A Semi-parametric Approach to Estimation of ROC Curves for Multivariate Binormal Mixtures

Sarat C. Dass[1] **and Seong W. Kim**[2]

Abstract

A Receiver Operating Characteristic (ROC) curve reflects the performance of a system which decides between two competing actions in a test of statistical hypothesis. This paper addresses the inference on ROC curves for the following problem: how can one statistically validate the performance of a system with a claimed ROC curve, ROC_0 say? Our proposed solution consists of two main components: First, a flexible family of distributions, namely the multivariate binormal mixtures, is proposed to account for intra-sample correlation and non-Gaussianity of the marginal distributions under both the null and alternative hypotheses. Second, a semi-parametric inferential framework is developed for estimating all unknown parameters based on a rank likelihood. Actual inference is carried out by running a Gibbs sampler until convergence, and subsequently constructing a highest posterior density (HPD) set for the true but unknown ROC curve based on the Gibbs output. Real data are analyzed to support out theoretical results.

Keywords: Bayesian computation, group invariance, mixture models, Semi-parametric inference, ROC band.

1 Introduction

The Receiving Operating Characteristics (ROC) curve is a popular tool for assessing the performance of a system which decides between two competing actions in a test of statistical hypotheses. A large variety of fields (for example, engineering, biology, genetics, finance and others) require the development of ROC curves for systems performance assessment. For this reason, the study of construction and inference on ROC curves has received a good deal of attention in the above-mentioned fields. Studies related to inference on ROC curves include DeLong et al. (1988), Alonzo and Pepe (2002), Cai (2004), and Braun and Alonzo (2008) in biological applications, and Kamitsuji and Kamatani (2006) in genetics.

[1] *Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA*

[2] *Division of Applied Mathematics, Hanyang University, Ansan, 426-791, South Korea*

The aim of this paper is to develop the inferential tool for validating ROC curves in engineering applications, with special emphasis on biometrics and biometric authentication systems. Biometric recognition or biometrics refers to the automatic authentication of a person based on his/her physiological or behavioral characteristics (see, for example, Jain et al. 1999, and Maltoni et al. 2003), and is gaining widespread use due to security and safety concerns. Biometric recognition offers many advantages over traditional PIN number or password and token-based approaches: the rightful owner of the biometric template can be easily identified, a biometric trait cannot be easily transferred, forgotten or lost, and it is difficult to duplicate a biometric trait. Some well-known examples of traits used in biometric recognition are fingerprint, iris, face, signature, voice, hand geometry, retina, and ear.

A number of commercial recognition systems based on biometric traits has been deployed. Often, it is necessary to ascertain the claim of a biometric vendor that the system in question has performance given by the ROC curve ROC_0. This calls for inference based on test samples, and in particular, the construction of confidence bands at a pre-specified level, say $100(1-\alpha)\%$ for some $0 < \alpha < 1$, for the true but unknown ROC curve. Once the ROC confidence bands are constructed, one can determine the validity of the vendor's claim at $100(1-\alpha)\%$ level by checking whether ROC_0 is within the derived confidence bands or not.

A number of challenges must be addressed when constructing confidence bands for the ROC curve. First, multiple values of the test statistic used to either accept or reject the null hypothesis is based on the same test sample, and therefore, are correlated; an example of this scenario is given in Section 5. Indeed, many earlier efforts to validate the performance of a biometric system assume that the multiple acquisitions of the test statistic are independent of each other (see, for example, Bolle et al. (2000)), and this assumption entails that the true coverage probability of the confidence bands is lower than $100(1-\alpha)\%$ (see Section 5). The second challenge is to construct confidence bands for a *curve*. There are many examples of previous methodologies that construct confidence *intervals* for pre-specified values of the false accept rates (FARs), and then combine these intervals to obtain a confidence band for the ROC curve. In presence of correlated observations, Bolle et al. (2004) introduced the subsets bootstrap approach to construct these confidence intervals whereas Schuckers (2003) proposed the beta-binomial family to model the correlation between the multiple biometric acquisitions as well as to account for varying false reject rate (FRR) and FAR values for different subjects. A well-known problem of combining confidence intervals in this way is that of multiple comparisons: The confidence level of the combined $100(1-\alpha)\%$ confidence intervals is not

$100(1-\alpha)\%$; in fact, it is much lower (see Section 5). A Bonferroni type correction can be done in this case. However, the resulting bands turn out to be unnecessarily large. Horv*á*th et *al.* (2008), for example, compute confidence bands for ROC curves based on smoothed bootstrap methods and the Bonferroni inequality.

Dass et *al.* (2006) presents a new method for constructing confidence regions for the ROC curve that alleviates the multiple comparison problem without resorting to the Bonferroni inequality. However, the essence of Dass et *al.* (2006) is still to combine confidence intervals corresponding to a number of pre-specified values of the FAR, and therefore, is not an approach that generates confidence *curves*. Another challenge that has to be faced is the problem of non-Gaussianity where the distributions of the test statistic under the null and alternative hypotheses are highly non-Gaussian. This calls for distributional models for the observations that are flexible, in that they can represent a wide range of distributional characteristics in a variety of contexts. It is important to note that the challenges mentioned here are not unique to biometric authentication, and in fact, are faced in many fields where inference on ROC curves is sought. The proposed methodology is thus relevant and applicable in a variety of disciplines.

The rest of the paper is organized as follows. In Section 2, we introduce the proposed models for deriving the ROC confidence bands, and notation that will be used throughout the paper. The Bayesian inferential framework is presented in Section 3, along with the appropriate likelihood and prior distributions. Section 4 presents the computational schemes and outlines steps to construct the $100(1-\alpha)\%$ highest posterior density (HPD) sets for the ROC curve based on samples from the Gibbs output. Section 5 provides experimental results for real data from fingerprint based authentication. A summary and concluding remarks are provided in Section 6, whereas theoretical results and their proofs are provided in the Appendix.

2 The statistical tests of hypotheses

We consider the following hypotheses testing problem in biometric authentication with the aim of recognizing the individual based on an input biometric trait. Let I_0 be the true (but unknown) identity of an individual who provides a biometric query Q and a claimed identity I_c. Consider the test of hypotheses

$$H_0 : I_0 \neq I_c \;\; \text{vs.} \;\; H_1 : I_0 = I_c, \tag{1}$$

where the null hypothesis H_0 states that the user is an impostor and the alternative H_1 implies that the user is a genuine. The test is proceeded with matching the query Q with the template T of the claimed identity I_c in the database based on a similarity measure $S(Q,T)$. Large (respectively, small) values of S implies that T and Q are similar (respectively, dissimilar) to each other. For a pre-specified threshold λ, the decision (or action) is to reject H_0 if $S(Q,T) > \lambda$ and accept H_0 otherwise. It is well known that there can be two types of errors associated with the hypotheses in (1): the false accept and the false reject rates, FAR and FRR, respectively. The FAR is the Type I error probability (i.e., the probability of rejecting H_0 when H_0 is true) which corresponds to erroneously concluding the user is genuine when in fact the user is an impostor. The FRR is the Type II error probability (i.e., the probability of accepting H_0 when H_1 is true) which concludes that the user is an impostor when in fact the user is genuine. Subsequently, the genuine accept rate (GAR) is $1 - FAR$ (which is the probability that the user is accepted given that he/she is genuine) corresponds to the power of the test based on $S(Q,T)$. The ROC curve is the plot of (FAR, GAR) for varying threshold values λ, and reflects the relationship between the FAR versus the GAR. In symbols,

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

for all values of $0 < \lambda < \infty$. The ROC is often used to assess the performance of a (biometric authentication) system: System 1 is said to perform better compared to System 2 in the range of thresholds $\lambda_1 \leq \lambda \leq \lambda_2$ if $ROC_1(\lambda) > ROC_2(\lambda_2)$ for all $\lambda_1 \leq \lambda \leq \lambda_2$. It is convenient to re-parameterize the ROC curve in terms of the $FAR = t$, say, with $0 < t < 1$. Solving for λ in equation (2) and assuming all distribution functions are strictly monotone, we get the following parametrization of the ROC curve

$$ROC(t) = (t, GAR(FAR^{-1}(t))) \tag{3}$$

in terms of t; in (3), $FAR^{-1}(t)$ is the unique value of λ for which $FAR(\lambda) = t$.

