IET Research Journals

Review Article

IET Journals
The Institution of Engineering and Technology

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Semiparametric Likelihood-ratio-based Biometric Score Level Fusion via Parametric Copula

Nanang Susyanto[1] ✉, *Raymond Veldhuis*[2], *Luuk Spreeuwers*[2], *Chris Klaassen*[3]

[1] *Department of Mathematics, Universitas Gadjah Mada, Sekip Utara BLS, Yogyakarta, Indonesia*
[2] *Faculty of EEMCS, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands*
[3] *Korteweg-de Vries Institute for Mathematics, University of Amsterdam, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands*
✉ *Email: nanang_susyanto@ugm.ac.id*

Abstract: We present a mathematical framework for modelling dependence between biometric comparison scores in likelihood-based fusion by copula models. The pseudo-maximum likelihood estimator (PMLE) for the copula parameters and its asymptotic performance are studied. For a given objective performance measure in a realistic scenario, a resampling method for choosing the best copula pair is proposed. Finally, the proposed method is tested on some public biometric databases from fingerprint, face, speaker, and video-based gait recognitions under some common objective performance measures: maximizing acceptance rate at fixed false acceptance rate, minimizing half total error rate, and minimizing discrimination loss.

1 Introduction

In a biometric verification system, biometric samples (images of faces, fingerprints, voices, gaits, etc.) of people are compared and classifiers indicate the level of similarity between any pair of samples by a comparison score. If two samples of the same person are compared, a genuine score is obtained. If a comparison concerns samples of different people, the resulting score is called an impostor score. Depending on the application, a biometric verification system may give either a *hard decision* or a *soft decision*. Hard decision means that the system decides whether two biometric samples (query and template) are from the same individual or not by comparing the score to a *threshold*. On the other hand, the soft decision can be used in a forensic scenario by only giving the likelihood ratio (LR) value as an evidential value and let the final decision to the judge [4]. A common performance measure for this scenario is the *cost of log likelihood ratio* that can be decomposed into discrimination and calibration performance [2].

When our biometric system has two or more classifiers, one has to fuse the resulting multiple scores into a new score, which is called score level fusion. It is convenient if the fused score is again an LR because: (1) it is optimal for standard biometric verification [20] and (2) it reflects evidential value in forensic individualization [4]. By assuming independence between scores, the fused LR can be computed as the product of the individual likelihood ratios of the classifiers (henceforth called PLR fusion). However, the score level fusion problem becomes difficult if the scores are dependent. In this paper, we propose a score level fusion method with the following advantages.

1. The fused score is an LR.
2. It can deal with dependent scores.
3. It is available as an open-source software framework, programmed in Matlab*, that implements the method.

This paper uses the copula concept to handle dependence between scores. Although the copula model has already been used in [11, 29–31] for some different scenarios, none of them provides analytically how this model is built and why the estimation of parameters determining the model is reliable. After explaining some related works in Section 2, this paper will explain how a copula model splits the LR computation for two or more classifiers into a product of the individual likelihood ratios and a correction factor in Section 3. Section 4 introduces a semiparametric model of LR-based fusion and subsequently provides an estimator of the proposed model with its convergence analysis. Detailed procedures to apply for several different applications of our method is given in Section 5. Finally, our conclusions are presented in Section 6.

2 Score Level Fusion

There are three categories of score level fusion: transformation-based [12], classifier-based [13], and density-based (henceforth called *likelihood-ratio-based*) [18]. The transformation-based fusion is done by mapping all components of the vector of comparison scores to a common domain and applying some simple rules such as sum, mean, max, med, etc. Apart from its simplicity, it is important that the training set is representative of the data. For instance, to normalize scores to the unit interval [0,1], one must have the minimum and maximum scores. However, if the training data has outlier(s) then this estimation will not be reliable and may destroy the fusion performance. The classifier-based fusion acts as a classifier of the vector of the comparison scores to distinguish between genuine and impostor scores. These two first categories cannot be used in a forensic scenario as the fused score is not always an LR value. The last category would be optimal for biometric verification if the underlying distributions were known according to the Neyman-Pearson lemma [20]. Moreover, the fused score, as an LR value, can be used for forensic evidence evaluation [4] in forensic individualization as a multiplicative factor for the information before analyzing the evidence (*prior odds*) to get the information after taking the evidence into account (*posterior odds*) via the Bayesian framework

$$O_{\text{posterior}} = \text{LR} \times O_{\text{prior}}. \tag{1}$$

The LR is defined as the ratio between the density functions of the genuine and impostor scores. There are two categories for computing the LR: (1) estimating the density functions of the genuine and impostor scores separately and (2) estimating the LR directly. The

* *http://scs.ewi.utwente.nl/downloads/show,Copula%20Fusion%20Framework/*

common approaches of the first category are modelling the underlying densities parametrically (assuming normal, Weibull, Gaussian mixture, etc.) and nonparametrically (kernel density estimation, histogram binning, etc.). Parametric models are usually chosen because of their simplicity and nonparametric models because of their flexibility. However, the main problem in using parametric models is the difficulty in choosing the appropriate model whereas nonparametric estimators are sensitive to the choice of the bandwidth or other smoothing parameters, especially for our multivariate case. A common parametric model to compute the likelihood ratio directly, is the logistic regression method (Logit) by assuming the LR having some parametric form such as linear, quadratic, and so on. Although this method can also be used for the multivariate case [16, 17], the same problem in choosing an appropriate model will appear. In [24] a number of fusion methods are compared among which several likelihood ratio based approaches. However, all of these LR based methods use a parametric model. On the other hand, the nonparametric approach of *Pool Adjacent Violators* (PAV), which seems promising because of its optimality in transforming scores into their LR values [2], is applicable only for 1-dimensional scores, which means that it cannot be used to compute the LR for fusion.

Many studies of score level fusion assume independence between classifiers; see [17, 32, 34]. However, the independence assumption is not realistic since the scores may rely on the same information. To incorporate the dependence between classifiers, we propose a semiparametric LR-based biometric fusion by modelling the marginal individual likelihood ratios nonparametrically and the dependence between them by parametric copulas, to trade off between the limitations of parametric and the flexibility of nonparametric models.

