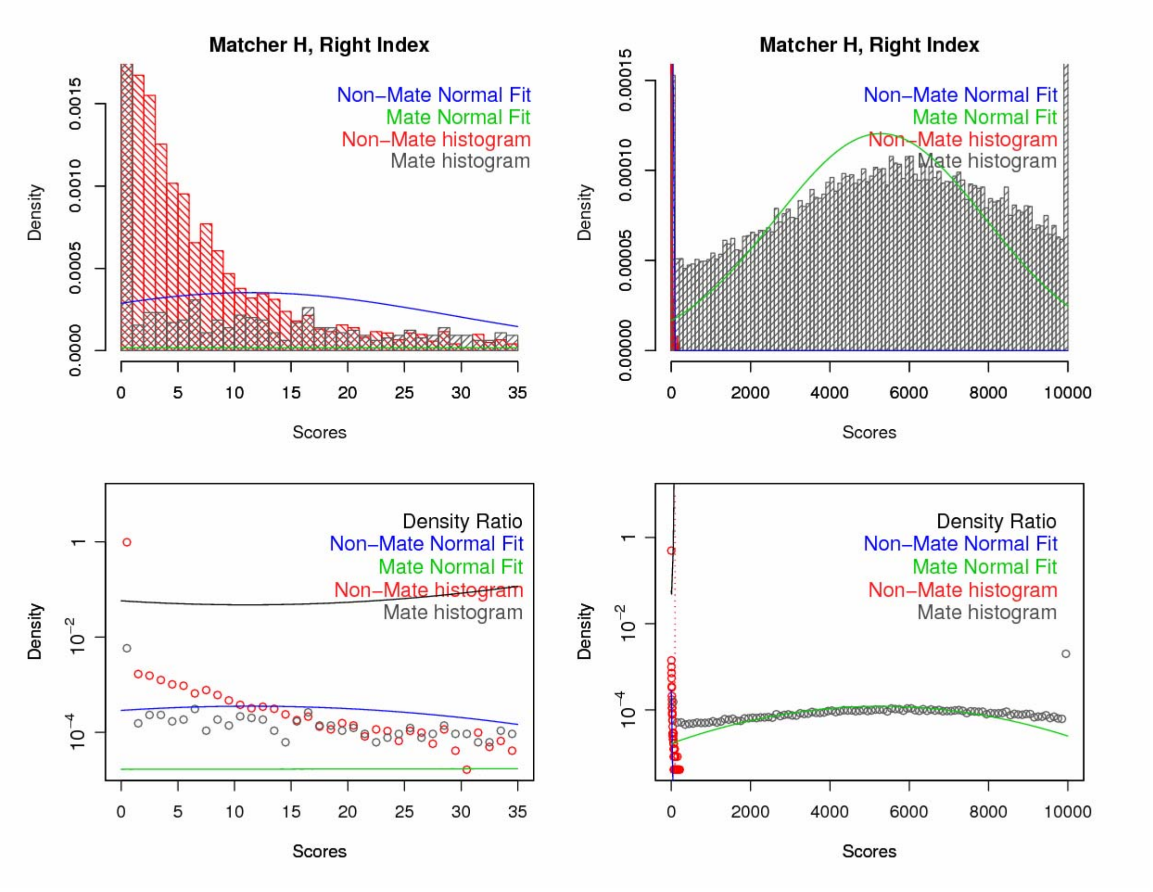

Figure 10: Density ratio modeling based on normal fits (fingerprint matcher H). Refer to Section 2.3 for information on how to read these charts.

