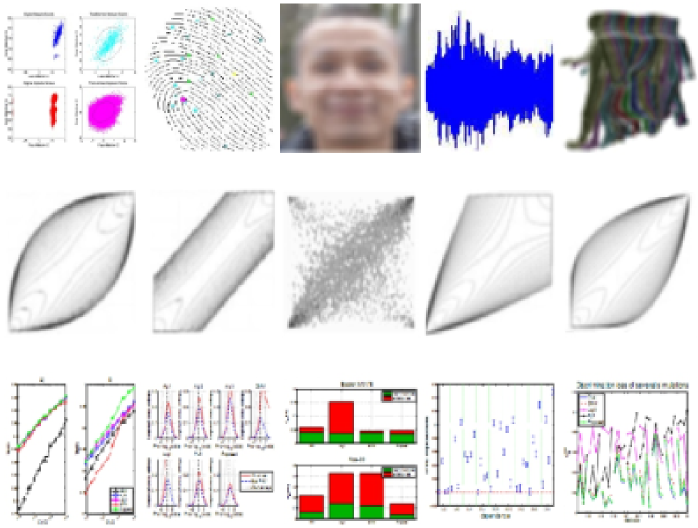
Semiparametric Copula Models for Biometric Score Level



Nanang Susyanto

SEMIPARAMETRIC COPULA MODELS FOR BIOMETRIC SCORE LEVEL FUSION

Semiparametric Copula Models for Biometric Score Level Fusion
PhD Thesis, Universiteit van Amsterdam

Layout	Nanang Susyanto
Printed by	Proefschriftmaken.nl & Uitgeverij BOXPress
ISBN	978-94-6295-513-4

SEMIPARAMETRIC COPULA MODELS FOR BIOMETRIC SCORE LEVEL FUSION

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties
ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 11 oktober 2016, te 14:00 uur
door

NANANG SUSYANTO
geboren te Temanggung, Indonesië

Promotiecommissie

Promotores:	Prof. dr. C.A.J. Klaassen	Universiteit van Amsterdam
	Prof. dr. ir. R.N.J. Veldhuis	Universiteit Twente
Copromotor:	dr. ir. L.J. Spreeuwers	Universiteit Twente
Overige leden:	dr. A.J. van Es	Universiteit van Amsterdam
	Prof. dr. M.R.H. Mandjes	Universiteit van Amsterdam
	Prof. dr. M.J. Sjerps	Universiteit van Amsterdam
	Prof. dr. J.H. van Zanten	Universiteit van Amsterdam
	Prof. dr. R.D. Gill	Universiteit Leiden
	Prof. dr. J. Kittler	University of Surrey
	dr. A. Ross	Michigan State University

Faculteit: Faculteit der Natuurwetenschappen, Wiskunde en Informatica

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

to my 3R:

Rini, Razka, Rasydan

SEMIPARAMETRIC COPULA MODELS FOR BIOMETRIC SCORE LEVEL FUSION

Nanang Susyanto

According to the Doctorate Regulations of the University of Amsterdam, the following list and explanation is provided.

This thesis is based on many intensive discussions between the author and his supervisors. The main body of this thesis consists of the following six papers written by the author and his supervisors.

[1] Klaassen, C.A.J. and Susyanto, N., 2015. *Semiparametrically efficient estimation of constrained Euclidean parameters.* arXiv:1508.03416.

[2] Klaassen, C.A.J. and Susyanto, N., 2016. *Semiparametrically efficient estimation of Euclidean parameters under equality constraints.* arXiv:1606.07749.

[3] Susyanto, N., Klaassen, C.A.J., Veldhuis, R.N.J. and Spreeuwers, L.J., 2015. *Semiparametric score level fusion: Gaussian copula approach.* In: Proceedings of the 36th WIC Symposium on Information Theory in the Benelux, 6-7 May 2015, Brussels. pp. 26-33.

[4] Susyanto, N., Veldhuis, R.N.J., Spreeuwers, L.J. and Klaassen, C.A.J., 2016. *Two-step calibration method for multi-algorithm score-based face recognition systems by minimizing discrimination loss.* In: Proceedings of the 9th IAPR International Conference on Biometrics, 13-16 June 2016, Halmstad.

[5] Susyanto, N., Veldhuis, R.N.J., Spreeuwers, L.J. and Klaassen, C.A.J., 2016. *Fixed FAR correction factor of score level fusion.* Accepted for publication at the 8th IEEE International Conference on Biometrics: Theory, Applications, and Systems , 6-9 September 2016, New York.

[6] Susyanto, N., Veldhuis, R.N.J., Spreeuwers, L.J. and Klaassen, C.A.J., 2016. *Semiparametric likelihood-ratio-based score level fusion via parametric copula.* In preparation.

Regarding co-authorship, co-authors confirm that the main contribution to these publications is by the author.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Preliminaries

1.1.1 Biometric verification

In biometric recognition systems, biometric samples (images of faces, fingerprints, voices, gaits, etc.) of people are compared and classifiers (matchers) indicate the level of similarity between any pair of samples by a 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. The scope of this thesis is about biometric *verification* (also known as authentication) in the sense that two biometric samples are compared to find out if they originate from the same person or stem from different people, without making any identity claim. Except when stated specifically, the random variables genuine score S_{gen} and impostor score S_{imp} are assumed to be continuous, taking values in the real line $\mathbb{R} = (-\infty, \infty)$ with distribution functions F_{gen} and F_{imp} and density functions f_{gen} and f_{imp}, respectively.

1.1.1.1 Standard biometric verification

A standard biometric verification system that takes *hard decisions*, will decide whether a matching score between two biometric samples is a genuine or an impostor score via a *threshold* Δ chosen in advance. A score greater than or equal to this threshold is classified as genuine score, while a score less than this threshold is classified as impostor score. Once this threshold has been chosen, the 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 Acceptance Rate* (FAR(Δ)) or *False Match Rate* (FMR(Δ)) with threshold Δ, while the probability of rejecting a genuine score is called the *False Rejection Rate* (FRR(Δ)) or *False Non-Match Rate* (FNMR(Δ)). The complement of the FRR(Δ) is called the *True Positive Rate* (TPR(Δ)) or *True Match Rate* (TMR(Δ)), 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 $\text{TPR}(\Delta) = 1 - \text{FRR}(\Delta)$. Theoretically, the FAR and the TPR can be computed as

$$\text{FAR}(\Delta) = 1 - F_{\text{imp}}(\Delta) \tag{1.1.1}$$

and

$$\text{TPR}(\Delta) = 1 - F_{\text{gen}}(\Delta) \tag{1.1.2}$$

for every threshold Δ. By varying Δ from $-\infty$ to ∞, we can plot the relation between FAR and TPR as a curve known as the *Receiver Operating Characteristic* (ROC) [1]. Mathematically, the function $\text{ROC} : [0,1] \to [0,1]$ maps FAR values to the corresponding TPR values via relation

$$\text{TPR}_\alpha = \text{ROC}(\alpha) = 1 - F_{\text{gen}}(F_{\text{imp}}^{-1}(1-\alpha)) \tag{1.1.3}$$

for every $\text{FAR} = \alpha \in [0,1]$ where F_{imp}^{-1} is the quantile function defined by

$$F_{\text{imp}}^{-1}(p) = \sup\{x \in \mathbb{R} \ : \ F_{\text{imp}}(x) \leq p\}, \quad \forall p \in [0,1].$$

The following performance measures are often used in biometric verification and will be used in Chapter 3.

- Area under ROC curve (AUC), i.e.,

$$\text{AUC} = \int_0^1 \text{ROC}(\alpha) d\alpha. \tag{1.1.4}$$

- Equal error rate EER: Let Δ^* be the threshold value at which FAR(Δ^*) and FRR(Δ^*) are equal. Then EER is defined as the common value

$$\text{EER} = \text{FAR}(\Delta^*) = \text{FRR}(\Delta^*). \tag{1.1.5}$$

- Total error rate TER(Δ): The sum of the FAR(Δ) and the FRR(Δ). One may also consider the half total error rate (HTER), which is one half of the

TER, to keep the error value between 0 and 1, i.e.,

$$\mathrm{TER}(\Delta) = \mathrm{FAR}(\Delta) + \mathrm{FRR}(\Delta) \quad \text{and} \quad \mathrm{HTER}(\Delta) = \mathrm{TER}(\Delta)/2. \tag{1.1.6}$$

- Weighted error rate WER(Δ): A weighted sum of the FAR(Δ) and the FRR(Δ), i.e.,

  $$\mathrm{WER}_\beta(\Delta) = \beta\mathrm{FAR}(\Delta) + (1-\beta)\mathrm{FRR}(\Delta), \ \beta \in [0,1]. \tag{1.1.7}$$

  the weights are usually called cost of false acceptance and cost of false rejection.

Here, Δ is the threshold to compute the FAR and FRR.

In practice, F_{gen} and F_{imp} are replaced by their empirical versions based on data. Let

$$W_1, \cdots, W_{n_{\text{gen}}} \tag{1.1.8}$$

$$B_1, \cdots, B_{n_{\text{imp}}} \tag{1.1.9}$$

be i.i.d copies of S_{gen} and S_{imp}, respectively. Then, all quantities given by (1.1.1) - (1.1.7) can be estimated by replacing F_{gen} and F_{imp} with their *left-continuous* empirical distribution functions $\hat{F}^-_{\text{gen}}$ and $\hat{F}^-_{\text{imp}}$ based on (1.1.8) and (1.1.9), respectively. The left-continuous empirical distribution function based on a sample $X_1, \ldots, X_n$ is defined by

$$\hat{F}^-_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i < x\}}, \quad \forall x \in \mathbb{R}. \tag{1.1.10}$$

In particular, the empirical version of the ROC function (1.1.3) is shown to be almost surely convergent to the true ROC [2].

1.1.1.2 Forensic biometric verification

In forensic science it is gradually accepted that, instead of giving a hard decision whether a matching score is a genuine or impostor score, a biometric system should give a *soft decision*, namely an evidential value in terms of a likelihood ratio (LR); see [3]. The LR of a matching score s is defined as the ratio between the densities of the genuine scores f_{gen} and the impostor scores f_{imp}, i.e.,

$$\mathrm{LR}(s) = \frac{f_{\text{gen}}(s)}{f_{\text{imp}}(s)}. \tag{1.1.11}$$

The hypothesis that the matching score s is a genuine score is called the hypothesis of the prosecutor and it is denoted by H_{p}. Similarly, H_{d} denotes the hypothesis of the defendant that the score is an impostor score. In forensic scenarios, the score s is usually known as *evidence* and the hypotheses H_{p} and H_{d} are defined as two mutually exclusive hypotheses supporting whether or not the suspect is the donor of the biometric trace. The LR(s) may be interpreted roughly as the probability that the evidence is s given the hypothesis H_{p} divided by this probability given H_{d}. It is computed by a forensic scientist and can be used to support the fact finder (judge/jury) in court to make an objective decision. The Bayesian framework explains elegantly how LR(s) supports the decision via

$$\frac{P(H_{\mathrm{p}})}{P(H_{\mathrm{d}})} \times \frac{f_{\mathrm{gen}}(s)}{f_{\mathrm{imp}}(s)} = \frac{P(H_{\mathrm{p}}|s)}{P(H_{\mathrm{d}}|s)}. \tag{1.1.12}$$

This means that the LR can be interpreted as a multiplicative factor to update the *prior odds* in favor of H_{p} versus H_{d} (before the evidence s has been taken into account) to the *posterior odds* (after the evidence has been taken into account).

Not all matchers in biometric recognition will give a LR value as matching score. Therefore, we need to transform such a matching score to its corresponding LR value; a process known as *calibration*, which uses definition (1.1.11). Several methods of computing the LR from a biometric comparison score have been proposed and evaluated in forensic scenarios. Briefly, there are four common calibration methods: Kernel Density Estimation (KDE), Logistic Regression (Logit), Histogram Binning (HB), and Pool Adjacent Violators (PAV) methods; see [4, 5] for a survey of these methods.

There are two types of measures for the reliability of calibration methods: *application-dependent* [6,7] and *application-independent* [8–11] measures. Since forensic scientists do not have access to the prior odds, this thesis will use the application-independent ones. The two following performance measures will be used in Chapter 4.

Cost of log likelihood ratio The cost of log likelihood ratio (C_{llr}) is introduced by Brümmer and du Preez [8] in the field of speaker recognition, is based on a generalization of cost evaluation metrics, and is used in forensic face scenarios in [12]. This measure is an expected cost

$$C = C_{\mathrm{fa}}\mathrm{FAR} + C_{\mathrm{fr}}\mathrm{FRR}$$

for any value of the cost of false acceptance C_{fa} and the cost of false rejection C_{fr}; see [8] for a detailed explanation. Given genuine scores (1.1.8), which correspond to the hypothesis of the prosecution, and impostor scores (1.1.9), which correspond to the hypothesis of the defense, the cost of log likelihood ratio C_{llr} is defined by

$$\begin{aligned} C_{\mathrm{llr}} = {} & \frac{1}{2n_{\mathrm{gen}}} \sum_{i=1}^{n_{\mathrm{gen}}} \log_2 \left(1 + \frac{1}{W_i} \right) \\ & + \frac{1}{2n_{\mathrm{imp}}} \sum_{j=1}^{n_{\mathrm{imp}}} \log_2 \left(1 + B_j \right). \end{aligned} \tag{1.1.13}$$

To explain the name of this measure we note that the scores are interpretable as LR values, more precisely as estimates of the LR and its inverse, and that they are rewritten in terms of the logarithm of 1+LR. Interestingly, this metric can be decomposed into a *discrimination* and *calibration* form via relation

$$C_{\mathrm{llr}} = C_{\mathrm{llr}}^{\mathrm{min}} + C_{\mathrm{llr}}^{\mathrm{cal}}. \tag{1.1.14}$$

Here, $C_{\mathrm{llr}}^{\mathrm{min}}$ and $C_{\mathrm{llr}}^{\mathrm{cal}}$ denote the discrimination and calibration loss, respectively. Discrimination loss is the opposite of discrimination power (the ability of the system to distinguish between genuine and impostor scores). The smaller the value of this quantity, the higher the discrimination power. The $C_{\mathrm{llr}}^{\mathrm{min}}$ is defined as the minimum C_{llr} value based on given scores (1.1.8) and (1.1.9), by preserving the discrimination power which is attained by the Pool-Adjacent-Violators (PAV) algorithm as proved in [8]. Therefore, we will have $0 \leq C_{\mathrm{llr}}^{\mathrm{min}} \leq 1$, where 0 represents the perfect system, i.e., it always gives ∞ for every genuine score and 0 for every impostor score, whereas 1 represents the neutral system, i.e., it always gives LR $= 1$ for every score. We also have $0 \leq C_{\mathrm{llr}}^{\mathrm{cal}} < \infty$ where 0 is for well-calibrated scores and grows without bound if the scores are miscalibrated.

ECE plot The Empirical Cross Entropy (ECE) plot is a generalization of the C_{llr} for measuring the reliability of calibration with an information theoretical interpretation [10]. The ECE is defined as the estimated value of the cross entropy for all possible values of the evidence E:

$$H_{Q\|P}(H|E) = -\sum_{i\in\{\mathrm{p,d}\}} Q(H_i) \int_{-\infty}^{\infty} q(e|H_i) \log_2 P(H_i|e) de, \tag{1.1.15}$$

where P denotes the posterior probability using a forensic system, Q is a distribution such that

$$\begin{cases} Q(H_\mathrm{p}|E) = 1, & \text{if } H_\mathrm{p} \text{ is true} \\ Q(H_\mathrm{d}|E) = 1, & \text{if } H_\mathrm{d} \text{ is true,} \end{cases}$$

called the *oracle* distribution with density function q. Note that this oracle distribution represents the posterior probability if the judge already knew the true hypotheses H_p and H_d. Therefore, the cross entropy $H_{Q\|P}(H|E)$ can be interpreted as an additional information loss because it was expected that the system computed Q, not P. So, the ECE at log prior odds $lp \in (-\infty, \infty)$ based on (1.1.8) and (1.1.9) can be computed as follows:

$$\begin{aligned} \mathrm{ECE}(lp) = {} & \frac{1}{2n_\mathrm{gen}} \sum_{i=1}^{n_\mathrm{gen}} \log_2 \left(1 + \frac{1}{W_i \times e^{lp}}\right) \\ & + \frac{1}{2n_\mathrm{imp}} \sum_{j=1}^{n_\mathrm{imp}} \log_2 \left(1 + B_j \times e^{lp}\right). \end{aligned} \qquad (1.1.16)$$

Clearly $C_\mathrm{llr} = \mathrm{ECE}(0)$ holds, which shows that the ECE generalizes the cost of log likelihood ratio. Figure 1.1 is an example of the ECE plot of the linear logistic regression when modelling gaussian mixture scores. The solid red curve represents the performance of the calibration, the dashed blue curve is the minimum ECE value under evaluation by preserving the discrimination power which is attained by PAV transformation, and the dashed black curve is the entropy of the neutral system without considering the evidence, i.e., all LR values equal to 1. The difference between the solid red and dashed blue curves is the calibration loss. We can see that the scores are miscalibrated for log prior odds greater than 2 under the linear logistic regression method.

1.1.2 Likelihood-ratio-based biometric fusion

Suppose we have d matchers, with $d > 1$. A fusion strategy is a function $\psi : \mathbb{R}^d \to \mathbb{R}$ that transforms a concatenated vector of d scores, which will just be called "score" for simplicity, to a scalar named a *fused score.* This process is called score level fusion. Let $\mathbf{S}_\mathrm{gen}$ and $\mathbf{S}_\mathrm{imp}$ denote the genuine and impostor scores with distribution functions $\mathbf{F}_\mathrm{gen}$ and $\mathbf{F}_\mathrm{imp}$ that correspond to density functions $\mathbf{f}_\mathrm{gen}$ and $\mathbf{f}_\mathrm{imp}$, respectively. We use the same notation as in Section 1.1.1, but now in bold face in order to emphasize that we are working in the multivariate case.

