Validating a Biometric Authentication System: Sample Size Requirements

Sarat C. Dass, *Member*, *IEEE*, Yongfang Zhu, and Anil K. Jain, *Fellow*, *IEEE*

Abstract—Authentication systems based on biometric features (e.g., fingerprint impressions, iris scans, human face images, etc.) are increasingly gaining widespread use and popularity. Often, vendors and owners of these commercial biometric systems claim impressive performance that is estimated based on some proprietary data. In such situations, there is a need to independently validate the claimed performance levels. System performance is typically evaluated by collecting biometric templates from n different subjects, and for convenience, acquiring multiple instances of the biometric for each of the n subjects. Very little work has been done in 1) constructing confidence regions based on the ROC curve for validating the claimed performance levels and 2) determining the required number of biometric samples needed to establish confidence regions of prespecified width for the ROC curve. To simplify the analysis that address these two problems, several previous studies have assumed that multiple acquisitions of the biometric entity are statistically independent. This assumption is too restrictive and is generally not valid. We have developed a validation technique based on multivariate copula models for correlated biometric acquisitions. Based on the same model, we also determine the minimum number of samples required to achieve confidence bands of desired width for the ROC curve. We illustrate the estimation of the confidence bands as well as the required number of biometric samples using a fingerprint matching system that is applied on samples collected from a small population.

Index Terms—Biometric authentication, error estimation, Gaussian copula models, bootstrap, ROC confidence bands.

1 Introduction

The purpose of a biometric authentication system is to validate the claimed identity of a user based on his/her physiological characteristics. In such a system operating in the verification mode, we are interested in accepting queries which are "close" or "similar" to the template of the claimed identity, and rejecting those that are "far" or "dissimilar." Suppose a user with true identity I_t supplies a biometric query Q and a claimed identity I_c. We are interested in testing the hypothesis

$$H_0 : I_t = I_c \quad \text{versus} \quad H_1 : I_t \neq I_c \tag{1}$$

based on the query Q and the template T of the claimed identity in the database, in (1), H_0 (respectively, H_1) is the null (alternative) hypothesis that the user is genuine (impostor). The testing in (1) is carried out by computing a similarity measure, $S(Q,T)$ where large (respectively, small) values of S indicate that T and Q are close to (far from) each other. A threshold, λ, is specified so that all similarity values lower (respectively, greater) than λ lead to the rejection (acceptance) of H_0. Thus, when a decision is made whether to accept or reject H_0, the testing procedure (1) is prone to two types of errors: the false reject rate (FRR) is the probability of rejecting H_0 when in fact the user is genuine, and the false accept rate (FAR) is the probability of accepting H_0 when in fact the user is an impostor. The genuine accept rate (GAR) is $1 - \mathrm{FRR}$, which is the probability that the user is accepted given that he/she is genuine. Both the FRR (and, hence, GAR) and the FAR are functions of the threshold value λ (see Fig. 1a). The Receiver Operating Curve (ROC) is a graph that expresses the relationship between the FAR versus GAR when λ varies, that is,

$$ROC(\lambda) = (FAR(\lambda), GAR(\lambda)), \tag{2}$$

and is commonly used to report the performance of a biometric authentication system (see Figs. 1a and 1b).

In marketing commercial biometric systems, it is often the case that error rates are either not reported or poorly reported (i.e., reported without giving details on how it was determined). In a controlled environment such as in laboratory experiments, one may achieve very high accuracies when the underlying biometric templates are of very good quality. However, these accuracies may not reflect the true performance of the biometric system in real field applications where uncontrolled factors such as noise and distortions can significantly degrade the system's performance. Thus, the problem we address in this paper is the validation of a claimed ROC curve, $\mathrm{ROC}_c(\lambda)$, by a biometric vendor. Of course, reporting just $\mathrm{ROC}_c(\lambda)$ does not give the complete picture. One should also report as much information as one can about the underlying biometric samples, such as the quality, the sample acquisition process, sample size, as well as a brief description of the subjects themselves. If the subjects used in the experiments for reporting $\mathrm{ROC}_c(\lambda)$ are not representative of the target population, then $\mathrm{ROC}_c(\lambda)$ is not very useful. But, assuming that the underlying samples are representative and can be replicated by other experimenters under similar conditions, one can then proceed to give margins of errors for validating $\mathrm{ROC}_c(\lambda)$.

- *S.C. Dass and Y. Zhu are with the Department of Statistics & Probability, Michigan State University, A-430 Wells Hall, East Lansing, MI 48824. E-mail: {sdass, zhuyongf}@msu.edu.*
- *A.K. Jain is with the Department of Computer Science & Engineering, Michigan State University, 3115 EB, East Lansing, MI 48824. E-mail: jain@cse.msu.edu.*

Manuscript received 3 Sept. 2004; revised 1 Apr. 2006; accepted 5 Apr. 2006; published online 12 Oct. 2006.
Recommended for acceptance by P.J. Phillips.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0463-0904.

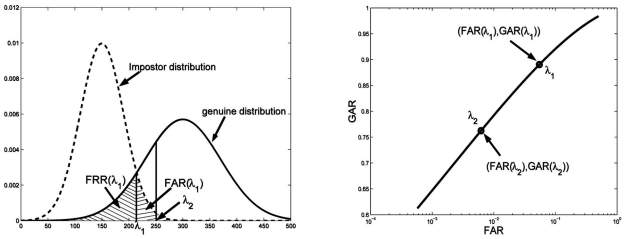

Fig. 1. Obtaining the ROC curve by varying the threshold λ. (a) Shows the FRR and FAR corresponding to a threshold λ_1. λ_2 is another threshold different from λ_1. (b) Shows the ROC curve obtained when λ varies. The values of (FAR, GAR) on the ROC curve corresponding to the thresholds λ_1 and λ_2 are shown.

The process of obtaining biometric samples usually involves selecting n individuals (or subjects) and using c different biometric instances[1] or from each individual. Additional biometric samples can be obtained by sampling each biometric multiple times, d, over a period of time. It is well-known that multiple acquisitions corresponding to each biometric exhibit a certain degree of dependence (or, correlation); see, for example, [1], [3], [10], [16], [17], [18], [19]). There have been several earlier efforts to validate the performance of a biometric system based on multiple biometric acquisitions. Bolle et al. [4] first obtained confidence intervals for the FRR and FAR assuming that the multiple biometric acquisitions were independent of each other. To account for correlation, Bolle et al. [2], [3] introduced the subsets bootstrap approach to construct confidence intervals for the FAR, FRR, and the ROC curve. Schuckers [16] proposed the beta-binomial family to model the correlation between the multiple biometric acquisitions as well as to account for varying FRR and FAR values for different subjects. He showed that the beta-binomial model gives rise to extra variability in the FRR and FAR estimates when correlation is present. However, a limitation of this approach is that it models correlation for a single threshold value. Thus, this method cannot be used to obtain a confidence region for the entire ROC curve. Further, Schucker's approach is strictly model-based; inference drawn from this model may be inappropriate when the true underlying model does not belong to the beta-binomial family.

To construct confidence bands for the ROC curve, Bolle et al. [3] select T threshold values, $\lambda_1, \lambda_2, \ldots, \lambda_T$ and compute the 90 percent confidence intervals for the associated FARs and GARs. At each threshold value λ_i, *combining* these 90 percent confidence intervals results in a confidence rectangle for $ROC(\lambda_i)$ (see (2)). Repeating this procedure for each $i = 1, 2, \ldots, T$ and *combining* the confidence rectangles obtained gives rise to a confidence region for $\mathrm{ROC}(\lambda)$. A major limitation of this approach is that the 90 percent confidence intervals for the FARs and GARs will neither automatically guarantee a 90 percent confidence rectangle at each λ_i nor a 90 percent confidence region for the ROC curve. In other words, ensuring a confidence level of 90 percent for each of the individual intervals cannot, in general, ensure a specific confidence level for the *combined* approach. This is the well-known problem of combining evidence from simultaneous hypothesis testing scenarios [9], [11], [12]: In essence, for each i, we are performing the tests

$$H_{0,i} : FAR(\lambda_i) = FAR_c(\lambda_i) \quad \text{versus} \quad H_{1,i} : \text{not } H_{0,i}, \quad (3)$$

and

$$H^*_{0,i} : GAR(\lambda_i) = GAR_c(\lambda_i) \quad \text{versus} \quad H^*_{1,i} : \text{not } H^*_{0,i}, \quad (4)$$

where $FAR(\lambda_i)$ (respectively, $FAR_c(\lambda_i)$) are the true but unknown (respectively, claimed) FAR at λ_i, and $GAR(\lambda_i)$ (respectively, $GAR_c(\lambda_i)$) are the true but unknown (respectively, claimed) GAR at λ_i. To test each $H_{0,i}$ (and $H^*_{0,i}$) individually, the 90 percent confidence interval for FAR (and GAR) can be used, and the resulting decision has a FRR of at most $100 - 90 = 10\%$. The confidence region for the ROC curve combines the $2T$ confidence intervals above and is used to test the hypothesis

$$H_0 : \cap_{i=1}^{T} \{H_{0,i} \cap H^*_{0,i}\} \qquad \text{versus} \qquad H_1 : \text{not } H_0. \quad (5)$$

However, the combined confidence region is not guaranteed to have a confidence level of 90 percent. In other words, the decision of whether to accept or reject H_0 does not have an associated FRR of 10 percent as in the case of the individual hypotheses. In fact, for a number α where $0 < \alpha < 1$, combining $2T$ $100(1-\alpha)\%$ level confidence intervals based on a-priori selected thresholds can only guarantee a lower bound of $100(1 - 2T\alpha)\%$ on the confidence level. This fact is based on Bonferroni's inequality and is well-known in the statistics literature. Instead of trying to derive this inequality, we point the reader to the relevant literature in statistics on simultaneous hypotheses testing procedures; see, for example, the following references [9], [11], [12]. The lower bound $100(1 - 2T\alpha)\%$ on the confidence level is not useful when T is large; in this case, $100(1 - 2T\alpha)\%$ is negative and we know that any confidence level should range between 0 percent and 100 percent. In Bolle et al.'s procedure, the value of T is large since the confidence rectangles are reported at various locations of the *entire* ROC curve.