3 Likelihood Ratio Computation via Copula

Suppose we have d classifiers and let $(s_1, \cdots, s_d)$ denote the concatenated vector of d similarity scores where s_k is the corresponding score from the k-th classifier for $k = 1, \ldots, d$. Let f_{gen} and f_{imp} be the densities of genuine and impostor scores, respectively. The likelihood ratio at a point $(s_1, \cdots, s_d)$ is defined by

$$\text{LR}(s_1, \cdots, s_d) = \frac{f_{\text{gen}}(s_1, \cdots, s_d)}{f_{\text{imp}}(s_1, \cdots, s_d)}. \tag{2}$$

Using the copula concept, the densities f_{gen} and f_{imp} will be split into their marginal densities and a factor modelling their dependence.

A d-variate copula is a distribution function on the unit cube $[0, 1]^d$, of which the marginals are uniformly distributed. Sklar [26] shows the existence of a copula for any multivariate distribution function.

Theorem 1 (Sklar (1959))**.** *Let $d \geq 2$, and suppose H is a distribution function on $\mathbb{R}^d$ with 1-dimensional continuous marginal distribution functions $F_1, \cdots, F_d$. Then there exists a unique copula C so that*

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)) \tag{3}$$

for every $(x_1, \ldots, x_d) \in \mathbb{R}^d$.

By taking the derivative of (3) with respect to x_i, we will get the joint density function

$$h(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \times \prod_{i=1}^{d} f_i(x_i) \tag{4}$$

where c is the copula density and f_i is the i-th marginal density for every $i = 1, \cdots, d$. This implies that (2) can be written as

$$\text{LR}(s_1, \cdots, s_d) = \frac{c_{\text{gen}}(F_{\text{gen},1}(s_1), \cdots, F_{\text{gen},d}(s_{\text{d}}))}{c_{\text{imp}}(F_{\text{imp},1}(s_1), \cdots, F_{\text{imp},d}(s_{\text{d}}))} \times \prod_{i=1}^{d} \frac{f_{\text{gen},i}(s_i)}{f_{\text{imp},i}(s_i)} \tag{5}$$

where c_{gen} and c_{imp} are the copula densities of genuine copula C_{gen} and impostor copula C_{imp}, respectively. The second factor of (5), which is the product of the individual likelihood ratios

$$\text{PLR}(\mathbf{s}) = \prod_{i=1}^{d} \frac{f_{\text{gen},i}(s_i)}{f_{\text{imp},i}(s_i)} = \prod_{i=1}^{d} \text{LR}_i(s_i), \tag{6}$$

will be called the *Naive Bayes part*, while the first factor, which is the copula density ratio

$$\text{CF}(\mathbf{s})^{(C_{\text{gen}}, C_{\text{imp}})} = \frac{c_{\text{gen}}(F_{\text{gen},1}(s_1), \cdots, F_{\text{gen},d}(s_{\text{d}}))}{c_{\text{imp}}(F_{\text{imp},1}(s_1), \cdots, F_{\text{imp},d}(s_{\text{d}}))} \tag{7}$$

will be called the *correction factor* where the superscript $(C_{\text{gen}}, C_{\text{imp}})$ means that the CF is modelled by copulas C_{gen} and C_{imp} for genuine and impostor scores, respectively. We call (7) a correction factor because it corrects the Naive Bayes part for dependence.

4 Semiparametric Model for Likelihood Ratio Computation

The LR as defined in (5) could be computed exactly if the marginal and copula densities of genuine and impostor scores were known. However, they have to be estimated from *training* data, which will be done semiparametrically, modelling the Naive Bayes part and distribution functions nonparametrically, and the dependence between them by parametric copulas. Note that we aim at incorporating dependence between classifiers. Therefore, the copula parameter is the main parameter that one is interested in, which is called the *parameter of interest*, while the marginal likelihood ratios and distribution functions are treated as *nuisance parameters* in the sense that they are less important than the copula parameter when modelling dependence between classifiers. However, in computing the LR itself, we need to estimate all parameters composing (5).

This section will present three main steps of computing the LR using our approach: (1) computing the Naive Bayes part, (2) computing the correction factor for a given copula pair of genuine and impostor scores, and (3) choosing the best copula pair for a specific performance measure. Let

$$\mathbf{W}_1, \ldots, \mathbf{W}_{n_{\text{gen}}} \tag{8}$$

and

$$\mathbf{B}_1, \ldots, \mathbf{B}_{n_{\text{imp}}} \tag{9}$$

be i.i.d copies of d-dimensional random variable of genuine scores $\mathbf{W} = (W_1, \cdots, W_d)$ and impostor scores $\mathbf{B} = (B_1, \cdots, B_d)$, respectively. Here, we will assume that the random variables of genuine and impostor scores are continuous.

4.1 Naive Bayes part

The Naive Bayes part is typically easy to be computed because there are several methods of computing the LR for 1-dimensional scores. The most common ways are Kernel Density Estimation (KDE), Logistic Regression (Logit), Histogram Binning (HB), and Pool Adjacent Violators (PAV) methods; see [1] for a brief explanation of these methods. In this paper, we choose the PAV method because of its optimality [34].

For every $k = 1, \cdots, d$, PAV sorts and assigns a posterior probability of 1 or 0 to the k-th component of genuine and impostor scores in a training set, respectively. It then finds the non-monotonic adjacent group of probabilities and replaces it with the average of that group. This procedure is repeated until the whole sequence is monotonically increasing which estimates the posterior probability $P(H_1|(\cdot))$ of the k-th component of (8) and (9) where H_1 correspond to a genuine score. By assuming

$$P(H_1) = \frac{n_{\text{gen}}}{n_{\text{gen}} + n_{\text{imp}}},$$

the corresponding LR_ks of (8) and (9) can be computed according to the Bayesian formula by

$$\widehat{\text{LR}}_k(\cdot) = \frac{P(H_1|(\cdot))}{1 - P(H_1|(\cdot))} \times \frac{n_{\text{imp}}}{n_{\text{gen}}} \tag{10}$$

so that we have a numerical function that maps a score to its $\widehat{\text{LR}}_k$. Finally, for every score for the k-th classifier, its corresponding $\widehat{\text{LR}}_k$ value can be computed by interpolation.