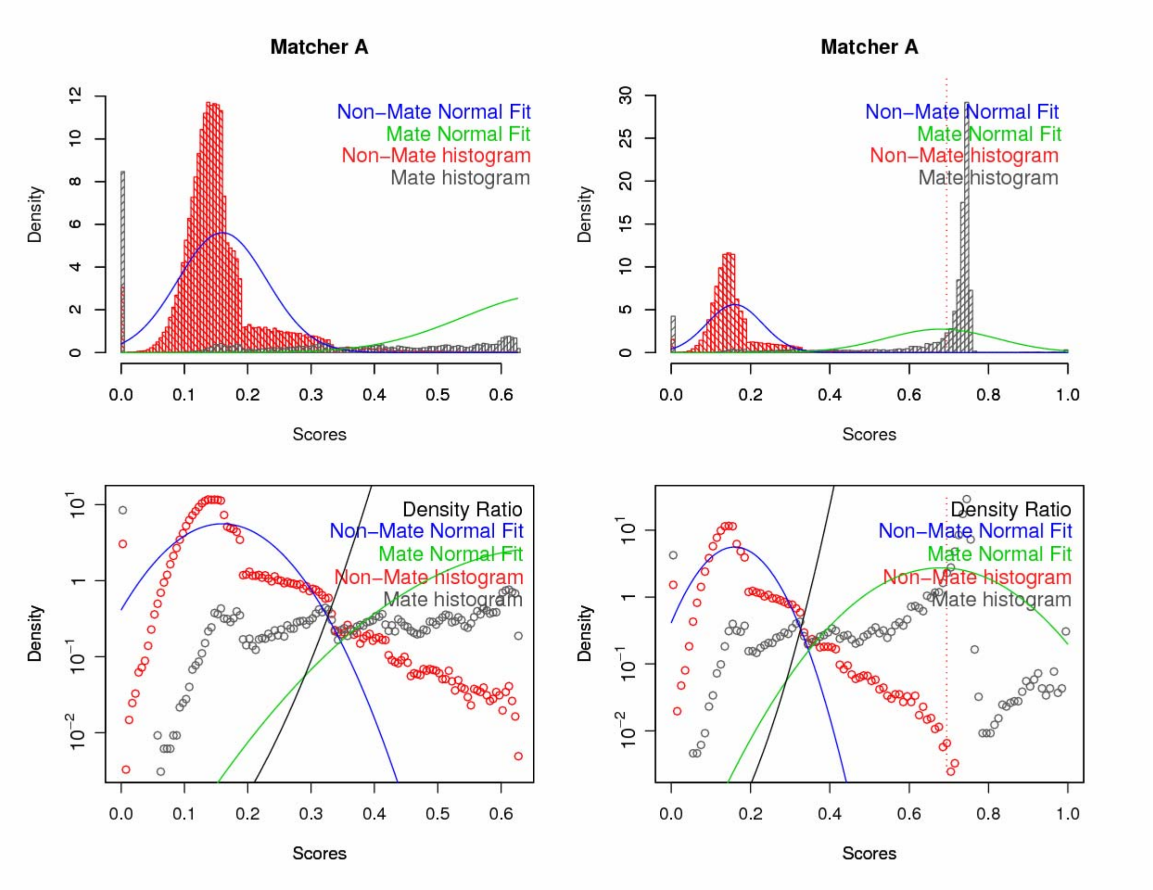

Figure 11: A: Density ratio modeling based on normal fits (face matcher A). Refer to Section 2.3 for information on how to read these charts.

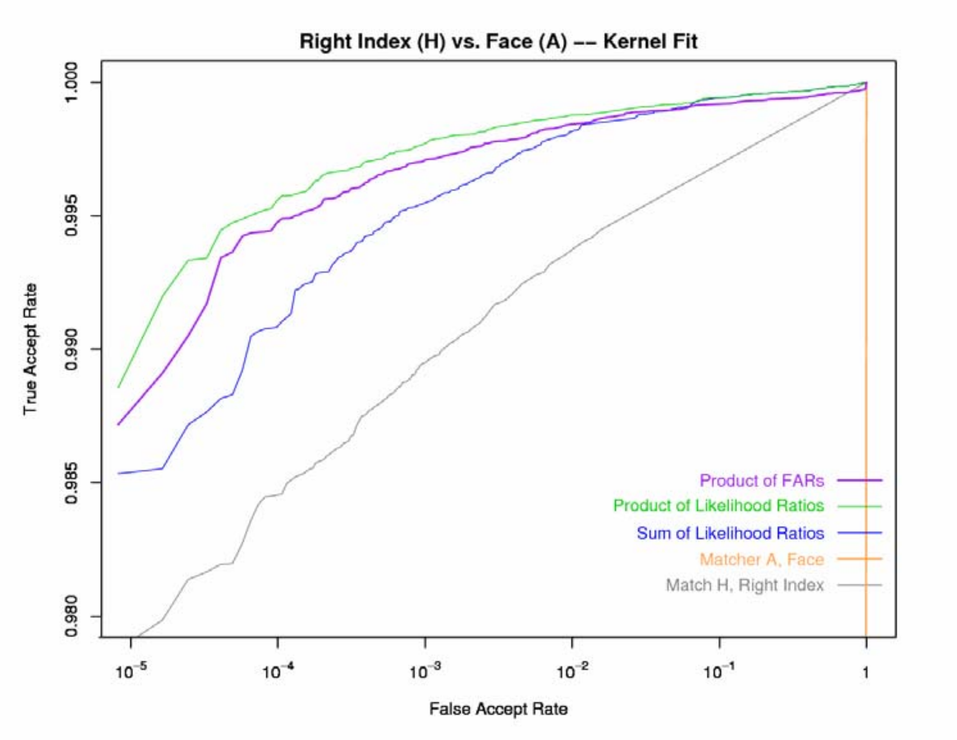

Figure 12: ROC curves for a face and right index fusion problem using variable bandwidth kernel density estimation.