In this paper, we present a new approach for constructing confidence regions for the ROC curve with a guaranteed prespecified confidence level. In fact, we are able to construct confidence regions for a *continuum* of threshold values, and not just for finite preselected threshold values. In contrast to the nonparametric bootstrap approach of [3], we develop a semiparametric approach for constructing confidence regions for $\mathrm{ROC}(\lambda)$. This is done by estimating the genuine and impostor distributions of similarity scores obtained from multiple biometric acquisitions of the n subjects where the marginals are first estimated nonparametrically (without any model assumptions) and then coupled together to form a multivariate joint distribution via a parametric family of Gaussian copula models [13]. The parametric form of the copula models enables us to investigate how correlation between the multiple biometric acquisitions affects the confidence regions. Confidence regions for the ROC are constructed using bootstrap resamples from our estimated semiparametric model. The main steps of our procedure are shown in Fig. 2. Note that our approach based on modeling the distribution of similarity scores is fundamentally different from that of [16], where binary (0 and 1) observations are used to construct confidence intervals for the FRRs and FARs.

Our approach also varies from that of [1], [3], [10], [16] in several respects. First, we explicitly model the correlation via

1. By instances or entities, we mean different fingers from each individual or iris images from the left and right eyes from each individual, etc.

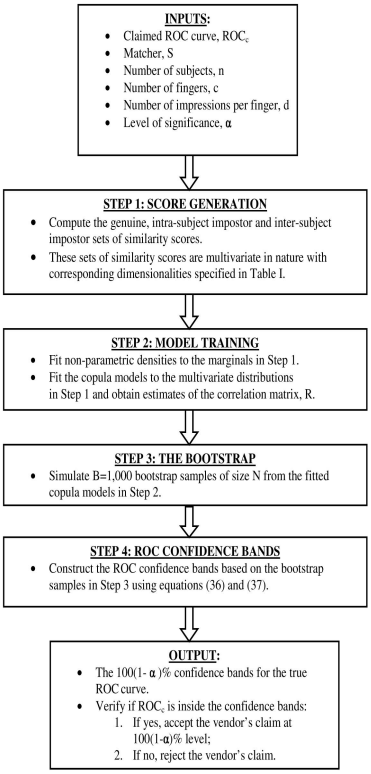

Fig. 2. The main steps involved in constructing the ROC confidence bands for validating the claim of a fingerprint vendor.

a parametric copula model and, thus, are able to demonstrate the effects of varying the correlation on the width of the ROC confidence regions. We also obtain a confidence *band*, rather than confidence rectangles as in [3], consisting of upper and lower bounds for the ROC curve. Further, the confidence bands come with a guaranteed confidence level for the *entire* ROC in the region of interest. Thus, we are able to perform tests of significance for the ROC curve and report error rates corresponding to our decision of whether to accept or reject the claimed ROC curve.

Another important issue that we address is that of the test sample size: How many subjects and how many biometric acquisitions per subject should be considered in order to obtain a confidence band for the ROC with a prespecified width? Based on the multivariate Gaussian copula model for correlated biometric acquisitions, we give the minimum number of subjects required to achieve the desired width. In presence of nonzero correlation, increasing the number of subjects is more effective in reducing the width of the confidence band compared to increasing the number of biometric acquisitions per subject. For achieving the desired confidence level, the required number of subjects based on our method is much smaller compared to the subset bootstrap. Rules of thumb such as the Rule of 3 [20] and the Rule of 30 [14] grossly underestimate the number of users required to obtain a specific width. The underestimation becomes more severe as the correlation between any two acquisitions of a subject increases.

The paper is organized as follows: Section 2 presents the problem formulation. Section 3 discusses the use of multivariate copula functions to model the correlation between multiple queries per subject for the genuine and impostor similarity score distributions. Section 4 presents the construction of confidence bands for the ROC curve. Section 5 discusses the minimum number of biometric samples required for obtaining confidence bands of a prespecified width for the ROC curve. Some of the more technical details and experimental results have been moved to the Appendix due to space restrictions (which can be found at http://computer.org/tpami/archives.htm); interested readers can also refer to the paper [6] which incorporates the relevant details into appropriate sections of the main text.

2 Preliminaries

Suppose we have n subjects available for validating a biometric authentication system. Often, during the data collection stage, multiple biometric entities (e.g., different fingers) from the same subject are used. We denote the number of biometric entities used per subject by c. To obtain additional data, each biometric of a subject is usually sampled a multiple number of times, d, over a period of time. Thus, at the end of the data collection stage, we acquire a total of ncd biometric samples from the n subjects. This collection of ncd biometric samples will be denoted by $\mathcal{B}$. To obtain similarity scores, a pair of biometric samples, B and B' with $B \neq B'$, are taken from $\mathcal{B}$ and a matcher S is applied to them, resulting in the similarity score $S(B, B')$. We will consider asymmetric matchers for S in this paper: The matcher S is asymmetric if $S(B, B') \neq S(B', B)$ for the pair of biometric samples (B, B') (a symmetric matcher implies that $S(B, B') = S(B, B')$).

In the subsequent text, we will use a fingerprint authentication system as the generic biometric system that needs to be validated. Thus, the c different biometric entities will be represented as c different fingers from each subject, and the d acquisitions will be represented by d impressions of each finger. When B and B' are multiple impressions of the same finger from the same user, the similarity score $S(B, B')$ is termed as a genuine similarity score, whereas when B and B' are impressions from either 1) different fingers from the same subject or 2) different subjects, the similarity score $S(B, B')$ is termed as an impostor score. The impostor scores arising from 1) (respectively, 2)) are termed as the intrasubject (respectively, intersubject) impostor scores.

We give some intuitive understanding of why similarity scores arising from certain pairs of fingerprint impressions in $\mathcal{B}$ are correlated (or dependent). During the fingerprint acquisition process, multiple impressions of a finger are obtained by successive placement of the finger onto the sensor. Therefore, given the first impression, B, and two subsequent impressions B_1 and B_2, the similarity scores $S(B, B_1)$ and $S(B, B_2)$ are most likely going to be correlated. Further, the fingerprint acquisition process is prone to many different types of uncontrollable factors such as fingertip pressure, fingertip moisture, and skin elasticity factor. These factors cause some level of dependence between fingerprint impressions of two different fingers of the same user. If this is the case, then we expect to see some level of correlation between the similarity scores $S(B_1, B_2)$ where B_1 and B_2 are impressions from different fingers. Also, as noted in [3], even the scores $S(B_1, B_2)$ from different fingers of different subjects could be correlated. All these facts lead us to statistically model the correlation for similarity scores in the

TABLE 1
Values of K for the Different Sets $\mathcal{G}_i$, $\mathcal{I}_i$, and $\mathcal{I}_{ij}$

Entities	$\mathcal{G}_i$	$\mathcal{I}_i$	$\mathcal{I}_{ij}$
Dimension, K	$cd(d-1)$	$c(c-1)d^2$	c^2d^2

Here, c is the number of fingers and d is the number of impressions per finger.

three major categories, namely, the genuine, intrauser impostor and interuser impostor similarity scores.

In order to develop the framework that incorporates correlation, we need to introduce some notation. We denote the set consisting of the d impressions of finger f, $f = 1, 2, \ldots, c$, from subject i by $\mathcal{M}_{i,f}$. The notation

$$\mathcal{S}(i, j, f, f') = \{S(B_u, B_v); B_u \in \mathcal{M}_{i,f}, B_v \in \mathcal{M}_{j,f'}, B_u \neq B_v\} \tag{6}$$

represents the set of all similarity scores available from matching the fingerprint impressions of finger f from subject i and those of finger f' from subject j. Three disjoint sets of (6) are of importance, namely, the set of genuine similarity scores (taking $i = j$ and $f = f'$ in (6)), the set of intrasubject impostor scores ($i = j$ and $f \neq f'$), and the set of intersubject impostor scores ($i \neq j$). We denote the genuine, intrasubject impostor and intersubject impostor score sets by

$$\mathcal{G}_i \equiv \bigcup_{f=1}^{c} \mathcal{S}(i, i, f, f), \quad \mathcal{I}_i \equiv \bigcup_{f=1}^{c} \bigcup_{\substack{f'=1 \\ f' \neq f}}^{c} \mathcal{S}(i, i, f, f'), \quad \text{and} \quad \mathcal{I}_{ij} \equiv \bigcup_{f=1}^{c} \bigcup_{f'=1}^{c} \mathcal{S}(i, j, f, f'), \tag{7}$$

where $i \neq j$, respectively.