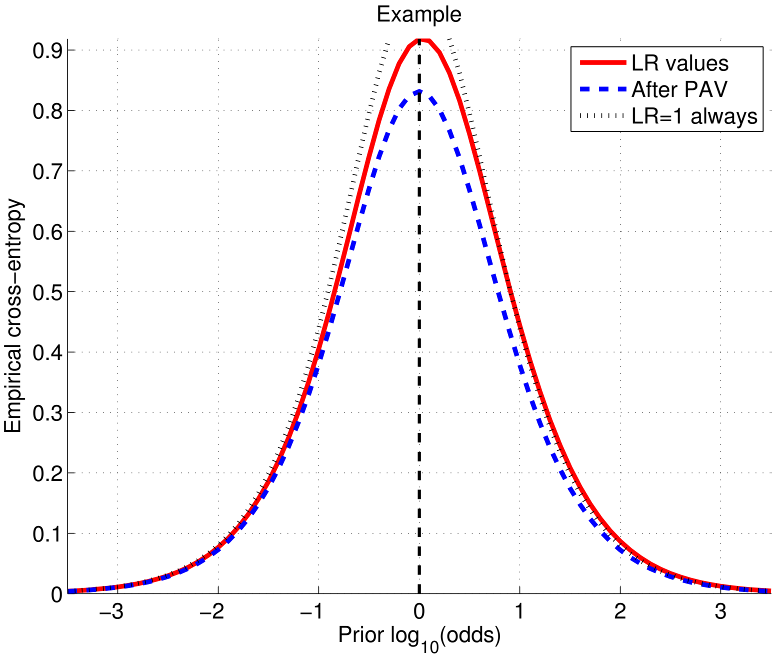


Figure 1.1: ECE plot of linear logistic regression

There are three categories in score level fusion: transformation-based [13], classifier-based [14], and density-based (henceforth called *likelihood-ratio-based*).

1. Transformation-based: the fusion strategy ψ maps all components of the vector of matching scores to a comparable domain and applying some simple rules such as sum, mean, max, med, etc.
2. Classifier-based fusion: the fusion strategy ψ acts as a classifier of the vector of the matching scores to distinguish between genuine and impostor scores.
3. likelihood ratio (LR)-based: the fusion strategy ψ computes the LR as defined by (1.1.11) for the multivariate case, i.e.,

$$\mathrm{LR}(s_1,\cdots,s_d)=\frac{\mathbf{f}_{\mathrm{gen}}(s_1,\cdots,s_d)}{\mathbf{f}_{\mathrm{imp}}(s_1,\cdots,s_d)}. \tag{1.1.17}$$

   for every score $(s_1,\cdots,s_d)$.

The LR-based fusion strategy is theoretically optimal according to the Neyman-Pearson lemma [15] in the sense that it gives the highest TPR at every FAR. Indeed, some experimental results [16,17] show that it consistently performs well compared to the transformation-based and classifier-based. Moreover, the use of the fused score of the LR-based fusion in forensic science is straightforward. In practice, the distributions $\mathbf{f}_{\text{gen}}$ and $\mathbf{f}_{\text{imp}}$ are unknown and have to estimated from data. Therefore, the performance of the LR-based fusion depends on the accuracy of the LR computation. This classical problem in statistics can be solved by parametric (e.g., normal distribution, Weibull distribution) and nonparametric (e.g., histogram, kernel density estimation) models. However, the choice of an appropriate parametric model is sometimes difficult while nonparametric estimators suffer from the difficulty that they are sensitive to the choice of the bandwidth or of other smoothing parameters, especially for our multivariate case. Therefore, it is natural to approach our estimation problem semiparametrically. Note that there are two types of data that will be used: a *training set* for estimating the underlying parameters of a fusion strategy and a disjoint set, which is called the *testing set*, for evaluating performance.

Gaussian copula approach Note that the scores contain two types of dependence. The basic dependence is between two comparisons that involve at least one common person. Even if these comparisons would be independent (e.g. because there are no comparisons that concern the same person), the different classifiers that attach a score to each comparison, may be dependent. If we model the joint distribution of all scores by a (semiparametric) Gaussian copula model, the resulting correlation matrix will be structured. It has many zeros and many correlations have a common value. Estimation of these parameters is a problem in constrained semiparametric estimation, a topic that we study in quite some generality in the Statistical Theory part of this thesis. The Biometric Application part of it focusses on score level fusion and models the dependence between classifiers also by semiparametric copula models.

1.1.3 Semiparametric copula model

A copula is a distribution function on the unit cube $[0,1]^m$, $m \geq 2$, of which the marginals are uniformly distributed. A classical result of Sklar [18] relates any continuous multivariate distribution function to a copula.

Theorem 1.1.1 (Sklar (1959)). *Let $m \geq 2$, and suppose H is a distribution function on $\mathbb{R}^m$ with one dimensional continuous marginal distribution functions $F_1, \cdots, F_m$. Then there is a unique copula C so that*

$$H(x_1, \ldots, x_m) = C(F_1(x_1), \ldots, F_m(x_m)) \tag{1.1.18}$$

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

The joint density function can be computed by taking the m-th derivative of (1.1.18):

$$\begin{aligned} h(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \\ \times \prod_{i=1}^{d} f_i(x_i) \end{aligned} \tag{1.1.19}$$

where c is the copula density and f_i is the i-th marginal density for every $i = 1, \cdots, d$.

Let

$$\mathbf{X}_1 = (X_{1,1}, \ldots, X_{1,m})^T, \ldots, \mathbf{X}_n = (X_{n,1}, \ldots, X_{n,m})^T$$

be i.i.d. copies of $\mathbf{X} = (X_1, \ldots, X_m)^T \sim H$. The key concept of the semiparametric copula model is the existence a parametric copula C_θ, with $\theta \in \Theta \subset \mathbb{R}^k$, Θ open and $k \geq 1$, such that

$$\mathbf{U} = (U_1, \ldots, U_m) = (F_1(X_1), \ldots, F_m(X_m))^T \sim C_\theta.$$

Here, θ is called *parameter of interest* and $G = (F_1(\cdot), \ldots, F_m(\cdot)$ is called *nuisance parameter*. Mathematically, the model is written as

$$\mathcal{P} = \{P_{\theta,G} \ : \ \theta \in \Theta \subset \mathbb{R}^k \ , G = (F_1(\cdot), \ldots, F_m(\cdot)) \in \mathcal{G}\}. \tag{1.1.20}$$

In practice, the marginal distribution functions are estimated by *modified empirical distribution functions*

$$\hat{F}_j(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbf{1}_{\{X_{ji} \leq x\}}, \quad \forall j = 1, \ldots, m,$$

and the parameter of interest θ is estimated by the maximum-likelihood estimator when their marginal distribution functions are replaced by their empirical versions. The resulting estimator is known as pseudo-maximum likelihood estimator (PMLE)

$$\hat{\theta}_n = \arg\min \frac{1}{n} \sum_{i=1}^{n} \log c_\theta \left(\hat{F}_1(X_{1i}), \ldots, \hat{F}_m(X_{mi}) \right). \tag{1.1.21}$$

Table 1.1: Distribution function of m-variate parametric copula. ParDim indicates the dimension of the copula parameter

Copula	**ParDim**	**Distribution function**
ind	0	$C_\theta(u_1,\dots,u_m) = \prod_{i=1}^m u_i$
GC	$m(m-1)/2$	$C_R(u_1,\dots,u_m) = \Phi_R(\Phi^{-1}(u_1),\dots,\Phi^{-1}(u_m))$
t	$m(m-1)/2+1$	$C_{\nu,R}(u_1,\dots,u_m) = t_{\nu,R}(t_\nu^{-1}(u_1),\dots,t_\nu^{-1}(u_m))$
Fr	1	$C_\theta(u_1,\dots,u_m) = -\frac{1}{\theta}\log\left(1+\frac{\prod_{i=1}^m \exp(-\theta u_i - 1)}{\exp -\theta - 1}\right)$
Cl	1	$C_\theta(u_1,\dots,u_m) = \left(\sum_{i=1}^m u_i^{-\theta} - 1\right)^{-\frac{1}{\theta}}$
fCl	1	$C_\theta(u_1,\dots,u_m) = \left(\sum_{i=1}^m (1-u_i)^{-\theta} - 1\right)^{-\frac{1}{\theta}}$
Gu	1	$C_\theta(u_1,\dots,u_m) = \exp\left[-\left(\sum_{i=1}^m (-\log u_i)^\theta\right)^{\frac{1}{\theta}}\right]$
fGU	1	$C_\theta(u_1,\dots,u_m) = \exp\left[-\left(\sum_{i=1}^m (-\log(1-u_i))^\theta\right)^{\frac{1}{\theta}}\right]$

This estimator has been shown to be asymptotically normal in [19], i.e.,

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \to \mathcal{N}(0,\Sigma) \tag{1.1.22}$$

for some positive definite covariance matrix Σ.

In this thesis, we will use the following parametric copulas: Independent copula (ind), Gaussian copula (GC), Student's t (t), Frank (Fr), Clayton (Cl), flipped Clayton (fCl), Gumbel (Gu), and flipped Gumbel (fGu). The density functions of these parametric copulas are given in Table 1.1.

1.2 Research Questions

1.2.1 Statistical Theory

Consider a quite arbitrary (semi)parametric model with a Euclidean parameter of interest and assume that an asymptotically (semi)parametrically efficient estimator of it is given. This thesis part aims at answering the following specific research questions:

- If the parameter of interest is known to lie on a general surface (image of a continuously differentiable vector valued function), what is the lower bound on the performance of estimators under this restriction and how can an efficient estimator be constructed?

- If the parameter of interest belongs to the zero set of a continuously differentiable function (for which it might be impossible to parametrize it as the image of a continuously differentiable vector valued function), what is the lower bound on the performance of estimators under this restriction and how can an efficient estimator be constructed?

1.2.2 Biometric Application

Suppose we have score-based multibiometric matchers, in which two or more different matchers compute a similarity score for any pair of two biometric samples. This thesis part aims at answering the following specific research questions:

- How can copula models handle dependence between matchers? How do we estimate the dependence parameters from training data? What are the performances of handling dependence compared to the simple independence assumption between matchers in applications?
- How can copula models be used in standard biometric verification? How can we compare copula-based biometric fusion to the simple independence assumption between matchers?
- How can copula models be used in forensic applications for combining multi-algorithm face recognition systems, which are usually dependent?

1.3 Contributions

1.3.1 Statistical Theory

The work carried out has several contributions to semiparametric estimation subject to restrictions:

- If the parameter of interest is known to lie on a general surface (image of a continuously differentiable vector valued function), we have a submodel in which the Euclidean parameter may be rewritten in terms of a lower-dimensional Euclidean parameter of interest. An estimator of this underlying parameter is constructed based on the original estimator, and it is shown to be (semi)parametrically efficient. It is proved that the efficient score function for the underlying parameter is determined by the efficient score function for the original parameter and the Jacobian of the function defining the general surface, via a chain rule for score functions. This general method is applied to linear regression and normal copula models, where it leads to natural results.

- For a given semiparametric model, quite frequently the elements of the parameter of interest are not mathematically independent but vanish on a vector-valued continuously differentiable function, thus resulting in a semiparametric model subject to equality constraints. We present an explicit method to construct (semi)parametrically efficient estimators of the Euclidean parameter in such equality constrained submodels and prove their efficiency. Our construction is based solely on the original efficient estimator and the constraining function.

1.3.2 Biometric Applications

Our work has the following contributions to the field of biometric fusion:

- We present a mathematical framework for modelling dependence between matchers 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 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.
- In standard biometric verification, we present two main contributions in score level fusion: (i) proposing a new method of measuring the performance of a fusion strategy at fixed FAR via Jeffreys credible interval analysis and (ii) subsequently providing a method to improve the fusion strategy under the independence assumption by taking the dependence into account via parametric families of copula models, which we call fixed FAR fusion. We test our method on some public databases, compare it to a Gaussian mixture model and linear logistic methods, which are also designed to handle dependence, and notice its significant improvement with respect to our evaluation method.
- We propose a new method for combining multi-algorithm score-based face recognition systems, which we call the two-step calibration method. The two-step method is based on parametric families of copula models to handle the dependence. Its goal is to minimize discrimination loss. We show that

our method is accurate and reliable on some real public databases using the cost of log likelihood ratio and the information-theoretical empirical cross-entropy (ECE).

1.4 Overview of the Thesis

The thesis contains, for the most part, published or submitted papers. Each chapter is preceded by a chapter introduction and closed by a chapter conclusion. The chapter introduction provides information where the repetitions (if any) are and how the chapter can be understood while the chapter conclusion summarizes the contents of the chapter. Each paper is inserted in a separate section and there is no modification in the contents besides small corrections such as typos.

Chapter 2 proposes a semiparametric estimation method for constrained Euclidean parameters. Two kinds of restrictions are considered and an efficient estimator for each case is provided in terms of the efficient estimator for the original parameter and the function defining the restriction. The first restriction, for which it is known that the parameter of interest lies on a general surface (image of a continuously differentiable vector valued function), is presented in Section 2.2. The second one is studied in Section 2.3, where the parameter of interest satisfies a functional equality constraint.

Chapter 3 introduces a semiparametric LR-based score level fusion strategy by splitting the marginal individual likelihood ratios and the dependence between matchers via the copula concept. A new quantity called the *Correction Factor* is defined, which incorporates the dependence between matchers to improve simple fusion under the independence assumption. While the individual likelihood ratios are computed nonparametrically using the PAV algorithm, a semiparametric model is proposed to compute the Correction Factor by proposing some well-known parametric copulas for genuine and impostor scores, and choosing the best pair by a resampling method. Finally, some experimental results on real databases are reported.

Chapter 4 implements the semiparametric LR-based fusion in forensic face scenarios that we called the *two-step method.* The best copula pair is chosen by minimizing the discrimination loss and the PAV algorithm is applied to make the fused score well calibrated. Some experiments using synthetic and real face databases are conducted to compare the performance of the two-step method to the performance of the other LR-based fusions (GMM and Logit) with respect to the cost of loglikelihood ratio and the ECE plot.

Chapter 5 concludes this thesis. It discusses how the research questions are answered by the work presented in the thesis. It also points out possibilities for future research, in particular it suggests to study some alternative methods for computing the Correction Factor.

Part I

Statistical Theory

Chapter 2

Semiparametric Reduced Parameter

2.1 Chapter Introduction

PURPOSE. This chapter presents efficient estimators of Euclidean parameters subject to restrictions, which are called *reduced parameters*. These estimators are based on estimators that are efficient within the model without restrictions. The restrictions are divided into two cases: the parameter has to be in the image of a continuously differentiable function of a lower dimensional parameter and the parameter has to belong to the zero set of a continuously differentiable function of the parameter.

CONTENTS. The main results for (semi)parametric models in which the parameter of interest is determined by a lower dimensional parameter, are given in Theorem 2.2.1 and Theorem 2.2.2. An explicit construction of an efficient estimator under this restriction is given and some examples are also given. If the parameter of interest satisfies an equality constraint, we propose another method to construct an efficient estimator in Theorem 2.3.1.

PUBLICATIONS. The manuscript presented in Section 2.2 has been published in [20] and the manuscript of Section 2.3 has been published in [21].

2.2 Semiparametrically Efficient Estimation of Constrained Euclidean Parameters

2.2.1 Abstract

Consider a quite arbitrary (semi)parametric model with a Euclidean parameter of interest and assume that an asymptotically (semi)parametrically efficient estimator of it is given. If the parameter of interest is known to lie on a general surface (image of a continuously differentiable vector valued function), we have a submodel in which this constrained Euclidean parameter may be rewritten in terms of a lower-dimensional Euclidean parameter of interest. An estimator of this underlying parameter is constructed based on the original estimator, and it is shown to be (semi)parametrically efficient. It is proved that the efficient score function for the underlying parameter is determined by the efficient score function for the original parameter and the Jacobian of the function defining the general surface, via a chain rule for score functions. Efficient estimation of the constrained Euclidean parameter itself is considered as well.

Our general estimation method is applied to location-scale, Gaussian copula and semiparametric regression models, and to parametric models under linear restrictions.

2.2.2 Introduction

Let $X_1, \dots, X_n$ be i.i.d. copies of X taking values in the measurable space $(\mathcal{X}, \mathcal{A})$ in a semiparametric model with Euclidean parameter $\theta \in \Theta$ where Θ is an open subset of $\mathbb{R}^k$. We denote this semiparametric model by

$$\mathcal{P} = \{P_{\theta,G} \ : \ \theta \in \Theta, \ G \in \mathcal{G}\} \, . \tag{2.2.1}$$

Typically, the nuisance parameter space $\mathcal{G}$ is a subset of a Banach or Hilbert space. This space may also be finite dimensional, thus resulting in a parametric model.

We assume an asymptotically efficient estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ is given of the parameter of interest θ, which under regularity conditions means that

$$\sqrt{n}\left(\hat{\theta}_n - \theta - \frac{1}{n}\sum_{i=1}^{n} \tilde{\ell}(X_i; \theta, G, \mathcal{P})\right) \to_{P_{\theta,G}} 0 \tag{2.2.2}$$

holds. Here $\tilde{\ell}(\cdot;\theta,G,\mathcal{P})$ is the efficient influence function at $P_{\theta,G}$ for estimation of θ within $\mathcal{P}$ and

$$\dot{\ell}(\cdot;\theta,G,\mathcal{P}) = \left(\int_{\mathcal{X}} \tilde{\ell}(x;\theta,G,\mathcal{P})\tilde{\ell}^T(x;\theta,G,\mathcal{P})dP_{\theta,G}(x)\right)^{-1} \tilde{\ell}(\cdot;\theta,G,\mathcal{P}) \quad (2.2.3)$$

is the corresponding efficient score function at $P_{\theta,G}$ for estimation of θ within $\mathcal{P}$.

The topic of this paper is asymptotically efficient estimation when it is known that θ lies on a general surface, or equivalently, when it is known that θ is determined by a lower dimensional parameter via a continuously differentiable function, which we denote by

$$\theta = f(\nu), \quad \nu \in N. \quad (2.2.4)$$

Here $f : N \subset \mathbb{R}^d \to \mathbb{R}^k$ with $d < k$ is known, N is open, the Jacobian

$$\dot{f}(\nu) = \left(\frac{\partial f_i(\nu)}{\partial \nu_j}\right)_{j=1,\ldots,d}^{i=1,\ldots,k} \quad (2.2.5)$$

of f is assumed to be of full rank on N, and ν is the unknown d-dimensional parameter to be estimated. Thus, we focus on the (semi)parametric model

$$\mathcal{Q} = \left\{P_{f(\nu),G} \ : \ \nu \in N, \ G \in \mathcal{G}\right\} \subset \mathcal{P}. \quad (2.2.6)$$

The first main result of this paper is that a semiparametrically efficient estimator of ν, the parameter of interest, has to be asymptotically linear with efficient score function for estimation of ν equal to

$$\dot{\ell}(\cdot;\nu,G,\mathcal{Q}) = \dot{f}^T(\nu)\dot{\ell}(\cdot;\theta,G,\mathcal{P}). \quad (2.2.7)$$

Such a semiparametrically efficient estimator of the parameter of interest can be defined in terms of $f(\cdot)$ and the efficient estimator $\hat{\theta}_n$ of θ; see equation (2.2.29) in Section 2.2.5. This is our second main result. How (2.2.7) is related to the chain rule for differentiation will be explained in Section 2.2.3, which proves this chain rule for score functions. The semiparametric lower bound for estimators of ν is obtained via the Hájek-LeCam Convolution Theorem for regular parametric models and without projection techniques in Section 2.2.4. In Section 2.2.5 efficient estimators within $\mathcal{Q}$ of ν and θ are constructed, as well as efficient estimators of θ under linear restrictions on θ. The generality of our approach facilitates the analysis of numerous statistical models. We discuss some of such parametric and semiparametric models and related literature in

Section 2.2.6. One of the proofs will be given in Appendix 1 in Subsection 2.2.7.