4.2 Semiparametric correction factor estimation

While the Naive Bayes part is modelled nonparametrically, the correction factor will be modelled semiparametrically by assuming C_{gen} and C_{imp} to be parametric copulas. Let θ_{gen} and θ_{imp} denote the dependence parameters determining C_{gen} and C_{imp}, respectively. Since the marginal distributions are treated as nuisance parameters as noted before, we will focus on the estimation of parameter of interest $\theta = \begin{pmatrix}\theta_{\text{gen}} \\ \theta_{\text{imp}}\end{pmatrix}$. Thus our correction factor model is defined by

$$\mathcal{CF} = \{\text{CF}_{\theta,F}^{(C_{\text{gen}},C_{\text{imp}})} \ : \ \theta \in \Theta, \ F \in \mathcal{F}\} \tag{11}$$

where $\Theta \subset \mathbb{R}^D$ is open and $\mathcal{F}$ is a collection of continuous marginal distribution functions. Here, D is the dimensionality of θ, which is the sum of dimensionalities of θ_{gen} and θ_{imp}, and

$$F = (F_{\text{gen},1}, \cdots, F_{\text{gen},d}, F_{\text{imp},1}, \cdots, F_{\text{imp},d}). \tag{12}$$

Note that if the marginal distributions $F_{\text{gen},k}$ and $F_{\text{imp},k}$ for $k = 1, \ldots, d$ are known then the log-likelihood of the combined samples (8) and (9) can be written as

$$L = \sum_{i=1}^{n_{\text{gen}}} \log c_{\theta_{\text{gen}}}\left(\mathbf{U}_{\text{gen},i}\right) + \sum_{j=1}^{n_{\text{imp}}} \log c_{\theta_{\text{imp}}}\left(\mathbf{U}_{\text{imp},j}\right) \tag{13}$$

where

$$\mathbf{U}_{\text{gen},i} = (F_{\text{gen},1}(W_{1,i}), \cdots, F_{\text{gen},d}(W_{d,i}))$$

and

$$\mathbf{U}_{\text{imp},i} = (F_{\text{imp},1}(B_{1,j}), \cdots, F_{\text{imp},d}(B_{d,j}))$$

for $i = 1, \ldots, n_{\text{gen}}$ and $j = 1, \ldots, n_{\text{imp}}$. Differentiating (13) with respect to θ gives

$$\sum_{i=1}^{n_{\text{gen}}} \frac{\partial c_{\theta_{\text{gen}}}\left(\mathbf{U}_{\text{gen},i}\right)/\partial\theta_{\text{gen}}}{c_{\theta_{\text{gen}}}\left(\mathbf{U}_{\text{gen},i}\right)} = 0$$

and

$$\sum_{j=1}^{n_{\text{imp}}} \frac{\partial c_{\theta_{\text{imp}}}\left(\mathbf{U}_{\text{imp},j}\right)/\partial\theta_{\text{imp}}}{c_{\theta_{\text{imp}}}\left(\mathbf{U}_{\text{imp},j}\right)} = 0.$$

As a consequence, if $F_{\text{gen},k}$ and $F_{\text{imp},k}$ are replaced by their empirical distribution functions based on samples

$$W_{k,1}, \ldots, W_{k,n_{\text{gen}}}$$

and

$$B_{k,1}, \ldots, B_{k,n_{\text{imp}}},$$

respectively, we will get *two-step* estimators $\hat{\theta}_{\text{gen},n_{\text{gen}}}$ and $\hat{\theta}_{\text{imp},n_{\text{imp}}}$ called *pseudo-maximum likelihood estimators* (PMLE) of θ_{gen} and θ_{imp}, respectively, as studied in [8] and extended in [33]. The terminology of the two-step estimator comes from the fact that one has to do two steps to obtain it: (1) transforming data to uniformly distributed and (2) finding the maximum likelihood estimator of the transformed data. We use a *modified* version of the empirical distribution function to avoid singularity problems; it is defined as

$$\hat{F}_n(x) = \frac{1}{n+1}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \le x\}}, \quad \forall x \in \mathbb{R} \tag{14}$$

for a given sample $X_1, \ldots, X_n$.

Under some regularity conditions, we can derive the convergence of

$$\hat{\theta}_n = \begin{pmatrix}\hat{\theta}_{\text{gen},n_{\text{gen}}} \\ \hat{\theta}_{\text{imp},n_{\text{imp}}}\end{pmatrix} \tag{15}$$

in the following theorem.

Theorem 2. *Write $n = n_{\text{gen}} + n_{\text{imp}}$ and assume $0 < \lim_{n\to\infty} n_{\text{gen}}/n < 1$. If copula C_{gen} and C_{imp} satisfy some regularity conditions; see Section 3 of [33],*

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \to \mathcal{N}(0, \Sigma) \tag{16}$$

holds as $n \to \infty$ for some positive definite covariance matrix Σ.

This theorem guarantees the convergence of $\hat{\theta}_n$ with order $1/\sqrt{n}$. In a weaker statement, it tells that the estimated LR tends to the true LR if our parametric copulas correctly specify the true copulas and the sample size is big enough.

4.3 Choosing the best copula pair

Note that the LR at score $\mathbf{s} = (s_1, \cdots, s_{\text{d}})$ under correction factor model (11) can be computed by a *rule of thumb*:

$$\text{LR}^{(C_{\text{gen}},C_{\text{imp}})}(\mathbf{s}) = \prod_{k=1}^{d} \widehat{\text{LR}}_k(s_k) \times \text{CF}_{\hat{\theta}_n,\hat{F}_n}^{(C_{\text{gen}},C_{\text{imp}})}(\mathbf{s}). \tag{17}$$

Here $\widehat{\text{LR}}_k(\cdot)$ and $\hat{\theta}_n$ are given by (10) and (15), respectively, while $\hat{F}_n$ is the modified empirical version of (12), which is obtained by replacing all components in F with their corresponding modified empirical distribution functions. However, the choice of the appropriate parametric copulas can be difficult in practice. Therefore, what we can do is assuming C_{gen} and C_{imp} belong to a family of parametric copulas and choosing the best copula pair. Interestingly, Theorem 2 is still valid if the copula pair $(C_{\text{gen}}, C_{\text{imp}})$ misspecified the true pair. It means that we will get a reasonable estimator of the dependence parameter whichever a copula pair is chosen.