We give the cardinality or dimension (the number of possibly distinct similarity scores) of each of the sets discussed above. The dimensions of $\mathcal{G}_i$, $\mathcal{I}_i$, and $\mathcal{I}_{ij}$ are $cd(d-1)$, $c(c-1)d^2$ and c^2d^2, respectively, when the matcher S is asymmetric. In all of these scenarios, we will denote the dimension corresponding to each set by K (see Table 1). The total number of sets of similarity scores arising from the genuine, intra and interimpostor cases will be denoted by N; we have that $N = n$, $N = n$ and $N = n(n-1)$, respectively, for the total number of sets of genuine, intrasubject impostor and intersubject scores.

When the matcher S is symmetric, the dimension associated with each of the genuine, intrasubject impostor and intersubject impostor sets of similarity scores gets reduced since many of the similarity scores in each of the three sets will be identical to each other. In the subsequent text, we outline the methodology for validating a vendor's claim for an asymmetric matcher. Our methodology for constructing the ROC confidence bands for a symmetric matcher can be handled in a similar fashion, keeping in mind the reduction in dimensions of each of the three sets of similarity scores discussed above.

Subsequently, N will denote the total number of independent sets of similarity scores and K will denote the dimension of each of these N sets. For $i = 1, 2, \ldots, N$, the ith set of similarity scores will be denoted by the K-dimensional vector

$$\underline{S}_i = (s(i,1), s(i,2), \ldots, s(i,K))^T, \tag{8}$$

where $s(i,k)$ is the generic score corresponding to the kth component of $\underline{S}_i$, for $k = 1, 2, \ldots, K$.

The ordered indices $1, 2, \ldots, K$ are associated to the elements of each of the sets $\mathcal{G}_i$, $\mathcal{I}_i$ and $\mathcal{I}_{ij}$ defined in (7) in the following way: Let $s(B_{f,u}, B_{f',v})$ denote the similarity score obtained when matching impression u of finger f, $B_{f,u}$, with impression v of finger f', $B_{f',v}$. In the case of a genuine set (that is, $\underline{S}_i \in \mathcal{G}_i$), the order of the genuine scores is taken as $\underline{s}(f) \equiv (s(B_{f,u}, B_{f,v}),\ v = 1, 2, \ldots, (u-1), (u+1), \ldots, d,\ u = 1, 2, \ldots, d)$, and $\underline{S}_i = (\underline{s}(1), \underline{s}(2), \ldots, \underline{s}(c))$. In the case when $\underline{S}_i \in \mathcal{I}_i$, the order of the scores is taken as $\underline{s}(f,f') \equiv (s(B_{f,u}, B_{f',v}),\ v = 1, 2, \ldots, d,\ u = 1, 2, \ldots, d)$, and $\underline{S}_i = (\underline{s}(f,f'),\ f' = 1, 2, \ldots, (f-1), (f+1), \ldots, c,\ f = 1, 2, \ldots, c)$. Finally, in the case when $\underline{S}_i$ is an intersubject impostor set (one of $\mathcal{I}_{ij}$), the order of the scores are taken as $\underline{s}(f,f') \equiv (s(B_{f,u}, B_{f',v}),\ v = 1, 2, \ldots, d, u = 1, 2, \ldots, d)$, and $\underline{S}_i = (\underline{s}(f,f'),\ f' = 1, 2, \ldots, c,\ f = 1, 2, \ldots, c)$.

If the scores $s(i,k)$ are bounded between two numbers a and b, the order preserving transformation

$$\mathcal{T}(s(i,k)) = \log\left(\frac{s(i,k) - a}{b - s(i,k)}\right) \tag{9}$$

converts each score onto the entire real line. This transformation yields better nonparametric density estimates for the marginal distribution of similarity scores. The transformed scores will be represented by the same notation $s(i,k)$. The distribution function for each $\underline{S}_i$ will be denoted by F, that is,

$$P\{s(i,k) \leq s_k, 1 \leq k \leq K\} = F(s_1, s_2, \ldots, s_k), \tag{10}$$

for real numbers $s_1, s_2, \ldots, s_K$. Note that 1) F is a multivariate joint distribution function on R^K and 2) we assume that F is the common distribution function for every $i = 1, 2, \ldots, N$. The distribution function F has K associated marginals, we denote the marginals by F_k, $k = 1, 2, \ldots, K$, where

$$P\{s(i,k) \leq s_k\} = F_k(s_k). \tag{11}$$

3 Copula Models for F

We propose a semiparametric family of Gaussian copula models as models for F. Let $H_1, H_2, \ldots, H_K$ be K continuous distribution functions on the real line. Suppose that H is a K-dimensional distribution function with the kth marginal given by H_k for $k = 1, 2, \ldots, K$. According to Sklar's theorem [13], there exists a unique function $C(u_1, u_2, \ldots, u_K)$ from $[0,1]^K$ to $[0,1]$ satisfying

$$H(s_1, s_2, \ldots, s_k) = C(H_1(s_1), H_2(s_2), \ldots, H_k(s_k)), \tag{12}$$

where $s_1, s_2, \ldots, s_K$ are K real numbers. The function C is known as a K-copula function that "couples" the one-dimensional distribution functions H_k, $k = 1, 2, \ldots, K$ to obtain H. Basically, K-copula functions are K-dimensional distribution functions on $[0,1]^K$ whose marginals are uniform. Equation (12) can also be used to construct K-dimensional distribution function H whose marginals are the prespecified distributions $H_k, k = 1, 2, \ldots, K$: choose a copula function C and define the function H as in (12). It follows that H is a K-dimensional distribution function with marginals $H_k, k = 1, 2, \ldots, K$.

The choice of C we consider in this paper is the K-dimensional Gaussian copulas [5] given by

$$C_R(u_1, u_2, \ldots, u_k) = \Phi_R^K(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_k)), \tag{13}$$

where each $u_k \in [0,1]$ for $k = 1, 2, \ldots, K$, $\Phi(\cdot)$ is the distribution function of the standard normal, $\Phi^{-1}(\cdot)$ is its inverse, and Φ_R^K is the K-dimensional distribution function of a normal random vector with component means and variances given by 0 and 1, respectively, and with correlation matrix R. Note that R is a positive definite matrix with diagonal entries equal to unity. The distribution function F will be assumed to be of the form (12) with $H_k = F_k$ for $k = 1, 2, \ldots, K$, and $C = C_R$; thus, we have

$$F(s_1, s_2, \ldots, s_k) = C_R(F_1(s_1), F_2(s_2), \ldots, F_K(s_k)). \tag{14}$$

We denote the observed genuine scores by $\mathcal{S}_0 \equiv \{s_0(i,k), k = 1, 2, \ldots, K_0, i = 1, 2, \ldots, N_0\}$ with $K_0 = cd(d-1)$ and $N_0 = n$. Each vector $(s_0(i,1), s_0(i,2), \ldots, s_0(i,K_0))$ is assumed to be independently distributed according to (14) with correlation matrix R_0 and marginals $F_{k,0}$, $k = 1, 2, \ldots, K_0$. Both R_0 and the K_0 marginals are unknown and have to be estimated from the observed scores. In Section 5, we show how this is done based on similarity scores obtained from a fingerprint matching system. The observed intrasubject and intersubject impostor similarity scores are denoted by

$$\mathcal{S}_{11} \equiv \{s_{11}(i,k), k = 1, 2, \ldots, K_{11}, i = 1, 2, \ldots, N_{11}\}$$

with $K_{11} = c(c-1)d^2$ and $N_{11} = n$, and

$$\mathcal{S}_{12} \equiv \{s_{12}(i,k), k = 1, 2, \ldots, K_{12}, i = 1, 2, \ldots, N_{12}\}$$

with $K_{12} = c^2d^2$ and $N_{12} = n(n-1)$, respectively. Each vector $(s_{11}(i,1), s_{11}(i,2), \ldots, s_{11}(i,K_{11}))$ (respectively, $(s_{12}(i,1), s_{12}(i,2), \ldots, s_{12}(i,K_{12}))$) is assumed to be independently distributed according to (14) with correlation matrix R_{11} (R_{12}) and marginals $F_{k,11}$, $k = 1, 2, \ldots, K_{11}$ ($F_{k,12}$, $k = 1, 2, \ldots, K_{12}$). The correlation matrices R_{11}, R_{12} and the associated marginals are estimated from the observed impostor scores in the same way as is done for the genuine case. Details of the estimation procedure for the impostor case are presented in the Appendix, which can be found at http://computer.org/tpami/archives.htm and [6].