The topic of this paper should not be confused with estimation of the parameter θ when it is known to lie in a subset of the original parameter space described by linear inequalities. A comprehensive treatment of such estimation problems may be found in [22]. Our model $\mathcal{Q}$ with its constrained Euclidean parameters also differs from the constraint defined models as studied by Bickel et al. (1993, 1998) (henceforth called BKRW), which are defined by restrictions on the distributions in $\mathcal{P}$.

2.2.3 The Chain Rule for Score Functions

The basic building block for the asymptotic theory of semiparametric models as presented in e.g. [23] is the concept of regular parametric model. Let $\mathcal{P}_\Theta = \{P_\theta : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^k$ open be a parametric model with all P_θ dominated by a σ-finite measure μ on $(\mathcal{X}, \mathcal{A})$. Denote the density of P_θ with respect to μ by $p(\theta) = p(\cdot;\theta,\mathcal{P}_\Theta)$ and the $L_2(\mu)$-norm by $\| \cdot \|_\mu$. If for each $\theta_0 \in \Theta$ there exists a k-dimensional column vector $\dot{\ell}(\theta_0, \mathcal{P}_\Theta)$ of elements of $L_2(P_{\theta_0})$, the so-called score function, such that the Fréchet differentiability

$$\begin{aligned} \| \sqrt{p(\theta)} - \sqrt{p(\theta_0)} - \tfrac{1}{2}\,(\theta-\theta_0)^T\,\dot{\ell}(\theta_0,\mathcal{P}_\Theta)\sqrt{p(\theta_0)}\ \|_\mu& \\ = o(|\theta-\theta_0|), \quad \theta \to \theta_0,& \qquad (2.2.8) \end{aligned}$$

holds and the $k \times k$ Fisher information matrix

$$I(\theta_0) = \int_{\mathcal{X}} \dot{\ell}(\theta_0,\mathcal{P}_\Theta)\dot{\ell}^T(\theta_0,\mathcal{P}_\Theta)dP_{\theta_0} \qquad (2.2.9)$$

is nonsingular, and, moreover, the map $\theta \mapsto \dot{\ell}(\theta,\mathcal{P}_\Theta)\sqrt{p(\theta)}$ from Θ to $L_2^k(\mu)$ is continuous, then $\mathcal{P}_\Theta$ is called a *regular parametric* model. Often the score function may be determined by computing the logarithmic derivative of the density with respect to θ; cf. Proposition 2.1.1 of [23]. We will call $\mathcal{P}$ from (2.2.1) a *regular semiparametric* model if for all $G \in \mathcal{G}$

$$\mathcal{P}_{\Theta,G} = \{P_{\theta,G} : \theta \in \Theta\} \qquad (2.2.10)$$

is a regular parametric model.

Fix $\theta_0 \in \Theta$ and $G_0 \in \mathcal{G}$, and write $P_{\theta_0,G_0} = P_0$. Let $\psi : \Theta \to \mathcal{G}$ with

$\psi(\theta_0) = G_0$ be such that

$$\mathcal{P}_\psi = \left\{ P_{\theta,\psi(\theta)} \,:\, \theta \in \Theta \right\} \tag{2.2.11}$$

is a regular parametric submodel of $\mathcal{P}$ with score function $\dot{\ell}(\theta_0, \mathcal{P}_\psi)$ at θ_0 and Fisher information matrix $I(\theta_0, \mathcal{P}_\psi)$, say. Let the density of $P_{\theta,\psi(\theta)}$ with respect to μ be denoted by $q(\theta)$. Since $\mathcal{P}_\psi$ is a regular parametric model the score function $\dot{\ell}(\theta_0, \mathcal{P}_\psi)$ for θ at θ_0 within $\mathcal{P}_\psi$ satisfies (cf. (2.2.8))

$$\begin{aligned} \| \sqrt{q(\theta)} - \sqrt{q(\theta_0)} - \tfrac{1}{2} \left(\theta - \theta_0\right)^T \dot{\ell}(\theta_0, \mathcal{P}_\psi) \sqrt{q(\theta_0)} \, \|_\mu & \\ = o(|\theta - \theta_0|), \quad \theta \to \theta_0. & \end{aligned} \tag{2.2.12}$$

Considering now the (semi)parametric submodel $\mathcal{Q}$ from(2.2.10) we fix ν_0 and write $f(\nu_0) = \theta_0$ and $f(\nu) = \theta$. Within $\mathcal{Q}$ the Fréchet differentiability (2.2.12) yields

$$\begin{aligned} \| \sqrt{q(f(\nu))} - \sqrt{q(f(\nu_0))} - \tfrac{1}{2} \left(f(\nu) - f(\nu_0)\right)^T \dot{\ell}(f(\nu_0), \mathcal{P}_\psi) \sqrt{q(f(\nu_0))} \, \|_\mu & \\ = o(|f(\nu) - f(\nu_0)|), \quad f(\nu) \to f(\nu_0), & \end{aligned} \tag{2.2.13}$$

and hence

$$\begin{aligned} \| \sqrt{q(f(\nu))} - \sqrt{q(f(\nu_0))} - \tfrac{1}{2}(\nu - \nu_0)^T \dot{f}^T(\nu_0) \dot{\ell}(\theta_0, \mathcal{P}_\psi) \sqrt{q(f(\nu_0))} \, \|_\mu & \\ = o(|\nu - \nu_0|), \quad \nu \to \nu_0, & \end{aligned} \tag{2.2.14}$$

in view of the differentiability of $f(\cdot)$. Since $\dot{f}(\cdot)$ is continuous, this means that

$$\mathcal{Q}_\psi = \left\{ P_{f(\nu),\psi(f(\nu))} \,:\, \nu \in N \right\} \tag{2.2.15}$$

is a regular parametric submodel of $\mathcal{Q}$ with score function

$$\dot{\ell}(\nu_0, \mathcal{Q}_\psi) = \dot{f}^T(\nu_0) \dot{\ell}(\theta_0, \mathcal{P}_\psi) \tag{2.2.16}$$

for ν at P_0 and Fisher information matrix

$$\dot{f}^T(\nu_0) I(\theta_0, \mathcal{P}_\psi) \dot{f}(\nu_0) = \dot{f}^T(\nu_0) \int_{\mathcal{X}} \dot{\ell}(\theta_0, \mathcal{P}_\psi) \dot{\ell}^T(\theta_0, \mathcal{P}_\psi) dP_0 \; \dot{f}(\nu_0). \tag{2.2.17}$$

We have proved

Proposition 2.2.1. *Let $\mathcal{P}$ as in (2.2.1) be a regular semiparametric model and let $\mathcal{Q}$ as in (2.2.10) be a regular semiparametric submodel with $f(\cdot)$ and $\dot{f}(\cdot)$ defined as in and below (2.2.4) and (2.2.5). If there exists a regular parametric submodel $\mathcal{P}_\psi$ of $\mathcal{P}$ with score function $\dot{\ell}(\theta_0, \mathcal{P}_\psi)$ for θ at $\theta_0 = f(\nu_0)$, then there* *exists a regular parametric submodel $\mathcal{Q}_\psi$ of $\mathcal{Q}$ with score function $\dot{\ell}(\nu_0, \mathcal{Q}_\psi)$ for ν at ν_0 satisfying (2.2.16).*

This Proposition is also valid for parametric models, as may be seen by choosing $\mathcal{G}$ finite dimensional or even degenerate. The basic version of the chain rule for score functions is for such a parametric model $\mathcal{P}_\Theta$. We have chosen the more elaborate formulation of Proposition 2.2.1 since we are going to apply the chain rule for such parametric submodels $\mathcal{P}_\psi$ of semiparametric models $\mathcal{P}$.

2.2.4 Convolution Theorem and Main Result

An estimator $\hat{\theta}_n$ of θ within the regular semiparametric model $\mathcal{P}$ is called (locally) regular at $P_0 = P_{\theta_0, G_0}$ if it is (locally) regular at P_0 within $\mathcal{P}_\psi$ for all regular parametric submodels $\mathcal{P}_\psi$ of $\mathcal{P}$ containing P_{Θ, G_0}. According to the Hájek-LeCam Convolution Theorem for regular parametric models (see e.g. Section 2.3 of [23]) this implies that such a regular estimator $\hat{\theta}_n$ of θ within $\mathcal{P}$ has a limit distribution under P_0 that is the convolution of a normal distribution with mean 0 and covariance matrix $I^{-1}(\theta_0, \mathcal{P}_\psi)$ and another distribution, for any regular parametric submodel $\mathcal{P}_\psi$ containing P_0. If there exists $\psi = \psi_0$ such that this last distribution is degenerate at 0, we call $\hat{\theta}_n$ (locally) efficient at P_0 and $\mathcal{P}_{\psi_0}$ a least favorable parametric submodel for estimation of θ within $\mathcal{P}$ at P_0. Then the Hájek-LeCam Convolution Theorem also implies that $\hat{\theta}_n$ is asymptotically linear in the efficient influence function $\tilde{\ell}(\theta_0, G_0, \mathcal{P}) = \tilde{\ell}(\cdot; \theta_0, G_0, \mathcal{P})$ satisfying

$$\tilde{\ell}(\theta_0, G_0, \mathcal{P}) = \tilde{\ell}(\theta_0, \mathcal{P}_{\psi_0}) = I^{-1}(\theta_0, \mathcal{P}_{\psi_0}) \dot{\ell}(\theta_0, \mathcal{P}_{\psi_0}), \tag{2.2.18}$$

which means

$$\sqrt{n} \left(\hat{\theta}_n - \theta_0 - \frac{1}{n} \sum_{i=1}^{n} \tilde{\ell}(X_i; \theta_0, G_0, \mathcal{P}) \right) \to_{P_0} 0. \tag{2.2.19}$$

The argument above can be extended to the more general situation that there exists a least favorable sequence of parametric submodels indexed by $\psi_j\,, j = 1, 2, \ldots$, such that the corresponding score functions $\dot{\ell}(\theta_0, \mathcal{P}_{\psi_j})$ for θ at θ_0 within model $\mathcal{P}_{\psi_j}$ converge in $L_2^k(P_0)$ to $\dot{\ell}(\theta_0, G_0, \mathcal{P}) = \dot{\ell}(\cdot; \theta_0, G_0, \mathcal{P})$, say. A regular estimator $\hat{\theta}_n$ of θ within $\mathcal{P}$ is called efficient then, if it is asymptotically linear as in (2.2.19) with efficient influence function $\tilde{\ell}(\theta_0, G_0, \mathcal{P}) =$

$\tilde{\ell}(\cdot;\theta_0, G_0, \mathcal{P})$ satisfying

$$\begin{aligned}\tilde{\ell}(\theta_0, G_0, \mathcal{P}) &= \left(\int_{\mathcal{X}} \dot{\ell}(\theta_0, G_0, \mathcal{P})\dot{\ell}^T(\theta_0, G_0, \mathcal{P})dP_0\right)^{-1} \dot{\ell}(\theta_0, G_0, \mathcal{P}) \\ &= I^{-1}(\theta_0, G_0, \mathcal{P})\dot{\ell}(\theta_0, G_0, \mathcal{P}). \quad (2.2.20)\end{aligned}$$

Indeed, by the Convolution Theorem for regular parametric models the convergence

$$\begin{pmatrix} \sqrt{n}\left(\hat{\theta}_n - \theta_0 - \frac{1}{n}\sum\limits_{i=1}^n \tilde{\ell}(X_i;\theta_0, \mathcal{P}_{\psi_j})\right) \\ \frac{1}{\sqrt{n}}\sum\limits_{i=1}^n \tilde{\ell}(X_i;\theta_0, \mathcal{P}_{\psi_j}) \end{pmatrix} \to_{P_0} \begin{pmatrix} R_j \\ Z_j \end{pmatrix} \qquad (2.2.21)$$

holds with the k-vectors R_j and Z_j independent and Z_j normal with mean 0 and covariance matrix $I^{-1}(\theta_0, \mathcal{P}_{\psi_j})$. Taking limits as $j \to \infty$ we see by tightness arguments and by the convergence of $\dot{\ell}(\theta_0, \mathcal{P}_{\psi_j})$ to $\dot{\ell}(\theta_0, G_0, \mathcal{P})$ in $L_2^k(P_0)$, that also

$$\begin{pmatrix} \sqrt{n}\left(\hat{\theta}_n - \theta_0 - \frac{1}{n}\sum\limits_{i=1}^n \tilde{\ell}(X_i;\theta_0, G_0, \mathcal{P})\right) \\ \frac{1}{\sqrt{n}}\sum\limits_{i=1}^n \tilde{\ell}(X_i;\theta_0, G_0, \mathcal{P}) \end{pmatrix} \to_{P_0} \begin{pmatrix} R_{\mathcal{P}} \\ Z_{\mathcal{P}} \end{pmatrix} \qquad (2.2.22)$$

holds with $R_{\mathcal{P}}$ and $Z_{\mathcal{P}}$ independent. If $R_{\mathcal{P}}$ is degenerate at 0, then $\hat{\theta}_n$ is locally asymptotically efficient at P_0 within $\mathcal{P}$ and the sequence of regular parametric submodels $\mathcal{P}_{\psi_j}$ is least favorable indeed.

Now, let us assume such a least favorable sequence and efficient estimator $\hat{\theta}_n$ exist at $P_0 = P_{\theta_0, G_0}$ with $\theta_0 = f(\nu_0)$ and $f(\cdot)$ from (2.2.4) and (2.2.5) continuously differentiable. By the chain rule for score functions from Proposition 2.2.1 the score function $\dot{\ell}(\nu_0, \mathcal{Q}_{\psi_j})$ for ν at ν_0 within $\mathcal{Q}_{\psi_j}$ satisfies

$$\dot{\ell}(\nu_0, \mathcal{Q}_{\psi_j}) = \dot{f}^T(\nu_0)\dot{\ell}(\theta_0, \mathcal{P}_{\psi_j}) \qquad (2.2.23)$$

and hence the corresponding influence function $\tilde{\ell}(\nu_0, \mathcal{Q}_{\psi_j})$ satisfies

$$\tilde{\ell}(\nu_0, \mathcal{Q}_{\psi_j}) = \left(\dot{f}^T(\nu_0)I(\theta_0, \mathcal{P}_{\psi_j})\dot{f}(\nu_0)\right)^{-1} \dot{f}^T(\nu_0)\dot{\ell}(\theta_0, \mathcal{P}_{\psi_j}). \qquad (2.2.24)$$

Let $\hat{\nu}_n$ be a locally regular estimator of ν at P_0 within the regular semiparametric model $\mathcal{Q}$. By the convergence of $\dot{\ell}(\theta_0, \mathcal{P}_{\psi_j})$ to $\dot{\ell}(\theta_0, G_0, \mathcal{P})$ in $L_2^k(P_0)$, the influence functions from (2.2.24) converge in $L_2^d(P_0)$ to

$$\tilde{\ell}(\nu_0, G_0, \mathcal{Q}) = \left(\dot{f}^T(\nu_0) I(\theta_0, G_0, \mathcal{P}) \dot{f}(\nu_0)\right)^{-1} \dot{f}^T(\nu_0) \dot{\ell}(\theta_0, G_0, \mathcal{P}) \tag{2.2.25}$$

and the argument leading to (2.2.22) yields the convergence

$$\begin{pmatrix} \sqrt{n}\left(\hat{\nu}_n - \nu_0 - \frac{1}{n}\sum_{i=1}^{n} \tilde{\ell}(X_i; \nu_0, G_0, \mathcal{Q})\right) \\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tilde{\ell}(X_i; \nu_0, G_0, \mathcal{Q}) \end{pmatrix} \to_{P_0} \begin{pmatrix} R_{\mathcal{Q}} \\ Z_{\mathcal{Q}} \end{pmatrix} \tag{2.2.26}$$

with $R_{\mathcal{Q}}$ and $Z_{\mathcal{Q}}$ independent. Note that $Z_{\mathcal{Q}}$ has a normal distribution with mean 0 and covariance matrix

$$I^{-1}(\nu_0, G_0, \mathcal{Q}) = \left(\dot{f}^T(\nu_0) I(\theta_0, G_0, \mathcal{P}) \dot{f}(\nu_0)\right)^{-1}. \tag{2.2.27}$$

Under an additional condition on $f(\cdot)$ we shall construct an estimator $\hat{\nu}_n$ of ν based on $\hat{\theta}_n$ for which $R_{\mathcal{Q}}$ is degenerate. This construction of $\hat{\nu}_n$ will be given in the next section together with a proof of its efficiency, and this will complete the proof of our main result formulated as follows.

Theorem 2.2.1. *Let $\mathcal{P}$ from (2.2.1) be a regular semiparametric model with $P_0 = P_{\theta_0, G_0} \in \mathcal{P}, \theta_0 = f(\nu_0)$, and $f(\cdot)$ from (2.2.4) and (2.2.5) continuously differentiable. Furthermore, let $f(\cdot)$ have an inverse on $f(N)$ that is differentiable with a bounded Jacobian. If there exists a least favorable sequence of regular parametric submodels $\mathcal{P}_{\psi_j}$ and an asymptotically efficient estimator $\hat{\theta}_n$ of θ satisfying (2.2.22) with $R_{\mathcal{P}} = 0$ a.s., then there exists a least favorable sequence of regular parametric submodels $\mathcal{Q}_{\psi_j}$ of the restricted model $\mathcal{Q}$ from (2.2.10) and an asymptotically efficient estimator $\hat{\nu}_n$ of ν satisfying (2.2.26) with $R_{\mathcal{Q}} = 0$ a.s. and attaining the asymptotic information bound (2.2.27).*

Note that the convolution result (2.2.26) and (2.2.25) also holds if the convergent sequence of regular parametric submodels $\mathcal{P}_{\psi_j}$ is not least favorable, and that it implies by the central limit theorem that the limit distribution of $\sqrt{n}\,(\hat{\nu}_n - \nu_0)$ is the convolution of a normal distribution with mean 0 and covariance matrix

$$I^{-1}(\nu_0, G_0, \mathcal{Q}) = \left(\dot{f}^T(\nu_0) I(\theta_0, G_0, \mathcal{P}) \dot{f}(\nu_0)\right)^{-1} \tag{2.2.28}$$

and the distribution of $R_{\mathcal{Q}}$.