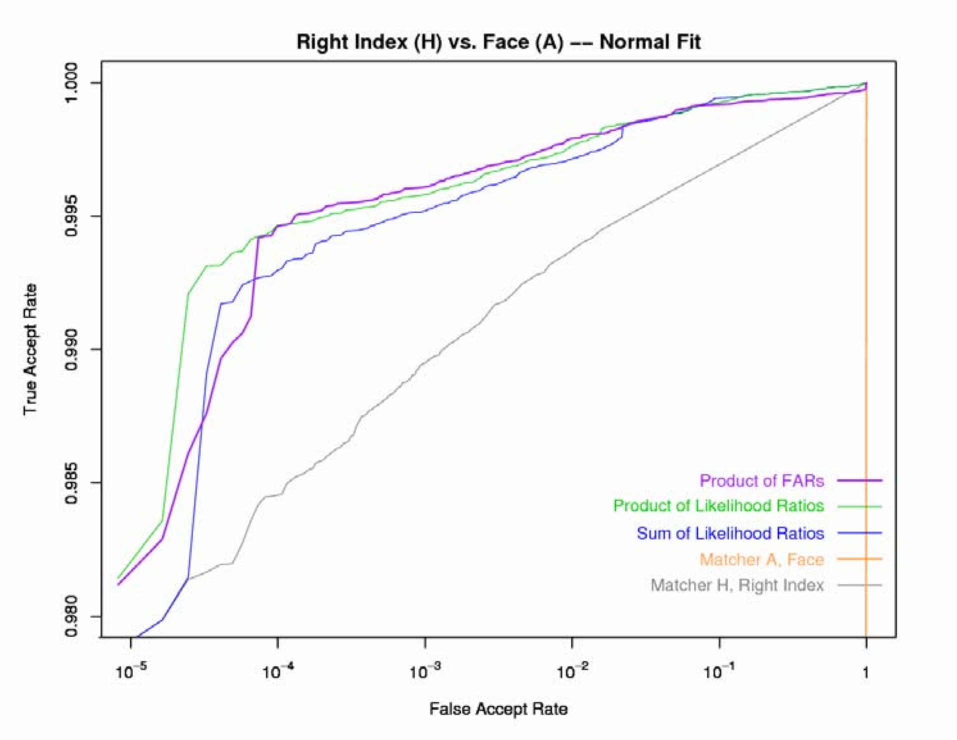

Figure 13: ROC curves for a face and right index fusion problem using normals for density estimation.

2.5 Density Estimation: Spikes

The Neyman-Pearson based Product of Likelihood Ratios method as implemented in this study is similar to that presented in [Dass-05]. One key difference is that those authors use a global bandwidth for the kernel density estimator, whereas this study used a variable bandwidth. Dass, *et al* note that "some parts of the score distributions can be discrete in nature. As a result, estimating the densities using continuous density functions can be inappropriate." They propose separate handling of each discrete component of the distribution. In this study, that advice was followed by removing spikes prior to density estimation, and handling those singular points discretely. Surprisingly, minimal benefit was derived from this special handling of spikes. This result was attributed to the very narrow bandwidths that were possible with such large samples, and the fact that the spikes occur at the extremes of the score ranges.

Figure 14 below compares the performance of the Product of Likelihood Ratios fusion method on various combinations of face (matchers A,B,C) and finger (matchers H,I,Q) data using four variant methods of density estimation:

- "Standard" refers to Product of Likelihood Ratios fusion, with density estimation based on variable bandwidth kernels, log-linear tail tapering and spike handling
- "No spikes" is the same as "Standard," except without spike handling
- "Linear Taper" is the same as "Standard," except a linear taper was used instead of log-linear
- "Normal" refers to Product of Likelihood Ratios fusion, with density estimation based simply on Gaussian models of the distributions; and spike handling.

As seen in the figure, handling spikes had very little effect on TAR when the distributions were otherwise well-fit using the kernel method. Proper fitting of the tail is important, as demonstrated by comparing the linear and log-linear fits. Modeling the densities as normal distributions, even with proper handling of spikes, will sometimes greatly reduce the benefits of fusion.

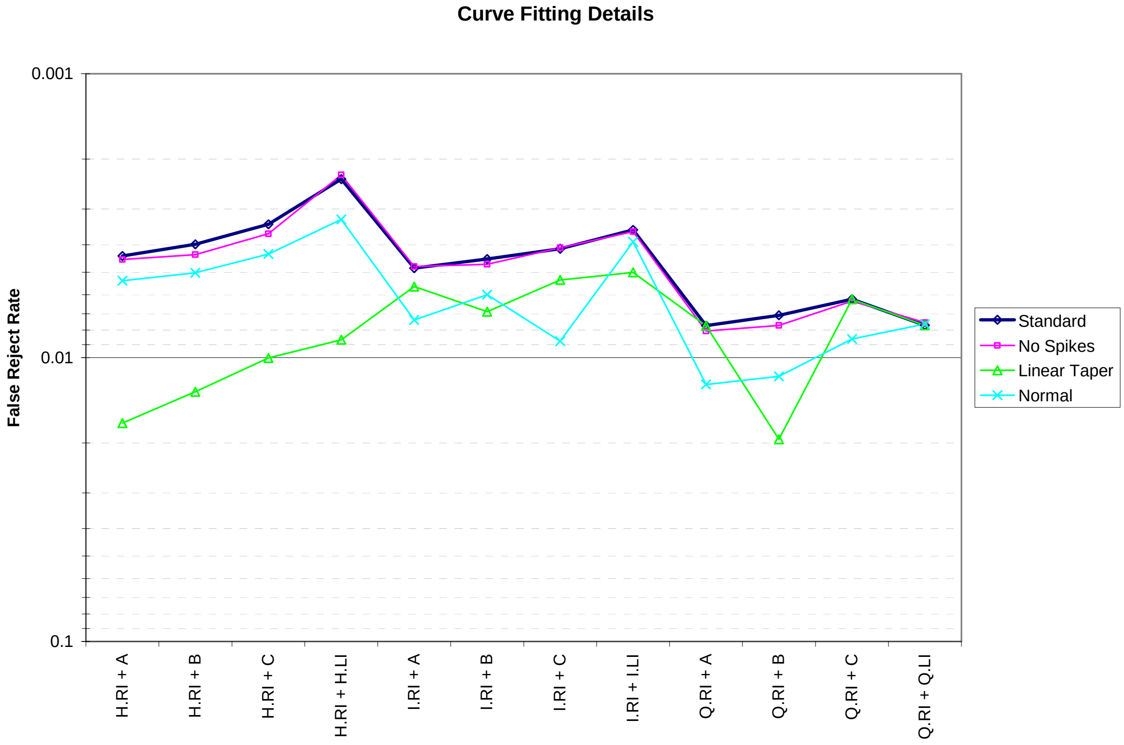

Figure 14: Explicit handling of spikes does not significantly improve performance. FRR is measured at FAR = 10^{-4}. RI and LI refer to right and left index fingers; A, B, and C refer to face matchers; H, I, and Q refer to fingerprint matchers.