4 Confidence Bands for the ROC Curve

The Receiver Operating Curve (ROC) is a graph that expresses the relationship between the Genuine Accept Rate (GAR) and the False Accept Rate (FAR), and is used to report the performance of a biometric authentication system. For the threshold λ, the empirical GAR and FAR can be computed using the formulas

$$GAR_e(\lambda) = \frac{1}{N_0 K_0} \sum_{i=1}^{N_0} \sum_{k=1}^{K_0} I\{s_0(i,k) > \lambda\}, \tag{15}$$

and

$$FAR_e(\lambda) = \frac{1}{N_1} \left\{ \sum_{i=1}^{N_{11}} \sum_{k=1}^{K_{11}} I\{s_{11}(i,k) > \lambda\} + \sum_{i=1}^{N_{12}} \sum_{k=1}^{K_{12}} I\{s_{12}(i,k) > \lambda\} \right\}, \tag{16}$$

where $I(A) = 1$ if property A is satisfied, and 0, otherwise, and $N_1 = N_{11}K_{11} + N_{12}K_{12}$ denotes the total number of impostor scores. The true but unknown values of $GAR(\lambda)$ and $FAR(\lambda)$ are the population versions of (15) and (16), the expression for the population $GAR(\lambda)$ is given by

$$\begin{aligned} E(GAR_e(\lambda)) &= \frac{1}{N_0 K_0} \sum_{i=1}^{N_0} \sum_{k=1}^{K_0} P\{s_0(i,k) > \lambda\} \\ &= \frac{1}{K_0} \sum_{k=1}^{K_0} P\{s_0(1,k) > \lambda\} \\ &\equiv G_0(\lambda), \end{aligned} \tag{17}$$

where each set $\{s_0(i,k), k = 1, 2, \ldots, K_0\}$ for $i = 1, 2, \ldots, N_0$ is independent and identically distributed according to the copula model (14). Subsequently, the probabilities in (17) are functions of the unknown genuine marginal distributions, $F_{k,0}, k = 1, 2, \ldots, K_0$, and the genuine correlation matrix, R_0. Also, the second equality in (17) is a consequence of the identically distributed assumption. In a similar fashion, the population $FAR(\lambda)$ is given by

$$\begin{aligned} E(FAR_e(\lambda)) &= \frac{1}{N_1} \left\{ \sum_{i=1}^{N_{11}} \sum_{k=1}^{K_{11}} P\{s_{11}(i,k) > \lambda\} + \sum_{i=1}^{N_{12}} \sum_{k=1}^{K_{12}} P\{s_{12}(i,k) > \lambda\} \right\} \\ &= \frac{N_{11}}{N_1} \sum_{k=1}^{K_{11}} P\{s_{11}(i,k) > \lambda\} + \frac{N_{12}}{N_1} \sum_{k=1}^{K_{12}} P\{s_{12}(i,k) > \lambda\} \\ &\equiv G_1(\lambda), \end{aligned} \tag{18}$$

where now, elements within each of the sets $\{s_{11}(i,k), k = 1, 2, \ldots, K_{11}\}$ for $i = 1, 2, \ldots, N_{11}$ and $\{s_{12}(i,k), k = 1, 2, \ldots, K_{12}\}$ for $i = 1, 2, \ldots, N_{12}$ are independent and identically distributed according to the copula model (14) with corresponding correlation matrices and marginals. The probabilities in (18) are functions of the unknown marginal distributions, $F_{k,11}$ for $k = 1, 2, \ldots, K_{11}$ and $F_{k,12}$ for $k = 1, 2, \ldots, K_{12}$, and the correlation matrices, R_{11} and R_{12}, for the intrasubject and intersubject impostor scores, respectively.

In light of the notations used for the population versions of FAR and GAR, (15) and (16) are sample versions of $G_0(\lambda)$ and $G_1(\lambda)$. Thus, we define

$$\hat{G}_0(\lambda) \equiv GAR_e(\lambda) \quad \text{and} \quad \hat{G}_1(\lambda) \equiv FAR_e(\lambda). \tag{19}$$

The empirical ROC curve can be obtained by evaluating the expressions for GAR and FAR in (15) and (16) at various values λ based on the observed similarity scores, and plotting the resulting curve $(\hat{G}_1(\lambda), \hat{G}_0(\lambda))$. However, there is an alternative way in which an ROC curve can be constructed. Note that the ROC expresses the relationship between the FAR and GAR, and the threshold values are necessary only at the intermediate step for linking the FAR and GAR values. Thus, another representation of the ROC curve can be obtained by the following reparameterization: We fix p as a value of FAR and obtain the threshold λ_* such that $\hat{G}_1(\lambda_*) = p$ or $\lambda_* \equiv \hat{G}_1^{-1}(p)$. Substituting λ_* in (15) gives the ROC curve in the form $(p, \hat{W}(p))$, where

$$\hat{W}(p) = \hat{G}_0(\lambda_*) \equiv \hat{G}_0\Big(\hat{G}_1^{-1}(p)\Big). \tag{20}$$

Note that, in the case when there is no λ_* such that $\hat{G}_1(\lambda_*) = p$, one can redefine the inverse, $\hat{G}_1^{-1}(p) \equiv \lambda_*$, where λ_* is the smallest λ satisfying $\hat{G}_1(\lambda) \leq p$. This definition of the inverse of $\hat{G}_1$ is more general and always yields a unique λ_*. The true but unknown ROC curve can be obtained in the same way as above by replacing the empirical versions with the corresponding population version; thus, we have

$$W(p) = G_0\big(G_1^{-1}(p)\big), \tag{21}$$

where $G_1^{-1}(p) \equiv \lambda_*$, where λ_* is the smallest λ satisfying $G_1(\lambda) \leq p$. The two representations of the ROC curves $(\hat{G}_1(\lambda), \hat{G}_0(\lambda))$ and $(p, \hat{W}(p))$, are close approximations of one another for large N_0, and therefore we use the latter representation for deriving the confidence bands. For fixed numbers C_0 and C_1 satisfying $0 \leq C_0 < C_1 \leq 1$, let us consider all $p = FAR$ values that fall in $[C_0, C_1]$. A confidence band for the true (claimed) ROC curve of a biometric system at confidence level $100(1-\alpha)\%$ gives two envelope functions, $e_L(p)$ and $e_U(p)$, so that for *all* p in $[C_0, C_1]$, the true ROC curve lies inside the interval $(\, e_L(p), e_U(p)\,)$ with probability of *at least* $100(1-\alpha)\%$. The numbers C_0 and C_1 form the lower and upper bounds of the range of FAR and will be chosen to cover typical reported values of FAR in biometric applications. If $C_0 = 0$ and $C_1 = 1$, the resulting ROC confidence band is constructed for the true ROC curve for all p in $(0, 1)$.

For a specific $p = FAR$, the corresponding value of GAR, $W(p)$, is a proportion which takes values in $[0, 1]$. For proportions, the transformation

$$\sqrt{N_0}\left(\sin^{-1}\sqrt{\hat{W}(p)} - \sin^{-1}\sqrt{W(p)}\right) \tag{22}$$

is a variance stabilizing transformation [15]; the quantity in (22) is asymptotically distributed as a normal with zero mean and constant variance (independent of p and $W(p)$) for large N_0. To obtain the envelopes, we first consider a continuum version of the absolute values of (22) for FAR values, p, in $[C_0, C_1]$, and take the maximum over $p \in [C_0, C_1]$. This gives the statistic

$$z \equiv \max_{p: C_0 \leq p \leq C_1} \sqrt{N_0}|\sin^{-1}\sqrt{\hat{W}(p)} - \sin^{-1}\sqrt{W(p)}|. \tag{23}$$

Assume for the moment that the distribution of z is known. If $z_{1-\alpha}$ denotes the $100(1-\alpha)\%$ percentile of z, the envelopes are given by

$$e_L(p) = (\sin(\sin^{-1}\sqrt{\hat{W}(p)} + z_{1-\alpha}/\sqrt{N_0}))^2$$

and

$$e_U(p) = (\sin(\sin^{-1}\sqrt{\hat{W}(p)} + z_{1-\alpha}/\sqrt{N_0}))^2. \tag{24}$$

However, the distribution of z is difficult to obtain analytically and, thus, we present two approaches to approximate the distribution of z in (23) based on 1) the bootstrap methodology and 2) an asymptotic representation of the distribution of z for large N_0.

4.1 The Semi and Nonparametric Bootstrap Approaches

The value $z_{1-\alpha}$ will be found based on bootstrap samples from the fitted semiparametric Gaussian copula models described in Section 3. This bootstrap procedure requires the simulation of scores from the estimated distribution functions in (14) and is described in detail in the Appendix, which can be found at http://computer.org/tpami/archives.htm. Thus, we denote by

$$\mathcal{S}_0^* \equiv \{s_0^*(i,k), k = 1, 2, \ldots, K_0, i = 1, 2, \ldots, N_0\},$$

$$\mathcal{S}_{11}^* \equiv \{s_{11}^*(i,k), k = 1, 2, \ldots, K_{11}, i = 1, 2, \ldots, N_{11}\},$$

and

$$\mathcal{S}_{12}^* \equiv \{s_{12}^*(i,k), k = 1, 2, \ldots, K_{12}, i = 1, 2, \ldots, N_{12}\}$$

to be the sets of genuine, intraimpostor and interimpostor similarity scores obtained by one simulation from the fitted copula models. Also, let

$$W^*(p) = G_0^*\Big(G_1^{*^{-1}}(p)\Big), \tag{25}$$

where $G_0^*(\lambda)$ (respectively, $G_1^*(\lambda)$) is obtained from (15) (respectively, (16)) with the bootstrap samples $s^*(i,k)$ used in place of the $s(i,k)$s. We form the quantity

$$z^* \equiv \max_{C_0 \leq p \leq C_1} \sqrt{N_0}|\sin^{-1}\sqrt{W^*(p)} - \sin^{-1}\sqrt{\hat{W}(p)}|, \tag{26}$$

with $\hat{W}(p)$ and $W^*(p)$ defined as in (20) and (25), respectively. By repeating the above procedure a large number of times, $B^* = 1,000$, we obtain 1,000 values of $z^*, z_1^*, z_2^*, \ldots, z_{1,000}^*$. The $100(1-\alpha)\%$ percentile of the distribution of z^* can be approximated by $z_{[1000(1-\alpha)]}^*$, which is the $[B^*(1-\alpha)]$th element in the ordered list of $z_1^*, z_2^*, \ldots, z_{1000}^*$. Thus, we approximate $z_{1-\alpha}$ by $z_{[1000(1-\alpha)]}^*$.