2.2.5 Efficient Estimator of the Parameter of Interest

There are many ways of constructing efficient estimators in (semi)parametric models. One of the common approaches is upgrading a $\sqrt{n}$-consistent estimator as in Sections 2.5 and 7.8 of [23]. A somewhat different upgrading approach is used in the following construction.

Theorem 2.2.2. *Consider the situation of Theorem 2.2.1. If the symmetric positive definite $k \times k$-matrix $\hat{I}_n$ is a consistent estimator of $I(\theta, G, \mathcal{P})$ within $\mathcal{P}$ and $\bar{\nu}_n$ is a $\sqrt{n}$-consistent estimator of ν within $\mathcal{Q}$, then*

$$\hat{\nu}_n = \bar{\nu}_n + \left(\dot{f}^T(\bar{\nu}_n)\hat{I}_n\dot{f}(\bar{\nu}_n)\right)^{-1} \dot{f}^T(\bar{\nu}_n)\hat{I}_n \left[\hat{\theta}_n - f\left(\bar{\nu}_n\right)\right] \tag{2.2.29}$$

is efficient, i.e., it satisfies (2.2.26) with $R_{\mathcal{Q}} = 0$ a.s.

Proof The continuity of $\dot{f}(\cdot)$ and the consistency of $\bar{\nu}_n$ and $\hat{I}_n$ imply that

$$\hat{K}_n = \left(\dot{f}^T(\bar{\nu}_n)\hat{I}_n\dot{f}(\bar{\nu}_n)\right)^{-1} \dot{f}^T(\bar{\nu}_n)\hat{I}_n \tag{2.2.30}$$

converges in probability under P_0 to

$$K_0 = \left(\dot{f}^T(\nu_0)I(\theta_0, G_0, \mathcal{P})\dot{f}(\nu_0)\right)^{-1} \dot{f}^T(\nu_0)I(\theta_0, G_0, \mathcal{P}). \tag{2.2.31}$$

This means that $\hat{K}_n$ consistently estimates K_0. In view of (2.2.29), (2.2.25), (2.2.20), and (2.2.22) with $R_{\mathcal{P}} = 0$ we obtain

$$\begin{aligned}
&\sqrt{n}\left(\hat{\nu}_n - \nu_0 - \frac{1}{n}\sum_{i=1}^n \tilde{\ell}(X_i; \nu_0, G_0, \mathcal{Q})\right) \\
&\quad = \sqrt{n}\left(\bar{\nu}_n - \nu_0 + \hat{K}_n\left[\hat{\theta}_n - f\left(\bar{\nu}_n\right)\right] - \frac{1}{n}\sum_{i=1}^n K_0\tilde{\ell}(X_i; \theta_0, G_0, \mathcal{P})\right) \\
&\quad = \sqrt{n}\left(\bar{\nu}_n - \nu_0 - \hat{K}_n\left[f\left(\bar{\nu}_n\right) - f(\nu_0)\right]\right) \\
&\qquad\qquad + \left[\hat{K}_n - K_0\right]\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\ell}(X_i; \theta_0, G_0, \mathcal{P}) + o_p(1).
\end{aligned} \tag{2.2.32}$$

By the consistency of $\hat{K}_n$ the second term at the right hand side of (2.2.32) converges to 0 in probability under P_0 in view of the central limit theorem. Because $f\left(\bar{\nu}_n\right) = f(\nu_0) + \dot{f}(\nu_0)\left(\bar{\nu}_n - \nu_0\right) + o_p\left(\bar{\nu}_n - \nu_0\right)$ holds and $K_0\dot{f}(\nu_0)$ equals the $d \times d$ identity matrix, the first part of the right hand side of (2.2.32) also converges to 0 in probability under P_0. □

To complete the proof of Theorem 2.2.1 with the help of Theorem 2.2.2 we will construct a $\sqrt{n}$-consistent estimator $\bar{\nu}_n$ of ν and subsequently a consistent estimator $\hat{I}_n$ of $I(\theta, G, \mathcal{P})$. Let $\| \cdot \|$ be a Euclidean norm on $\mathbb{R}^k$. We choose $\bar{\nu}_n$ in such a way that

$$\| f\left(\bar{\nu}_n\right) - \hat{\theta}_n \| \leq \inf_{\nu \in N} \| f(\nu) - \hat{\theta}_n \| + \frac{1}{n} \tag{2.2.33}$$

holds. Of course, if the infimum is attained, we choose $\bar{\nu}_n$ as the minimizer. By the triangle inequality and the $\sqrt{n}$-consistency of $\hat{\theta}_n$ we obtain

$$\begin{aligned} \| f(\bar{\nu}_n) - f(\nu_0) \| &\leq \inf_{\nu \in N} \| f(\nu) - \hat{\theta}_n \| + \frac{1}{n} + \| f(\nu_0) - \hat{\theta}_n \| \\ &\leq 2 \| \hat{\theta}_n - f(\nu_0) \| + \frac{1}{n} = O_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned} \tag{2.2.34}$$

The assumption from Theorem 2.2.1 that $f(\cdot)$ has an inverse on $f(N)$ that is differentiable with a bounded Jacobian, suffices to conclude that (2.2.34) guarantees $\sqrt{n}$-consistency of $\bar{\nu}_n$.

In constructing a consistent estimator of the Fisher information matrix based on the given efficient estimator $\hat{\theta}_n$, we split the sample in blocks as follows. Let $(k_n), (\ell_n)$, and (m_n) be sequences of integers such that $k_n = \ell_n m_n, k_n/n \to \kappa, 0 < \kappa < 1$, and $\ell_n \to \infty, m_n \to \infty$ hold as $n \to \infty$. For $j = 1, \ldots, \ell_n$ let $\hat{\theta}_{n,j}$ be the efficient estimator of θ based on the observations $X_{(j-1)m_n+1}, \ldots, X_{jm_n}$ and $\hat{\theta}_{n,0}$ be the efficient estimator of θ based on the remaining observations $X_{k_n+1}, \ldots, X_n$. Consider the "empirical" characteristic function

$$\hat{\phi}_n(t) = \frac{1}{\ell_n} \sum_{j=1}^{\ell_n} \exp\left\{it\sqrt{m_n}\left(\hat{\theta}_{n,j} - \hat{\theta}_{n,0}\right)\right\},\ t \in \mathbb{R}^k, \tag{2.2.35}$$

which we rewrite as

$$\begin{aligned} \hat{\phi}_n(t) &= \exp\left\{-it\sqrt{m_n}\left(\hat{\theta}_{n,0} - \theta_0\right)\right\} \frac{1}{\ell_n} \sum_{j=1}^{\ell_n} \exp\left\{it\sqrt{m_n}\left(\hat{\theta}_{n,j} - \theta_0\right)\right\} \\ &= \exp\left\{-it\sqrt{m_n}\left(\hat{\theta}_{n,0} - \theta_0\right)\right\} \tilde{\phi}_n(t). \end{aligned} \tag{2.2.36}$$

In view of $m_n/(n - k_n) \to 0$ and (2.2.22) with $R_{\mathcal{P}} = 0$ a.s. we see that the first factor at the right hand side of (2.2.36) converges to 1 as $n \to \infty$. The efficiency of $\hat{\theta}_n$ in (2.2.22) with $R_{\mathcal{P}} = 0$ a.s. also implies

$$\begin{aligned} E\left(\tilde{\phi}_n(t)\right) &= E\left(\exp\left\{it\sqrt{m_n}\left(\hat{\theta}_{n,1} - \theta_0\right)\right\}\right) \\ &\to E\left(\exp\left\{itZ_{\mathcal{P}}\right\}\right) \end{aligned} \tag{2.2.37}$$

as $n \to \infty$, with $Z_{\mathcal{P}}$ normally distributed with mean 0 and covariance matrix $I^{-1}(\theta_0, G_0, \mathcal{P})$. Some computation shows

$$\begin{aligned} &E\left(\left|\tilde{\phi}_n(t) - E\left(\tilde{\phi}_n(t)\right)\right|^2\right) \\ &\quad = \frac{1}{\ell_n}\left(1 - \left|E\left(\exp\left\{it\sqrt{m_n}\left(\hat{\theta}_{n,1} - \theta_0\right)\right\}\right)\right|^2\right) \le \frac{1}{\ell_n}. \end{aligned} \tag{2.2.38}$$

It follows by Chebyshev's inequality that $\tilde{\phi}_n(t)$ and hence $\hat{\phi}_n(t)$ converges under $P_0 = P_{\theta_0,G_0}$ to the characteristic function of $Z_{\mathcal{P}}$ at t,

$$\hat{\phi}_n(t) \to_{P_0} E\left(\exp\left\{itZ_{\mathcal{P}}\right\}\right) = \exp\left\{-\tfrac{1}{2}t^T I^{-1}(\theta_0, G_0, \mathcal{P})t\right\}. \tag{2.2.39}$$

For every $t \in \mathbb{R}^k$ we obtain

$$-2\log\left(\Re\left(\hat{\phi}_n(t)\right)\right) \to_{P_0} t^T I^{-1}(\theta_0, G_0, \mathcal{P})t. \tag{2.2.40}$$

Choosing $k(k+1)/2$ appropriate values of t we may obtain from (2.2.40) an estimator of $I^{-1}(\theta_0, G_0, \mathcal{P})$ and hence of $I(\theta_0, G_0, \mathcal{P})$. Indeed, with t equal to the unit vectors u_i we obtain estimators of the diagonal elements of $I^{-1}(\theta_0, G_0, \mathcal{P})$ and an estimator of its (i,j) element is obtained via

$$\log\left(\Re\left(\hat{\phi}_n(u_i)\right)\right) + \log\left(\Re\left(\hat{\phi}_n(u_j)\right)\right) - \log\left(\Re\left(\hat{\phi}_n(u_i + u_j)\right)\right).$$

When needed, the resulting estimator of $I(\theta_0, G_0, \mathcal{P})$ can be made positive definite by changing appropriate components of it by an asymptotically negligible amount, while the symmetry is maintained.

Under a mild uniform integrability condition it has been shown by [24], that existence of an efficient estimator $\hat{\theta}_n$ of θ in $\mathcal{P}$ implies the existence of a consistent and $\sqrt{n}$-unbiased estimator of the efficient influence function $\tilde{\ell}(\cdot; \theta, G, \mathcal{P})$. Basing this estimator on one half of the sample and taking the average of this estimated efficient influence function at the observations from the other half of the sample, we could have constructed another estimator of the efficient Fisher information. However, this estimator would have been more involved, and, moreover, it needs this extra uniformity condition.

With the help of Theorem 2.2.2, the estimator $\bar{\nu}_n$ of ν from (2.2.33), and the construction via (2.2.40) of an estimator $\hat{I}_n$ of the efficient Fisher information we have completed our construction of an efficient estimator $\hat{\nu}_n$ as in (2.2.29) of ν. This estimator can be turned into an efficient estimator of $\theta = f(\nu)$ within the model $\mathcal{Q}$ from (2.2.10) by

$$\tilde{\theta}_n = f(\hat{\nu}_n) \tag{2.2.41}$$

with efficient influence function

$$\begin{aligned}\tilde{\ell}(\theta_0, G_0, \mathcal{Q}) &= \dot{f}(\nu_0)\tilde{\ell}(\nu_0, G_0, \mathcal{Q}) \\ &= \dot{f}(\nu_0)\left(\dot{f}^T(\nu_0)I(\theta_0, G_0, \mathcal{P})\dot{f}(\nu_0)\right)^{-1}\dot{f}^T(\nu_0)\dot{\ell}(\theta_0, G_0, \mathcal{P}) \end{aligned} \tag{2.2.42}$$

and asymptotic information bound

$$I^{-1}(\theta_0, G_0, \mathcal{Q}) = \dot{f}(\nu_0)\left(\dot{f}^T(\nu_0)I(\theta_0, G_0, \mathcal{P})\dot{f}(\nu_0)\right)^{-1}\dot{f}^T(\nu_0). \tag{2.2.43}$$

Indeed, according to [23] Section 2.3, $\tilde{\theta}_n$ is efficient for estimation of θ under the additional information $\theta = f(\nu)$.

Remark 2.2.1. If $f(\cdot)$ is a linear function, i.e., $\theta = L\nu + \alpha$ holds with the $k \times d$-matrix L of maximum rank d, then

$$\bar{\nu}_n = (L^T L)^{-1} L^T (\hat{\theta}_n - \alpha) \tag{2.2.44}$$

attains the infimum at the right hand side of (2.2.33). So, the estimator (2.2.29) becomes

$$\hat{\nu}_n = \left(L^T \hat{I}_n L\right)^{-1} L^T \hat{I}_n \left[\hat{\theta}_n - \alpha\right] \tag{2.2.45}$$

with efficient influence function (2.2.25) and asymptotic information bound (2.2.27) with $\dot{f}(\nu_0) = L$, and the estimator from (2.2.41)

$$\tilde{\theta}_n = L\left(L^T \hat{I}_n L\right)^{-1} L^T \hat{I}_n \left[\hat{\theta}_n - \alpha\right] + \alpha. \tag{2.2.46}$$

Note that $\tilde{\theta}_n$ is the projection of $\hat{\theta}_n$ on the flat $\{\theta \in \mathbb{R}^k \,:\, \theta = L\nu + \alpha, \nu \in \mathbb{R}^d\}$ under the inner product determined by $\hat{I}_n$ (cf. Appendix 1 in Subsection 2.2.7) and that the covariance matrix of its limit distribution equals the asymptotic information bound

$$I^{-1}(\theta_0, G_0, \mathcal{Q}) = L\left(L^T I(\theta_0, G_0, \mathcal{P}) L\right)^{-1} L^T. \tag{2.2.47}$$

Another way to describe this submodel $\mathcal{Q}$ with $\theta = L\nu + \alpha$ is by linear restrictions

$$\mathcal{Q} = \{P_{L\nu+\alpha} \, : \, \nu \in N, G \in \mathcal{G}\} = \left\{P_{\theta,G} \, : \, R^T\theta = \beta, \theta \in \Theta, G \in \mathcal{G}\right\}, \quad (2.2.48)$$

where $R^T\alpha = \beta$ holds and the $k \times d$-matrix L and the $k \times (k-d)$-matrix R are matching such that the columns of L are orthogonal to those of R and the $k \times k$-matrix $(L\, R)$ is of rank k. Note that the open subset N of $\mathbb{R}^d$ determines the open subset Θ of $\mathbb{R}^k$ and vice versa. See [25], [26], [27], and [28] for some examples of estimation under linear restrictions.

In terms of the restrictions described by R and β the efficient estimator $\tilde{\theta}_n$ of θ from(2.2.46) within the submodel $\mathcal{Q}$ can be rewritten as

$$\tilde{\theta}_n = \hat{\theta}_n - \hat{I}_n^{-1} R \left(R^T \hat{I}_n^{-1} R\right)^{-1} \left(R^T \hat{\theta}_n - \beta\right), \quad (2.2.49)$$

with asymptotic information bound

$$L(L^T I L)^{-1} L^T = I^{-1} - I^{-1} R (R^T I^{-1} R)^{-1} R^T I^{-1}, \; I = I(\theta_0, G_0, \mathcal{P}), \quad (2.2.50)$$

as will be proved in Appendix 1 in Subsection 2.2.7.

2.2.6 Examples

In this section we present five examples, which illustrate our construction of (semi)parametrically efficient estimators. We shall discuss location-scale, Gaussian copula, and semiparametric regression models, and parametric models under linear restrictions.

Example 2.2.1. Coefficient of variation known

Let $g(\cdot)$ be an absolutely continuous density on $(\mathbb{R}, \mathcal{B})$ with mean 0, variance 1, and derivative $g'(\cdot)$, such that $\int [1+x^2](g'/g(x))^2 g(x)dx$ is finite. Consider the location-scale family corresponding to $g(\cdot)$. Let there be given efficient estimators $\bar{\mu}_n$ and $\bar{\sigma}_n$ of μ and σ, respectively, based on $X_1, \ldots, X_n$, which are i.i.d. with density $\sigma^{-1} g((\cdot - \mu)/\sigma)$. By I_{ij} we denote the element in the $i$the row and jth column of the matrix $I = \sigma^2 I(\theta, G, \mathcal{P})$, where the Fisher information matrix $I(\theta, G, \mathcal{P})$ is as defined in (2.2.20) with $\theta = (\mu, \sigma)^T$ and $\mathcal{G} = \{g(\cdot)\}$. Some computation shows $I_{11} = \int (g'/g)^2 g, I_{12} = I_{21} = \int x(g'/g(x))^2 g(x)dx$, and $I_{22} = \int [xg'/g(x) + 1]^2 g(x)dx$ exist and are finite; cf. Section I.2.3 of [29].

We consider the submodel with the coefficient of variation known to be equal to a given constant $c = \sigma/\mu$ and with $\nu = \mu$ the parameter of interest. Since in a parametric model the model itself is always least favorable, the conditions of Theorem 2.2.2 are satisfied and the estimator $\hat{\nu}_n = \hat{\mu}_n$ of μ from (2.2.29) with $\bar{\nu}_n = \bar{\mu}_n$, $\hat{\theta}_n = (\bar{\mu}_n, \bar{\sigma}_n)^T$, and $\hat{I}_n = \bar{\sigma}_n^{-2} I$ is efficient and some computation shows

$$\hat{\mu}_n = \left(I_{11} + 2cI_{12} + c^2 I_{22}\right)^{-1} \left[\left(I_{11} + cI_{12}\right)\bar{\mu}_n + \left(I_{12} + cI_{22}\right)\bar{\sigma}_n\right]. \qquad (2.2.51)$$

In case the density $g(\cdot)$ is symmetric around 0, the Fisher information matrix is diagonal and $\hat{\mu}_n$ from (2.2.51) becomes

$$\hat{\mu}_n = \left(I_{11} + c^2 I_{22}\right)^{-1} \left[I_{11}\bar{\mu}_n + cI_{22}\bar{\sigma}_n\right]. \qquad (2.2.52)$$

In the normal case with $g(\cdot)$ the standard normal density $\hat{\mu}_n$ reduces to

$$\hat{\mu}_n = (1 + c^2)^{-1} \left[\bar{\mu}_n + 2c\bar{\sigma}_n\right] \qquad (2.2.53)$$

with $\bar{\mu}_n$ and $\bar{\sigma}_n$ equal to e.g. the sample mean and the sample standard deviation, respectively; cf. [30], [31], and [32].