As noted in Section 1, combining classifiers in biometric recognition may have different goals depending on the application or scenario. For a given classifier K, let $e(K)$ be a performance measure of classifier K. Assume that the smaller value of $e(K)$, the better performance of the classifier K. For instances, $e(K)$ can be the equal error rate, total error rate, false rejection rate at certain false accept rate, 1 minus area under ROC curve, and so on.

Let

$$\mathcal{C} = \{C_1 \cdots, C_{n_c}\}$$

be a family of n_c candidate copulas. Since a goodness-of-fit test as provided in [6] will only give the copula pair that is closest to the pair $(c_{\text{gen}}, c_{\text{imp}})$, but whose ratio is not necessarily closest to the ratio $c_{\text{gen}}/c_{\text{imp}}$, we propose to choose the best copula pair as follows. Let $K(i,j)$ be the classifier using copula pair (C_i, C_j) in its correction factor model for $1 \leq i, j, \leq n_c$, which is defined by

$$K_{i,j}(\mathbf{s}) = \text{LR}^{(C_i,C_j)}(\mathbf{s}), \quad \forall \mathbf{s} \in \mathbb{R}^d$$

as given in (17). Given a performance measure e, our model selection will choose (C_x, C_y) as the best copula pair with respect e if $e(K_{x,y})$ has the smallest value among other pairs, i.e.,

$$e(K_{x,y}) \leq e(K_{i,j}), \quad \forall 1 \leq i, j, \leq n_c.$$

If there are two or more best pairs then we choose one of them at random. Note that it is always useful to include the independence copula in $\mathcal{C}$ to guarantee that the chosen fused classifier performs at least as good as fusion under independence.

5 Applications

We will present how our correction factor works and improves the PLR fusion in some practical scenarios: in a standard biometric verification and in forensic scenarios. To approximate the correction factor, we will use the following parametric copulas: independence copula (ind), Gaussian copula (GC), Student's t (t), Frank (Fr), Clayton (Cl), flipped Clayton (fCl), Gumbel (Gu), and flipped Gumbel (fGu). Therefore, the copulas C_{gen} and C_{imp} are chosen from the copula family

$$\mathcal{C} = \{\text{ind}, \text{GC}, \text{t}, \text{Fr}, \text{Cl}, \text{Gu}, \text{fCl}, \text{fGu}\}.$$

These parametric copulas are the same as used in [30, 31].

Suppose that we have genuine and impostor scores as given in (8) and (9), respectively, that will be used to train our method with respect to an evaluation measure e. Our procedure to choose the best copula pair is simple. We randomize the genuine (impostor) scores and take two disjoint subsets with size

$$n_w = \min\{10000, \lfloor n_{\text{imp}}/2 \rfloor\}$$

and

$$n_b = \min\{10000, \lfloor n_{\text{gen}}/2 \rfloor\}.$$

This re-sampling method is aimed at increasing the computation speed because it will be repeated 100 times to see the consistency. After all 64 fused classifiers

$$\mathcal{LR} = \{\text{LR}^{(C_{\text{gen}},C_{\text{imp}})} \ : \ C_{\text{gen}}, C_{\text{imp}} \in \mathcal{C}\}$$

are trained using the first subset, their evaluation measures are then computed on the second subset. Of the $n_c \times n_c$ resulting different fused classifiers we choose the one that minimizes the performance measure e. We then compare the performance of the chosen pair to the PLR method using the paired t-test at significance level 0.01 to see whether the difference is significant or not. If the performance of the chosen pair is significantly different from the PLR method then we use this copula pair in computing the correction factor. Otherwise, we take (ind,ind) as the best pair or in other words we simply use the PLR method. For the Logit and GMM methods, we employ the linear logistic regression as used in [17] for the Logit method while the parameters in the GMM method are fitted by the algorithm proposed in [7], which automatically estimates the number of mixture components using the minimum message length criterion with the minimum and maximum numbers of components being 1 and 20, respectively. Once all fusion strategies have been trained based on the training set, their performances with respect to the performance measure e are computed based on the testing set. The sample sizes of genuine and impostor scores for both training and testing sets of all databases are given in Table 1.

Table 1 Sample size of training and testing sets

Databases	training		testing	
	genuine	impostor	genuine	impostor
NIST-finger	1,000	999,000	5,000	2,4995,000
Face-3D	106,762	21,987,938	46,912	16,005,130
BSS1	968	936,056	968	1,089,000
BSS2P1	1,853	3,431,756	1,853	3,431,756
BSS2P2	1,853	3,431,756	1,853	3,431,756
BSS3	7,252	2,460,980	7,629	2,612,826
XM2VTS	600	40,000	400	111,800

5.1 Maximizing TMR at Fixed FMR

In standard biometric verification, one has to set a threshold Δ such that a score greater than or equal to the threshold is recognized as genuine score while a score less than the threshold is recognized as impostor score. Therefore, a biometric recognition system can make two different errors: accept an impostor score as genuine score and reject a genuine score. The probability of accepting an impostor score is called the *False Match Rate* ($\text{FMR}(\Delta)$) with threshold Δ, while the probability of rejecting a genuine score is called the *False Non-Match Rate* ($\text{FNMR}(\Delta)$). The complement of the $\text{FNMR}(\Delta)$ is called the *True Match Rate* ($\text{TMR}(\Delta)$), which is defined as the probability of accepting a genuine score as genuine score. Since every genuine score will be either accepted or rejected by the system, we have $\text{TMR}(\Delta) = 1 - \text{FNMR}(\Delta)$. The most common method to see the performance of a biometric person verification system is by plotting the relation between $\text{FMR}(\Delta)$ and $\text{TMR}(\Delta)$ for all $\Delta \in (-\infty, \infty)$, which is known as *Receiver Operating Characteristic* (ROC) [5].