Several factors may account for the lack of sensitivity to handling spikes:

- When distributions are modeled with high bias (e.g., normal distribution, or wide kernel bandwidth), the effects of the spikes are very broad, i.e., much of the estimated distribution is affected. This might be more likely to occur when modeling distributions based on small sample datasets. This study involved large datasets and the kernel bandwidth was controlled to limit bias.
- Spikes commonly occur at score extremes, away from the region of interest (e.g., the density cross-over point, or some specific low FAR). This tends to limit their effect on the decision boundaries.
- As density estimation was implemented in this study, score extremes routinely received special handling.
  - o Kernel modeling does not automatically handle bounded distributions well. In this implementation, the kernel density estimates were sometimes computed to points near the extreme of each score distribution, then linearly extrapolated to the extreme.[4]
  - o The initial kernel density model was adjusted by enforcing a log-linear taper of the right tail of non-mate distributions (from FAR=0.00005 to the right).
- These empirical findings are based on only one dataset and six matchers. This data is not fully representative of score distributions encountered in score-level biometric fusion.

[4] An alternative approach might have been to reflect the sample data across each extreme before fitting the curve.

Under different circumstances, spike handling may matter more. The impact of spikes should depend on several factors:

- The magnitude and distribution of spikes in the score distributions.
- The region of interest, i.e., the system operating point.
- The type of fusion technique:
  - Parametric techniques such as linear combinations do not involve density estimation.
  - Product of FARs is affected only by the non-mate right-tail integral of the density estimate. Spikes are not common in this tail region (i.e., at low FAR)
  - Likelihood ratios are affected by poor density estimates in a region of interest.
- The size of the dataset: higher bias is generally required to model smaller datasets; this exacerbates the problem of spikes for density estimation.
- On small training datasets, there is a risk of overfitting the density estimates at the spikes. In such cases, it might be beneficial not to handle spikes discretely.

3 Standard Modeling Procedure

This section summarizes the modeling procedure used in the Product of Likelihood Ratios, Product of FARs, Max of FARs, and Min of FARs techniques. The FAR techniques differ from the Ratios only in the computation of an integral (by trapezoidal approximation). Effectiveness of these techniques is sensitive to some of these modeling decisions.

3.1 Outline of Steps

This section describes the standard modeling procedure that was used throughout these papers to obtain the quantitative results. The modeling process was performed in two steps: first, the parameters of Table 1 were determined manually; the remainder of the process was fully automated.

1. Spikes were manually identified. Decisions are recorded in Table 1. There were no firm criteria, but in most cases spikes were distinctive anomalies.[5] The spike densities noted in the table are the best explanation of why they were selected. Some data is highly discrete and much of the non-mate score range could have been handled as spikes.

2. Kernel fitting. As shown in Table 1, very simple bandwidth functions were selected, partly to avoid overfitting concerns. The selection procedure was ad hoc, but the objective was to select functions that produced satisfying results in the “4-charts” (introduced in Figure 9).

3. Horizontal extension. As kernel fits do not model bounded distributions well, the kernel method was used over less than the entire range of the data. R’s KernSec function (GenKern library) supports fitting to a range of the distribution (“Fit Range” in Table 1). All scores except spikes were provided to that function. This initial fit was simply extended horizontally at each end to the True Range. Although not the best approach, this proved sufficient and was readily automated.

[5] [Dass-05] provides a criterion, but it is implicitly based on their sample size.

Dataset	Mating	Bandwidth function	Spikes	Spike Density	True range	Fit range
H index	NonMate	$((x + 40)/35)^2 - 1$	0 9999	0.9842787 0	0 215	1 139
H index	Mate	min(350, 2 + (5000 - abs(x - 5000))/2)	0 9999	0.00554691 0.1939107	0 9999	2 9960
I index	NonMate	$5 + x^{1.5}/450$	0 2000	0.00135246 0	-1 962	9 728
I index	Mate	min(65, 20 + x/10, 510 - x/4)	0 2000	0.00147918 0.8452874	0 2000	20 1929
Q index	NonMate	min(0.05, 0.002 + x/3)	-1 0 1	0.00019672 0.9948852 0	-1 0.55	0.01 .12
Q index	Mate	min(0.07, 0.01 + x/3, 1.01 - x)	-1 0 1	0.00020030 0.01838184 0.9331597	-1 1	0.02 0.98
A	NonMate	max(0.25, 2*x - 0.2)/100	0	0.01520492	0 0.719831	0 1
A	Mate	max(0.5, x + 0.3)/100	0	0.04235682	0 1	0 1
B	NonMate	min(2, x*0.05 + 0.02)	0 .105334	0.2153607 0.04598361	-1.173e-16 4.65605	-1.173e-16 3.723024
B	Mate	min(x*0.03 + 0.03, 2.5)	0 .105334	0.02026163 0.0030970	0 179.693	0 179.693
C	NonMate	(x*0.045) - 0.7			12.1152 83.4224	12.1152 79.15603
C	Mate	0.7			12.7615 100.004	12.7615 100.004

Table 1: Parameters used for curve fitting to match score distributions

4. Log-linear tapering [section 2.2]. This procedure was automated by extending non-mate distributions from the right end of Fit Range based on the slope just approaching that region.
5. FAR. The kernel density estimate, tabular in representation, was integrated by trapezoidal approximation. Spikes were worked back into the integral as a separate step.
6. Joint density estimates. These were modeled simply as the product of the univariate estimates.