In the nonparametric bootstrap approach, the set $\mathcal{S}_0^*$ is obtained as follows: Sample with replacement one K_0 dimensional vector from the N_0 sets in $\mathcal{S}_0$ and repeat this sampling N_0 times. The sets $\mathcal{S}_{11}^*$ and $\mathcal{S}_{12}^*$, respectively, are obtained from the sets $\mathcal{S}_{11}$ and $\mathcal{S}_{12}$ in a similar fashion. The nonparametric bootstrap confidence bands are then constructed using the methodology outlined in the preceding paragraph.

4.2 An Asymptotic Representation of z

We approximate the distribution of z asymptotically when N_0 is large. Let $C_0 \equiv p_1 < p_2 < \ldots < p_m < p_{m+1} < \ldots < p_M \equiv C_1$ be a partition of the interval $[C_0, C_1]$. In the Appendix, which can be found at http://computer.org/tpami/archives.htm, we show that

$$\begin{aligned} z &\equiv \max_{C_0 < p < C_1} \sqrt{N_0}|\sin^{-1}\sqrt{\hat{W}(p)} - \sin^{-1}\sqrt{W(p)}| \\ &\approx \max_{1 \leq m \leq M} |D_M \cdot \hat{G}_{0,M} + D_M \cdot \hat{G}_{1,M}|, \end{aligned} \tag{27}$$

where D_m is a diagonal matrix with the (m,m)-th entry given by $1/\sqrt{4W(p_m)(1-W(p_m))}$, $D_M \cdot \hat{G}_{0,M}$ and $D_M \cdot \hat{G}_{1,M}$ are independent of each other, the distribution of $D_M \cdot \hat{G}_{0,M}$ (respectively, $D_M \cdot \hat{G}_{1,M}$) is approximately a M-dimensional multivariate normal with mean 0 (respectively, 0) and covariance matrix given by Γ_0 (respectively, Γ_1) given in (58)

in the Appendix, which can be found at http://computer.org/tpami/archives.htm. The maximum in $[C_0, C_1]$ is approximated by the component of the multivariate normal that takes on the maximum absolute value. We define

$$\max_{1\le m\le M} |D_M \cdot \hat{G}_{0,M} + D_M \cdot \hat{G}_{1,M}| \equiv z_M. \quad (28)$$

The distribution of z is approximated by the distribution of z_M for large M. Denoting the $100(1-\alpha)\%$ percentile of z_M by $z_{1-\alpha,M}$, the $100(1-\alpha)\%$ confidence interval for $W(p)$ is given by $(e_L(p), e_U(p))$, where

$$e_L(p) = \left(\sin\left(\sin^{-1}\sqrt{\hat{W}(p)} - z_{1-\alpha,M}/\sqrt{N_0}\right)\right)^2$$

and

$$e_U(p) = \left(\sin\left(\sin^{-1}\sqrt{\hat{W}(p)} - z_{1-\alpha,M}/\sqrt{N_0}\right)\right)^2. \quad (29)$$

4.3 Testing the Claim of a Biometric Vendor

Suppose that a vendor of a biometric authentication system claims that his/her biometric authentication system has a ROC curve given by $ROC_c = (p, W_c(p))$, for p in some interval $[C_0, C_1]$. Based on acquisitions from n subjects, we can test the validity of this claim by generating our own genuine and impostor similarity scores and obtaining the $100(1-\alpha)\%$ confidence band for the true ROC curve, $(p, W(p))$, for $p \in [C_0, C_1]$. We assume that the subjects as well as the scores generated from the subjects in the vendor's database are a representative sample from the underlying population of subjects and the corresponding distributions of genuine and impostor scores derived from this population. If this assumption is true, then the confidence bands constructed from the previous section can be used for validating the vendor's claim. We perform the test

$$H_0 : W(p) = W_c(p) \quad \text{versus} \quad H_1 : W(p) \neq W_c(p), \quad (30)$$

for some p, and will accept H_0 (the claimed ROC curve) if

$$e_L(p) \le W_c(p) \le e_U(p) \quad (31)$$

for all $p \in (C_0, C_1)$; otherwise, we will reject it. We can also perform a test for claims of specific values of FRR and FAR, FRR_c, and FAR_c. At $p_c = FAR_c$, we obtain the upper and lower limits of $GAR(p_c)$, $GAR_L(p_c)$, and $GAR_U(p)$. We will accept the claimed error rates if

$$GAR_L(p_c) \le GAR_c \le GAR_U(p_c), \quad (32)$$

where $GAR_c = 1 - FRR_c$ and reject it otherwise.

5 Experimental Results

We evaluate the methodology developed in the previous sections for biometric authentication systems based on fingerprints. For evaluation purposes, it is necessary that the fingerprint databases consist of multiple impressions of a finger as well as impressions from several different fingers for each subject. Many publicly available databases do not meet these requirements and as a result, we focused on two databases that were appropriate for our purpose, namely, a database consisting of fingerprint impressions collected in our laboratory, and a different database obtained from West Virginia University.

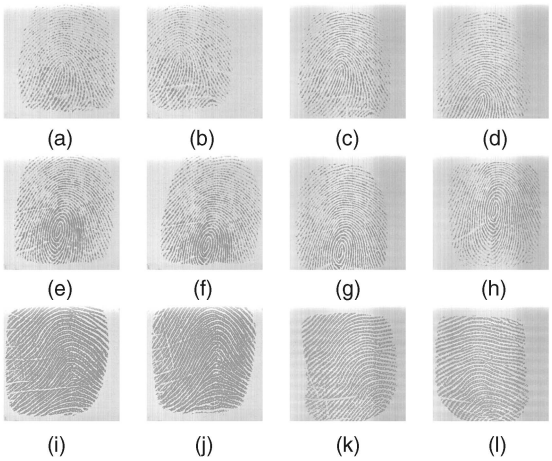

Fig. 3. Examples of fingerprint impressions from [8]: Each row gives the four impressions per finger collected. The first two rows are different fingers from the same subject, whereas the last row contains fingerprint impressions from a different subject.

The Michigan State University (MSU) database [8] consists of fingerprint impressions from four different fingers (the right index, right middle, left index, and left middle fingers) of 160 users. A total of four impressions per finger were obtained; two impressions were obtained on the first day and the remaining two after a period of a week. The fingerprint images were acquired using a solid state sensor manufactured by Veridicom, Inc, with image sizes 300×300 and resolution 500 dpi. Fig. 3 show all four impressions of three fingers in this database. The first two fingers (first two rows) are from the same subject, whereas the images in the last row are from a different subject. A fingerprint similarity score was generated using an asymmetric matcher, described in [17]. All raw scores ranged between 0 and 1,000 and, thus, the transformation (9) with $a = 0$ and $b = 1,000$ was used to convert the scores onto the real line. All subsequent analysis was performed on the transformed similarity scores. Thus, we have the following values for N and K (with $n = 160$, $c = d = 4$): $N = 160$ and dimensionality $K = 4 \times 4 \times 3 = 48$ for the set of genuine scores, $N = 160$ and $K = 4 \times 3 \times 4^2 = 192$ for the set of intrasubject impostor scores, and $N = 160 \times 159 = 25,440$ and $K = 4^2 \times 4^2 = 256$ for the set of intersubject impostor scores. The number of parameters in the correlation matrices that need to be estimated for the genuine, intrasubject impostor and intersubject impostor scores are, respectively, $(48\times47)/2 = 1,128$, $(192\times191)/2 = 18,336$, and $(256\times255)/2 = 32,640$. The number of parameters far exceeds the total number of observations in each of the three sets of scores. In order to avoid overfitting, we reduce the value of K in each case. Instead of selecting all four fingers, we choose only $c = 2$, namely, the right index and right middle fingers, and use the $d = 2$ impressions per finger obtained on the first day. In this case, the number of parameters that need to be estimated are 6, 28, and 120 for the genuine, intrasubject and intersubject impostor sets of scores, respectively.