Performance measure: The threshold can also be determined by putting a FMR value in advance. For a given fixed $\text{FMR} = \alpha$, the corresponding TMR value can be estimated based on data. Let

$$W_1, \cdots, W_{n_{\text{gen}}} \tag{18}$$

$$B_1, \cdots, B_{n_{\text{imp}}} \tag{19}$$

be 1-dimensional genuine and impostor scores, respectively. In our case, these are the fused scores of the testing set. According to [9], the TMR_α can be estimated by

$$1 - \hat{F}^-_{\text{gen}}(\hat{Q}_{\hat{F}^-_{\text{imp}}}(1-\alpha))$$

where $\hat{F}^-_{\text{gen}}$ and $\hat{F}^-_{\text{imp}}$ are *left-continuous* empirical distribution functions based on (18) and (19), respectively while $\hat{Q}_{\hat{F}^-_{\text{imp}}}$ is the empirical quantile function with respect to $\hat{F}^-_{\text{imp}}$. Slightly different from (14), the left-continuous empirical distribution function based on a sample $X_1, \cdots, X_n$ is defined by

$$\hat{F}^-_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i < x\}}, \quad \forall x \in \mathbb{R} \tag{20}$$

and its corresponding quantile function is defined by

$$\hat{Q}_{\hat{F}^-_n}(p) = \sup\{y \ : \ \hat{F}^-_n(y) \leq p\}, \quad \forall p \in [0,1]. \tag{21}$$

Since higher TMR leads to a better classifier, the performance measure in this standard verification scenario is $e = 1 - \text{TMR}_\alpha$.

Databases: We use NIST-finger [19] and Face-3D [22, 25, 27, 28] data to simulate fingerprint and face authentication, respectively.

Table 2 Best copula pair at several FMRs of our method on the NIST-finger and Face-3D databases

Databases	Best pair at FMR		
	10^{-5}	10^{-4}	10^{-3}
NIST-finger	{Gu,t}	{Gu,t}	{ind,ind}
Face-3D	{ind,Fr}	{ind,Fr}	{ind,Fr}

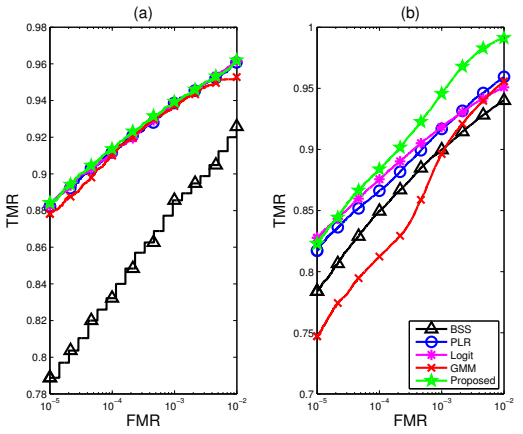

Fig. 1: ROC curves of different fusion strategies on (a)NIST-finger (b)face-3D

• NIST-finger: NIST-finger contains fingerprint similarity scores from one system run on images of 6000 subjects. Each subject has one left index and one right index fingerprint both in the gallery and probe sets. All comparison scores of all pairs of left index fingerprints and all pairs of right index fingerprints are then computed. Here, we can consider the comparison scores based on left and right index fingerprints to be the first and second classifiers that will be combined. We use the first 1000 subjects for training and the rest for testing.
• Face-3D: Face-3D is used in [27, 28] for 3D face recognition. The training and the testing set are already defined and contain very different images (taken with different cameras, backgrounds, poses, expressions, illuminations and time). In his papers, the author proposes 30 classifiers operating on 30 different facial regions. We only use 5 regions out of these 30: similarity of the full face, the left half, the right half, the bottom part, and the upper part of the face. This choice is made to have dependent classifiers.

Results: We train our method at FMR 10^{-5}, 10^{-4}, and 10^{-3}. The best copula pairs and the TMRs for all scenarios are given by Table 2 and Table 3, respectively. It is shown that on the NIST-finger database the improvement of our method compared to the PLR method is relatively small and that all fusion methods have almost the same performance as seen in Figure 1, which shows that the ROC curves of all fusion methods almost coincide. On the other hand, the improvement of our method compared to the PLR method can be clearly seen by the Face-3D database. This phenomenon occurs because the left and right index fingerprints are almost independent while the overlapping regions on the Face-3D database are dependent. Interestingly, the dependence on the Face-3D database cannot be captured by the GMM method and this GMM method even performs worse than the best single matcher (BSM). This happens because the estimated number of components in the GMM method is equal to the maximum value (20) that we chose. This suggests that the number of components might be more than 20. However, if we increase the number of components then the estimator becomes less reliable.

Table 3 The TMRs of different fusion strategies on the NIST-finger and Face-3D databases. The bold number in every column is the best one.

Methods	NIST-finger			Face-3D		
	TMR at FMR					
	10^{-5}	10^{-4}	10^{-3}	10^{-5}	10^{-4}	10^{-3}
BSM	0.793	0.835	0.887	0.784	0.849	0.900
PLR	0.882	0.912	**0.939**	0.817	0.866	0.917
Logit	0.883	0.911	**0.939**	**0.828**	0.876	0.918
GMM	0.878	0.910	0.937	0.747	0.812	0.896
Proposed	**0.884**	**0.914**	**0.939**	0.823	**0.884**	**0.946**

5.2 *Minimizing HTER*

Performance measure: Besides maximizing the TMR at certain FMR, one may also be interested in minimizing some types of error:

• Equal error rate EER: Let Δ^* be the threshold value at which $\mathrm{FMR}(\Delta^*)$ and $\mathrm{FNMR}(\Delta^*)$ are equal. Then EER is defined as the common value $\mathrm{EER} = \mathrm{FMR}(\Delta^*) = \mathrm{FNMR}(\Delta^*)$.
• Total error rate $\mathrm{TER}(\Delta)$: The sum of the $\mathrm{FMR}(\Delta)$ and the $\mathrm{FNMR}(\Delta)$, i.e., $\mathrm{TER}(\Delta) = \mathrm{FMR}(\Delta) + \mathrm{FNMR}(\Delta)$. One may also consider the half total error rate $\mathrm{HTER}(\Delta) = \mathrm{TER}(\Delta)/2$, to keep the error value between 0 and 1.
• Weighted error rate $\mathrm{WER}_\beta(\Delta)$: A weighted sum of the $\mathrm{FMR}(\Delta)$ and the $\mathrm{FNMR}(\Delta)$, i.e., $\mathrm{WER}_\beta(\Delta) = \beta\mathrm{FMR}(\Delta) + (1-\beta)\mathrm{FNMR}(\Delta)$, $\beta \in [0,1]$. The weights are usually called cost of false acceptance and cost of false rejection.
• Area under ROC curve (AUC).