3.2 Known Deficiencies

Several deficiencies of these procedures and the resulting models are recognized. These are summarized in Table 2.

Modeling Step	Concern	Response
Product (independence assumption)	In general, the scores are not strictly independent (see chapters *VI* and *VII*). For algorithm fusion and instance fusion, the scoresets are clearly correlated. The independence assumption was often less appropriate at higher accuracies.	Removing this assumption might produce better fusion results: this analysis demonstrates what can be achieved with this simplifying assumption. Experience with two techniques, logistic regression and Best Linear, suggest (but do not prove) that this assumption did not substantially hurt performance.
Tail modeling	Limited sample data precludes precise and confident modeling in the most critical regions of difficult discrimination. Incorrect modeling at low FAR might lead to poor operational results at low FAR.	The log-linear extension seems to fit the sample data well as seen in the "4-charts," but this modeling assumption clearly goes well beyond the supporting data. Instance fusion is not expected to be very sensitive to this assumption because the multiple distributions are produced by the same matcher. This is a potential source of severe modeling inaccuracies particularly in cases where operational decisions will be made at very low FAR (relative to size of training data) or where the benefits of fusion are marginal (e.g., algorithm fusion).
Overfitting	The kernel method and spike handling both invite overfitting, as does separate fitting of each distribution (vs. modeling the ratio directly). In general, the entire training set was used for both fitting and evaluation.	The full sets of scores were used to study the benefits of fusion at very low FAR. Overfitting occurs when the scale parameters for kernels are too small, or the degree of the fitted polynomial is too high. Because of the large volume of data, coupled with a strong belief in the inherent smoothness of the distributions (apart from spikes), overfitting is unlikely for this application. With the kernel method, the "4-charts" were used during the modeling process to visually assess variance and bias, and limit the risk of overfitting. Similar performance of the kernel fits and logistic regression suggest that overfitting was not a great concern. Separate validation runs were performed in which data was partitioned into training and evaluation sets [see section 3.4].

Table 2: Known modeling deficiencies

3.3 Tail Fitting Lesson

When the FRR gain tables were generated for algorithm fusion (Part IX), one instance occurred where the TAR actually decreased when two algorithms were fused. This was for Right Index fingers of Matchers H and Q. Closer examination revealed that several other ROCs — Best Linear, Sum of Ratios, Max of FARs -- were superior to the Product of Ratios in this case. Similar underperformance, though less severe, was noted for H and Q Left Index fingers.[6]

[6] The Best Linear technique correctly classified 148 more Right Index finger mates than did the Product of Ratios at FAR = 10^{-4} (TAR = .98573 vs. .98345). This corresponds to an FRR gain of 8 (rather than -7).

The underlying problem was determined to be a poor fit to the tail of the Q non-mate distribution. In this study, the curve fitting process was largely automated. At least two curve fits were manually inspected for each matcher (typically the index finger and thumb) to determine the manually specified parameters (Table 1). Inspection was performed using the "4-charts." Curve fits for most of the other fingers were not inspected.

The problem was corrected simply by changing the slope of the log-linear extension used to model the tail. Specifically, Q had too much influence at low FAR, so the rate of descent of the tail was decreased. The best results were obtained through trial-and-error. After a few trials accuracy approximated that of Best Linear. When making these final manual adjustments, it was not directly apparent how to optimize. The original fits looked reasonable on the "4-charts."

Lessons learned:

- Further improvements to the reported TARs are still possible simply by tuning the curve fits of the non-mate tails.
- Theoretical guidance is needed to optimize these curve fits.

3.4 Cross-validation

In general, the results of this study are based on measurements using the full sets of scores. That is, all of the data was used for both training and evaluation. This raises an important concern about the validity of the results, namely, the risk of overestimating the benefits of one or more fusion techniques due to overfitting the sample data. This section addresses that concern, and provides evidence that such overfitting did not significantly influence the results.

In general, overfitting is a concern when the model contains a large number of parameters relative to the size of the dataset. Despite the large size of the NBDF06 dataset, this problem might manifest in either of two ways: overfitting might occur specifically in the critical region of discrimination at low FAR where there is relatively little data; and the kernel method in particular might define a highly complex boundary that overfits the data.

Logistic Regression: as implemented and described in Part V, this method retains from the training process only a low-order polynomial description of the log odds. The method of maximum likelihood estimation implements a global fit to the log odds. It is not specifically tuned to the region of low FAR, but the use of higher order polynomial fits can largely overcome this limitation.

Best Linear: this method retains only one parameter from the training process (the slope), but it is trained to a specific FAR. As discussed in Section 4 (and shown in Figure 17), the value of this parameter is not highly sensitive to the choice of FAR, nor is performance (TAR) highly sensitive to errors in its estimation.

Kernel fits: the various methods that rely on kernel fits are subject to overfitting. As discussed in section 2, care was taken to control variance to guard against overfitting. Preventative steps included monitoring variance and bias, modeling the tail of the non-mate distribution as log-linear, and specifying a limited number of spikes. Modeling decision boundaries as products of independent variables further limits the opportunity to overfit.