The West Virginia University (WVU) fingerprint database consists of fingerprint impressions from 263 different users. We used the first two impressions of the right index finger to obtain similarity scores with the same matcher as above; thus, $c = 1$ and $d = 2$ for the WVU database. Consequently, there is

TABLE 2
Values of n, c, and d for the MSU and WVU Databases Used in the Experiments

Databases	n	c	d
MSU	160	2	2
WVU	263	1	2

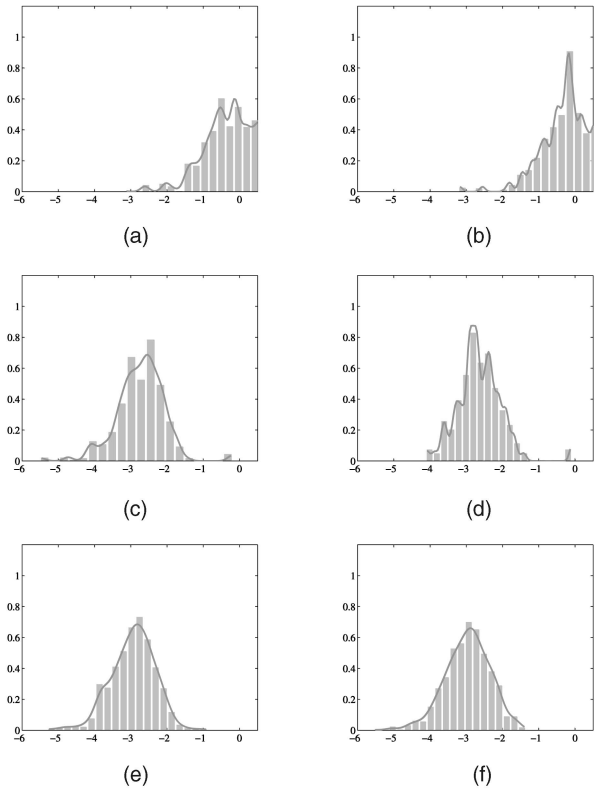

Fig. 4. Fitted density functions (solid line) for the (a) and (b) genuine, (c) and (d) intrasubject, and (e) and (f) intersubject marginal distributions. (a), (c), and (e) $k = 1$ and (b), (d), and (f) $k = 2$.

only one kind of impostor score, namely, the intersubject impostor score for this database. Table 2 gives the number of subjects (n), as well as the values of c (number of different fingers per subject) and d (number of impressions per finger) for the MSU and WVU databases.

5.1 Estimating the Joint Distribution of Similarity Scores

In order to estimate the joint distribution, F, of similarity scores corresponding to the genuine, intrasubject and intersubject impostor sets, we first need to estimate each marginal F_k, $k = 1, 2, \ldots, K$ and correlation matrix R from observed data. The estimation of F_k and R are described in detail in the Appendix, which can be found at http://computer.org/tpami/archives.htm and in [6]. We show the results of the nonparametric estimation procedure for the first two marginal distributions corresponding to each of the genuine, intrasubject impostor and intersubject impostor scores for the MSU database (see Fig. 4). Note the very good agreement between the observed density histogram and the fitted density curve for each figure, especially at the tails of the distributions. A good fit at the tails is essential for the construction of a valid ROC curve that accurately reflects the authentication performance based on the observed data of similarity scores.

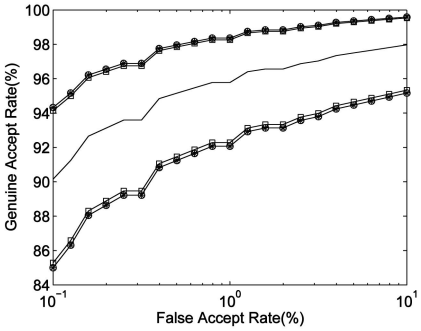

Fig. 5. Upper and lower ROC envelopes obtained using the three different methods: The nonparametric, semiparametric bootstrap, and asymptotic envelopes are represented by the symbols ∘, □, and ∗, respectively. The middle solid line is the nonparametric ROC curve.

The estimate of the genuine correlation matrix (of dimension 4×4) is given by

$$\hat{R}_0 = \begin{pmatrix} 1.00 & 0.99 & 0.15 & 0.16 \\ 0.99 & 1.00 & 0.15 & 0.16 \\ 0.15 & 0.15 & 1.00 & 0.99 \\ 0.16 & 0.16 & 0.99 & 1.00 \end{pmatrix}. \quad (33)$$

The ordered row (and column) dimensions 1, 2, 3, and 4, respectively, represents the scores $s(B_{1,1}, B_{1,2})$, $s(B_{1,2}, B_{1,1})$, $s(B_{2,1}, B_{2,2})$, and $s(B_{2,2}, B_{2,1})$; recall that $c = 2$ and $d = 2$. Consequently, the off-diagonal entries of (33) give the correlation between the corresponding row and column dimensions. For example, the entry 0.15 in the second row and third column of matrix $\hat{R}_0$ is the correlation between between $s(B_{1,1}, B_{1,2})$ and $s(B_{2,1}, B_{2,2})$. The off-diagonal entries of $\hat{R}_0$ indicate that there is a significant amount of correlation in the set of genuine similarity scores. We also obtained estimates of the intrasubject (of dimension 8×8) and intersubject (of dimension 16×16) correlation matrices in a similar fashion (see the Appendix, which can be found at http://computer.org/tpami/archives.htm). We also developed an assessment of fit of the copula functions to the observed data and found that the estimated Gaussian copula functions are a good fit to each of the genuine, intrasubject and intersubject impostor sets of similarity scores. The methodology and related plots are presented in the Appendix, which can be found at http://computer.org/tpami/archives.htm.

5.2 Construction of the ROC Confidence Bands

The 95 percent ROC confidence bands are constructed based on the semiparametric bootstrap, asymptotic, and the nonparametric bootstrap approaches for the MSU and WVU databases. The resulting upper and lower bounds of all the three approaches closely match with each other for the two databases; due to space restrictions, we only show the bands for the MSU database in Fig. 5. Fig. 5 shows that the semi-parametric bootstrap and the asymptotic approaches give good approximations to the true upper and lower confidence bands even for moderate sample sizes.

TABLE 3
Different Values of $\hat{\rho}_1$ and $\hat{\rho}_2$ for the Genuine, Intrasubject Impostor, and Intersubject Impostor Similarity Scores, as Well as the Different Dimensions of $R(\rho_1)$ and $R(\rho_2)$ for an Asymmetric Matcher

Sets/Estimates	$\hat{\rho}_1$	$\hat{\rho}_2$	$\dim R_*(\rho_1)$	$\dim R_*(\rho_2)$
Genuine	0.15	0.99	c	$d(d-1)$
Intra-Subject Impostor	0.80	0.27	$c(c-1)$	d^2
Inter-Subject Impostor	0.26	0.55	c^2	d^2

5.3 Effects of Correlation on the ROC Confidence Bands

Our next set of experiments consist of studying the effect of correlation among the multiple impressions of a user on the width of the ROC confidence band. Since this requires varying the correlation, this experiment is not possible using real data since real data would give only one estimate of correlation for each of the sets of genuine, intrasubject and intersubject impostor similarity scores. Instead, our experiment is based on simulated sets of genuine, intersubject impostor and intrasubject impostor similarity scores from the multivariate Gaussian K-copula models with Toeplitz forms for the correlation matrix. Let

$$R_*(\rho) = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix} \tag{34}$$

denote the correlation matrix with all off-diagonal entries equal to ρ. The dimension of $R_*(\rho)$ will be different according to whether the sets of scores are genuine, intrasubject or intersubject impostor scores.

For a genuine set, the parameterization of the correlation matrix as $R \equiv R_*(\rho_1) \otimes R_*(\rho_2)$ implies that the correlation between any two components of $\underline{s}(f)$ corresponding to finger f is ρ_2 and the correlation between a component of $\underline{s}(f)$ and a component of $\underline{s}(f')$ for two different fingers, $f \neq f'$, is $\rho_1 \cdot \rho_2$. For an intrasubject impostor set, the paramterization of the correlation matrix implies that the correlation between any two components of $\underline{s}(f, f')$ for each pair (f, f') is ρ_2 and the correlation between a component of $\underline{s}(f, f')$ and a component of $\underline{s}(g, g')$ for two different pairs, $(f, f') \neq (g, g')$, is $\rho_1 \cdot \rho_2$. For an intersubject impostor set, the parameterization implies that the correlation between any two pairs of components in $\underline{s}(f, f')$ is ρ_2 and the correlation between a component of $\underline{s}(f, f')$ and a component of $\underline{s}(g, g')$ for two different pairs, $(f, f') \neq (g, g')$, is $\rho_1 \cdot \rho_2$.

One advantage of selecting correlation matrices to be of the form $R \equiv R_*(\rho_1) \otimes R_*(\rho_2)$ is that the matrices can be determined from specifying only two real numbers, ρ_1 and ρ_2, and is therefore, easy to use for illustrative purposes. For a given estimated correlation matrix $\hat{R}$, we find the values of ρ_1 and ρ_2 that minimize the sum of Euclidean distances between the entries of $\hat{R}$ and $R_*(\rho_1) \otimes R_*(\rho_2)$,

$$\| \hat{R} - R_*(\rho_1) \otimes R_*(\rho_2) \|^2, \tag{35}$$

where $R_*(\rho_1)$ and $R_*(\rho_2)$ are as in (34) with ρ_1 and ρ_2 plugged in for ρ, respectively, and $\otimes$ is the Kronecker delta product. The minimizers of ρ_1 and ρ_2, $\hat{\rho}_1$ and $\hat{\rho}_2$, for each of the genuine, intrasubject impostor and intersubject impostor sets of scores, as well as the dimensions of each of $R_*(\rho_1)$ and $R_*(\rho_2)$ are given in Table 3 for the MSU database. For the WVU database, the estimated values of ρ_2 was found to be 0.99 and 0.39, respectively, for the genuine and impostor sets of similarity scores.