3 Multivariate Binormal Mixture Distributions

The aim of this section is to develop a parametric form for the ROC curve while taking into account the highly non-Gaussian distributions of $S(Q,T)$ under H_0 and H_1. One such preliminary assumption is that of binormality. Let X and Y denote random variables representing the values of $S(Q,T)$ under H_0 and H_1, respectively. The assumption of binormality entails that there exists a monotone increasing transformation τ such that: (1) $\tau(X)$ follows a standard normal distribution, and (2) $\tau(Y)$ follows a normal distribution with mean μ and variance σ^2. The attractiveness of the binormality assumption is that it alleviates the problem of non-Gaussian distributions while allowing us to retain a parametric form for the ROC curve. This can be seen as follows: the distribution of X, represented by the df F say, is transformed to a standard normal distribution by choosing $\tau(x) = \Phi^{-1} \circ F(x)$ where Φ is the df of the standard normal distribution. If, further, the distribution function of Y, G say, satisfies

$$G \circ \tau^{-1}(y) \equiv \Phi((y - \mu)/\sigma), \tag{4}$$

then the binormality assumption becomes valid. McClish (1989) shows that the ROC curve under the binormality assumption is given by

$$R(t) = \Phi(a + b\Phi^{-1}(t)), \tag{5}$$

where $a = \mu/\sigma$, $b = 1/\sigma$. Inference under the binormality assumption has been carried out in a number of earlier studies based on the likelihood of the data; see, for example, Swets (1986), Hanley (1989), Hsieh and Turnbull (1996), and Metz et *al.* (1998) for the related studies. A considerable amount of work is concentrated on a rank-based likelihood and various methodologies are proposed on this likelihood (cf. Dabrowska and Doksum, 1998; Zou and Hall, 2000; Alonzo and Pepe, 2002; Cai and Moskowitz, 2004). Recently, Gu and Ghosal (2008) propose a Bayesian method based on a rank likelihood to obtain consistent estimators under mild conditions.

A major limitation of the binormality assumption is equation (4), which states that after transformation, the Y-data is also normally distributed with some mean and variance. In this paper, we aim to retain the attractive parametrization of the ROC curve in (5) under the binormality assumption while doing away with the limitation in (4). A flexible family of distributions that can represent a variety of distributional forms is the mixture of normal distributions. Thus, we impose a less stringent requirement that

$$G \circ \tau^{-1}(y) \equiv \sum_{k=1}^{K} p_k \, \Phi((y - \mu_k)/\sigma_k), \tag{6}$$

where K is the number of mixture components, p_k $(k = 1, 2, \ldots, K)$ is the mixing probability with $0 < p_k < 1$ and $\sum_{k=1}^{K} p_k = 1$, and μ_k and σ_k are, respectively, the mean and standard deviation of the k-th normal component for $k = 1, 2, \ldots, K$. To maintain identifiability with respect to the labeling of the K components, the constraint

$$\mu_1 < \mu_2 < \cdots < \mu_K \tag{7}$$

is maintained. We state

Theorem 1 Under the assumption in (6), the ROC curve has a parametric form given by

$$ROC(t) = \sum_{k=1}^{K} p_k \Phi(a_k + b_k \Phi^{-1}(t)), \tag{8}$$

where $a_k = \mu_k/\sigma_k$,$b_k = 1/\sigma_k$, independent of the monotone increasing function τ.

The reader is referred to the proof in the Appendix.

Correlated observations arise in the derivation of the ROC curve in at least two ways: first, a common biometric input Q can be matched with two or more templates, T and T', to obtain the similarity measures $S(Q,T)$ and $S(Q,T')$, in which case there is a significant correlation between $S(Q,T)$ and $S(Q,T')$. Second, there is a significant correlation when matching multiple impressions of the same finger: $S(Q,T)$ and $S(Q',T')$ are correlated if Q and Q' are impressions from one finger and T and T' are impressions from another (possibly different) finger. In the latter case, the genuine correlation (corresponding to the same finger) is usually significantly larger than the impostor (corresponding to different fingers) case. Any multivariate distribution elicited for the similarity measures $S(Q,T)$ must also be exchangeable, and therefore, these similarity measures should possess common marginals. This can be argued as follows: in the first scenario above, whether Q is first matched with T and then T' to obtain the vector $(S(Q,T), S(Q,T'))^T$ should be equivalent to matching with T' first and then T. One can also make a similar argument in the second scenario for the vector $(S(Q,T), S(Q,T'), S(Q',T), S(Q',T'))^T$.

The binormality assumption can be generalized to the multivariate case to incorporate correlation as follows: the random vectors $\mathbf{X} = (X_1, X_2, \cdots, X_r)'$ and $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_s)'$, with possibly different integers r and s are said to possess the multivariate bi-normal mixture property if there exists a common componentwise monotone increasing transformation τ such that

$$\tau(\mathbf{X}) \equiv \begin{pmatrix} \tau(X_1) \\ \tau(X_2) \\ \vdots \\ \tau(X_r) \end{pmatrix} \equiv \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_r \end{pmatrix} \equiv \mathbf{Z} \quad \sim \quad \mathrm{N}_r(0, B_0), \tag{9}$$

where $\mathrm{N}_r(0, B_0)$ denotes a multivariate normal distribution (of dimension r) with mean vector 0 and correlation matrix B_0, and simultaneously,

$$\tau(\mathbf{Y}) \equiv \begin{pmatrix} \tau(Y_1) \\ \tau(Y_2) \\ \vdots \\ \tau(Y_s) \end{pmatrix} \equiv \begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_s \end{pmatrix} \equiv \mathbf{W}, \tag{10}$$

with $\mathbf{W}$ having a mixture of multivariate normal pdfs:

$$g(\mathbf{w}) = \sum_{k=1}^{K} p_k \phi_s(\cdot|\boldsymbol{\nu}_k, \Sigma_k), \tag{11}$$

where $\phi_s(\cdot\,|\,\boldsymbol{\nu}_k, \Sigma_k)$ is the pdf of an s-variate normal distribution with mean vector $\boldsymbol{\nu}_k$ and covariance matrix Σ_k. Here $\boldsymbol{\nu}_k = (\mu_k, \mu_k, \ldots, \mu_k)'$ and $\Sigma_k = \sigma_k^2 \cdot B_1$, where B_1 is a correlation matrix. Exchangeability necessitates that the correlation matrices B_0 and B_1 have the forms of

$$B_j = (1 - \rho_j)\mathbf{I}_{s_j} + \rho_j \, \mathbf{1}_{s_j} \mathbf{1}_{s_j}' \tag{12}$$

for $j = 0, 1$ where in (12), $\mathbf{I}_{s_j}$ is the identity matrix of dimension $s_j \times s_j$ with $s_0 = r$ and $s_1 = s$, and $\mathbf{1}_{s_j}$ is the unit vector of dimension $s_j \times 1$. Note that ρ_j is restricted to be in the range $-1/(s_j - 1) \leq \rho_j \leq 1$ in order for B_j to be positive definite.

3.1 An Invariance Property

Assume that the set of all observations consists of m independent and identically distributed (iid) copies of $\mathbf{X}$, denoted by $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ where $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ir})'$, and n iid copies of $\mathbf{Y}$, $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$, where $\mathbf{Y}_l = (Y_{l1}, Y_{l2}, \ldots, Y_{ls})'$. The distributions of $\mathbf{X}$ and $\mathbf{Y}$ are assumed to satisfy the multivariate binormal mixture assumption given in equations (9) and (11). The group of transformations

$$\mathcal{G} = \{\, \tau \; : \; \tau \text{ is a monotone increasing function}\} \tag{13}$$

leaves the ROC curve (see Theorem 1) invariant under the group action of function composition. To find the maximal invariant statistic under this group action, we first develop some notation. For $i = 1, 2, \ldots, m$, let $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ir})'$ denote the vector of the transformed X-variables $\mathbf{X}_i$, that is $\mathbf{Z}_i = \tau(\mathbf{X}_i)$. Similarly, define $\mathbf{W}_l = (W_{l1}, \ldots, W_{ls})'$ to be the vector of the transformed Y-variable $\mathbf{Y}_l$, $\mathbf{W}_l = \tau(\mathbf{Y}_l)$, for $l = 1, 2, \ldots, n$. Let $\mathbf{D} = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_m', \mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_n')'$ denote the concatenated vector of observed data, and $\mathbf{q} = (\mathbf{z}_1', \mathbf{z}_2', \ldots, \mathbf{z}_m', \mathbf{w}_1', \mathbf{w}_2', \ldots, \mathbf{w}_n')'$ denote the corresponding transformed variables. Both $\mathbf{D}$ and $\mathbf{q}$ are of length $N = rm + sn$.