In order to directly measure any overfitting associated with the Product of Likelihood Ratios technique, **cross-validation** runs were performed. For these runs, the mate and non-mate data were each randomly partitioned into two equal sized sets. Density ratios were modeled on the first set, and performance was evaluated on the second set. This process was fully automated using the tuning parameters given in Table 1. Figure 15 shows the results of one set of ten runs. At least one pairwise validation was performed involving each matcher. No evidence of significant training bias was observed.

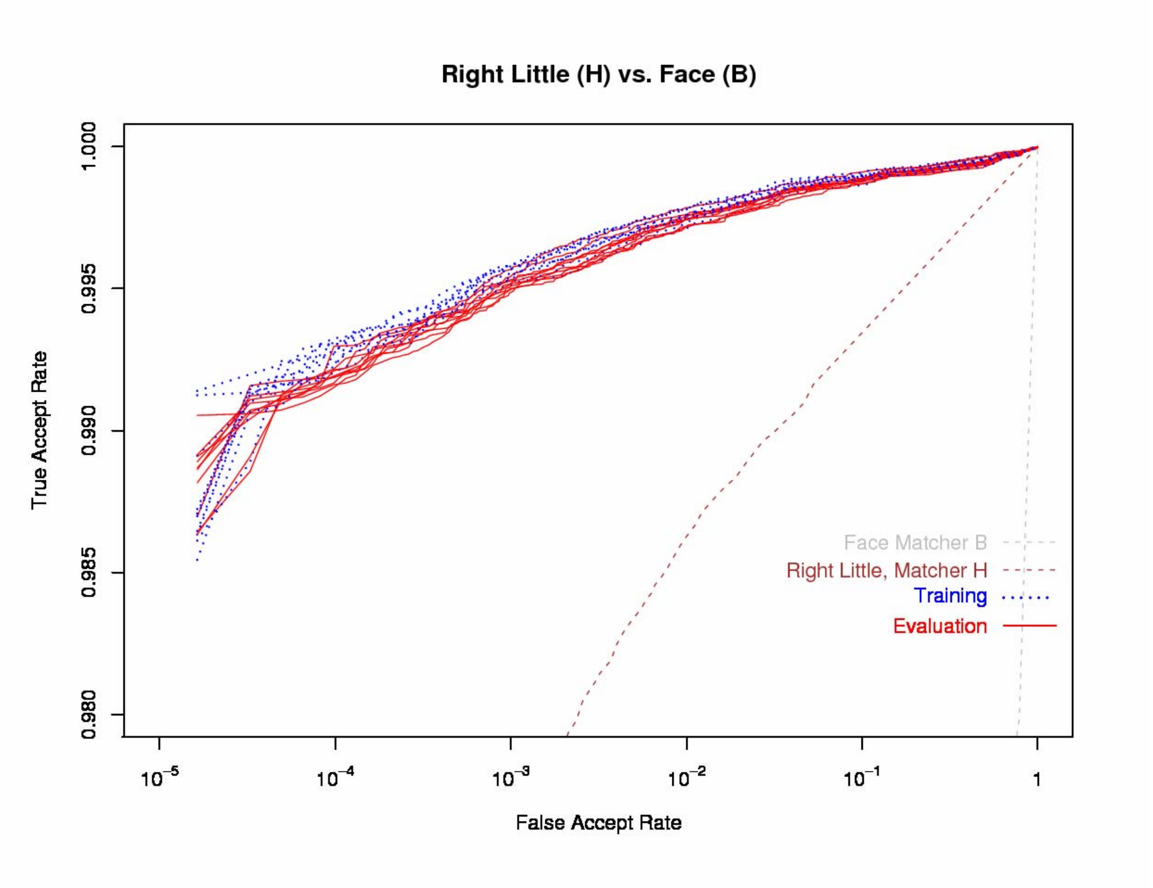

Figure 15: Example cross-validation results.

These reports contain further evidence that overfitting is not a great concern:

- consistently similar results across techniques
- consistent results across face, finger positions, and matchers
- a comparative example showing effects of overfitting using Parzen technique (Figure 6)
- ROCs that are "well-behaved" (smooth, roughly log-linear to a low FAR)

It is also worth noting the context in which the overfitting concern arises. First, these empirical results are strongly influenced by the source of the biometric data and the matching algorithms. Thus many of the absolute measures will differ from those of any specific operational system. Second, many of the comparative results are based on a common approach. Thus, for those comparisons, overfitting may be less of a concern.

4 Best Linear Fusion Technique

The Product of Likelihood Ratios and FAR-based techniques require the modeling of score distributions, as described above. Linear methods are conceptually much easier, because they are simply the weighted sum of the scores to be fused. This section discusses one method of optimizing linear score fusion: the Best Linear method.

In the Best Linear fusion technique, scores from each pair of matchers are combined using a weighted sum, i.e., $Z = weight \times X + Y$, where X and Y are the raw scores from each matcher, and weight is selected empirically to maximize TAR at a given FAR — in this case, FAR = 10^{-4}.

The weighting coefficient accounts for differences in raw score scales, as well as differences in matcher strengths. Thus, normalization and fusion are performed as a single operation. For example, in Figure 16, the "Best Linear" combination, $0.0062 \times A + G = A/\sigma_A + 2.7 \times G/\sigma_G$, i.e., the 0.0062 multiplier is mathematically equivalent to z-normalizing the scores[7], then applying a 1:2.7 weighting favoring the stronger matcher.