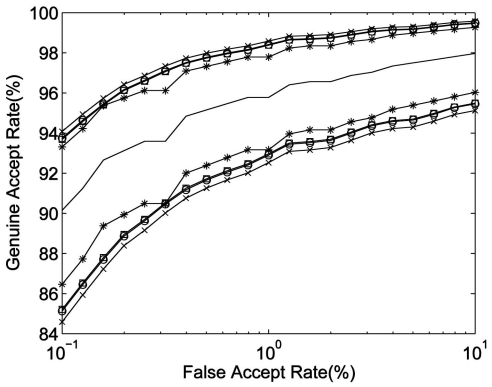

Fig. 6. Effects of correlation on the ROC confidence bands. The lines with "*," □, ○, and ×, respectively, denote the four different combinations of intrafinger and interfinger correlations 1, 2, 3, and 4.

In order to show the effects of increasing correlation on the ROC confidence bands, four combinations of (ρ_1, ρ_2) were selected. The first three combinations are:

1. $(\rho_1 = 0, \rho_2 = 0)$,
2. $(\rho_1 = 0, \rho_2 = \hat{\rho}_2)$, and
3. $(\rho_1 = \hat{\rho}_1, \rho_2 = \hat{\rho}_2)$, where $\hat{\rho}_1$ and $\hat{\rho}_2$ are selected according to the entries of Table 3 for each set of genuine, intrasubject impostor and intersubject impostor similarity scores.
4. This combination is obtained by setting the genuine ρ_1 to 0.6 and the remaining ρ_1s and ρ_2s selected according to the entries in Table 3.

The 95 percent ($\alpha = 0.05$) level confidence bands for the ROC curve were constructed based on $B^* = 1,000$ bootstrap resamples. Fig. 6 gives the ROC confidence bands based on the semiparametric bootstrap. Note that the width of the confidence bands generally increases as we move from combination 1 to 4. The median widths of the confidence bands for the four combinations are 4.62, 5.41, 5.51, and 6.06, respectively. The effects of correlation on the confidence bands using the asymptotic approach and for the WVU database were similar to the bootstrap approach and, therefore, are not presented here.

5.4 Validation of the ROC Confidence Bands

We conducted several tests to validate the ROC confidence bands at a specified confidence level. Recall that the $100(1-\alpha)\%$ ROC confidence bands, by definition, cover the true ROC curve with a probability of *at least* $100(1-\alpha)\%$ on repeated sampling from the underlying population of similarity scores. Treating the entire MSU database with $n = 160$ subjects as the underlying population, we selected a subset of 120 subjects from this population for constructing the semiparametric bootstrap ROC confidence bands; a subset of 120 subjects (as opposed to smaller subsets of the data) is selected so that estimation of the nonparametric marginal distributions can

be performed reliably. We then determined if the population ROC curve (the empirical ROC curve for the 160 subjects) was within the constructed confidence bands. This procedure was repeated 200 times (with different subsets of 120 subjects from the population of 160), and each time, we determined if the population ROC curve was within the constructed ROC confidence bands. The percentage of coverage based on this validation procedure should be at least $100(1-\alpha)\%$. In our experiments, we selected $\alpha = 0.05$ for the 95 percent ROC confidence bands and obtained a coverage proportion of 99.5 percent. For the WVU database, validation of the ROC confidence bands was carried out with subsamples of 198 users. The procedure of constructing the ROC confidence bands was repeated 500 times. The empirical ROC curve (ROC curve based on the 263 users) was found to be inside the 95 percent confidence bands in 497 (out of the 500) trials, resulting in a coverage probability of 99.4 percent.

5.5 Sample Size Requirements

As correlated multiple biometric observations affect the width of the ROC confidence bands, we now proceed to determine the number of users, n^*, required by a system to report a $100(1-\alpha\%)$ ROC confidence band with a width of at most w. We take $w = 1\%$. Our results are based on simulation with correlations selected according to combinations 1, 2, 3, and 4 in Section 5.3. Thus, the results reported here can be generalized to real data which exhibit different degrees of intrafinger and interfinger correlations. The values of n^* are given for different combinations of c and d and, therefore, varying dimensionality of the genuine, intrasubject and intersubject sets of similarity scores. Consequently, we assume a common marginal for each of the three sets given by the mixture over component scores. We selected $C_0 = 0.1\%$, $C_1 = 10\%$, and $M = 21$ here, and $p_m = 10^{(-1+0.1(m-1))}, m = 1, 2, \ldots, M$. For each $m = 1, 2, \ldots, M$, the width of the ROC confidence band at each $FAR = p_m$ (see (29)) is given by

$$\begin{aligned} w(p_m) &= e_U(p_m) - e_L(p_m) \\ &= \frac{4z_{1-\alpha,M}\sqrt{W(p_m)(1-W(p_m))}}{\sqrt{n}} \end{aligned} \tag{36}$$

for large $n(= N_0)$, where $z_{1-\alpha,M}$ is the $100(1-\alpha)\%$ percentile of the distribution of z_M defined in (28); the second equality is from applying the delta method [15] to $e_U(p_m) - e_L(p_m)$ in (29). In order to determine $z_{1-\alpha,M}$, we first estimate the covariance matrices Γ_0 and Γ_1 (see (59) in the Appendix, which can be found at http://computer.org/tpami/archives.htm) as accurately as possible. This estimation is performed based on 1,000 simulated samples from each of the correlation combinations 1, 2, 3, and 4 for $n = 1,000$ subjects. To achieve a width of w for the confidence band implies that $w(p_m) \leq w$ for all p_m, $m = 1, 2, \ldots, M$. Thus, the minimum number of users required is given by the formula $n^* = n_0 + 1$ where n_0 is the greatest integer less than or equal to

$$\max_{1\leq m\leq M}\left(\frac{4z_{1-\alpha,M}\sqrt{W(p_m)(1-W(p_m))}}{w(p_m)}\right)^2 + 1. \tag{37}$$

We also compare the minimum sample size requirements given by our method to that of the subset bootstrap approach

TABLE 4
Mean n^* and n^*_{sb} Values for Achieving a Width of 1 Percent for the 95 Percent Confidence Band

	Values of c and d					
	$c=1, d=2$		$c=2, d=2$		$c=2, d=3$	
Correlations (ρ_1, ρ_2)	n^* mean (sd)	n^*_{sb} mean (sd)	n^* mean (sd)	n^*_{sb} mean (sd)	n^* mean (sd)	n^*_{sb} mean (sd)
(0,0)	11,443 (246) 22,885 (492)	48,674 (600) 97,350 (1,200)	5,809 (148) 23,235 (590)	24,201 (373) 96,810 (1,493)	1,967 (31) 11,801 (190)	8,143 (136) 48,860 (814)
$(0, \hat{\rho}_2)$	20,439 (790) 40,877 (1,581)	90,725 (315) 181,450 (630)	10,476 (279) 41,905 (1,115)	46,209 (837) 184,840 (3,346)	9,505 (263) 57,028 (1,580)	43,500 (455) 261,000 (2,729)
$(\hat{\rho}_1, \hat{\rho}_2)$	21,403 (1,004) 42,806 (2,008)	90,477 (407) 180,950 (813)	11,056 (346) 44,223 (1,382)	47,855 (430) 191,420 (1,720)	9,749 (163) 58,492 (977)	46,269 (968) 277,620 (5,811)
$(0.6, \hat{\rho}_2)$	19,015 (503) 38,029 (1,006)	89,993 (429) 179,990 (858)	13,321 (506) 53,285 (2,026)	61,394 (884) 245,570 (3,536)	11,558 (423) 69,346 (2,540)	56,723 (826) 340,340 (4,956)

The total number of observations, n^*cd and $n^*_{sb}cd$, are given below the n^* and n^*_{sb} entries, respectively. Entries are calculated as the means of 10 simulation runs. The corresponding standard deviations are given in parenthesis.

[3]. One important point is that [3] obtains confidence rectangles, and not confidence bands, at each threshold value on the ROC curve. In order to perform a valid band to band comparison of the two methods, we applied the subset bootstrap procedure to the alternative parametrization of the ROC curve given in (20). As mentioned earlier, the subset bootstrap is not able to give an overall confidence level of $100(1-\alpha)\%$ using M individual $100(1-\alpha)\%$ confidence intervals. To guarantee a $100(1-\alpha)\%$ confidence level, the level of each individual confidence interval would have to be $100(1-\alpha/M)\%$ using Bonferroni's inequality. For $m = 1, 2, \ldots, M$, the minimum sample size requirement, $n_{sb}(m)$, for the mth confidence interval can be obtained using similar asymptotic arguments as in Section 4.2 with $C_0 = C_1 = p_m$. It follows that the minimum sample size required to achieve the prespecified width for all M confidence intervals is given by

$$n^*_{sb} = \max_{1\leq m\leq M} n_{sb}(m). \tag{38}$$

Table 4 reports the average n^* and n^*_{sb} over 10 simulation runs with the numbers below n^* (respectively, n^*_{sb}) representing the average total number of observations n^*cd ($n^*_{sb}cd$). The numbers in the parenthesis are the corresponding standard deviations over the 10 runs. If a biometric authentication system was tested based on n users, where n is chosen according to the n^* entries in Table 4, we will be 95 percent certain that the true ROC *curve* will lie in the interval $[\hat{W} - 0.5, \hat{W} + 0.5]$. Table 4 indicates that as the correlation among the multiple impressions of a finger increases for each fixed c and d, the total number of observations needed to achieve the width w for the confidence band increases. The same holds true when c and d values are increased for each correlation combination. Thus, in the presence of nonzero correlation, we are better off selecting a larger number of