Subsequently, we define two reduced statistics in terms of the observation vector $\mathbf{D}$. Let $\mathbf{R}(\mathbf{D}) \equiv (R_{N1}, \ldots, R_{NN})'$ be the ranks of the observations in the vector $\mathbf{D}$ with the convention that rank 1 (respectively, N) denoting the smallest (respectively, largest) observation in the set. Also, denote the vector of labels $\mathbf{L}^{(j)}(\mathbf{D}) \equiv (L_1^{(j)}, L_2^{(j)}, \cdots, L_N^{(j)})'$ for $j = 1, 2$ where the entries of $\mathbf{L}^{(1)}$ are 0 or 1 according to whether the entry corresponds to a X or Y observation. The entries of $\mathbf{L}^{(2)}$ take values in the label set $\{1, 2, \cdots, \max(r, s)\}$ representing the component of the X (or Y) observation the entry came from. Both $\mathbf{L}^{(1)}(\mathbf{q})$ and $\mathbf{L}^{(2)}(\mathbf{q})$ have explicit expressions given as follows: $\mathbf{L}^{(1)}(\mathbf{q}) = (\underbrace{0, 0, 0, \cdots, 0}_{rm \text{ times}}, \underbrace{1, 1, \cdots, 1}_{sn \text{ times}})'$ and

$$\mathbf{L}^{(2)}(\mathbf{q}) = (\underbrace{1, 2, \cdots, r, 1, 2, \cdots, r, \cdots, 1, 2, \cdots, r}_{\text{length } rm}, \underbrace{1, 2, \cdots, s, 1, 2, \cdots, s, \cdots, 1, 2, \cdots, s}_{\text{length } sn})'.$$

Under the multivariate binormality mixture assumption, the ranks and labels of $\mathbf{q}$, $\mathbf{R}(\mathbf{q})$, $\mathbf{L}^{(1)}(\mathbf{q})$ and $\mathbf{L}^{(2)}(\mathbf{q})$, preserve those of $\mathbf{D}$. So, we can define a collection of invariant statistics under the group action of $\mathcal{G}$ given by $\mathcal{D}_{\text{obs}} \equiv (\mathbf{R}(\mathbf{D}), \mathbf{L}^{(1)}(\mathbf{D}), \mathbf{L}^{(2)}(\mathbf{D}))$. We state

Theorem 2 The statistic $\mathcal{D}_{\text{obs}}$ is maximal invariant under the group action of $\mathcal{G}$.

We refer the readers to the proof in the Appendix.

4 Bayesian Inference for ROC curves

4.1 The partial likelihood and prior specifications

For the K-component mixture of multivariate normals, let $\boldsymbol{\mu}_K = (\mu_1, \mu_2, \cdots, \mu_K)$, $\boldsymbol{\sigma}_K = (\sigma_1, \sigma_2, \cdots, \sigma_K)$, and $\boldsymbol{p}_K = (p_1, \ldots, p_{K-1})$ denote the vectors of means, standard deviations, and mixing probabilities. For fixed K, we also denote the set of all the parameters by $\Theta = (\boldsymbol{\mu}_K, \boldsymbol{\sigma}_K, \boldsymbol{p}_K, \rho_0, \rho_1)$. Note that $\Theta \equiv \Theta(K)$ is a function of K with dimension of the parameter space being $3K + 1$. The partial likelihood of $\mathcal{D}_{\text{obs}}$ given Θ can be thus written as

$$\ell(\mathcal{D}_{\text{obs}}|\Theta) = \int \cdots \int_{\mathbf{R}(\mathbf{q})=\mathbf{R}(\mathbf{D})} \ell_0(\mathbf{z}|\rho_0)\, \ell_1(\mathbf{w}\,|\,\boldsymbol{\mu}_K, \boldsymbol{\sigma}_K, \boldsymbol{p}_K, \rho_1)\, d\mathbf{q}, \tag{14}$$

where $\mathbf{q} = (\mathbf{z}', \mathbf{w}')'$ is an N-dimensional concatenated vector in R^N with $\mathbf{z} = (\mathbf{z}_1', \mathbf{z}_2', \ldots, \mathbf{z}_m')'$ and $\mathbf{w} = (\mathbf{w}_1', \mathbf{w}_2', \ldots, \mathbf{w}_n')'$. Further,

$$\ell_0(\mathbf{z}|\,\rho_0) = \prod_{i=1}^{m} \phi_r(\mathbf{z}_i|\,0, 1, B_0) \quad \text{and} \quad \ell_1(\mathbf{w}\,|\,\boldsymbol{\mu}_K, \boldsymbol{\sigma}_K, \boldsymbol{p}_K, \rho_1) = \prod_{j=1}^{n} \sum_{k=1}^{K} p_k \phi_s(\mathbf{w}_j\,|\boldsymbol{\nu}_k, \Sigma_k\,) \quad \text{(see (11))} \tag{15}$$

are, respectively, the multivariate normal and mixture normal likelihoods associated with the $\mathbf{z}$ and $\mathbf{w}$ variables. Since it is quite difficult to work with the integral as well as the sum over k corresponding to the mixture in (14), we treat these integrals as missing data as is typically done in these situations. Augmenting the missing components to the likelihood in (14) gives the complete likelihood as

$$L(\mathbf{q}\,|\,\Theta) = \ell_0(\mathbf{z}|\rho_0)\left(\prod_{l=1}^{n}\prod_{k=1}^{K}\left(\,p_k\phi_s(\mathbf{w}_l\,|\boldsymbol{\nu}_k,\Sigma_k\,)\right)^{I\{C_l=k\}}\right) I\{\,\mathbf{q}\,:\,\mathbf{R}(\mathbf{q})\;=\mathbf{R}(\mathbf{D})\,\}; \qquad (16)$$

two indicator functions are introduced in (16): $I\{C_l = k\}$ takes the value 1 or 0 according to whether the mixture label corresponding to $\mathbf{y}_l$, C_l, is equal to k, and the indicator function $I\{\,\mathbf{q}\,:\,\mathbf{R}(\mathbf{q})\;=\mathbf{R}(\mathbf{D})\,\}$ to account for all missing data $\mathbf{q}$ whose ranks coincide with those from the observed data $\mathbf{D}$.

The following prior specifications are used for the Bayesian inferential framework: Under independent a *priori*, we use the following proper priors on Θ:

$$\mu_k \sim N(\eta, \kappa^{-1}), \qquad (17)$$

subject to the increasing constraints on $\boldsymbol{\mu}_K$, say, $\mu_1 < \mu_2 \cdots < \mu_K$,

$$\sigma_k^2 \overset{iid}{\sim} \text{igamma}(\alpha_0,\beta_0),\;\boldsymbol{p}_K \sim \text{dirichlet }(\delta,\delta,\ldots,\delta), \text{and}\,\rho_j \sim \text{unif}\left(-\frac{1}{s_j-1},1\right) \text{ for } j=0,1, (18)$$

where 'igamma', 'dirichlet' and 'unif' stand for the inverse gamma (with shape and scale parameters α_0 and β_0, respectively), Dirichlet, and uniform distributions on the respective parameter spaces. The full and marginal posterior distributions are given by

$$\pi(\Theta,\,\mathbf{q}\,|\,\mathbf{R}(\mathbf{D}),\mathbf{L}^{(1)}(\mathbf{D}),\mathbf{L}^{(2)}(\mathbf{D})) = \frac{L(\mathbf{q}\,|\Theta)\cdot\pi_0(\Theta)}{\int_\Theta\,L(\mathbf{q}\,|\Theta)\cdot\pi_0(\Theta)\,d\Theta} \qquad (19)$$

and

$$\pi(\Theta\,|\,\mathbf{R}(\mathbf{D}),\mathbf{L}^{(1)}(\mathbf{D}),\mathbf{L}^{(2)}(\mathbf{D})) = \int_{\{\,\mathbf{q}\,:\,\mathbf{R}(\mathbf{q})=\mathbf{R}(\mathbf{D})\,\}}\pi(\Theta,\,\mathbf{q}\,|\,\mathbf{R}(\mathbf{D}),\mathbf{L}^{(1)}(\mathbf{D}),\mathbf{L}^{(2)}(\mathbf{D}))\,d\mathbf{q}, \quad (20)$$

where $\pi_0(\Theta)$ is the prior specified in (17) and (18). The prior components of π_0 are chosen for their conjugacy properties with respect to the conditional likelihood and entails convenient updating steps for the Gibbs sampler (see Section 4.3). The selection of the hyperparameters α_0, β_0 and δ are discussed in the experimental results section in Section 5. Note that the posteriors in (19) and (20) are based on a fixed (yet unknown) value of K.

4.2 Estimation of K

This section outlines how one can construct the consistent estimator, $\hat{K}_N$. The transformation τ is unknown but as mentioned earlier, the choice $\tau = \Phi^{-1} \circ F$ allows for normality of the $\mathbf{x}$ observations after transformation by τ. To estimate F consistently, we use the model based clustering algorithm of Fraley and Raftery (2006) (`mbclust`) which assumes that the underlying distribution is a mixture of Gaussians. The algorithm `mbclust` performs a model selection procedure to obtain the number of mixture components and associated component parameter estimates using the Bayes Information Criteria (BIC). The steps are as follows:

1. The marginal distribution of each component of $\mathbf{x}_i$, $(i = 1, 2, \cdots, m)$ is F. Considering the data consisting of the first component of each $\mathbf{x}_i$, `mbclust` gives the estimate of F, $\hat{F}$, in terms of a mixture of univariate Gaussian densities. It follows that $\hat{F}$ is exactly consistent (as $m \to \infty$) for any true F that is represented by a mixture of Gaussian densities. Further, an attractive property of Gaussian mixtures is that they can approximate any arbitrary density to a desired accuracy. Hence it follows that $\hat{F}$ can be made arbitrarily close to any F (with a density) as $m \to \infty$.