Here, Δ is the threshold to compute the FMR and FNMR. Note that for a given $\beta \in [0,1]$, we can set the performance measure $e = \inf_\Delta \mathrm{WER}_\beta(\Delta)$ and follow our procedure to get the best copula pair. To give an illustration, we will put $\beta = 0.5$, which leads to $\inf_\Delta \mathrm{HTER}(\Delta)$. Frequently, the minimum value of HTER is approximated by the EER. This EER is used to report a fusion performance in [14, 32]. However, as pointed out in [23], the corresponding Δ^* is only a decision threshold and hence EER should not be used to measure performance. To report fusion performance itself, they suggest to set the threshold Δ^* using the training set and to report the final performance by computing the $\mathrm{HTER}(\Delta^*)$ on the testing set. Therefore, we train our method by following this procedure but adapted as follows. Once all 64 copula based fusion strategies have been trained on the first subset of the training set, they are applied to the first subset of the training set to determine the threshold Δ^* and to the second subset of the training set to compute the $\mathrm{HTER}(\Delta^*)$. For all our benchmark methods, the threshold Δ^* is determined using the fused scores of training data and the $\mathrm{HTER}(\Delta^*)$ is computed using the fused scores of the testing data.
Databases: We use the same publicly available scores databases as used in [14] from video-based gait biometrics. There are 4 databases in which their training and testing sets are already clearly defined.

• BSS1: This database contains three-dimensional scores based on the gait energy image (GEI), gait period, and height of the subject [10].
• BSS2P1: This database is composed of three-dimensional scores based on the GEI and 1- and 2-times frequency elements in the frequency-domain feature [10].
• BSS2P2: This database is almost the same as the BSS2P1 database but the scores are computed based on the GEI, chrono-gait image, and gait low image [10].
• BSS3: This database is composed of two-dimensional scores from a wearable accelerometer and a gyroscope sensor [21].

Results: The HTERs of different fusion strategies are reported in Table 4. We do not present the performance of other fusion strategies used in [14] because it is already shown there that the pseudo likelihood ratio method is always among the first or second best results for all databases. Hence we are mostly interested in how our method can improve the PLR method. Interestingly, we can see that

Table 4 The HTERs of different fusion strategies on the video-based gait databases. The bold number in every column is the best one.

Methods	BSS1	BSS2P1	BSS2P2	BSS3
BSM	0.042	0.050	0.048	0.150
PLR	0.035	0.056	0.042	0.132
Logit	**0.034**	0.052	0.041	0.136
GMM	**0.034**	**0.047**	**0.034**	0.134
Proposed	**0.034**	**0.047**	0.035	**0.131**
Best pair	{Gu,fCl}	{Gu,GC}	{t,t}	{Gu,Cl}

the PLR method performs worse than the best single classifier on the BSS2P1 database. This may be because ignoring dependence of dependent classifiers will degrade the performance of the fusion. This is confirmed by the performance of our method, which does take the dependence into account. It is better than the best single classifier. We can also see that the GMM method is comparable to our method for all databases. Apparently, the GMM method fits the dependence structures quite well.

5.3 Minimizing discrimination loss

The last application of our method concerns forensic biometric scenarios. Unlike the standard biometric verification that gives a *hard decision* whether a score is genuine or impostor, the likelihood ratio value in the forensic case only provides a *soft decision*, which can be used to support the judge in court to make an objective decision [4].
Performance measure: Fusion is hoped to integrate the complementary information from the individual classifiers. In a forensic scenario one aims at increasing the discrimination power (the ability of distinguishing between genuine and impostor scores). Brümmer and du Preez [2] introduce a measure called the *cost of log likelihood ratio* (C_{llr}) in the field of speaker recognition, which may be interpreted as a summary statistic for a LR computation [3]. This measure is also used in forensic face scenarios in [15]. Note that the scores are interpretable as likelihood ratios when computing this measure. Given 1-dimensional genuine scores (18), which correspond to the hypothesis of the prosecution, and impostor scores (19), which correspond to the hypothesis of the defense, the cost of log likelihood ratio C_{llr} is defined by

$$C_{\text{llr}} = \frac{1}{2n_{\text{gen}}} \sum_{i=1}^{n_{\text{gen}}} \log_2 \left(1 + \frac{1}{W_i}\right) + \frac{1}{2n_{\text{imp}}} \sum_{j=1}^{n_{\text{imp}}} \log_2 \left(1 + B_j\right). \quad (22)$$

To explain the name of this measure we note that our fused scores are LR values and may be rewritten in terms of the logarithm of LR. The minimum value of the C_{llr} (denoted by $C_{\text{llr}}^{\text{min}}$), which is obtained by plugging the scores after PAV transformation into (22), is called the *discrimination loss*. This measure can be seen as the opposite of discrimination power. The smaller the value of this quantity, the higher the discrimination power. The difference between the C_{llr} and the $C_{\text{llr}}^{\text{min}}$ is called the *calibration loss*

$$C_{\text{llr}}^{\text{cal}} = C_{\text{llr}} - C_{\text{llr}}^{\text{min}}. \quad (23)$$

Calibration is transforming a biometric comparison score to its LR value. It means that the calibration loss $C_{\text{llr}}^{\text{cal}}$ tends to zero if the scores are well-calibrated and grows without bound if the scores are miscalibrated. Since we are interested in having better discrimination power, we put the performance measure $e = C_{\text{llr}}^{\text{min}}$. Nevertheless, the C_{llr} and the discrimination loss will also be reported.
Databases: We use the following databases:

• XM2VTS: There are 8 classifiers in this database: 5 face classifiers and 3 speech classifiers. In order to have an application in the field of speaker recognition, we only take the speech classifiers. Moreover, only the LFCC-GMM and SSC-GMM classifiers are used in this experiment because they have the highest correlation value among all pairs. The training and testing sets are already defined [23].
• Face-3D: The same database as used for the standard verification in Section 5.1.