TABLE 5
Mean n^* and n^*_{sb} Values for Achieving a Width of 1 Percent for the 95 Percent Confidence Band Based on the West Virginia University Database

	Values of c and d					
	$c=1,d=2$		$c=1,d=3$		$c=1,d=4$	
Correlations $(\rho_2^{gen}, \rho_2^{imp})$	n^* mean (sd)	n^*_{sb} mean (sd)	n^* mean (sd)	n^*_{sb} mean (sd)	n^* mean (sd)	n^*_{sb} mean (sd)
(0,0)	12,875 (283) 25,749 (477)	47,526 (655) 95,050 (1,310)	4,251 (77) 12,754 (231)	16,170 (280) 48,510 (841)	2,103 (37) 8,412 (148)	8,144 (169) 32,580 (676)
(0.49, 0.19)	15,215 (513) 30,430 (1,025)	61,195 (1,074) 122,390 (2,148)	7,719 (215) 23,158 (645)	35,053 (697) 105,160 (2,091)	6,200 (299) 24,799 (1,197)	29,149 (940) 116,600 (3,761)
(0.99, 0.39)	23,802 (886) 47,604 (1,772)	90,334 (170) 180,670 (304)	20,898 (414) 62,693 (1,244)	86,357 (400) 259,070 (1,200)	18,748 (698) 74,991 (2,793)	84,478 (766) 337,910 (3,064)

The total number of observations, n^*cd and $n^*_{sb}cd$, are given below the n^* and n^*_{sb} entries, respectively. Entries are calculated as the means of 10 simulation runs. The corresponding standard deviations are given in parenthesis.

users rather than increasing the number of acquisitions per user. Note that the sample sizes required by our method, n^*, is smaller compared to n^*_{sb} for achieving the same overall confidence level.

We also obtained the minimum sample sizes determined by the "Rule of 3" [20] and the "Rule of 30" [14] (see Appendix, which can be found at http://computer.org/tpami/archives.htm for their derivation). For the fingerprint database [8], n_3 was approximately 150 for all pairs of correlation combination, c and d, while n_{30} was approximately 770. Comparing the values of n_3 and n_{30} with n^*cd, we see that both n_3 and n_{30} grossly underestimate the total number of biometric acquisitions required to achieve a desired width. The underestimation becomes more prominent when significant correlation is present between multiple acquisitions of the biometric templates from a subject.

To illustrate the effects of correlation on the sample size requirement for the WVU database, we choose three combinations of the genuine and impostor within finger correlations, namely, $(\rho_2^{gen}, \rho_2^{imp}) = (0,0), (0.49, 0.19)$ and (0.99, 0.39) to reflect the no correlation (or, independence), intermediate and high correlation states. Table 5 reports the average n^* and n^*_{sb} over 10 simulation runs for the width $w = 1\%$, with the average total number of observations, n^*d and $n^*_{sb}d$ given by the entries directly below the n^*s. The numbers in the parenthesis are the corresponding standard deviations over the 10 runs. Note here, again, that n^* is smaller compared to n^*_{sb} for achieving the same overall confidence level.

6 Conclusion

With the growing deployment of biometric systems in several government and commercial applications, it has become even more important to validate the performance levels of a system claimed by a vendor. We present a flexible semiparametric approach for estimating both the genuine and impostor distributions of similarity scores using multivariate Gaussian copula functions with nonparametric marginals. Confidence bands for the ROC curve are constructed using 1) semiparametric bootstrap re-samples and 2) asymptotic approximations derived from the estimated models. We also determine the minimum required number of subjects needed to achieve a desired width for the confidence band of the ROC curve. Currently, the implementation of the ROC validation procedure and the estimation of required number of samples are based on fingerprint databases with a small number of subjects. We plan to test our methodology on larger databases as they become available. We will also focus on extending the current framework to validate reported performances of multimodal systems.

Acknowledgments

The authors would to thank Karthik Nandakumar, Arun Ross, Umut Uludag, and Yi Chen for their help when conducting the experiments. This research is partially supported by the US National Science Foundation ITR grant 0312646.

References

[1] J.R. Beveridge, K. She, and B.A. Draper, "A Nonparametric Statistical Comparison of Principal Component and Linear Discriminant Subspaces for Face Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* Dec. 2001.
[2] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior, *Guide to Biometrics.* Springer, 2004.
[3] R. Bolle, N. Ratha, and S. Pankanti, "Error Analysis of Pattern Recognition Systems: The Subsets Bootstrap," *Computer Vision and Image Understanding,* vol. 93, no. 1, pp. 1-33, Jan. 2004.
[4] R.M. Bolle, S. Pankanti, and N.K. Ratha, "Evaluation Techniques for Biometrics-Based Authentication Systems (FRR)," *Proc. 14th Int'l Conf. Pattern Recognition,* pp. 2831-2837, Aug. 2000.
[5] U. Cherubini, E. Luciano, and W. Vecchiato, *Copula Methods in Finance.* Wiley, 2004.
[6] S.C. Dass, Y. Zhu, and A.K. Jain, "Validating a Biometric Authentication System: Sample Size Requirements," Technical Report MSU-CSE-05-23, Dept. of Computer Science and Eng., Michigan State Univ., East Lansing, Aug. 2005.
[7] A.K. Jain, L. Hong, and R. Bolle, "On-Line Fingerprint Verification," *IEEE Trans. Pattern Recognition and Machine Intelligence,* vol. 19, no. 4 pp. 302-314, 1997.
[8] A.K. Jain, S. Prabhakar, and A. Ross, "Fingerprint Matching: Data Acquisition and Performance Evaluation," Technical Report TR99-14, Michigan State Univ., East Lansing, 1999.
[9] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis.* Prentice Hall, 1988.
[10] R.J. Micheals and T.E. Boult, "Efficient Evaluation of Classification and Recognition Systems," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* Dec. 2001.
[11] R.G. Miller, *Simultaneous Statistical Inference.* Springer-Verlag, 1981.
[12] D.F. Morrison, *Multivariate Statistical Methods.* McGraw-Hill, 1990.
[13] R.B. Nelsen, *An Introduction to Copulas.* Springer, 1999.
[14] J. Porter, "On the "30 Error Criterion," *Nat'l Biometric Center Collected Works,* J. Wayman, ed., pp. 51-56, 2000.
[15] C.R. Rao, *Linear Statistical Inference and Its Applications.* Wiley, 1991.
[16] M.E. Schuckers, "Using the Beta-Binomial Distribution to Assess Performance of a Biometric Identification Device," *Int'l J. Image and Graphics,* special issue on biometrics, vol. 3, no. 3, pp. 523-529, July 2003.

[17] "Best Practices in Testing and Reporting Performance of Biometric Devices," *U.K. Biometrics Working Group,* www.cesg.gov.uk/technology/biometrics, 2000.
[18] J. Wayman, "Technical Testing and Evaluation of Biometric Identification Devices," *Biometrics: Personal Identification in Networked Society,* A.K. Jain, R. Bolle, and S. Pankanti, eds. Kluwer Academic, 1999.
[19] J. Wayman, "Confidence Interval and Test Size Estimation for Biometric Data," *Nat'l Biometric Center Collected Works,* J. Wayman, ed., pp. 91-95, 2000.
[20] J. Wayman, "Technical Testing and Evaluation of Biometric Identification Devices," *Nat'l Biometric Center Collected Works,* J. Wayman, ed., pp. 67-89, 2000.

Sarat C. Dass received the MSc degree and the PhD degree in Statistics from Purdue University, in 1995 and 1998, respectively. He is currently an assistant professor in the Department of Statistics and Probability and an adjunct assistant professor in the Department of Computer Science and Engineering at Michigan State University. His research and teaching interests include statistical image processing and pattern recognition, shape analysis, spatial statistics, Bayesian computational methods, foundations of statistics, and nonparametric statistical methods. He is a member of the IEEE and the IEEE Computer Society.

Yongfang Zhu received the BS degree in 2002 from Nankai University, China. She is currently a PhD student in the Department of Statistics & Probability and a master's student in the Department of Computer Science and Engineering at Michigan State University. Her research interests are in the areas of statistical pattern recognition, biometric authentication, and data mining. She is currently working on various aspects of statistical modeling in biometric authentication systems with applications to assessing the extent of uniqueness of fingerprints and developing performance evaluation platforms for biometric systems.

Anil K. Jain received the BTech degree from the Indian Institute of Technology, Kanpur in 1969 and the MS and PhD degrees from the Ohio State University in 1970 and 1973, respectively. He is a University Distinguished Professor in the Departments of Computer Science & Engineering, Electrical & Computer Engineering and Statistics & Probability at Michigan State University. His research interests are in the areas of cluster analysis, feature selection and extraction, Markov random field models for texture, 3D object recognition, and deformable models for object representation and matching. In addition to his theoretical contributions, he has been active in solving a variety of practical pattern recognition problems, including medical image analysis, document image understanding, remote sensing, and biometric authentication. He is a fellow of the IEEE.