2. Using the monotone transformation $\hat{\tau} = \Phi^{-1} \circ \hat{F}$, the original $\mathbf{x}$ and $\mathbf{y}$ observations are transformed to $\mathbf{x}^* = \hat{\tau}(\mathbf{x})$ and $\mathbf{y}^* = \hat{\tau}(\mathbf{y})$ by applying $\hat{\tau}$ componentwise.

3. Again we apply `mbclust` to the dataset comprising of the first component of each $\mathbf{y}_l^*$, $l = 1, 2, \cdots, n$. The assumption of multivariate binormality mixture in (11) entails that observations in this dataset are iid from the univariate Gaussian mixture

$$\sum_{k=1}^{K} p_k \phi_1(\cdot|\mu_k, \sigma_k^2), \tag{21}$$

and $\hat{K}_N$ is taken to be the number of components in the mixture estimated by `mbclust`. Once $\hat{K}_N$ is obtained, the unknown value of K is fixed at $\hat{K}_N$ for subsequent inference on Θ.

4. Initial estimate of Θ: Estimates of $\boldsymbol{p}_k$, $\boldsymbol{\mu}_k$, and $\boldsymbol{\sigma}_k$ for $k = 1, \ldots, \hat{K}$ are automatically determined by `mbclust` in the previous step. For running the MCMC procedure, these values are taken to be the initial estimate of Θ. The initial values of the correlations, ρ_0 and ρ_1, are obtained by calculating the sample correlations between the first and second components of $\mathbf{x}_i^*$, $i = 1, 2, \cdots, m$, and of $\mathbf{y}_l^*$, $l = 1, 2, \cdots, n$, respectively.

Remark 1: Note that the above procedure needs not be restricted to the first and second component of $\mathbf{x}$ and $\mathbf{y}$ observations. In fact, in Section 5, we consider the estimate of Θ based on an average of the estimates obtained by the above procedure for every pair of components of $\mathbf{x}$ and $\mathbf{y}$ (respectively, $r(r-1)/2$ and $s(s-1)/2$ many pairs).

Remark 2: Once $\hat{K}_N$ is determined, we fix the value of the unknown $K = \hat{K}_N$ for subsequent analysis. In particular, the inference on Θ is now based on the posterior (19) with $\hat{K}_N$ plugged in for the unknown K. We remark that regardless of the prior specification, the posterior is always consistent at the true value of the parameters $(K_0, \Theta_0(K_0))$. We state

Theorem 3 Let $m/(m+n) \to \lambda$ as $m, n \to \infty$. The posterior in (20) is consistent for Θ_0, that is, for any neighborhood $\mathcal{U}_0$ of Θ_0,

$$\lim_{N\to\infty} \pi(\Theta \in \mathcal{U}_0 \mid \mathbf{R}(\mathbf{D}), \mathbf{L}^{(1)}(\mathbf{D}), \mathbf{L}^{(2)}(\mathbf{D})) = 1 \quad \text{a.s.} \quad \left[P^{\infty}_{(K_0,\Theta_0)}\right], \tag{22}$$

where $\left[P^{\infty}_{(K_0,\Theta_0)}\right]$ denotes the joint distribution of all $\boldsymbol{X}$ and $\boldsymbol{Y}$ under the K_0-normality model.

The reader is referred to the proof in the Appendix.

4.3 A Bayesian computational procedure

We utilize the Gibbs sampler to obtain inference from the full posterior in (19) based on the six main updating steps given below:

1. **Update μ_k:** Each μ_k is updated from its conditional distribution:

$$\pi(\mu_k \mid \cdots) = N\left(\frac{a_k A_k + b_k \eta}{a_k + b_k}, \frac{1}{a_k + b_k}\right)$$

subject to the restriction $\mu_{k-1} < \mu_k < \mu_{k+1}$ with the convention that $\mu_0 \equiv -\infty$ and $\mu_{\hat{K}_N+1} = +\infty$; in the expression of the conditional posterior above,

$$a_k = \frac{\mathbf{1}'_{s_1} B_1^{-1} \mathbf{1}_{s_1}\, n_k}{\sigma_k^2}, \qquad b_k = \kappa, \qquad A_k = \frac{\sum\limits_{l\,:\,C_l = k} \mathbf{1}'_{s_1} B_1^{-1} \mathbf{w}_l}{n_k \mathbf{1}'_{s_1} B_1^{-1} \mathbf{1}_{s_1}},$$

and $n_k = \#\{l \,:\, C_l = k\}$.

2. **Update σ_k^2:** The variances σ_k^2's are updated from the conditional distribution

$$\pi(\sigma_k^2 \mid \cdots) = \text{igamma}\left(\frac{n_k}{2} + \alpha_0, \left[\frac{\sum_{l\,:\,C_l=k} (\mathbf{w}_l - \mu_k)' B_1^{-1} (\mathbf{w}_l - \mu_k)}{2} + \frac{1}{\beta_0}\right]^{-1}\right).$$

3. **Update ρ_0 and ρ_1:** The correlations are updated from the following conditional distributions:

$$\pi(\rho_0 | \cdots) \propto \frac{1}{|B_0|^{m/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{m} \mathbf{z}_i^{'} B_0^{-1} \mathbf{z}_i \right\},$$

and

$$\pi(\rho_1 | \cdots) \propto \frac{1}{|B_1|^{n/2}} \exp\left\{ -\frac{1}{2} \sum_{k=1}^{K} \sum_{l\,:\,C_l=k} \left[(\mathbf{w}_l - \mu_k)^{'} \Sigma_k^{-1} (\mathbf{w}_l - \mu_k) \right] \right\}.$$

The conditional distributions for ρ_0 and ρ_1 above do not have closed forms. For these two parameters, a histograming technique is utilized for generating samples.

4. **Update $\mathbf{p}_K$:** We update the weights $\boldsymbol{p}_K$ from the conditional distribution

$$\boldsymbol{p}_K \sim dirichlet(\delta + n_1, \delta + n_2, \cdots, \delta + n_{\hat{K}_N}).$$

5. **Update C_l:** The allocation variables C_ls are updated as follows. We first compute

$$P(C_l = k) = \frac{p_k \phi_s(\mathbf{w}_l \,|\, \boldsymbol{\nu}_k, \Sigma_k\,)}{\sum_{k=1}^{K} p_k \phi_s(\mathbf{w}_l \,|\, \boldsymbol{\nu}_k, \Sigma_k\,)} \quad \text{for} \quad k = 1, \ldots, \hat{K}_N,$$

and use a discrete inverse cdf method for generating the label C_l.

6. **Update z and w:** Fix $u \in \{1, 2, \cdots, r\}$. We denote the rank of z_{iu}, $\mathbf{R}(z_{iu})$, by r_0. To update z_{iu} (which is the u-th component of the vector $\mathbf{z}_i$ for $i = 1, 2, \cdots, m$), we consider the partition of the covariance matrix B_0 into four parts, namely,

$$B_0 = \begin{pmatrix} B_{00} & b_{0u} & B_{00} \\ b_{0u}^{'} & 1 & b_{0u}^{'} \\ B_{00} & b_{0u} & B_{00} \end{pmatrix} \tag{23}$$

where b_{0u} is the u-th column of B_0 excluding the (u, u)-th entry and B_{00} is the submatrix formed by deleting the u-th row and u-th column from B_0. The conditional distribution of z_{iu} is given by

$$\pi(z_{iu} \,|\, \cdots) = N(b_{0u}^{'} B_{00}^{-1} \mathbf{z}_{i,-u}, 1 - b_{0u}^{'} B_{00}^{-1} b_{0u})$$

subject to the constraint that $z_L \leq z_{iu} \leq z_U$; in the formulas above, $\mathbf{z}_{i,-u}$ is the vector $\mathbf{z}_i$ with z_{iu} removed, and z_L and z_U are those variables in $\mathbf{z}$ whose ranks correspond to $r_0 - 1$ and $r_0 + 1$, respectively (again, the convention $z_L = -\infty$ if $r_0 = 1$ and $z_U = \infty$ if $r_0 = N$ is adopted). In a similar fashion, the update of w_{lv} is carried out based on the conditional distribution

$$\pi(w_{lv} \,|\, \cdots) = N\left(\mu_k + b_{1v}^{'} B_{11}^{-1} (\mathbf{w}_{l,-v} - \mu_k),\, \sigma_k^2\, (1 - b_{1v}^{'} B_{11}^{-1} b_{1v}\,) \right)$$

subject to the constraint $w_U \leq w_{lv} \leq w_U$; w_l and w_U retain the same interpretation as z_L and z_U given that the rank of w_{lv} equals r_0, b_{1v} and B_{11} constitute a similar partition of B_1 as in (23), and k denotes the mixture label for $\mathbf{w}_l$ (i.e., $C_l = k$). Cycling through $i = 1, 2, \cdots, m$ and $u = 1, 2, \cdots, r$ for $\mathbf{z}$ and $l = 1, 2, \cdots, n$ and $v = 1, 2, \cdots, s$ for $\mathbf{w}$ completes this updating step.