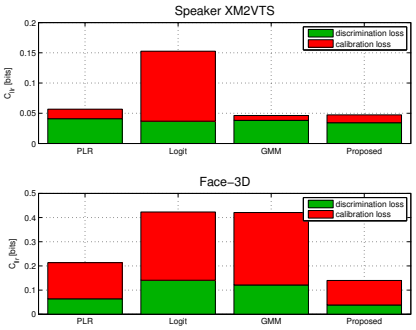

Fig. 2: The discrimination and calibration loss of different fusion strategies on the XM2VTS and Face-3D databases

Table 5 The $C_{\text{llr}}^{\text{min}}$ and C_{llr} values of different fusion strategies on the XM2VTS and Face-3D. The bold number in every column is the best one.

Methods	XM2VTS		Face-3D	
	$C_{\text{llr}}^{\text{min}}$	C_{llr}	$C_{\text{llr}}^{\text{min}}$	C_{llr}
BSM	0.044	0.587	0.072	1.596
PLR	0.041	0.057	0.064	0.214
Logit	0.037	0.153	0.141	0.423
GMM	0.038	**0.046**	0.121	0.421
Proposed	**0.034**	0.047	**0.038**	**0.140**
Best Pair	{Fr,fGu}		{ind,t}	

Results: The $C_{\text{llr}}^{\text{min}}$ and C_{llr} values of different fusion strategies and the best copula pair of our method on the XM2VTS and Face-3D databases are presented in Table 5. Our method outperforms other methods with respect to the performance measure $C_{\text{llr}}^{\text{min}}$. Moreover, the C_{llr} of our method on the XM2VTS database is only slightly higher than the GMM method and on the Face-3D database our method even has by far the smallest C_{llr} among all methods. As before, the GMM method performs poorly even if it is compared to the best single classifier. Surprisingly, even though the Logit method performs better than the PLR method for the standard biometric verification scenario in Section 5.1, its performance is also worse than the best single classifier. This means that the Logit method can discriminate genuine and impostor scores quite well in the tails, but it fails in the middle. Another interesting thing is that if we use the best copula pair {ind,Fr} chosen in Section 5.1 then the corresponding $C_{\text{llr}}^{\text{min}}$ is 0.040 which is higher than the 0.038 for the copula pair {ind,t}, which is trained to minimize the $C_{\text{llr}}^{\text{min}}$ here. It tells us that the copula pair {ind,t} handles dependence on the whole scores better than the copula pair {ind,Fr}, which is trained to handle dependence in the tail. Finally, we also notify that the calibration loss of all fusion strategies (including our method) is pretty high on the Face-3D database as seen in Figure 2. In order to reduce this calibration loss, we proposed in our previous work [31] a method called the *two-step calibration method*. Briefly, the first step of this method is computing both training and testing sets to their fused scores once the best copula pair has been found and the second step is calibrating the fused scores by the PAV algorithm trained based on the fused scores of the training set. Readers who are interested in the detailed explanation of the two-step calibration method may refer to [31].

6 Conclusion

We have presented the mathematical framework of a semiparametric LR-based score level fusion method to improve via parametric copula families the PLR fusion strategy. Estimators of the dependence parameters have been provided and subsequently their convergence has been analyzed. It has also been shown in detail how our LR-based method is used and how the best copula pair is chosen that depends on a specific application. Finally, application to standard biometric verification and forensic scenarios has been demonstrated on real databases from fingerprint, face, speaker, and video-based gait recognition, and it has been confirmed that our LR-based method outperforms the GMM and Logit fusion methods, which are also designed to handle dependence.

7 Acknowledgements

This research is supported by the Netherlands Organisation for Scientific Research (NWO) via the project Forensic Face Recognition, 727.011.008.