Example 2.2.2. Gaussian copula models

Let

$$\mathbf{X}_1 = (X_{1,1}, \ldots, X_{1,m})^T, \ldots, \mathbf{X}_n = (X_{n,1}, \ldots, X_{n,m})^T$$

be i.i.d. copies of $\mathbf{X} = (X_1, \ldots, X_m)^T$. For $i = 1, \ldots, m$, the marginal distribution function of X_i is continuous and will be denoted by F_i. It is assumed that $(\Phi^{-1}(F_1(X_1)), \ldots, \Phi^{-1}(F_m(X_m)))^T$ has an m-dimensional normal distribution with mean 0 and positive definite correlation matrix $C(\theta)$, where Φ denotes the one-dimensional standard normal distribution function. Here the parameter of interest θ is the vector in $\mathbb{R}^{m(m-1)/2}$ that summarizes all correlation coefficients ρ_{rs}, $1 \leq r < s \leq m$. We will set this general Gaussian copula model as our semiparametric starting model $\mathcal{P}$, i.e.,

$$\mathcal{P} = \{P_{\theta,G} \;:\; \theta = \left(\rho_{12}, \ldots, \rho_{(m-1)m}\right)^T, G = (F_1(\cdot), \ldots, F_m(\cdot)) \in \mathcal{G}\}. \qquad (2.2.54)$$

The unknown continuous marginal distributions are the nuisance parameters collected as $G \in \mathcal{G}$.

Theorem 3.1 of [33] shows that the normal scores rank correlation coefficient is semiparametrically efficient in $\mathcal{P}$ for the 2-dimensional case with normal marginals with unknown variances constituting a least favorable parametric submodel. As [34] explain at the end of their Section 1 and in their Section 4, their Theorem 4.1 proves that normal marginals with unknown, possibly unequal variances constitute a least favorable parametric submodel, also for the general m-dimensional case. Since the maximum likelihood estimators are efficient for the parameters of a multivariate normal distribution, the sample correlation coefficients are efficient for estimation of the correlation coefficients based on multivariate normal observations. But each sample correlation coefficient and hence its efficient influence function involve only two components of the multivariate normal observations. Apparently, the other components of the multivariate normal observations carry no information about the value of the respective correlation coefficient. Effectively, for each correlation coefficient we are in the 2-dimensional case and invoking again Theorem 3.1 of [33] we see that also in the general m-dimensional case the normal scores rank correlation coefficients are semiparametrically efficient. They are defined as

$$\hat{\rho}_{rs}^{(n)} = \frac{\frac{1}{n}\sum_{j=1}^{n} \Phi^{-1}\left(\frac{n}{n+1}\mathbb{F}_r^{(n)}(X_{j,r})\right)\Phi^{-1}\left(\frac{n}{n+1}\mathbb{F}_s^{(n)}(X_{j,s})\right)}{\frac{1}{n}\sum_{j=1}^{n}\left[\Phi^{-1}\left(\frac{j}{n+1}\right)\right]^2} \tag{2.2.55}$$

with $\mathbb{F}_r^{(n)}$ and $\mathbb{F}_s^{(n)}$ being the marginal empirical distributions of F_r and F_s, respectively, $1 \leq r < s \leq m$. The Van der Waerden or normal scores rank correlation coefficient $\hat{\rho}_{rs}^{(n)}$ from (3.3.9) is a semiparametrically efficient estimator of ρ_{rs} with efficient influence function

$$\begin{aligned}\tilde{\ell}_{\rho_{rs}}(X_r, X_s) = {} & \Phi^{-1}\left(F_r(X_r)\right)\Phi^{-1}\left(F_s(X_s)\right) \\ & - \tfrac{1}{2}\rho_{rs}\left\{\left[\Phi^{-1}\left(F_r(X_r)\right)\right]^2 + \left[\Phi^{-1}\left(F_s(X_s)\right)\right]^2\right\}.\end{aligned} \tag{2.2.56}$$

This means that

$$\hat{\theta}_n = (\hat{\rho}_{12}^{(n)}, \dots, \hat{\rho}_{(m-1)m}^{(n)})^T \tag{2.2.57}$$

efficiently estimates θ with efficient influence function

$$\tilde{\ell}(\mathbf{X}; \theta, G, \mathcal{P}) = (\tilde{\ell}_{\rho_{12}}(X_1, X_2), \dots, \tilde{\ell}_{\rho_{(m-1)m}}(X_{m-1}, X_m))^T. \tag{2.2.58}$$

Subexample 2.2.2.1. Exchangeable Gaussian copula

The exchangeable m-variate Gaussian copula model

$$\mathcal{Q} = \{P_{\mathbf{1}_k\rho, G} \,:\, \rho \in (-1/(m-1), 1),\ G \in \mathcal{G}\} \subset \mathcal{P} \tag{2.2.59}$$

is a submodel of the Gaussian copula model $\mathcal{P}$ with a one-dimensional parameter of interest $\nu = \rho$. In this submodel all correlation coefficients have the same value ρ. So, $\theta = \mathbf{1}_k\rho$ with $\mathbf{1}_k$ indicating the vector of ones of dimension $k = m(m-1)/2$. In order to construct an efficient estimator of ρ within $\mathcal{Q}$ along the lines of Section 2.2.5, in particular Remark 2.2.1, we first apply (2.2.44) with $\alpha = 0$ and $L = \mathbf{1}_k$ to obtain the (natural) $\sqrt{n}$-consistent estimator

$$\bar{\rho}_n = \bar{\nu}_n = \frac{1}{k} \sum_{r=1}^{m-1} \sum_{s=r+1}^{m} \hat{\rho}_{rs}^{(n)}. \tag{2.2.60}$$

For $\theta = \mathbf{1}_k \rho$ we get by simple but tedious calculations (see Appendix 2 in Subsection 2.2.7)

$$E\tilde{\ell}_{\rho_{rs}} \tilde{\ell}_{\rho_{tu}} = \begin{cases} (1-\rho^2)^2 & \text{if} \quad |\{r,s\} \cap \{t,u\}| = 2, \\ \frac{1}{2}(1-\rho)^2 \rho(2+3\rho) & \text{if} \quad |\{r,s\} \cap \{t,u\}| = 1, \\ 2(1-\rho)^2 \rho^2 & \text{if} \quad |\{r,s\} \cap \{t,u\}| = 0. \end{cases} \tag{2.2.61}$$

It makes sense to estimate $I(\mathbf{1}_k, G, \mathcal{P})$ by substituting $\bar{\rho}_n$ for ρ in (2.2.61), to compute the inverse of the resulting matrix, and to choose this matrix as the estimator $\hat{I}_n$. To this end, we note that for every pair $\{r,s\}$, $1 \le r \neq s \le m$, there are $2(m-2)$ pairs of $\{t,u\}$'s having one element in common and there are $\frac{1}{2}(m-2)(m-3)$ pairs of $\{t,u\}$'s having no elements in common. Hence, the sum of the components of each column vector of $I^{-1}(\mathbf{1}_k\rho, G, \mathcal{P})$ is $(1-\rho)^2(1+(m-1)\rho)^2$. Each matrix with the components of each column vector adding to 1 has the property that the sum of all row vectors equals the vector with all components equal to 1, and hence the components of each column vector of its inverse also add up to 1. This implies

$$\mathbf{1}_k^T \hat{I}_n = (1-\bar{\rho}_n)^{-2} \left(1+(m-1)\bar{\rho}_n\right)^{-2} \mathbf{1}_k^T$$

and hence by (2.2.45)

$$\hat{\rho}_n = \left(\mathbf{1}_k^T \hat{I}_n \mathbf{1}_k\right)^{-1} \mathbf{1}_k^T \hat{I}_n \hat{\theta}_n = \frac{1}{k} \mathbf{1}_k^T \hat{\theta}_n = \binom{m}{2}^{-1} \sum_{r=1}^{m-1} \sum_{s=r+1}^{m} \hat{\rho}_{rs}^{(n)} = \bar{\rho}_n \tag{2.2.62}$$

attains the asymptotic information bound (cf. (2.2.27))

$$\left(\mathbf{1}_k^T I\left(\mathbf{1}_k \rho, G, \mathcal{P}\right) \mathbf{1}_k\right)^{-1} = \binom{m}{2}^{-1} (1-\rho)^2 (1+(m-1)\rho)^2. \tag{2.2.63}$$

Hoff et al. [34] proved the efficiency of the pseudo-likelihood estimator for ρ in dimension $m = 4$. Segers et al. [35] extended this result to general m and presented the efficient lower bounds for $m = 3$ and $m = 4$ in their Example 5.3. However, their maximum pseudo-likelihood estimator is not as explicit as our (2.2.62).

Subexample 2.2.2.2. Four-dimensional circular Gaussian copula

A particular, one-dimensional parameter type of four-dimensional circular Gaussian copula model has been studied by [34] and [35]. It is defined by its correlation matrix

$$\begin{pmatrix} 1 & \rho & \rho^2 & \rho \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho & \rho^2 & \rho & 1 \end{pmatrix}. \tag{2.2.64}$$

Our semiparametric starting model $\mathcal{P}$ is the same as in (2.2.54) with $m = 4$, but with the components of θ rearranged as follows

$$\theta = (\rho_{12}\ ,\ \rho_{14}\ ,\ \rho_{23}\ ,\ \rho_{34}\ ,\ \rho_{13}\ ,\ \rho_{24})^T.$$

Now, with $f(\rho) = (\rho\ ,\ \rho\ ,\ \rho\ ,\ \rho\ ,\ \rho^2\ ,\ \rho^2)^T$ the present circular Gaussian submodel $\mathcal{Q}$ may be written as

$$\mathcal{Q} = \{P_{f(\rho),G}\ :\ \rho \in (-\tfrac{1}{3}, 1)\,,\ G \in \mathcal{G}\}.$$

In order to construct an efficient estimator of ρ within $\mathcal{Q}$ along the lines of Theorem 2.2.2, we propose as a $\sqrt{n}$-consistent estimator of ρ

$$\begin{aligned} \bar{\rho}_n &= \tfrac{2}{3}\bar{\rho}_{n,1} + \tfrac{1}{3}\,\mathrm{sign}\,(\bar{\rho}_{n,1})\,\bar{\rho}_{n,2}, \\ \bar{\rho}_{n,1} &= \tfrac{1}{4}\left(\hat{\rho}_{12}^{(n)} + \hat{\rho}_{14}^{(n)} + \hat{\rho}_{23}^{(n)} + \hat{\rho}_{34}^{(n)}\right),\ \bar{\rho}_{n,2} = \tfrac{1}{2}\left(\sqrt{\hat{\rho}_{13}^{(n)}} + \sqrt{\hat{\rho}_{24}^{(n)}}\right) \end{aligned} \tag{2.2.65}$$

As in (2.2.61) we get by simple but tedious calculations (see Appendix 2 in Subsection 2.2.7)

$$I^{-1}(f(\rho), G, \mathcal{P}) = \tfrac{1}{2}\left(1-\rho^2\right)^2 \tag{2.2.66}$$

$$\begin{pmatrix} 2 & \rho^2 & \rho^2 & 2\rho^2 & \rho\left(2+\rho^2\right) & \rho\left(2+\rho^2\right) \\ \rho^2 & 2 & 2\rho^2 & \rho^2 & \rho\left(2+\rho^2\right) & \rho\left(2+\rho^2\right) \\ \rho^2 & 2\rho^2 & 2 & \rho^2 & \rho\left(2+\rho^2\right) & \rho\left(2+\rho^2\right) \\ 2\rho^2 & \rho^2 & \rho^2 & 2 & \rho\left(2+\rho^2\right) & \rho\left(2+\rho^2\right) \\ \rho\left(2+\rho^2\right) & \rho\left(2+\rho^2\right) & \rho\left(2+\rho^2\right) & \rho\left(2+\rho^2\right) & 2\left(1+\rho^2\right)^2 & 4\rho^2 \\ \rho\left(2+\rho^2\right) & \rho\left(2+\rho^2\right) & \rho\left(2+\rho^2\right) & \rho\left(2+\rho^2\right) & 4\rho^2 & 2\left(1+\rho^2\right)^2 \end{pmatrix},$$

which has inverse

$$I(f(\rho), G, \mathcal{P}) = \tfrac{1}{2}\left(1-\rho^2\right)^{-4} \tag{2.2.67}$$

$$\begin{pmatrix}
\rho^4+2 & 3\rho^2 & 3\rho^2 & \rho^4+2\rho^2 & -\left(\rho^3+2\rho\right) & -\left(\rho^3+2\rho\right)\\
3\rho^2 & \rho^4+2 & \rho^4+2\rho^2 & 3\rho^2 & -\left(\rho^3+2\rho\right) & -\left(\rho^3+2\rho\right)\\
3\rho^2 & \rho^4+2\rho^2 & \rho^4+2 & 3\rho^2 & -\left(\rho^3+2\rho\right) & -\left(\rho^3+2\rho\right)\\
\rho^4+2\rho^2 & 3\rho^2 & 3\rho^2 & \rho^4+2 & -\left(\rho^3+2\rho\right) & -\left(\rho^3+2\rho\right)\\
-\left(\rho^3+2\rho\right) & -\left(\rho^3+2\rho\right) & -\left(\rho^3+2\rho\right) & -\left(\rho^3+2\rho\right) & 2\frac{\rho^6+\rho^4+1}{\rho^4+1} & 2\frac{\rho^6+2\rho^2}{\rho^4+1}\\
-\left(\rho^3+2\rho\right) & -\left(\rho^3+2\rho\right) & -\left(\rho^3+2\rho\right) & -\left(\rho^3+2\rho\right) & 2\frac{\rho^6+2\rho^2}{\rho^4+1} & 2\frac{\rho^6+\rho^4+1}{\rho^4+1}
\end{pmatrix}.$$

Substituting $\bar{\rho}_n$ into (2.2.67) we obtain a $\sqrt{n}$-consistent estimator of $I(f(\rho), G, \mathcal{P})$. In view of $\dot{f}(\rho) = (1,1,1,1,2\rho,2\rho)^T$ we have

$$\dot{f}^T(\rho) I(f(\rho), G, \mathcal{P}) = \left(1-\rho^2\right)^{-3}\left(1+\rho^2, 1+\rho^2, 1+\rho^2, 1+\rho^2, -2\rho, -2\rho\right).$$

Consequently the asymptotic lower bound for estimation of ρ within $\mathcal{Q}$ equals

$$\left[\dot{f}(\rho)^T I(f(\rho), G, \mathcal{P}) \dot{f}(\rho)\right]^{-1} = \tfrac{1}{4}\left(1-\rho^2\right)^2. \tag{2.2.68}$$

Substituting $\bar{\rho}_n$ for ρ we obtain as the efficient estimator from Theorem 2.2.2

$$\hat{\rho}_n = \bar{\rho}_n + \frac{1+\bar{\rho}_n^2}{1-\bar{\rho}_n^2}\left(\bar{\rho}_{n,1} - \bar{\rho}_n\right) - \frac{\bar{\rho}_n}{1-\bar{\rho}_n^2}\left(\tfrac{1}{2}\left(\hat{\rho}_{13}^{(n)} + \hat{\rho}_{24}^{(n)}\right) - \bar{\rho}_n^2\right). \tag{2.2.69}$$

Hoff et al. [34] have shown that the pseudo-likelihood estimator is not efficient in this case. Segers et al. [35] have established the asymptotic lower bound (2.2.68) and have constructed an alternative, efficient, one-step updating estimator suggesting the pseudo-maximum likelihood estimator as the preliminary estimator.

Example 2.2.3. Partial spline linear regression

Here the observations are realizations of i.i.d. copies of the random vector $X = (Y, Z^T, U^T)^T$ with Y, Z, and U 1-dimensional, k-dimensional, and p-dimensional random vectors with the structure

$$Y = \theta^T Z + \psi(U) + \varepsilon, \tag{2.2.70}$$

where the measurement error ε is independent of Z and U, has mean 0, finite variance, and finite Fisher information for location, and where $\psi(\cdot)$ is a real valued function on $\mathbb{R}^p$. Schick [36] calls this partly linear additive regression, [23] mention it as partial spline regression, whereas [37] are talking about the partial smoothing spline model. Under the regularity conditions of his Theorem 8.1 [36] presents an efficient estimator of θ and a consistent estimator of $I(\theta, G, \mathcal{P})$. Consequently our Theorem 2.2.2 may be applied directly in order to obtain an efficient estimator of ν in appropriate submodels with $\theta = f(\nu)$ without our construction of an estimator of $I(\theta, G, \mathcal{P})$ via characteristic functions. Note that for submodels with θ restricted to a linear subspace, $\theta = L\nu$ say, our approach is not needed, since the reparametrization $Y = \nu^T L^T Z + \psi(U) + \varepsilon$ brings the estimation problem back to its original (2.2.70).

Example 2.2.4. Multivariate normal with common mean

Let $\mathcal{G}$ be the collection of nonsingular $k \times k$-covariance matrices and let the parametric starting model be the collection of nondegenerate normal distributions with mean vector θ and covariance matrix Σ,

$$\mathcal{P} = \left\{ P_{\theta,\Sigma} \,:\, \theta \in \mathbb{R}^k,\ \Sigma \in \mathcal{G} \right\}. \tag{2.2.71}$$

Efficient estimators of θ and Σ are the sample mean $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ and the sample covariance matrix $\hat{\Sigma}_n = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)^T$, respectively. Note that $\bar{X}_n$ attains the finite sample Cramér-Rao bound and the asymptotic information bound with $I(\theta, \Sigma, \mathcal{P}) = \Sigma^{-1}$.

The parametric submodel we consider is

$$\mathcal{Q} = \left\{ P_{\mathbf{1}_k\mu,\Sigma} \,:\, \mu \in \mathbb{R},\ \Sigma \in \mathcal{G} \right\}. \tag{2.2.72}$$

In view of (2.2.45) and (2.2.28)

$$\hat{\mu}_n = \left(\mathbf{1}_k^T \hat{\mathbf{\Sigma}}_n^{-1} \mathbf{1}_k \right)^{-1} \mathbf{1}_k^T \hat{\mathbf{\Sigma}}_n^{-1} \bar{\mathbf{X}}_\mathbf{n} \tag{2.2.73}$$

is an efficient estimator of μ within $\mathcal{Q}$ that attains the asymptotic lower bound $\left(\mathbf{1}_k^T \Sigma^{-1} \mathbf{1}_k\right)^{-1}$. In case the covariance matrix Σ is diagonal with its variances denoted by $\sigma_1^2, \dots, \sigma_k^2$, we are dealing with the Graybill-Deal model as presented by [22] on her page 88. With $\bar{X}_{i,n} = \frac{1}{n}\sum_{j=1}^n X_{j,i}$, $S_{i,n}^2 = \frac{1}{n}\sum_{j=1}^n (X_{j,i} - \bar{X}_{i,n})^2$, and $\hat{\Sigma}_n = \operatorname{diag}(S_{1,n}^2, \dots, S_{k,n}^2)$ we obtain the Graybill-Deal estimator

$$\hat{\mu}_n = \frac{\sum_{i=1}^k \bar{X}_{i,n}/S_{i,n}^2}{\sum_{i=1}^k 1/S_{i,n}^2} \tag{2.2.74}$$

with asymptotic lower bound $\left(\mathbf{1}_k^T \Sigma^{-1} \mathbf{1}_k\right)^{-1} = 1/\sum_{i=1}^k 1/\sigma_i^2$.