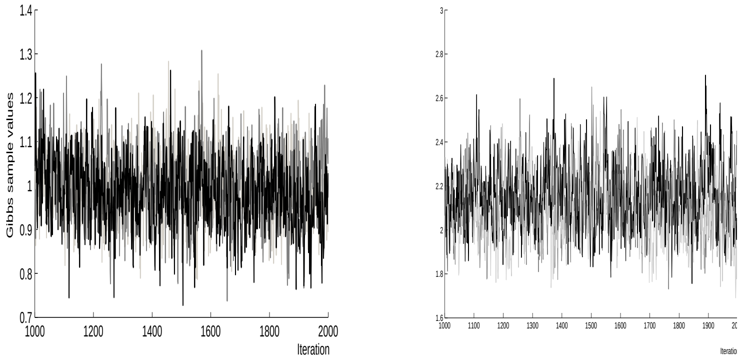
A cycle is defined to be one sweep through the updating steps 1-6 above.

4.4 Convergence diagnostics

The assessment of convergence of the Gibbs sampler is carried out based on the methodology of Gelman and Rubin (1992). A total of 3 chains are run from different starting values for $(\Theta, \mathbf{q})$. The monitoring statistic is taken to be the complete log-likelihood (see (16)) whose value is evaluated based on the current $(\Theta, \mathbf{q})$ output at the completion of each cycle. Gelman and Rubin (1992) propose the use of the PSRF (potential scale reduction factor) ratio as a measure to check for convergence. Roughly speaking, the PSRF ratio measures the ratio of between to within variances of the monitoring statistic from the chains. Thus, a PSRF ratio of close to 1 indicates that the chains have sufficiently mixed and is close to the stationary (or, target) distribution.

4.5 Inference on ROC curves

This section describes the construction of a $100(1-\alpha)\%$ highest posterior density (HPD) region for the true ROC curve based on samples from the Gibbs output after convergence is established. The HPD set is usually constructed for a range of small FAR values, say $[t_L, t_U]$, where $t_L = 0.01$ and $t_U = 0.1$, for example. Ideally, the HPD region for the ROC curve can be obtained by determining the corresponding region in Θ-space (i.e., which corresponds to the highest value of marginal posterior density in (20)), and subsequently mapping this region back to the space of ROC curves. However, the main challenge here is the presence of N integrals in the expression of the marginal posterior density which makes any analytical simplification impossible. We utilize the Gibbs output to obtain the height of the posterior density at specific Θ-values as well as to construct the HPD region. After convergence has been established, we run the Gibbs chain further to obtain a current value of Θ, say Θ_b. Given Θ_b, the missing values $\mathbf{z}$ and $\mathbf{w}$ are sampled M times using Updating Step 6 in Section 4.3. This gives rise to



(a) (b)

Figure 1: Trace plots for (a) μ_1 and (b) σ_2^2 for simulated data. The light grey, dark grey and black lines represent three different chains.

the samples $\mathbf{q}_f \equiv (\mathbf{z}_f, \mathbf{w}_f)$, $f = 1, 2, \cdots, M$. The marginal posterior density is evaluated as

$$\pi(\Theta_b \,|\, \mathbf{R}(\mathbf{D}), \mathbf{L}^{(1)}(\mathbf{D}), \mathbf{L}^{(2)}(\mathbf{D})\,) = \frac{1}{M} \sum_{f=1}^{M} \pi(\Theta_b, \mathbf{q}_f \,|\, \mathbf{R}(\mathbf{D}), \mathbf{L}^{(1)}(\mathbf{D}), \mathbf{L}^{(2)}(\mathbf{D})\,).$$

This process is repeated for $b = 1, 2, \cdots, B$ in the Gibbs chain for a large value of B. Subsequently, the proportion

$$p(\gamma) = \frac{1}{B} \sum_{b=1}^{B} I\{\pi(\Theta_b \,|\, \mathbf{R}(\mathbf{D}), \mathbf{L}^{(1)}(\mathbf{D}), \mathbf{L}^{(2)}(\mathbf{D})\,) \geq \gamma\}$$

is computed for different γ values and γ^* is selected as the γ value such that $p(\gamma)$ is closest to $100(1-\alpha)\%$. All Θ_b samples that satisfy $\pi(\Theta_b \,|\, \mathbf{R}(\mathbf{D}), \mathbf{L}^{(1)}(\mathbf{D}), \mathbf{L}^{(2)}(\mathbf{D})\,) \geq \gamma^*$ fall in the $100(1-\alpha)\%$ HPD region in the Θ-space.

5 Experimental Results

We analyze a publicly available database namely FVC2002, consisting of three sub-databases DB1, DB2 and DB3 (cf. `http://bias.csr.unibo.it/fvc2002/`). Each DB database contains fingerprint images of $F = 100$ different fingers and $L = 8$ impressions per finger obtained using different sensors. See Figure 2 for examples of fingerprint impressions from this database. The white squares in Figure 2 denote the location of the fingerprint feature called minutiae (a minutiae is a ridge anomaly consisting of either a ridge bifurcation or a ridge ending; see Zhu et al. (2007) for more details). The white lines in Figure 2 denote the minutiae orientation: this is the direction of the ridge flow at that minutiae location. The minutiae location and orientation information consists of fingerprint features that are subsequently used for matching. The similarity measure $S(Q,T)$ is based on the observed number of feature matches, w_0, between Q and T based on a matching algorithm. Figure 3 gives two examples of the matching procedure. The pair Q and T in panels (a,b) are an impostor pair corresponding to the null hypothesis H_0 in (1) whereas panels (c,d) represent a genuine pair (corresponding to H_1 in (1)). Subsequently, a matching score is obtained as

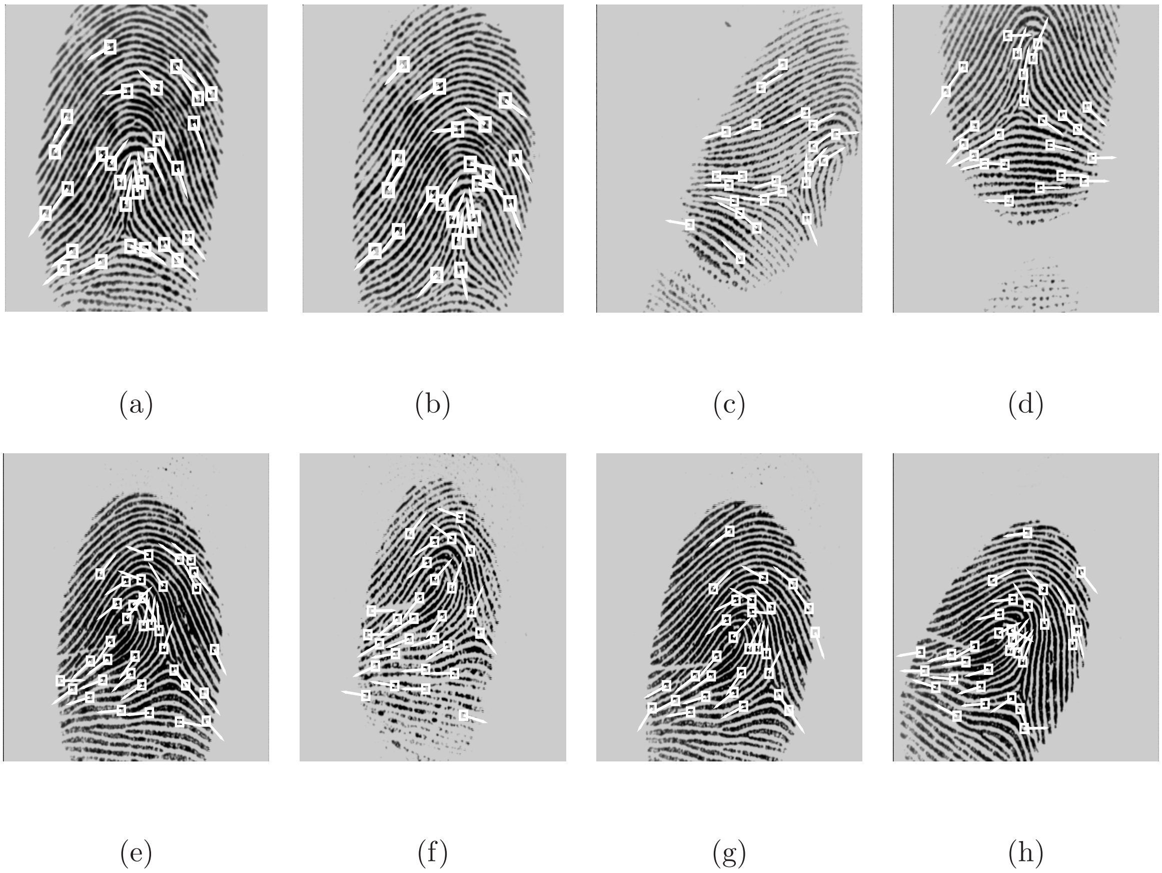
$$S_0 = \frac{w_0}{\min(m_0, n_0)} \times 1000, \tag{24}$$

where m_0, n_0 and w_0 are, respectively, the number of minutiae in Q, in T, and the number of minutiae matches between Q and T. Higher values of S_0 indicate a higher degree of similarity between Q and T, thus leading to the rejection of H_0.