The following graphs illustrate how the weighting coefficient was determined. Figure 16 shows a scatterplot of z-normalized scores from two FpVTE matchers, with four parallel decision boundaries (corresponding to different decision thresholds). Figure 17 shows the effect of the angle of linear decision boundaries (x-axis) on TAR (y-axis) for various FAR values (curves).

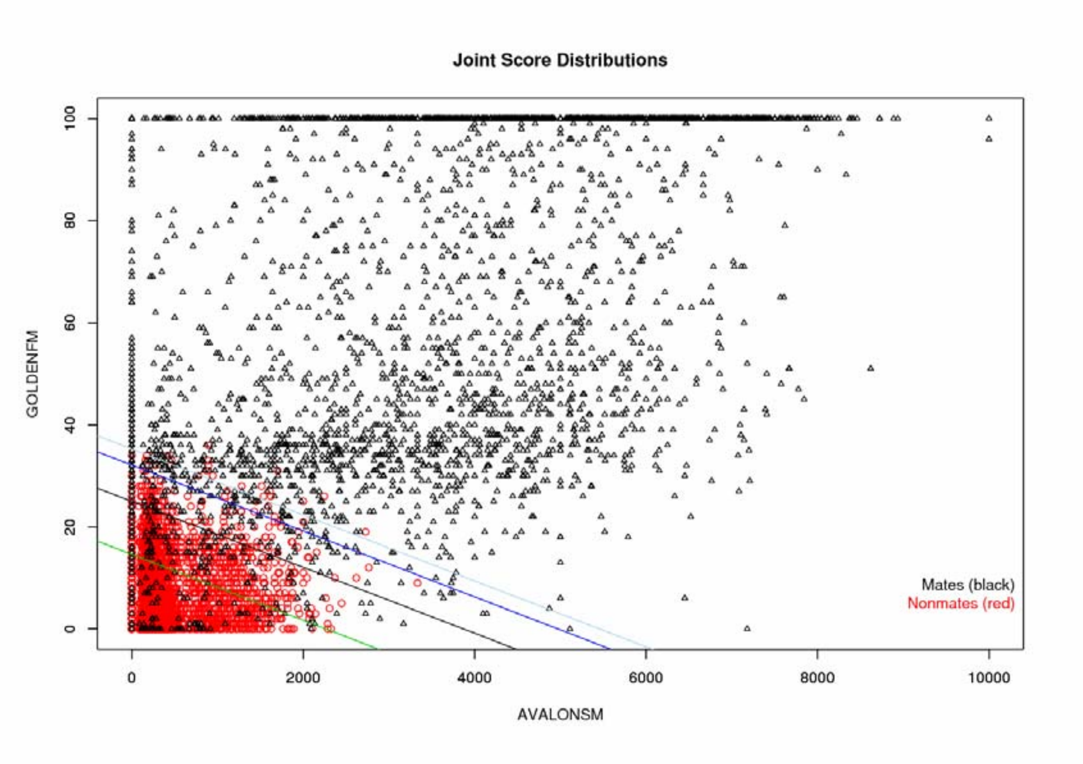

Figure 16: Joint matcher score distributions for the Avalon and Golden Finger matchers (FpVTE MST). Four linear decision boundaries shown (FAR=10^{-1},10^{-2},10^{-3},10^{-4}), each having the "Best Linear" slope as determined from Figure 17. An angle of 90° (vertical) corresponds to Avalon alone; 0° (horizontal) corresponds to Golden Finger alone.

[7] A denotes an individual Avalon score; σ_A denotes the standard deviation of all non-mate Avalon scores. The term for recentering scores (subtracting the mean score) is omitted because it has no effect on the ROC.