Example 2.2.5. Restricted maximum likelihood estimator

Maximum likelihood estimation of the generalized linear model under linear restrictions on the parameters is done in [27] via an iterative procedure using a penalty function. Kim and Taylor [28] introduce the restricted EM algorithm for maximum likelihood estimation under linear restrictions. Our approach as described in Remark 2.2.1 with $\hat{\theta}_n$ a(n unrestricted) maximum likelihood estimator avoids such iterative procedures.

2.2.7 Appendices

Appendix 1: Additional Proofs

In this appendix proofs will be presented of (2.2.49) and (2.2.50).

Since $\hat{I}_n$ has been chosen to be symmetric and positive definite, $x^T\hat{I}_n y$, $x, y \in \mathbb{R}^k$, is an inner product on $\mathbb{R}^k$. Define the $k \times k$-matrices $\Pi_{n,L}$ and $\Pi_{n,R}$ by

$$\begin{aligned}\Pi_{n,L} &= L\left(L^T\hat{I}_nL\right)^{-1}L^T\hat{I}_n,\\ \Pi_{n,R} &= \hat{I}_n^{-1}R\left(R^T\hat{I}_n^{-1}R\right)^{-1}R^T.\end{aligned} \tag{2.2.75}$$

With the above inner product these matrices are projection matrices on the linear subspaces spanned by the columns of L and $\hat{I}_n^{-1}R$, respectively. Indeed, $\Pi_{n,L}\Pi_{n,L} = \Pi_{n,L}$, $\Pi_{n,R}\Pi_{n,R} = \Pi_{n,R}$, $(x - \Pi_{n,L}x)^T\hat{I}_n\Pi_{n,L}x = 0$, $x \in \mathbb{R}^k$, $(y - \Pi_{n,R}y)^T\hat{I}_n\Pi_{n,R}y = 0$, $y \in \mathbb{R}^k$, $\Pi_{n,L}Lx = Lx$, $x \in \mathbb{R}^d$, and $\Pi_{n,R}\hat{I}_n^{-1}Ry = \hat{I}_n^{-1}Ry$, $y \in \mathbb{R}^{k-d}$ hold. The linear subspaces spanned by the columns of L and $\hat{I}_n^{-1}R$ have dimensions d and $k-d$, respectively, since the matrices (L, R) and $\hat{I}_n$ are nonsingular. Moreover, these linear subspaces are orthogonal in view of $L^T\hat{I}_n\hat{I}_n^{-1}R = L^TR = 0$. This implies

$$\Pi_{n,L}x + \Pi_{n,R}x = x, \quad x \in \mathbb{R}^k. \tag{2.2.76}$$

Combining (2.2.75), (2.2.76), and (2.2.46) we obtain (2.2.49) and, by the consistency of $\hat{I}_n$, (2.2.50).

Appendix 2: Computational Details

We present the computational details for (2.2.61) and (2.2.66) presented in Example 2.2.2. Since our computations will be based on fourth moments of multivariate normal random variables, we consider

$$Z = \begin{pmatrix} Z_a \\ Z_b \\ Z_c \\ Z_d \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{ab} & \rho_{ac} & \rho_{ad} \\ \rho_{ba} & 1 & \rho_{bc} & \rho_{bd} \\ \rho_{ca} & \rho_{cb} & 1 & \rho_{cd} \\ \rho_{da} & \rho_{db} & \rho_{dc} & 1 \end{pmatrix}\right).$$

The following fourth moments of Z can be obtained by straightforward computations:

- $E(Z_a^4) = 3$
- $E(Z_a^3 Z_b) = 3\rho_{ab}$
- $E(Z_a^2 Z_b^2) = 1 + 2\rho_{ab}^2$
- $E(Z_a^2 Z_b Z_c) = \rho_{bc} + 2\rho_{ab}\rho_{ac}$
- $E(Z_a Z_b Z_c Z_d) = \rho_{ab}\rho_{cd} + \rho_{ac}\rho_{bd} + \rho_{ad}\rho_{bc}.$

For every $i, j = 1, \ldots, \binom{n}{2}$ let M_{ij} be the element in the i-th row and j-th column of the efficient lower bound $I^{-1}(\theta, G, \mathcal{P})$. Because of $\theta_i = \rho_{ab}$, $\theta_j = \rho_{cd}$ for some a, b, c, and d, we have

$$M_{ij} = E\left(Z_a Z_b - \tfrac{1}{2}\rho_{ab}\left[Z_a^2 + Z_b^2\right]\right)\left(Z_c Z_d - \tfrac{1}{2}\rho_{cd}\left[Z_c^2 + Z_d^2\right]\right).$$

We have three cases:

- $|\{a, b\} \cap \{c, d\}| = 2$

$$\begin{aligned} M_{ii} &= E\left(Z_a Z_b - \tfrac{1}{2}\rho_{ab}\left[Z_a^2 + Z_b^2\right]\right)^2 \\ &= E\left(Z_a^2 Z_b^2\right) - \rho_{ab} E\left(Z_a^3 Z_b + Z_b^3 Z_a\right) + \tfrac{1}{4}\rho_{ab}^2 E\left(Z_a^4 + 2Z_a^2 Z_b^2 + Z_b^4\right) \\ &= \left(1 + 2\rho_{ab}^2\right) - \rho_{ab}\left(3\rho_{ab} + 3\rho_{ab}\right) + \tfrac{1}{4}\rho_{ab}^2\left(3 + 2\left[1 + 2\rho_{ab}^2\right] + 3\right) \\ &= \left(1 - \rho_{ab}^2\right)^2 \end{aligned}$$

- $|\{a,b\}\cap\{c,d\}| = 1$ (*without lost of generality* assume $d = a$)

$$
\begin{aligned}
M_{ij} &= E\left(Z_aZ_b - \tfrac{1}{2}\rho_{ab}\left[Z_a^2 + Z_b^2\right]\right)\left(Z_aZ_c - \tfrac{1}{2}\rho_{ac}\left[Z_a^2 + Z_c^2\right]\right) \\
&= E\left(Z_a^2Z_bZ_c\right) - \tfrac{1}{2}\rho_{ab}E\left(Z_a^3Z_c + Z_b^2Z_aZ_c\right) \\
&\quad -\tfrac{1}{2}\rho_{ac}E\left(Z_a^3Z_b + Z_c^2Z_aZ_b\right) \\
&\quad +\tfrac{1}{4}\rho_{ab}\rho_{ac}E\left(Z_a^4 + Z_a^2Z_b^2 + Z_a^2Z_c^2 + Z_b^2Z_c^2\right) \\
&= \left(\rho_{bc} + 2\rho_{ab}\rho_{ac}\right) - \tfrac{1}{2}\rho_{ab}\left(3\rho_{ac} + \left[\rho_{ac} + 2\rho_{ab}\rho_{bc}\right]\right) \\
&\quad -\tfrac{1}{2}\rho_{ac}\left(3\rho_{ab} + \left[\rho_{ab} + 2\rho_{ac}\rho_{bc}\right]\right) \\
&\quad +\tfrac{1}{4}\rho_{ab}\rho_{ac}\left(3 + \left[1 + 2\rho_{ab}^2\right] + \left[1 + 2\rho_{ac}^2\right] + \left[1 + 2\rho_{bc}^2\right]\right) \\
&= \tfrac{1}{2}\left(1 - \rho_{ab}^2 - \rho_{ac}^2\right)\left(2\rho_{bc} - \rho_{ab}\rho_{ac}\right) + \tfrac{1}{2}\rho_{ab}\rho_{ac}\rho_{bc}^2
\end{aligned}
$$

- $|\{a,b\}\cap\{c,d\}| = 0$

$$
\begin{aligned}
M_{ij} &= E\left(Z_aZ_b - \tfrac{1}{2}\rho_{ab}\left[Z_a^2 + Z_b^2\right]\right)\left(Z_cZ_d - \tfrac{1}{2}\rho_{cd}\left[Z_c^2 + Z_d^2\right]\right) \\
&= E\left(Z_aZ_bZ_cZ_d\right) - \tfrac{1}{2}\rho_{ab}E\left(Z_a^2Z_cZ_d + Z_b^2Z_cZ_d\right) \\
&\quad -\tfrac{1}{2}\rho_{cd}E\left(Z_c^2Z_aZ_b + Z_d^2Z_aZ_b\right) \\
&\quad +\tfrac{1}{4}\rho_{ab}\rho_{cd}E\left(Z_a^2Z_c^2 + Z_b^2Z_c^2 + Z_a^2Z_d^2 + Z_b^2Z_d^2\right) \\
&= \rho_{ab}\rho_{cd} + \rho_{ac}\rho_{bd} + \rho_{ad}\rho_{bc} - \tfrac{1}{2}\rho_{ab}\left(\left[\rho_{cd} + 2\rho_{ac}\rho_{ad}\right] + \left[\rho_{cd} + 2\rho_{bc}\rho_{bd}\right]\right) \\
&\quad -\tfrac{1}{2}\rho_{cd}\left(\left[\rho_{ab} + 2\rho_{ac}\rho_{bc}\right] + \left[\rho_{ab} + 2\rho_{ad}\rho_{bd}\right]\right) \\
&\quad +\tfrac{1}{4}\rho_{ab}\rho_{cd}\left(\left[1 + 2\rho_{ac}^2\right] + \left[1 + 2\rho_{bc}^2\right] + \left[1 + 2\rho_{ad}^2\right] + \left[1 + 2\rho_{bd}^2\right]\right) \\
&= \rho_{ac}\rho_{bd} + \rho_{ad}\rho_{bc} - \left(\rho_{ab}\rho_{ac}\rho_{ad} + \rho_{ba}\rho_{bc}\rho_{bd} + \rho_{ca}\rho_{cb}\rho_{cd} + \rho_{da}\rho_{db}\rho_{dc}\right) \\
&\quad +\tfrac{1}{2}\rho_{ab}\rho_{cd}\left(\rho_{ac}^2 + \rho_{bc}^2 + \rho_{ad}^2 + \rho_{bd}^2\right)
\end{aligned}
$$

Finally, substitution of the correlation structures in Subexample 2.2.2.1 and Subexample 2.2.2.2 give (2.2.61) and (2.2.66), respectively.

2.3 Semiparametrically Efficient Estimation of Euclidean Parameters under Equality Constraints

2.3.1 Abstract

Assume a (semi)parametrically efficient estimator is given of the Euclidean parameter in a (semi)parametric model. A submodel is obtained by constraining this model in that a continuously differentiable function of the Euclidean parameter vanishes. We present an explicit method to construct (semi)parametrically efficient estimators of the Euclidean parameter in such

equality constrained submodels and prove their efficiency. Our construction is based solely on the original efficient estimator and the constraining function.

Only the parametric case of this estimation problem and a nonparametric version of it have been considered in literature.

2.3.2 Introduction

Let $X_1, \ldots, X_n$ be i.i.d. copies of X taking values in the measurable space $(\mathcal{X}, \mathcal{A})$ in a regular semiparametric model with Euclidean parameter $\theta \in \Theta$, where Θ is an open subset of $\mathbb{R}^k$. We denote this semiparametric model by

$$\mathcal{P} = \{P_{\theta,G} \ : \ \theta \in \Theta, \ G \in \mathcal{G}\} . \tag{2.3.1}$$

Typically, the nuisance parameter space $\mathcal{G}$ is a subset of a Banach or Hilbert space. If this space is finite dimensional, we are dealing with a parametric model.

We assume an asymptotically efficient estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ is given of the parameter of interest θ, which under regularity conditions means that

$$\sqrt{n}\left(\hat{\theta}_n - \theta - \frac{1}{n}\sum_{i=1}^{n} \tilde{\ell}(X_i; \theta, G, \mathcal{P})\right) \to_{P_{\theta,G}} 0 \tag{2.3.2}$$

holds. Here $\tilde{\ell}(\cdot; \theta, G, \mathcal{P})$ is the efficient influence function for estimation of θ within $\mathcal{P}$ and

$$I^{-1}(\theta, G, \mathcal{P}) = \int_{\mathcal{X}} \tilde{\ell}(x; \theta, G, \mathcal{P})\tilde{\ell}^T(x; \theta, G, \mathcal{P}) dP_{\theta,G}(x) \tag{2.3.3}$$

is the information bound, which corresponds to the efficient information matrix $I(\theta, G, \mathcal{P})$.

Quite frequently the elements of the parameter of interest $\theta = (\theta_1, \ldots, \theta_k)$ are not mathematically independent but satisfy d functional relationships $S_i(\theta) = 0, \ i = 1, \ldots, d$, with $d < k$. Formally, this can be described as

$$S(\theta) = 0, \quad \theta \in \Theta, \tag{2.3.4}$$

where S is a function from $\mathbb{R}^k$ to $\mathbb{R}^d$. We will assume that the $d \times k$ Jacobian matrix $\dot{S}(\cdot)$ exists, is continuous in θ on Θ, and has full rank d. Thus, we have constrained the semiparametric model $\mathcal{P}$ to a semiparametric submodel of it, namely

$$\mathcal{Q} = \{P_{\theta,G} \,:\, S(\theta) = 0,\ \theta \in \Theta,\ G \in \mathcal{G}\}\,. \tag{2.3.5}$$

Given the constraint $S(\theta) = 0$, we will adapt the semiparametrically efficient estimator $\hat{\theta}_n$ of θ within $\mathcal{P}$ in such a way that the adapted estimator is semiparametrically efficient within the constrained model $\mathcal{Q}$. Of course, it has to have at least as small asymptotic variance as the original estimator $\hat{\theta}_n$ and to be at least as close to the true value stochastically.

Efficient estimation of Euclidean parameters under equality constraints for *nonparametric* models has been studied in [38], [39], [40], [41], in Example 1.3.6, 3.2.3, and 3.3.3 of Bickel et al. (1993), henceforth called [23], in [42], and in [43]. In [23] nonparametric models under equality constraints are called constraint defined models. Let the semiparametric model $\mathcal{P}$ be embedded into a nonparametric model $\tilde{\mathcal{P}}$ and let the map $\nu : \tilde{\mathcal{P}} \to \mathbb{R}^k$ be such that $\nu(P_{\theta,G}) = \theta$ holds for all $P_{\theta,G} \in \mathcal{P}$. In view of $\mathcal{P} \subset \tilde{\mathcal{P}}$ estimation of $\nu(P)$ within $\mathcal{P}$ is easier than within $\tilde{\mathcal{P}}$. This relation between these models also holds under the equality constraint $S(\nu(P)) = 0$. Consequently, the results for nonparametric models under constraints are not directly applicable to our semiparametric situation.

For the constrained parametric estimation problem so-called restricted maximum likelihood estimators have been studied. Aitchison and Silvey [44] have used Lagrange multipliers with an iterative computation method. An alternative iterative construction has been proposed by [45], who also presents a long list of examples of constrained parametric estimation problems. To prove efficiency of these restricted maximum likelihood estimators additional regularity conditions are needed. Our method does not need these additional conditions, provided an efficient estimator for the original unconstrained parametric model is given. Finite sample Cramér-Rao bounds for the constrained parametric case have been derived by e.g. [46], [47], and [48].

To the best of our knowledge the semiparametric version of the topic of the present paper has not been studied in literature yet.

Estimation of the Euclidean parameters constrained by equalities is quite different from estimation of parameters constrained by *inequalities*. A comprehensive treatment of the latter estimation problems may be found in [22].

If $\mathcal{Q}$ can be reparametrized as

$$\mathcal{Q} = \left\{P_{f(\nu),G} \,:\, \nu \in N,\ G \in \mathcal{G}\right\}, \tag{2.3.6}$$

where N is open and $f : N \to \Theta$ is injective and continuously differentiable with full rank Jacobian, then ν can be estimated semiparametrically efficiently as in [20] and, as noted there, θ can be estimated efficiently as well by applying $f(\cdot)$ to the efficient estimator of ν. However, it may be hard or even impossible to find such a reparametrization. A simple, formal example is estimation of the mean vector of a bivariate normal distribution where it is known that this mean vector lies on the unit circle. The unit circle cannot be parametrized as in (2.3.6) with N open and $f(\cdot)$ continuous and injective. Indeed, assume $\nu_n \in N$ converge to a point at the boundary of N. Then $f(\nu_n)$ converge to a point on the unit circle $f(\nu_0)$, say, with $\nu_0 \in N$. But by the continuity of $f(\cdot)$ this implies that there exist a point in N close to the boundary of N and a point in N close to ν_0 that are mapped on the same point of the circle by $f(\cdot)$, which contradicts its injectivity. On the other hand there are submodels $\mathcal{Q}$ of the type (2.3.6) that cannot be viewed as a submodel of the type (2.3.5). Again consider estimation of the mean vector of a bivariate normal distribution where it is known now that this mean vector lies on the unit circle with one point removed. This unit circle with one point removed can be parametrized as in (2.3.6) with $f(\cdot)$ continuous and N open, but it cannot be described via (2.3.4), since the preimage of the closed set $\{0\}$ under a continuous function $S(\cdot)$ has to be closed and the unit circle with one point removed is not. In the present paper, everything will be done directly to the original parameter subject to equality constraints without reparametrizing it.

The outline of the paper is as follows. In Section 2.3.3, we will present a lower bound to the efficient information bound for estimating the parameter of interest within the constrained model $\mathcal{Q}$. This lower bound will be formulated in terms of the efficient information bound of the original model $\mathcal{P}$ and the Jacobian of the constraining function $S(\cdot)$. An explicit estimator that is efficient within the constrained model, will be given in Section 2.3.4. It attains the lower bound from Section 2.3.3, which shows that both this information bound and the estimator are efficient within the constrained model. Examples are discussed in Section 2.3.5. Our conclusions are presented in Section 2.3.6.