The performance of a fingerprint based authentication system depends on the matching algorithm used. Often, biometric vendors claim that their matching algorithms are superior than currently available systems and thus, it is necessary to validate their claims. Deriving ROC confidence bands is one way of validating (or refuting) the claim. We obtained the matching scores based on the algorithm described in Zhu et al. (2007). The two sets of minutiae locations and directions (corresponding to Q and T) are rotated and translated to find the highest matching number of minutiae within a pre-specified bounding box. The highest matching number is taken to be w_0 and the matching score S_0 is calculated as in (24). In Figure 3, the matching numbers w_0 based on this algorithm is 7 and 16, respectively.

Impostor scores are obtained by considering all image pairs (Q,T) arising from different fingers while genuine scores are obtained from image pairs from the same finger. Thus, for the FVC2002 DB1 database there are $\binom{100}{2} \times 8 \times 8 = 316,800$ impostor and $\binom{8}{2} \times 100 = 5,600$ genuine scores. The $\binom{8}{2}$ genuine scores for each finger are highly correlated based on the discussion presented after equation (8); for example, the score from impressions 1 and 2, and the score from impressions 1 and 3 are highly correlated since one of the input image is common to both scores. For each of the 100 fingers, we randomly select a baseline impression followed by two other impressions to get the first pair of a bivariate score. The procedure is repeated again with the remaining impressions to obtain a second pair of the bivariate score, and finally repeated for all the 100 fingers in each DB database to get $n = 3 \times 2 \times 100 = 600$ genuine bivariate scores. This constitutes the $\mathbf{Y}$ data.

For the $\mathbf{X}$ data, the procedure is as follows. Since the number of impostor scores was too large, we selected a random sample of 100 pairs of different fingers out of $\binom{100}{2}$ in each DB



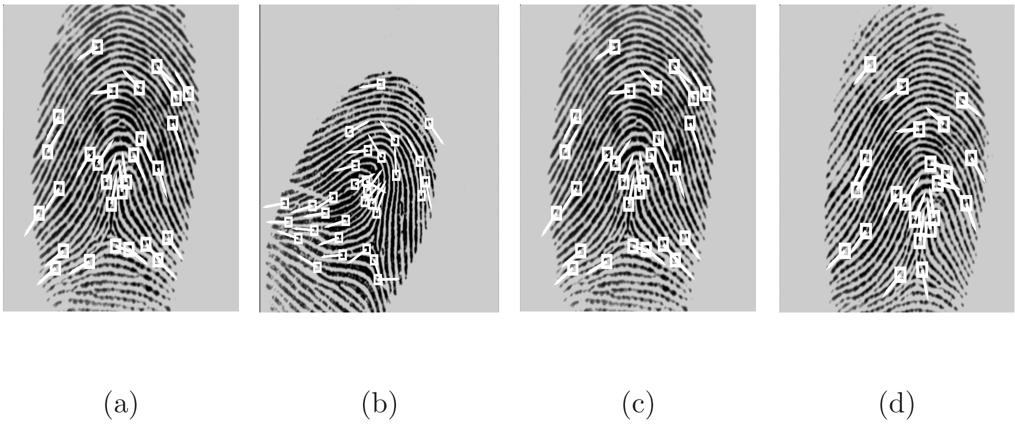
(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2: Sample fingerprint images from the FVC2002 database. Panels (a-d) show 4 impressions of one finger whereas panels (e-h) show 4 impressions of a different finger.

database. For each finger pair, two impressions were randomly selected, one from 8 impressions of the first finger and the other from 8 impressions of the second finger. This procedure is repeated once more to give a bivariate impostor score. Finally, repeating this procedure for all the 100 fingers in each DB database gives $m = 3 \times 2 \times 100 = 600$ bivariate impostor scores.

The Gibbs sampler was run for $10,000$ iterations. Figure 4 gives two examples of these trace plots corresponding to three different chains (represented by light grey, grey and black lines). Convergence was established after $9,000$ iterations using Gelman and Rubin's R-statistic. The 90% confidence bands for the ROC curve is given in Figure 6 for $t = FAR$ values from 10^{-4} to 10^{-1} (see (3)) with the same specifications for B and M as in the simulation experiments. To illustrate the need for dependent bivariate modeling, we give the posterior distributions of ρ_0 and ρ_1 in Figure 5. The 90% HPD set for ρ_0 and ρ_1 is $[0.01, 0.33]$ and $[0.33, 0.56]$, respectively, indicating that the correlation is significant for both the impostor and genuine cases, with higher value for the latter.



(a) (b) (c) (d)

Figure 3: Examples of impostor and genuine matching based on (Q, T) pairs: Panels (a,b) and (c,d) give an impostor and genuine (Q, T) pair for matching with $S(Q, T)$ equalling 7 and 16 matches, respectively.

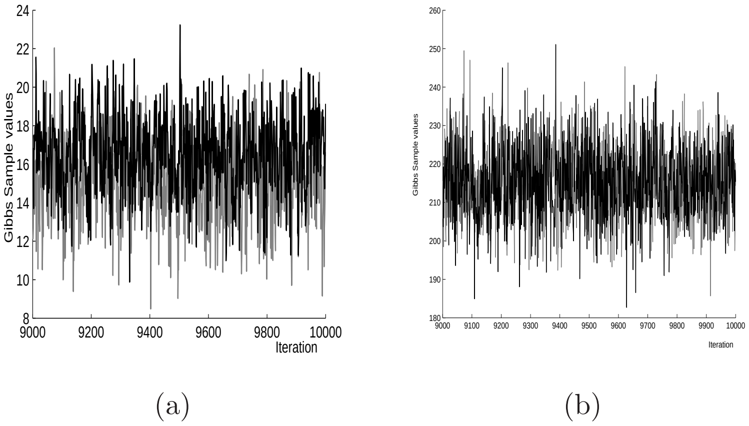


Figure 4: Trace plots for (a) μ_1 and (b) σ_1^2 for the fingerprint matching score data. The light grey, dark grey and black lines represent three different chains.

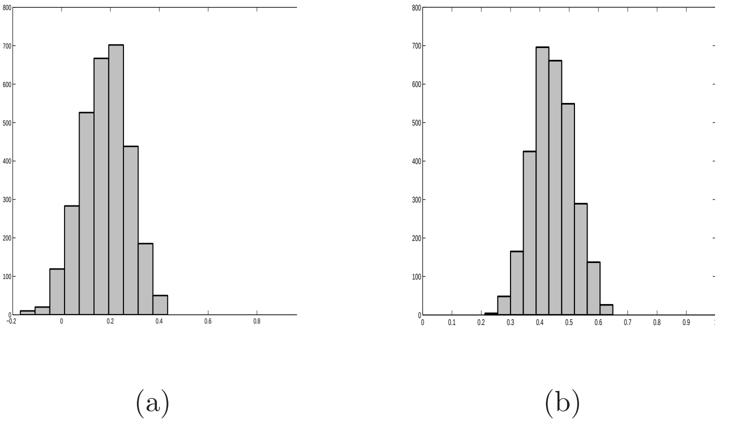


Figure 5: Posterior distribution of (a) ρ_0 and (b) ρ_1.

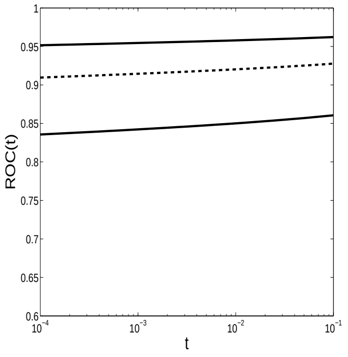


Figure 6: The 90% confidence band for the true ROC curve based on the fingerprint data.

6 Concluding Remarks

We have outlined a methodology for validating system performance based on HPD confidence bands for the ROC curve under multivariate dependence. The resulting methodology utilizes rank information and mixture distributions, and is therefore, flexible. The approach can be used reliably on datasets that exhibit various forms of joint dependence as well as marginal distributions. The inferential framework for deriving the ROC bands is Bayesian which is developed via Gibbs sampling. Our future work will be to extend the normal mixture model to more general mixture distributions.

ACKNOWLEDGEMENT

The first author's research is partially supported by the NSF DMS grant 0706385.