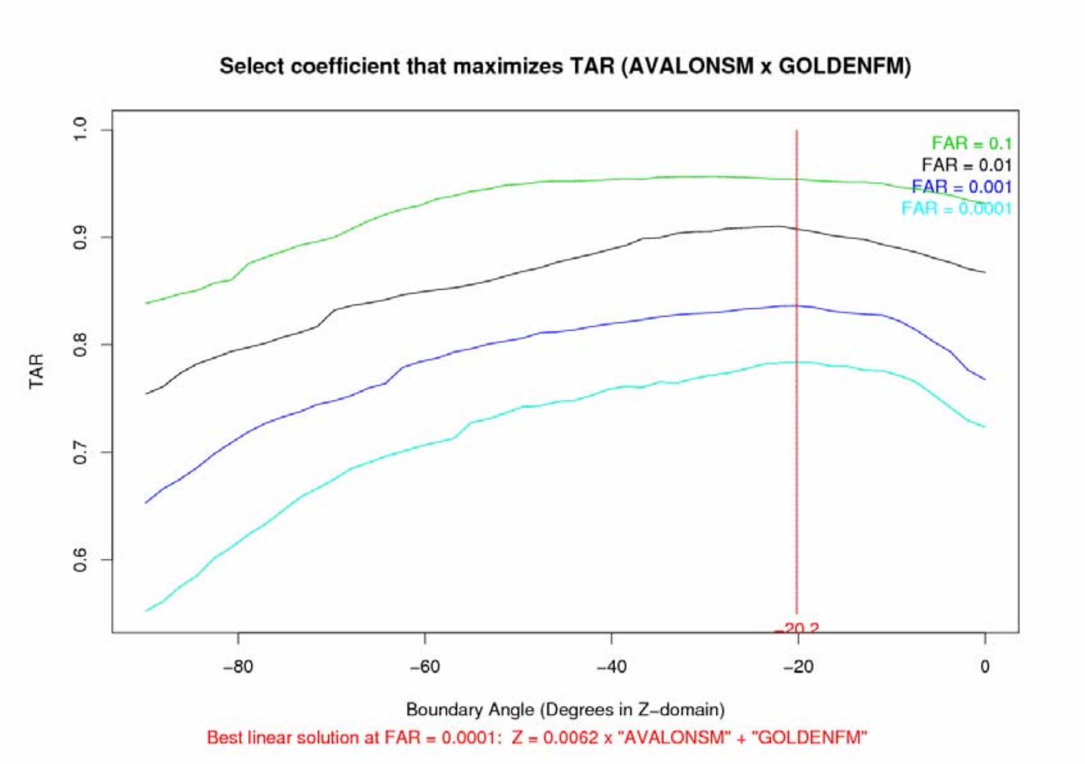

Figure 17: TAR as a function of linear weighting coefficient, based on the boundary angle from the scatterplot in Figure 16. Optimal angle (20°) corresponds to decision boundary in depicted in Figure 16.

The weighting coefficient is simply the (negative) slope of the decision boundary back in the raw score domain (i.e., not z-normalized). It is selected as that slope which maximizes the TAR at a given FAR (10^{-4} in this analysis). This chart is typical in that a minor error in estimating the optimal slope from sample data has only a minor detrimental effect on the TAR achieved. It is also typical in that the optimal slope is not highly sensitive to the FAR[8]. This implies that in most cases a large (>> 1/FAR) dataset is not required to obtain a good estimate of the weighting coefficient.

5 Conclusions

The modeling procedure described in this paper led to higher accuracy results than the alternatives to which it was compared (see Part-IV), but it is also quite complex. The intent of the complexity was to investigate the practical importance of various modeling steps in the context of this study.

Variable bandwidth kernel estimation is a useful technique for arriving at an accurate fit to irregular score distributions, but it is neither necessary nor sufficient. The kernel estimates provide a good initial model that must be adjusted in subsequent steps.

Special handling of "spikes" was — somewhat surprisingly — found to have little benefit. Two explanations are offered: spikes generally occur at the extremes of distributions, away from the region

[8] Often, the optimal slope gradually increases or decreases as a function of FAR, but typically without a substantial effect on TAR.

where low FAR decisions are made; and the large datasets and kernel estimation techniques used in this study allowed very low bias estimates. Special handling of spikes may be important in other contexts.

Standards ([BioAPI] and [FIF]) encourage transforming raw scores to FAR for the purpose of fusion. It was shown that there are alternative interpretations of FAR on discrete data, and none was consistently preferred.

Accurate modeling of the right tail of the non-mate distribution is both difficult and important. In several contexts, poor tail fits resulted in dramatic drops in accuracy at low FAR. **This finding raises a serious concern about the efficacy of fusion for systems that operate at very low FAR: if a system operates at a FAR setting that is orders of magnitude lower than where the training data supports modeling, then one can have little or no confidence in the relevant region of fit.** This may result in fused performance ranging anywhere from the accuracy of the weakest contributor to optimal performance.

So what performance is likely to occur at very low FAR? In the case of *multiple algorithms*, where benefits to fusion may be small, the risks of underperforming the better input is considerable. In the case of *multiple instances* processed by the same algorithm, the risks are small: optimal fusion is quite effective, so suboptimal performance still can be highly effective; and simply balancing the contribution of the multiple inputs (as with Simple Sum) may suffice to produce good results. Operating a *multi-modal* system at very low FAR would seem to be very risky following this basic procedure if one of the modes is much more accurate than the other. Perhaps a more complex fusion architecture that "locks in" definitive match decisions and only fuses uncertain scores could partially overcome this challenge.

6 References

[BioAPI] BioAPI Consortium; "BioAPI Specification Version 1.1"; March 16, 2001.

[Dass-05] S. Dass, K. Nandakumar and A. Jain, "A Principled Approach to Score Level Fusion in Multimodal Biometric Systems", *Proc. of Audio- and Video-based Biometric Person Authentication (AVBPA) 2005*, pp. 1049-1058, Rye Brook, NY, July 2005.

[FIF] "Fusion Information Format for Data Interchange"; First Working Draft; September 12, 2005.

[Griffin-05] P. Griffin; "Optimal Fusion for Multi-Biometric Identity Verification"; Identix Research Preprint RDNJ-05-0001, Jan. 2005.

[Jain-00] A. Jain, R. Duin, J. Mao; "Statistical Pattern Recognition: A Review"; *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, 2000, 4-37.

[Jain-05] Jain, Nandakumar, Ross; "Score Normalization in Multimodal Biometric Systems"; *Pattern Recognition 38* (2005) 2270-2285; Oct. 2004.

[Neyman-33] J. Neyman and E. S. Pearson; "On the problem of the most efficient tests of statistical hypotheses"; *Philosophical Transactions of the Royal Society, Series A, Containing Papers of a Mathematical or Physical Character*, 231, p. 289-337; 1933.

[Scott-05] C. Scott and R. Nowak; "A Neyman-Pearson Approach to Statistical Learning"; 2005 (http://www.ece.wisc.edu/~nowak/np.pdf)