2.3.3 Efficient Influence Functions and Projection

In the situation of Section 2.3.2 we denote the so-called *efficient score function* for θ by

$$\ell^*(\cdot;\theta,G,\mathcal{P}) = I(\theta,G,\mathcal{P})\tilde{\ell}(\cdot;\theta,G,\mathcal{P}). \tag{2.3.7}$$

We will restrict attention to regular semiparametric models for which at every $P_0 = P_{\theta_0,G_0} \in \mathcal{P}$ the parameter θ is pathwise differentiable, the tangent space $\dot{\mathcal{P}}$ is the sum of the tangent space $\dot{\mathcal{P}}_1$ for θ and the tangent space $\dot{\mathcal{P}}_2$ for G, and the efficient score function $\ell^*(\cdot;\theta,G,\mathcal{P})$ for θ is the projection of the (ordinary) score function $\dot{\ell}(\cdot;\theta,G,\mathcal{P})$ for θ on the orthocomplement of $\dot{\mathcal{P}}_2$ within $\dot{\mathcal{P}}$ in the sense of componentwise projection within $L_2^0(P_0) = \{f \in L_2(P_0) : E_{P_0} f(X) = 0\}$; for details see Chapter 3 of [23] and Chapter 25 of [49].

By Proposition 3.3.1 of [23] the efficient influence function $\tilde{\ell}(\cdot;\theta,G,\mathcal{Q})$ for θ within the submodel $\mathcal{Q}$ can be obtained by projecting the efficient influence function $\tilde{\ell}(\cdot;\theta,G,\mathcal{P})$ for θ within $\mathcal{P}$ onto the tangent space $\dot{\mathcal{Q}}$ of $\mathcal{Q}$ or onto an appropriate subspace of this tangent space.

Let $\{\theta_\eta : \theta_\eta \in \mathbb{R}^k,\ \eta \in \mathbb{R},\ |\eta| < \epsilon\}$ for sufficiently small $\epsilon > 0$ be a path through θ_0 in $\mathbb{R}^k$ in the direction $r \in \mathbb{R}^k$, which means that $|\theta_\eta - \theta_0 - \eta r| = o(|\eta|)$. If this path satisfies $S(\theta_\eta) = 0$, $|\eta| < \epsilon$, then the differentiability of $S(\cdot)$ at θ_0 implies $|S(\theta_\eta) - S(\theta_0) - \eta \dot{S}(\theta_0) r| = o(|\eta|)$, meaning $|\eta \dot{S}(\theta_0) r| = o(|\eta|)$, and hence $\dot{S}(\theta_0) r = 0$. In other words, such a path within the parameter set $\{\theta : S(\theta) = 0, \theta \in \mathbb{R}^k\}$, has a direction r at θ_0 that belongs to the orthocomplement of the d-dimensional linear space within $\mathbb{R}^k$ spanned by the d row vectors of the Jacobian matrix $\dot{S}(\theta_0)$. In fact, to each element of this orthocomplement $[\dot{S}(\theta_0)]^\perp$ corresponds such a path, as is proved in detail in Appendix 2.3.7 with the help of the implicit function theorem.

With $P_0 \in \mathcal{Q}$ let L be a $k \times (k-d)$-matrix, whose columns span this $(k-d)$-dimensional orthocomplement. Since $\mathcal{P}$ is a regular semiparametric model, the parametric submodel $\mathcal{P}_1 = \{P_{\theta,G_0} : \theta \in \Theta\}$ is regular. With $s(\theta)$ denoting the square root of the density of P_{θ,G_0} with respect to an appropriate dominating measure μ, this regularity implies

$$||s(\theta_\eta) - s(\theta_0) - \tfrac{1}{2} s(\theta_0)(\theta_\eta - \theta_0)^T \dot{\ell}(\theta_0)||_\mu = o(|\theta_\eta - \theta_0|), \quad \theta_\eta \to \theta_0, \quad (2.3.8)$$

where $||\cdot||_\mu$ is the norm of $L_2(\mu)$ and $\dot{\ell}(\theta_0) = \dot{\ell}(\cdot;\theta_0,G_0,\mathcal{P})$ is the score function for θ at θ_0; cf. Definition 2.1.1 and formula (2.1.4) of [23]. For a path $\{\theta_\eta : \theta_\eta \in \mathbb{R}^k,\ \eta \in \mathbb{R},\ |\eta| < \epsilon\}$ with direction r at θ_0 as above, this implies

$$||s(\theta_\eta) - s(\theta_0) - \tfrac{1}{2} \eta s(\theta_0) r^T \dot{\ell}(\theta_0)||_\mu = o(|\eta|), \quad \eta \to 0. \quad (2.3.9)$$

Consequently, we are dealing here with a 1-dimensional regular parametric model with score function $r^T \dot{\ell}(\theta_0)$ for η at $\eta = 0$. It follows that the closed linear span $\left[L^T \dot{\ell}(\theta_0)\right]$ of all such score functions $r^T \dot{\ell}(\theta_0)$ is the tangent space $\dot{\mathcal{Q}}_1$ of $\mathcal{Q}_1 = \{P_{\theta,G_0} : S(\theta) = 0,\ \theta \in \Theta\}$ at P_0. This implies that the tangent space

$\dot{\mathcal{Q}}$ of $\mathcal{Q}$ at P_0 contains both $\left[L^T\dot{\ell}(\theta_0)\right]$ and $\dot{\mathcal{P}}_2$. Writing $\ell^*(\theta_0)$ for $\ell^*(\cdot;\theta_0,G_0,\mathcal{P})$ we have for every tangent $t \in \dot{\mathcal{P}}_2$

$$r^T\dot{\ell}(\theta_0) + t = r^T\ell^*(\theta_0) + t + r^T\left(\dot{\ell}(\theta_0) - \ell^*(\theta_0)\right). \tag{2.3.10}$$

Since $\ell^*(\theta_0)$ is the componentwise projection of $\dot{\ell}(\theta_0)$ on the orthocomplement of $\dot{\mathcal{P}}_2$, each component of $\dot{\ell}(\theta_0) - \ell^*(\theta_0)$ belongs to $\dot{\mathcal{P}}_2$ and we obtain from (2.3.10)

$$\dot{\mathcal{Q}} \supset \left[L^T\dot{\ell}(\theta_0)\right] + \dot{\mathcal{P}}_2 = \left[L^T\ell^*(\theta_0)\right] + \dot{\mathcal{P}}_2 \supset \left[L^T\ell^*(\theta_0)\right]. \tag{2.3.11}$$

Taking $\theta = \theta_0$ in formula (2.3.7) and suppressing θ_0 and $\mathcal{P}$ from the notation we rewrite (2.3.11) as

$$\dot{\mathcal{Q}} \supset \left[L^T\dot{\ell}\right] + \dot{\mathcal{P}}_2 = \left[L^T\ell^*\right] + \dot{\mathcal{P}}_2 \supset \left[L^T\ell^*\right] = \left[L^TI\tilde{\ell}\right]. \tag{2.3.12}$$

We shall denote the componentwise inner product within $L_2^0(P_0)$ by $<\cdot,\cdot>_0$ and the projection within $L_2^0(P_0)$ of the efficient influence function $\tilde{\ell}$ into $\left[L^TI\tilde{\ell}\right]$ by

$$\Pi_0\left(\tilde{\ell} \mid \left[L^TI\tilde{\ell}\right]\right) = AL^TI\tilde{\ell}, \tag{2.3.13}$$

where A is a $k\times(k-d)$-matrix. Since $\tilde{\ell} - \Pi_0\left(\tilde{\ell} \mid \left[L^TI\tilde{\ell}\right]\right)$ has to be orthogonal to $\left[L^TI\tilde{\ell}\right]$, i.e., since

$$\left\langle \tilde{\ell} - AL^TI\tilde{\ell}\,,\, \tilde{\ell}^TIL\right\rangle_0 = I^{-1}IL - AL^TII^{-1}IL = 0 \tag{2.3.14}$$

holds, we have

$$\Pi_0\left(\tilde{\ell} \mid \left[L^TI\tilde{\ell}\right]\right) = L\left(L^TIL\right)^{-1}L^TI\tilde{\ell}. \tag{2.3.15}$$

In order to write this projection in terms of $\dot{S} = \dot{S}(\theta_0)$ we note that according to the Appendix 1 in Subsection 2.2.7 of [20] $L(L^TIL)^{-1}L^TI + I^{-1}\dot{S}^T(\dot{S}I^{-1}\dot{S}^T)^{-1}\dot{S}$ is the identity map, which implies

$$\Pi_0\left(\tilde{\ell} \mid \left[L^TI\tilde{\ell}\right]\right) = \tilde{\ell} - I^{-1}\dot{S}^T(\dot{S}I^{-1}\dot{S}^T)^{-1}\dot{S}\tilde{\ell}. \tag{2.3.16}$$

By Theorem 3.3.2.A of [23] and formula (3.3.27) in particular, this implies that the limit distribution under P_0 of any properly normalized regular estimator of θ within the submodel $\mathcal{Q}$ is the convolution of a normal distribution with mean 0 and covariance matrix

$$L\left(L^TIL\right)^{-1}L^T = I^{-1} - I^{-1}\dot{S}^T(\dot{S}I^{-1}\dot{S}^T)^{-1}\dot{S}I^{-1} \tag{2.3.17}$$

and some other distribution. In the next Section we shall construct an estimator of θ within $\mathcal{Q}$ that is asymptotically linear in the influence function from (2.3.16). Consequently, it is asymptotically normal with minimal covariance matrix, i.e.,

$$\sqrt{n}\left(\tilde{\theta} - \theta_0\right) \to_{P_0} \mathcal{N}\left(0, I^{-1} - I^{-1}\dot{S}^T(\dot{S}I^{-1}\dot{S}^T)^{-1}\dot{S}I^{-1}\right) \tag{2.3.18}$$

holds.

2.3.4 Efficient Estimator under Equality Constraints

Note that $S(\theta) = S(\theta) - S(\theta_0) = \dot{S}(\theta_0)(\theta - \theta_0) + o(|\theta - \theta_0|)$ holds for θ_0 with $S(\theta_0) = 0$. Since an efficient estimator $\hat{\theta}_n$ within $\mathcal{P}$ is asymptotically linear in the efficient influence function $\tilde{\ell}(\cdot;\theta, G, \mathcal{P})$, this implies that $S(\hat{\theta}_n)$ is asymptotically linear in the influence function $\dot{S}(\theta_0)\tilde{\ell}(\cdot;\theta_0, G_0, \mathcal{P})$ under θ_0. In order to construct an efficient estimator of θ within $\mathcal{Q}$ we will use this asymptotic linearity.

Our main result reads as follows.

Theorem 2.3.1. *Consider the regular semiparametric model $\mathcal{P}$ and its submodel $\mathcal{Q}$ given by* (2.3.1) *and* (2.3.5)*, respectively. Assume that $S : \mathbb{R}^k \to \mathbb{R}^d, d < k$, is continuously differentiable with Jacobian matrix $\dot{S}(\cdot)$ of full rank d, and that the tangent spaces satisfy the conditions mentioned in the first paragraph of Section 2.3.3. Let $X_1, \ldots, X_n$ be i.i.d. with distribution $P \in \mathcal{P}$ and suppose that $\hat{\theta}_n$ is an efficient estimator of the parameter of interest θ within $\mathcal{P}$ based on $X_1, \ldots, X_n$ with efficient influence function $\tilde{\ell}(\cdot;\theta, G, \mathcal{P})$ and that $\hat{I}_n$ is a consistent estimator of $I(\theta, G, \mathcal{P})$ from* (2.3.3)*. Write*

$$\theta_n^* = \hat{\theta}_n - \hat{I}_n^{-1}\dot{S}^T(\hat{\theta}_n)\left(\dot{S}(\hat{\theta}_n)\hat{I}_n^{-1}\dot{S}^T(\hat{\theta}_n)\right)^{-1}S(\hat{\theta}_n) \tag{2.3.19}$$

and define

$$\tilde{\theta}_n = \underset{\zeta,\, S(\zeta)=0}{\arg\min} \parallel \zeta - \theta_n^* \parallel \tag{2.3.20}$$

with $\parallel \cdot \parallel$ the Euclidean norm or a topologically equivalent norm. Then $\tilde{\theta}_n$

efficiently estimates θ within the submodel $\mathcal{Q}$ with efficient influence function

$$\begin{aligned} \tilde{\ell}(\cdot;\theta,G,\mathcal{Q}) &= \tilde{\ell}(\cdot;\theta,G,\mathcal{P}) \\ &\quad -I^{-1}(\theta,G,\mathcal{P})\dot{S}^T(\theta)\left(\dot{S}(\theta)I^{-1}(\theta,G,\mathcal{P})\dot{S}^T(\theta)\right)^{-1}\dot{S}(\theta)\tilde{\ell}(\cdot;\theta,G,\mathcal{P}) \end{aligned} \tag{2.3.21}$$

and hence it satisfies (2.3.18). *Furthermore,*

$$\sqrt{n}(\tilde{\theta}_n - \theta_n^*) \to_{P_{\theta,G}} 0 \tag{2.3.22}$$

holds.

Proof. In view of the convolution result proved in Section 2.3.3 (cf. (2.3.17)) it suffices to show that $\tilde{\theta}_n$ is asymptotically linear in the influence function from (2.3.21), since this yields both sharpness of the convolution bound and efficiency of the estimator. Fix θ_0 with $S(\theta_0) = 0$ and $P_0 = P_{\theta_0,G_0} \in \mathcal{Q}$, and write

$$\begin{aligned} &\sqrt{n}\left(\theta_n^* - \theta_0 - \tfrac{1}{n}\sum_{i=1}^n \tilde{\ell}(X_i;\theta_0,G_0,\mathcal{Q})\right) \\ &\quad = \sqrt{n}\left(\hat{\theta}_n - \theta_0 - \tfrac{1}{n}\sum_{i=1}^n \tilde{\ell}(X_i;\theta_0,G_0,\mathcal{P})\right) \\ &\qquad - \hat{I}_n^{-1}\dot{S}^T(\hat{\theta}_n)\left(\dot{S}(\hat{\theta}_n)\hat{I}_n^{-1}\dot{S}^T(\hat{\theta}_n)\right)^{-1} \\ &\qquad\qquad \times\sqrt{n}\left(S(\hat{\theta}_n) - \tfrac{1}{n}\sum_{i=1}^n \dot{S}(\theta_0)\tilde{\ell}(X_i;\theta_0,G_0,\mathcal{P})\right) \\ &\qquad - \left(\hat{I}_n^{-1}\dot{S}^T(\hat{\theta}_n)\left(\dot{S}(\hat{\theta}_n)\hat{I}_n^{-1}\dot{S}^T(\hat{\theta}_n)\right)^{-1}\right. \\ &\qquad\qquad \left. -I^{-1}(\theta_0,G_0,\mathcal{P})\dot{S}^T(\theta_0)\left(\dot{S}(\theta_0)I^{-1}(\theta_0,G_0,\mathcal{P})\dot{S}^T(\theta_0)\right)^{-1}\right)\dot{S}(\theta_0) \\ &\qquad\qquad \times\tfrac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\ell}(X_i;\theta_0,G_0,\mathcal{P}) \\ &\quad = R_{1,n} - R_{2,n} - R_{3,n}. \end{aligned} \tag{2.3.23}$$

The asymptotic linearity of $\hat{\theta}_n$ from (2.3.2) implies that $R_{1,n}$ converges to 0 in probability under P_0. By the central limit theorem the second factor of $R_{3,n}$ is asymptotically normal with mean 0 and covariance matrix $I^{-1}(\theta_0,G_0,\mathcal{P})$ from (2.3.3). Since $\dot{S}(\cdot)$ is continuous and $\hat{I}_n$ and $\hat{\theta}_n$ are consistent in estimating $I(\theta_0,G_0,\mathcal{P})$ and θ_0, respectively, this implies that $R_{3,n}$ converges to 0 in probability under P_0 as well. We also conclude that the first factor of $R_{2,n}$ is bounded in probability. Together with the asymptotic linearity of $S(\hat{\theta}_n)$, as noted at the start of this Section, this yields the convergence of $R_{2,n}$ to 0 in probability under P_0.

It remains to be shown that (2.3.22) holds. In view of $S(\theta_0) = 0$ and Appendix 2.3.7 we may parametrize a part of the zero set of $S(\cdot)$ near θ_0 by

$$\mathcal{S}_0 = \{\theta \,|\, \theta = \theta_0 + L\eta + r(\eta),\ \eta \in H\}\,, \tag{2.3.24}$$

where the $k - d$ columns of the matrix L span the orthocomplement of $[\dot{S}(\theta_0)]$, $r(\eta) = o(\| \eta \|)$ holds as $\| \eta \|$ tends to 0, and H is an appropriate neighborhood of 0 within $\mathbb{R}^{k-d}$. Note that $n^{-1}\sum_{i=1}^n \tilde{\ell}(X_i;\theta_0,G_0,\mathcal{Q})$ is of the order $O_p(1/\sqrt{n})$ under P_0 and takes its values in $[L]$ in view of (2.3.15). Together with (2.3.24) this shows that there exists a random k-vector $\tilde{R}_n = o_p(1/\sqrt{n})$ such that

$$\theta_0 + \tfrac{1}{n}\sum_{i=1}^n \tilde{\ell}(X_i;\theta_0,G_0,\mathcal{Q}) + \tilde{R}_n \in \mathcal{S}_0 \tag{2.3.25}$$

holds with probability tending to 1. Because of the definition of $\tilde{\theta}_n$, the triangle inequality, and the asymptotic linearity of θ_n^* in the efficient influence function as proved above, this yields

$$\begin{aligned}
&\| \tilde{\theta}_n - \theta_n^* \| \\
&\quad \leq \| \theta_0 + \tfrac{1}{n}\sum_{i=1}^n \tilde{\ell}(X_i;\theta_0,G_0,\mathcal{Q}) + \tilde{R}_n - \theta_n^* \| \\
&\quad \leq \| \theta_0 + \tfrac{1}{n}\sum_{i=1}^n \tilde{\ell}(X_i;\theta_0,G_0,\mathcal{Q}) - \theta_n^* \| + \| \tilde{R}_n \| \\
&\quad = o_p\left(\tfrac{1}{\sqrt{n}}\right),
\end{aligned} \tag{2.3.26}$$

which proves (2.3.22). □

Remark 2.3.1. Consistent estimators $\hat{I}_n$ of $I(\theta, G, \mathcal{P})$ may be constructed from $\hat{\theta}_n$ as in Section 4 of [20]. In regular parametric cases the Fisher information $I(\theta) = I(\theta, G, \mathcal{P})$ depends on θ only and is continuous in it. Consequently, $\hat{I}_n = I(\hat{\theta}_n)$ is consistent in estimating $I(\theta)$ then.

Remark 2.3.2. According to Theorem 2.3.1 the estimators θ_n^* and $\tilde{\theta}_n$ have the same asymptotic performance to first order. However, only $\tilde{\theta}_n$ is guaranteed to be efficient within $\mathcal{Q}$, since θ_n^* need not be a zero of $S(\cdot)$. In order to compute $\tilde{\theta}_n$ to the desired order of precision one typically needs an iterative numerical procedure, like Newton-Raphson.