Appendix: Theoretical Results

Proof of Theorem 1: The random variables $\tau(X)$ and $\tau(Y)$ have distribution functions $F_\tau \equiv F \circ \tau^{-1}$ and $G_\tau \equiv G \circ \tau^{-1}$, respectively. The expression for the ROC curve is $ROC_\tau(t) = 1 - G_\tau(\lambda)$ where $1 - F_\tau(\lambda) = t$. Note that there is no ambiguity in the definition of the ROC curve with regard to which components of X or Y are taken since all of $X_1, X_2, \cdots, X_r$ (and $Y_1, Y_2, \cdots, X_s$) have the same marginals. Solving for λ from the second equation and substituting in the first gives $\lambda = \tau(F^{-1}(1-t))$ and $ROC_\tau(t) = 1 - G_\tau(\tau(F^{-1}(1-t))) = 1 - G \circ F^{-1}(1-t)$, independent of τ. Specializing to the τ that gives binormality, we get $F = \Phi$ and $G = \sum_{k=1}^{K} \pi_k \Phi((\cdot - \mu_k)/\sigma_k)$. For fixed $0 < t < 1$, it follows that $\lambda = \Phi^{-1}(1-t) = -\Phi^{-1}(t)$ and

$$ROC(t) \quad = \quad 1 - G(\lambda) = 1 - \sum_{k=1}^{K} \pi_k \Phi((\lambda - \mu_k)/\sigma_k) = 1 - \sum_{k=1}^{K} \pi_k \Phi((-\Phi^{-1}(t) - \mu_k)/\sigma_k)$$

$$= \sum_{k=1}^{K} \pi_k(1 - \Phi((-\Phi^{-1}(t) - \mu_k)/\sigma_k)) = \sum_{k=1}^{K} \pi_k \Phi(a_k + b_k \Phi^{-1}(t)),$$

where $a_k = \mu_k/\sigma_k$ and $b_k = 1/\sigma_k$. QED.

Proof of Theorem 2: Let $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{x}', \mathbf{y}')$ be two different sets of observations. It is easy to see that the vector of ranks $\mathbf{R}(\mathbf{x}, \mathbf{y})$ and labels $\mathbf{L}^{(j)}(\mathbf{x}, \mathbf{y})$ for $j = 1, 2$ are invariant under any monotone increasing transformation τ. Conversely, suppose $\mathbf{L}^{(j)}(\mathbf{x}, \mathbf{y}) = \mathbf{L}^{(j)}(\mathbf{x}', \mathbf{y}')$ for $j = 1, 2$ and $\mathbf{R}(\mathbf{x}, \mathbf{y}) = \mathbf{R}(\mathbf{x}', \mathbf{y}')$. By applying standard arguments (see, for example, Hájek and Šidák (1967)) for two sets of observations having the same ranks, there exist a monotone increasing function such that $(\mathbf{x}', \mathbf{y}') = \tau(\mathbf{x}, \mathbf{y})$. QED.

Next, we need two lemmas to prove Theorem 3.

Lemma 1 **Doob's Theorem** (Ghosal and Van der Vaart (2009)): Let $\Xi^{(n)}$ be observations whose distribution depends on a parameter θ, both of which take values in Polish spaces. Assume that θ is equivalent to a $\Xi^{(\infty)}$-measurable random variable, i.e., there exists a $\Xi^{(\infty)}$ measurable function f on $\Xi^{(\infty)}$ such that $\theta = f(\omega^\infty)$ a.e. $[\Pi \times P_\theta^{(\infty)}]$. Then the posterior $\Pi(\cdot|\Xi^{(n)})$ is strongly consistent at θ for $[\Pi]$-almost every θ.

Lemma 2 **Spearman's rank correlation coefficient** (see Nelson (2006)): Let (X, Y) be a pair of continuous random variables with a joint distribution associated to the copula C. The Spearman's rank correlation coefficient of (X, Y) is given by

$$\rho_C = 12 \int\int_{\boldsymbol{I}^2} uv dC(u, v) - 3 = 12 \int\int_{\boldsymbol{I}^2} C(u, v) du dv - 3, \tag{25}$$

where $\boldsymbol{I}^2 = \boldsymbol{I} \times \boldsymbol{I}$ is the product of the unit closed interval $\boldsymbol{I} = [0, 1]$.

Proof of Theorem 3: Let (K_0, Θ_0) be the true value of the pair (K, Θ). Let ν_K denote the Lebesgue measure on $R^{2K} \times \boldsymbol{I}^{K-1}$, the range of the parameter space $\Theta \equiv \Theta(K)$. We consider the case where the true value K_0 lies in the range $\mathcal{K} = \{\, K \,:\, K_{min} \leq K \leq K_{max} \,\}$ with known integers K_{min} and K_{max}. Suppose $\pi_0(K, \Theta)$ is a prior on $\Omega \equiv \cup_{K \in \mathcal{K}} R^{2K} \times \boldsymbol{I}^{K-1}$ satisfying $\pi_0(K, \Theta) = \pi_0(\Theta \,|\, K) \cdot \pi_0(K)$ where $\pi_0(K)$ follows a discrete uniform between K_{min} and K_{max}, and $\pi_0(\Theta \,|\, K)$ is the prior elicitation given in equations (17) and (18) for every fixed K. Note that $\pi_0(K, \Theta) > 0$ for every (K, Θ). Let ν be the measure defined by $\nu(\cup_{K \in \mathcal{A}} B_K) = \sum_{K \in \mathcal{A}} \nu_K(B_K)$. An application of Doob's theorem in Lemma 1 guarantees posterior consistency in the following way: for (K_0, Θ_0) a.e. $[\nu]$, and for any neighborhood $\mathcal{N}_0 = \cup_{K \in \mathcal{U}_0} V_K$ of (K_0, Θ_0),

$$\lim_{N\to\infty} \pi((K,\Theta) \in \mathcal{N}_0 \mid \mathbf{R}(\mathbf{D}), \mathbf{L}^{(1)}(\mathbf{D}), \mathbf{L}^{(2)}(\mathbf{D})) = 1 \;\text{ a.s. } \left[P^{\infty}_{(K_0,\Theta_0)}\right], \tag{26}$$

where $\left[P^{\infty}_{(K_0,\Theta_0)}\right]$ denotes the joint distribution of all $\boldsymbol{X}$ and $\boldsymbol{Y}$ under the K_0-normality model. Note that statistics that are consistent for (K_0, Θ_0) must be obtained for the above conclusion to hold which we will demonstrate later. For the moment, taking the neighborhood $\mathcal{U}_0 = \{K_0\}$, it follows from (26) that for (K_0, Θ_0) a.e. $[\nu]$,

$$\begin{aligned} & \lim_{N\to\infty} \pi(K = K_0 \mid \mathbf{R}(\mathbf{D}), \mathbf{L}^{(1)}(\mathbf{D}), \mathbf{L}^{(2)}(\mathbf{D})\,) \\ = & \lim_{N\to\infty} \left[\pi(K = K_0,\ \Theta_{K_0} \in B_{K_0} | \cdots) + \pi(K = K_0,\ \Theta_{K_0} \in B^c_{K_0} | \cdots)\right] \\ = & \ 1 + 0 = 1 \;\text{ a.s. } \left[P^{\infty}_{(K_0,\Theta_0)}\right]; \end{aligned} \tag{27}$$

the latter set in (27) $(K_0, B^c_{K_0}) \subset \mathcal{N}^c_0$ and hence the probability tends to zero. Consequently,

$$\lim_{N\to\infty} \pi(\Theta_0 \in B_{K_0} \,|\, K = K_0,\, \mathbf{R}(\mathbf{D}), \mathbf{L}^{(1)}(\mathbf{D}), \mathbf{L}^{(2)}(\mathbf{D})\,) \tag{28}$$

$$= \lim_{N\to\infty} \frac{\pi(K = K_0,\ \Theta_{K_0} \in B_{K_0} | \cdots)}{\pi(K = K_0 \mid \cdots)} = 1/1 = 1 \;\text{ a.s. } \left[P^{\infty}_{(K_0,\Theta_0)}\right]. \tag{29}$$

Denoting the set

$$\mathcal{A} \equiv \{\, \omega \,:\, \lim_{N\to\infty} \pi(\Theta_0 \in B_{K_0} \,|\, K = K_0,\, \mathbf{R}(\mathbf{D}), \mathbf{L}^{(1)}(\mathbf{D}), \mathbf{L}^{(2)}(\mathbf{D})\,) = 1\,\},$$

it follows that $P^{\infty}_{(K_0,\Theta_0)}(\mathcal{A}) = 1$ from (28). Since $\hat{K}_N \to K_0$ a.s. $\left[P^{\infty}_{(K_0,\Theta_0)}\right]$, the set

$$\mathcal{B} = \{\, \omega \,:\, \hat{K}_N = K_0 \quad \text{for all but finitely many } N\text{s}\}$$

also has $P^{\infty}_{(K_0,\Theta_0)}(\mathcal{B}) = 1$. Now taking $\omega \in \mathcal{A} \cap \mathcal{B}$ (note that $P^{\infty}_{(K_0,\Theta_0)}(\mathcal{A} \cap \mathcal{B}) = 1$), we have

$$\lim_{N\to\infty} \pi(\Theta(\hat{K}_N) \in B_{K_0} \,|\, K = \hat{K}_N,\, \mathbf{R}(\mathbf{D}), \mathbf{L}^{(1)}(\mathbf{D}), \mathbf{L}^{(2)}(\mathbf{D})\,) = 1.$$

The conditional posterior distribution above is exactly equal to the marginal posterior distribution of Θ in (20).