8 References

1 Ali, T., Spreeuwers, L.J., Veldhuis, R.N.J.: 'Forensic face recognition: A survey'. In A. Quaglia and C. M. Epifano, editors, *Face Recognition: Methods, Applications and Technology*, 2012, Computer Science, Technology and Applications, page 9. Nova Publishers
2 Brümmer, N., du Preez, and J.: 'Application-independent evaluation of speaker detection', *Computer Speech & Language*, 2006, **20**, (2), pp. 230–275
3 Brümmer, N., de Villiers, E.: 'The BOSARIS toolkit: Theory, algorithms and code for surviving the new DCF', 2013, *CoRR*, abs/1304.2865
4 Facey, O.E, Davis, R.J.: 'Re: Expressing evaluative opinions; a position statement',*Science & Justice*, 2011, **51**, (4), pp. 212 – 212
5 Fawcett, T.: 'An introduction to roc analysis', *Pattern Recogn. Lett.*, 2006, **27**, (8), pp. 861–874
6 Fermanian, J.D.: 'Goodness-of-fit tests for copulas', *Journal of Multivariate Analysis*, 2005, **95**, (1), pp. 119 – 152
7 Figueiredo, M.A.T, Jain, A.K: 'Unsupervised learning of finite mixture models', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, **24**, (3), pp. 381–396
8 Genest, C., Ghoudi, K., Rivest, L.-P.: 'A semiparametric estimation procedure of dependence parameters in multivariate families of distributions', *Biometrika*, 1995, **82**, (3), pp. 543–552
9 Hsieh, F., Turnbull, B.W.: 'Nonparametric and semiparametric estimation of the receiver operating characteristic curve'., *Ann. Statist.*, 1996, **24**, (1), pp. 25–40
10 Iwama, H., Okumura, M., Makihara, Y., Yagi, Y.: 'The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition', *Information Forensics and Security, IEEE Transactions on*, 2012, **7**, (5), pp. 1511–1521
11 Iyengar, S.G., Varshney, P.K., Damarla, T.: 'A parametric copula-based framework for hypothesis testing using heterogeneous data', *IEEE Transactions on Signal Processing*, 2011, **59**, (5), pp. 2308–2319
12 Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: 'On combining classifiers', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, **20**, (3), pp. 226–239
13 Ma, Y., Cukic, B., Singh, H.: 'A classification approach to multi-biometric score fusion', In *Proceedings of the 5th International Conference on Audio- and Video-Based Biometric Person Authentication*, AVBPA'05, 2005, pp. 484–493, Berlin, Heidelberg, Springer-Verlag
14 Makihara, Y., Muramatsu, D., Iwama, H., Ngo, T.T., Yagi, Y., Hossain, M.A.: 'Score-level fusion by generalized delaunay triangulation', In *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pp. 1–8
15 Mandasari, M.I., Gulnther, M., Wallace, R., Saeidi, R., Marcel, S., van Leeuwen, D.A.: 'Score calibration in face recognition', *IET Biometrics*, 2014, **3**, (4), pp. 246–256
16 Morrison, G.S.: 'A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus gaussian mixture model and universal background model (GMM-UBM)', *Speech Communication*, 2011, **53**, (2), pp. 242 – 256
17 Morrison, G.S.: 'Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio', *Australian Journal of Forensic Sciences*, 2013, **45**, (2), pp. 173–197
18 Nandakumar, K., Chen, Y., Dass, S.C, Jain, A.K.: 'Likelihood ratio-based biometric score fusion', *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2008, **30**, (2), pp. 342–347
19 National Institute of Standards and Technology, NIST biometric scores set - release 1, 2004, Available at http://www.itl.nist.gov/iad/894.03/biometricscores.
20 Neyman, J., Pearson, E.S.: 'On the problem of the most efficient tests of statistical hypotheses', *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 1933, **231**, (694-706), pp. 289–337
21 Nagahara, H., Sagawa, R. , Mukaigawa, Y., Yagi, Y., Trung, N.T., Makihara, Y.: 'Performance evaluation of gait recognition using the largest inertial sensor-based gait database', In *The 5th IAPR International Conference on Biometrics (ICB 2012)*, 2012
22 Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: 'Overview of the face recognition grand challenge', In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, Washington, DC, USA, 2005, pp. 947–954
23 Poh, N., and Bengio, S.: 'Database, protocols and tools for evaluating score-level fusion algorithms in biometric authentication', *Pattern Recogn.*, 2006, **39**, (2), pp. 223–233
24 Poh, N., Bourlai, T., Kittler, J., Allano, L., Alonso-Fernandez, F., Ambekar, O., Baker, J., Dorizzi, B., Fatukasi, O., Fierrez, J., Ganster, H., Ortega-Garcia, J., Maurer, D., Salah, A.A., Scheidat, T., Vielhauer, C.,:'Benchmarking Quality-Dependent and Cost-Sensitive Score-Level Multimodal Biometric Fusion Algorithms', *IEEE Transactions on Information Forensics and Security*, 2009, **4**, (4), pp. 849–866
25 Savran, A., Alyüz, N, Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L.: 'Biometrics and identity management', chapter Bosphorus Database for 3D Face Analysis, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 47–56
26 Sklar, M., *Fonctions de Répartition À N Dimensions Et Leurs Marges*, Université Paris 8, 1959.
27 Spreeuwers, L.: 'Breaking the 99% barrier: optimisation of three-dimensional face recognition', *Biometrics, IET*, 2015, **4**, (3), pp. 169–178
28 Spreeuwers, L.: 'Fast and accurate 3D face recognition', *International Journal of Computer Vision*, 2011, **93**, (3), pp. 389–414
29 Susyanto, N., Klaassen, C.A.J., Veldhuis, R.N.J, Spreeuwers, L.J.: 'Semiparametric score level fusion: Gaussian copula approach', In *Proceedings of the 36th WIC Symposium on Information Theory in the Benelux, Brussels*, Brussels, Université Libre de Bruxelles, 2015, pp. 26–33
30 Susyanto, N., Veldhuis, R.N.J, Spreeuwers, L.J., Klaassen, C.A.J.: 'Fixed far correction factor of score level fusion', In *Proceedings of the 8th BTAS2016*, Buffalo, NY, USA, 2016, pp. 1–8
31 Susyanto, N., Veldhuis, R.N.J, Spreeuwers, L.J., Klaassen, C.A.J.: 'Two-step calibration method for multi-algorithm score-based face recognition systems by minimizing discrimination loss, In *Proceedings of the 9th ICB2016*, Halmstad, Sweden, 2016, pp. 1–8
32 Tao, Q., Veldhuis, R.N.J.: 'Robust biometric score fusion by naive likelihood ratio via receiver operating characteristics', *IEEE Transactions on Information Forensics and Security*, 2013, **8**, (2), pp. 305–313
33 Fan, Y., Chen, X.: 'Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection', *The Canadian Journal of Statistics*, 2005, **33**, (3), pp. 389–414
34 Zadrozny, B., Elkan, E.: 'Transforming classifier scores into accurate multiclass probability estimates', In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, New York, NY, USA, 2002. ACM, pp. 694–699

9 Appendix

Proof of Theorem 2

According to Proposition 2 of Chen and Fan [33], we have

$$\sqrt{n_{\text{gen}}}\left(\hat{\theta}_{\text{gen},n_{\text{gen}}} - \theta_{\text{gen}}\right) \to \mathcal{N}(0, \Sigma_{\text{gen}})$$

and

$$\sqrt{n_{\text{imp}}}\left(\hat{\theta}_{\text{imp},n_{\text{imp}}} - \theta_{\text{imp}}\right) \to \mathcal{N}(0, \Sigma_{\text{imp}})$$

for some positive definite matrices Σ_{gen} and Σ_{imp}. Define $\lambda_n = n_{\text{gen}}/n$ with $\lim_{n\to\infty} \lambda_n = \lambda$. Since $\hat{\theta}_{\text{gen},n_{\text{gen}}}$ and $\hat{\theta}_{\text{imp},n_{\text{imp}}}$ are independent then

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}_n - \theta\right) &= \begin{pmatrix} \sqrt{n_{\text{gen}}/\lambda_n}\left(\hat{\theta}_{\text{gen},n_{\text{gen}}} - \theta_{\text{gen}}\right) \\ \sqrt{n_{\text{imp}}/(1-\lambda_n)}\left(\hat{\theta}_{\text{imp},n_{\text{imp}}} - \theta_{\text{imp}}\right) \end{pmatrix} \\ &\to \mathcal{N}(0, \Sigma)\end{aligned}$$

where

$$\Sigma = \begin{pmatrix} \Sigma_{\text{gen}}/\lambda & 0 \\ 0 & \Sigma_{\text{imp}}/(1-\lambda) \end{pmatrix}.$$

IET Research Journals, pp. 1–8