To complete the proof, we need to construct consistent estimators of Θ. We have already proved that $\hat{K}_N \to K$. Gu and Ghosal (2008) demonstrate the existence of rank-based statistics that are consistent for Θ when $K = 1$. Since our case is multivariate with $K > 1$, we focus on the marginal distributions of the first component of $\boldsymbol{X}$ and $\boldsymbol{Y}$ (that is, all indices with $\mathbf{L}^{(2)}(\mathbf{D}) = 1$). Following Gu and Ghosal (2008), the pooled observations $\boldsymbol{S} = (S_1, S_2, \cdots, S_{N_0})$ $(N_0 = m + n)$ are iid with $S_i \sim (1 - \lambda)F + \lambda G$. It follows that

$U_i = [(1-\lambda)F + \lambda G]\,(S_i) = \left[(1-\lambda)\Phi + \lambda \sum_{k=1}^{K} \pi_k\, \Phi((\cdot - \mu_k)/\sigma_k)\right](N_i)$ (where $N_i = \tau(S_i)$) are iid uniform $(0,1)$. Considering the indices of $\mathbf{L}^{(1)}(\mathbf{D}) = 0$ (corresponding to the x observations only), the subset (denoted by indices i_js) consists of iid observations with $N_{i_j} \sim N(0,1)$ and $U_{i_j} \sim (1-\lambda)\Phi(N_{i_j}) + \lambda \sum_{k=1}^{K} \pi_k\, \Phi((N_{i_j} - \mu_k)/\sigma_k)$. Subsequently, the random variables U_{i_j} are iid from a univariate mixture density g_Θ, say. Fixing K at $\hat{K}_N$, regularity conditions guarantee that the MLE (in terms of the U_{i_j}s), for example, is consistent for Θ. Next, Gu and Ghosal (2008) show that the U_{i_j} are a limit of a subsequence of ranks (which necessarily range from 1 to N_0) based on the observations in $\boldsymbol{S}$. In our case, the ranks of $\boldsymbol{S}$ are a subset of $\mathbf{R}(\mathbf{D})$ thus ranging from 1 to $N = rm + sn$. However, one can obtain the reduced ranks of Gu and Ghosal (2008) by re-ranking the ranks in this subset of $\mathbf{R}(\mathbf{D})$ (for example, the smallest rank in this subset gets new rank 1, the second smallest gets rank 2, and so on).

To prove the consistency at ρ_0, we look at the subset of ranks in $\mathbf{R}(\mathbf{D})$ that correspond to $\mathbf{L}^{(1)}(\mathbf{D}) = \{0\}$ and $\mathbf{L}^{(2)}(\mathbf{D}) = \{1, 2\}$ (i.e. the first and second components of the x observations). Using the re-ranking procedure discussed above, we re-rank the first and second components separately and calculate Spearman's rank correlation coefficient, $\hat{\rho}_C$, based on the reduced ranks. Note that $\hat{\rho}_C$ is consistent for the population value ρ_C as given in Lemma 2. The copula corresponding to the x observations is $C(u,v) = \Phi_2(\Phi^{-1}(u), \Phi^{-1}(v) \,|\, \rho\,)$ where $\Phi_2(a, b \,|\, \rho\,)$ is the cdf of a bivariate normal distribution with means 0, standard deviations 1, and correlation ρ. Plugging in $\hat{\rho}_C$ in place of ρ_C and solving for ρ in (25) gives a consistent estimator of ρ_0, $\hat{\rho}_0$. A similar argument can be made for ρ_1 based on the copula for the first and second components of the y observations,

$$C_1(u,v) \;=\; \sum_{k=1}^{K} p_k \Phi_2(\Phi_k^{-1}(u), \Phi_k^{-1}(v) | \mu_k, \sigma_k, \rho\,), \tag{30}$$

where $\Phi_2(a, b \,|\, \mu_k, \sigma_k, \rho\,)$ is the cdf of $N(\boldsymbol{\mu}_k, \Sigma_1)$ and Φ_k^{-1} is the inverse cdf of a normal random variable with mean μ_k and variance σ_k^2. We plug in the consistent estimators of Θ from the previous paragraph in the expression for C_1 above and the Spearman rank correlation for ρ_C in (25). Solving for ρ in (25) gives the consistent estimator of ρ_1, $\hat{\rho}_1$.

References

Alonzo, T. A., and Pepe, M. S., (2002), Distribution-free ROC analysis using binary regression techniques. *Biostatistics*, **3**, 421-432.

Bolle, R., Connell, J., Pankanti, S., Ratha N., and Senior A., (2004), Guide to

Biometrics.

Bolle, R. M., Pankanti, S., and Ratha, N. K., (2000), Evaluation techniques for biometrics-based authentication systems (FRR), In Proceedings of the 14th International Conference on Pattern Recognition, ICPR, 2831-2837.

Bolle, R., Ratha, N., and Pankanti, S., (2004), Error analysis of pattern recognition systems: The subsets bootstrap. *Computer Vision and Image Understanding*, **93**(1), 1-33.

Braun, T. M., and Alonzo, T. A., (2008). A modified sign test for comparing paired ROC curves. *Biostatistics*, **9**, 364-372.

Cai, T., (2004). Semi-parametric ROC regression analysis with placement values. *Biostatistics*, **5**, 45-60.

Cai, T., and Moskowitz, C., (2004), Semiparametric estimation of the binormal ROC curve. *Biostatistics*, **5**(4), 573-586.

Dass, S. C., Zhu, Y., and Jain, A. K., (2006), Validating a biometric authentication system: Sample Size Requirements. *IEEE Trans. on PAMI*, **28**(12), 1902-1319.

Dabrowska, D. M., and Doksum, K. A., (1998), Estimation and testing in a two-sample generalized odds-rate model. *Journal of the American Statistical Association*, *83*, 744-749.

De Long, E. R., De Long, D. M., and Clarke-Pearson, D. L., (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44**(3), 837-845.

Fraley, C., and Raftery, A. E., (2006), MCLUST: Model-based cluster analysis. R package version 2.1-12. Department of Statistics and University of Washington, Seattle.

Gelman, A., and Rubin, D. B., (1992), Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, **7**(4), 457-472

Ghosal, S., Van der Vaart, A. W., (2009), Theory of Nonparametric Bayesian Inference. Cambridge University Press, Cambridge.

Gu, J. and Ghosal, S.(2008), Bayesian ROC curve estimation under binormality using a rank likelihood. *Journal of Statistical Planning and Inference*, **139**(6), 2076-2083.

Hájek, J., and *Š*idák, Z., (1967), Theory of Rank Tests. Academic press, New York.

Hanley, J. A., (1989), Receiver operating characteristic (ROC) methodology: the state of the art. *Critical Rev. Diagnostic Imaging*, **29**(3), 307-335.

Horv*á*th, L., Horv*á*th, Z., and Zhou, W., (2008), Confidence bands for ROC curves. *Journal of Statistical Planning and Inference*, **138**, 1894-1904.

Hsieh, F. S., Turnbull, B. W., (1996), Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Ann. Statist.*, **24**, 25-40.

Jain, A. K., Bolle, R., and Pankanti, S., (1999), (eds) BIOMETRICS: Personal Identification in Networked Society. Kluwer Academic Publishers, Boston.

Kamitsuji, S., and Kamatani, N., (2006). Estimation of haplotype associated with several quantitative phenotypes based on maximization of area under a receiver operating characteristic (ROC) curve. *Journal of Human Genetics*, **51**(4), 314-325.

Maltoni, D., Maio, D., Jain, A. K., and Prabhakar, S., (2003), Handbook of Fingerprint Recognition, Springer-Verlag.

McClish, D.K., 1989. Analyzing a portion of the ROC curve. *Medical Decision Making*, **9**, 190-195.

Metz, C. E., Herman, B. A., and Shen, J., (1998), Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*, **17**, 1033-1053.

Nelson, R. E., (2006), *An Introduction to copulas.* The second edition, Springer-Verlag.

Schuckers, M. E., (2003), Using the beta-binomial distribution to assess performance of a biometric identification device. *International Journal of Image and Graphics (Special Issue on Biometrics)*, **3**(3), 523-529.

Swets, J. A., (1986), Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychol. Bull.*, **99**, 100-117.

Zhu, Y., Dass, S. C., and Jain, A. K., (2007), Statistical models for assessing the individuality of fingerprints. *IEEE Transactions on Information Forensics and Security*, **2**(3), 391-401.

Zou, K. H., and Hall, W. J., (2000), Two transformation models for estimating an ROC curve derived from continuous data. *Journal of Applied Statistics*, **27**, 621-631.