Remark 2.3.3. Parametrize the linear case by $S(\theta) = R^T(\theta - \alpha)$ with R a $d \times k$-matrix and α a fixed k-vector. Now, $\dot{S}(\theta) = R^T$ holds and the estimator from (2.3.20) reduces to

$$\tilde{\theta}_n = \hat{\theta}_n - \hat{I}_n^{-1} R \left(R^T \hat{I}_n^{-1} R \right)^{-1} R^T \left(\hat{\theta}_n - \alpha \right). \tag{2.3.27}$$

In terms of a $k \times (k-d)$-matrix L, whose columns span the orthocomplement of $[\dot{S}^T(\theta)] = [R]$, this estimator may be written as

$$\tilde{\theta}_n = \alpha + L \left(L^T \hat{I}_n L \right)^{-1} L^T \hat{I}_n \left(\hat{\theta}_n - \alpha \right) \tag{2.3.28}$$

according to the Appendix 1 in Subsection 2.2.7 of [20]. Note that this estimator attains the asymptotic information bound

$$L \left(L^T I(\theta, G, \mathcal{P}) L \right)^{-1} L^T. \tag{2.3.29}$$

Comparing their formula (4.18) to (2.3.28) above we note that the approaches of the present paper and of [20] yield exactly the same estimator in the linear case, although the approaches differ in the general case.

Remark 2.3.4. The estimators $\tilde{\theta}_n$ and $\hat{\theta}_n$ are efficient within the models $\mathcal{Q}$ and $\mathcal{P}$, respectively. Since $\mathcal{Q}$ is a submodel of $\mathcal{P}$, it is easier to estimate θ within $\mathcal{Q}$ than within $\mathcal{P}$. This is visible in the respective limit distributions by comparing (2.3.2) and (2.3.3) to (2.3.18). The difference between the two limit covariance matrices is $I^{-1}\dot{S}^T(\dot{S}I^{-1}\dot{S}^T)^{-1}\dot{S}I^{-1}$, which is positive semidefinite because of the nonsingularity of the symmetric information matrix I, the maximum rank of $\dot{S}$, and the fact that the inverse of a symmetric positive definite matrix is also symmetric positive definite.

By Theorem 2.3.1, (2.3.15), and (2.3.16) we have

$$\tilde{\theta}_n - \theta_0 = L \left(L^T I L \right)^{-1} L^T I \left(\hat{\theta}_n - \theta_0 \right) + o_p\left(\tfrac{1}{\sqrt{n}} \right). \tag{2.3.30}$$

This means that $\tilde{\theta}_n - \theta_0$ may be viewed as a projection of $\hat{\theta}_n - \theta_0$ into $[L]$, approximately. In other words, $\tilde{\theta}_n$ tends to be closer to the true value θ_0 than $\hat{\theta}_n$ in the metric induced by I.

2.3.5 Examples

Our construction of (semi)parametrically efficient estimators will be illustrated in this section by some examples, all of which have been discussed also in Section 5 of the companion paper [20].

Example 2.3.1. Coefficient of variation known

Let $g(\cdot)$ be an absolutely continuous density on $(\mathbb{R},\mathcal{B})$ with mean 0, variance 1, distribution function G, and derivative $g'(\cdot)$, such that $\int[1+x^2](g'/g(x))^2 g(x)dx$ is finite. Consider the location-scale family corresponding to $g(\cdot)$. Let there be given efficient estimators $\bar{\mu}_n$ and $\bar{\sigma}_n$ of μ and σ, respectively, based on $X_1,\ldots,X_n$, which are i.i.d. with density $\sigma^{-1}g((\cdot-\mu)/\sigma)$. By I_{ij} we denote the element in the $i$the row and jth column of the matrix $I=\sigma^2 I(\theta,G,\mathcal{P})$, where the Fisher information matrix $I(\theta,G,\mathcal{P})$ is as defined in (2.3.3) with $\theta=(\mu,\sigma)^T$. Some computation shows

$$I_{11}=\int(g'/g)^2 g, I_{12}=I_{21}=\int x(g'/g(x))^2 g(x)dx,$$

and

$$I_{22}=\int[xg'/g(x)+1]^2 g(x)dx$$

exist and are finite; cf. Section I.2.3 of [29].

We consider the submodel with the coefficient of variation σ/μ known to be equal to a given constant c. We may put this constraint in a linear form by choosing $S(\theta)=c\theta_1-\theta_2$. By Remark 2.3.3 and Example 5.1 of [20] this implies that the efficient estimator $\tilde{\theta}_n$ of θ within the constraint model $\mathcal{Q}$ from Theorem 2.3.1 equals

$$\tilde{\theta}_n=(\hat{\mu}_n,c\hat{\mu}_n)^T \tag{2.3.31}$$

with

$$\hat{\mu}_n=\left(I_{11}+2cI_{12}+c^2I_{22}\right)^{-1}\left[\left(I_{11}+cI_{12}\right)\bar{\mu}_n+\left(I_{12}+cI_{22}\right)\bar{\sigma}_n\right]. \tag{2.3.32}$$

Similar relations hold for the symmetric and normal cases as discussed in Example 5.1 of [20]. Note that one gets another, but still efficient estimator of θ, if one formulates the constraint in a nonlinear way. Choosing e.g. $S(\theta)=\theta_2/\theta_1-c$, we arrive by Theorem 2.3.1 at $\theta_n^*=(\mu_n^*,\sigma_n^*)^T$, where straightforward computations with $\bar{c}_n = \bar{\sigma}_n / \bar{\mu}_n$ yield

$$\mu_n^* = \left(I_{11} + 2\bar{c}_n I_{12} + \bar{c}_n^2 I_{22}\right)^{-1} \left[\left(I_{11} + \{2\bar{c}_n - c\} I_{12}\right) \bar{\mu}_n + \left(I_{12} + \{2\bar{c}_n - c\} I_{22}\right) \bar{\sigma}_n\right] \tag{2.3.33}$$

and

$$\sigma_n^* = \left(I_{11} + 2\bar{c}_n I_{12} + \bar{c}_n^2 I_{22}\right)^{-1} \left[\left(c I_{11} + c\bar{c}_n I_{12}\right) \bar{\mu}_n + \left(\bar{c}_n I_{12} + \bar{c}_n^2 I_{22}\right) \bar{\sigma}_n\right]. \tag{2.3.34}$$

Indeed, this estimator is asymptotically equivalent to the one from (2.3.31), but the corresponding coefficient of variation does not equal c. The projection from (2.3.20) of $\theta_n^* = (\mu_n^*, \sigma_n^*)^T$ yields $\tilde{\theta}_n = (\tilde{\mu}_n, c\tilde{\mu}_n)^T$ with

$$\tilde{\mu}_n = \left(I_{11} + 2\bar{c}_n I_{12} + \bar{c}_n^2 I_{22}\right)^{-1} \left[\left(I_{11} + \tfrac{2\bar{c}_n - c + c^2 \bar{c}_n}{1+c^2} I_{12}\right) \bar{\mu}_n + \left(\tfrac{1 + c\bar{c}_n}{1+c^2} I_{12} + \tfrac{2\bar{c}_n - c + c\bar{c}_n^2}{1+c^2} I_{22}\right) \bar{\sigma}_n\right], \tag{2.3.35}$$

which is asymptotically equivalent to $\hat{\mu}_n$, but differs from it.

Example 2.3.2. Exchangeable Gaussian copula model

Let

$$\mathbf{X}_1 = (X_{1,1}, \dots, X_{1,m})^T, \dots, \mathbf{X}_n = (X_{n,1}, \dots, X_{n,m})^T$$

be i.i.d. copies of $\mathbf{X} = (X_1, \dots, X_m)^T$. For $i = 1, \dots, m$, the marginal distribution function of X_i is continuous and will be denoted by F_i. It is assumed that $(\Phi^{-1}(F_1(X_1)), \dots, \Phi^{-1}(F_m(X_m)))^T$ has an m-dimensional normal distribution with mean 0 and positive definite correlation matrix $C(\theta)$, where Φ denotes the one-dimensional standard normal distribution function. Here the parameter of interest θ is the vector in $\mathbb{R}^{m(m-1)/2}$ that summarizes all correlation coefficients ρ_{rs}, $1 \le r < s \le m$. We will set this general Gaussian copula model as our semiparametric starting model $\mathcal{P}$, i.e.,

$$\mathcal{P} = \{P_{\theta,G} \ : \ \theta = (\rho_{12}, \dots, \rho_{(m-1)m})^T \ , G = (F_1(\cdot), \dots, F_m(\cdot)) \in \mathcal{G}\}. \tag{2.3.36}$$

As argued in [20] the Van der Waerden or normal scores rank correlation coefficient

$$\hat{\rho}_{rs}^{(n)} = \frac{\frac{1}{n} \sum_{j=1}^{n} \Phi^{-1}\left(\frac{n}{n+1} \mathbb{F}_r^{(n)}(X_{j,r})\right) \Phi^{-1}\left(\frac{n}{n+1} \mathbb{F}_s^{(n)}(X_{j,s})\right)}{\frac{1}{n} \sum_{j=1}^{n} \left[\Phi^{-1}\left(\frac{j}{n+1}\right)\right]^2} \tag{2.3.37}$$

with $\mathbb{F}_r^{(n)}$ and $\mathbb{F}_s^{(n)}$ being the marginal empirical distributions of F_r and F_s, respectively, $1 \leq r < s \leq m$, is a semiparametrically efficient estimator of ρ_{rs} with efficient influence function

$$\begin{aligned}\tilde{\ell}_{\rho_{rs}}(X_r, X_s) = \Phi^{-1}\left(F_r(X_r)\right)\Phi^{-1}\left(F_s(X_s)\right) \\ - \tfrac{1}{2}\rho_{rs}\left\{\left[\Phi^{-1}\left(F_r(X_r)\right)\right]^2 + \left[\Phi^{-1}\left(F_s(X_s)\right)\right]^2\right\}.\end{aligned} \tag{2.3.38}$$

This means that

$$\hat{\theta}_n = (\hat{\theta}_{n1}, \ldots, \hat{\theta}_{nk})^T = (\hat{\rho}_{12}^{(n)}, \ldots, \hat{\rho}_{(m-1)m}^{(n)})^T, \quad k = m(m-1)/2, \tag{2.3.39}$$

efficiently estimates θ within $\mathcal{P}$ with efficient influence function

$$\tilde{\ell}(\mathbf{X}; \theta, G, \mathcal{P}) = (\tilde{\ell}_{\rho_{12}}(X_1, X_2), \ldots, \tilde{\ell}_{\rho_{(m-1)m}}(X_{m-1}, X_m))^T. \tag{2.3.40}$$

The submodel

$$\mathcal{Q} = \{P_{\theta,G} \,:\, \theta = \mathbf{1}_k\rho,\ \rho \in (-1/(m-1), 1),\ G \in \mathcal{G}\} \subset \mathcal{P} \tag{2.3.41}$$

with $\mathbf{1}_k$ indicating the vector of ones of dimension k is the exchangeable m-variate Gaussian copula model. In this submodel all correlation coefficients have the same value ρ.

With J_k the $k \times k$ identity matrix we choose $R = J_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T$ and $\alpha = 0$ in Remark 2.3.3 and obtain

$$\tilde{\theta}_n = \mathbf{1}_k\bar{\theta}_n = \mathbf{1}_k\bar{\rho}_n, \quad \bar{\theta}_n = \bar{\rho}_n = \frac{1}{k}\sum_{r=1}^{m-1}\sum_{s=r+1}^{m}\hat{\rho}_{rs}^{(n)} = \frac{1}{k}\sum_{j=1}^{k}\hat{\theta}_{nj}, \tag{2.3.42}$$

as efficient estimator of θ within submodel $\mathcal{Q}$.

Example 2.3.3. Partial spline linear regression

As in Example 5.3 of [20] the observations are realizations of i.i.d. copies of the random vector $X = (Y, Z^T, U^T)^T$ with Y, Z, and U 1-dimensional, k-dimensional, and p-dimensional random vectors with the structure

$$Y = \theta^T Z + \psi(U) + \varepsilon, \tag{2.3.43}$$

where the measurement error ε is independent of Z and U, has mean 0, finite variance, and finite Fisher information for location, and where $\psi(\cdot)$ is a real valued function on $\mathbb{R}^p$. The distribution function of Z, U, and ε and the function $\psi(\cdot)$ together constitute the nuisance parameter G whereas θ is the parameter of interest. Schick [36] presents an efficient estimator of θ and a consistent estimator of $I(\theta, G, \mathcal{P})$ in his Theorem 8.1. Consequently our Theorem 2.3.1 may be applied directly in order to obtain an efficient estimator of θ in appropriate submodels $\mathcal{Q}$ without our construction of an estimator of $I(\theta, G, \mathcal{P})$ via characteristic functions. In the linear case of Remark 2.3.3 the parameter of interest θ within the submodel $\mathcal{Q}$ may be reparametrized by $\theta = \alpha + L\nu$ with the vector α and the matrix L known. Now ν is the parameter of interest and we return to the situation of (2.3.43) with $X = (Y - \alpha^T Z, Z^T L, U^T)^T$.

Example 2.3.4. Multivariate normal with common mean

Let $\mathcal{G}$ be the collection of nonsingular $k \times k$-covariance matrices and let the parametric starting model be the collection of nondegenerate normal distributions with mean vector θ and covariance matrix Σ,

$$\mathcal{P} = \left\{ P_{\theta,\Sigma} \,:\, \theta \in \mathbb{R}^k,\ \Sigma \in \mathcal{G} \right\}. \tag{2.3.44}$$

Efficient estimators of θ and Σ are the sample mean $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ and the sample covariance matrix $\hat{\Sigma}_n = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)^T$, respectively. Note that $\bar{X}_n$ attains the finite sample Cramér-Rao bound and the asymptotic information bound with $I(\theta, \Sigma, \mathcal{P}) = \Sigma^{-1}$.

The parametric submodel we consider is

$$\mathcal{Q} = \left\{ P_{\theta,\Sigma} \,:\, \theta \in \mathbb{R}^k,\ \theta = \mathbf{1}_k \frac{1}{k} \sum_{j=1}^k \theta_j,\ \Sigma \in \mathcal{G} \right\}, \tag{2.3.45}$$

in which all marginals of each distribution have the same mean. In view of (2.3.28) with $L = \mathbf{1}_k$

$$\tilde{\theta}_n = \mathbf{1}_k \left(\mathbf{1}_k^T \hat{\mathbf{\Sigma}}_n^{-1} \mathbf{1}_k \right)^{-1} \mathbf{1}_k^T \hat{\mathbf{\Sigma}}_n^{-1} \bar{\mathbf{X}}_{\mathbf{n}} \tag{2.3.46}$$

is an efficient estimator of θ within $\mathcal{Q}$, which attains the asymptotic information bound $\left(\mathbf{1}_k^T \Sigma^{-1} \mathbf{1}_k\right)^{-1} \mathbf{1}_k \mathbf{1}_k^T$. See also Example 5.4 of [20].

Example 2.3.5. Restricted maximum likelihood estimator

Maximum likelihood estimation of the generalized linear model under linear restrictions on the parameters is done in [27] via an iterative procedure using a penalty function. Kim and Taylor [28] introduce the restricted EM algorithm for maximum likelihood estimation under linear restrictions. Jamshidian [45] compares the performance of the gradient projection and of the expectation-restricted-maximization (ERM) method under linear restrictions. Our approach as described in Remark 2.3.3 with $\hat{\theta}_n$ a(n unrestricted) maximum likelihood estimator avoids such iterative procedures, provided $\hat{\theta}_n$ can be computed without iterations. Moreover, Theorem 2.3.1 is not constrained to linear restrictions.

2.3.6 Conclusion

In this paper, we have shown that the efficient influence function for estimation of θ within the semiparametric model

$$\mathcal{Q} = \{P_{\theta,G} \ : \ S(\theta) = 0, \theta \in \Theta, G \in \mathcal{G}\}$$

can be obtained by projecting the efficient influence function for estimation of θ within the unconstrained model

$$\mathcal{P} = \{P_{\theta,G} \ : \ \theta \in \Theta, \ G \in \mathcal{G}\}.$$

It follows that these influence functions are related by

$$\begin{aligned}&\tilde{\ell}(\theta, G, \mathcal{Q}) \\ &\quad = \left(J - I^{-1}(\theta, G, \mathcal{P})\dot{S}(\theta)^T \left(\dot{S}(\theta)I^{-1}(\theta, G, \mathcal{P})\dot{S}(\theta)^T\right)^{-1} \dot{S}(\theta)\right) \tilde{\ell}(\theta, G, \mathcal{P})\end{aligned}$$

and hence the corresponding efficient lower bounds by

$$\begin{aligned}I^{-1}(\theta, G, \mathcal{Q}) &= I^{-1}(\theta, G, \mathcal{P}) \\ &\quad -I^{-1}(\theta, G, \mathcal{P})\dot{S}(\theta)^T \left(\dot{S}(\theta)I^{-1}(\theta, G, \mathcal{P})\dot{S}(\theta)^T\right)^{-1} \dot{S}(\theta)I^{-1}(\theta, G, \mathcal{P}).\end{aligned}$$

Furthermore, Theorem 2.3.1 provides a simple method to upgrade an asymptotically efficient estimator for θ within the unconstrained model to an efficient estimator within the constrained model.

2.3.7 Additional Proof: Existence of a Path with a Given Direction

Given a continuously differentiable function $S : \Theta \subset \mathbb{R}^k \mapsto \mathbb{R}^d$ with $k > d$. Define

$$\mathcal{M} = \{\theta \in \Theta : S(\theta) = 0\}$$

and let $\theta_0 \in \mathcal{M}$ be such that the Jacobian of the function S at θ_0, say $\dot{S}(\theta_0)$, has full-rank d. Suppose that $r \in \mathbb{R}^k$ with $\dot{S}(\theta_0)r = 0$. We would like to construct a path through θ_0 with direction r.

Note that according to the Implicit Function Theorem, there exists an open subset $U \subset \mathbb{R}^{k-d}$, $0 \in U$, and a unique continuously differentiable function $\phi : U \to \mathcal{M}$ with $\phi(0) = \theta_0$ (usually, called parametrization). If $\dot{\phi}_0$ denotes the Jacobian of the function ϕ at 0, then the chain rule gives

$$\dot{S}(\theta_0)\dot{\phi}_0 = 0$$

in view of $S(\phi(u)) = 0$ for every $u \in U$. This implies

$$im(\dot{\phi}_0) \subset \dot{S}(\theta_0)^\perp.$$

Since $dim(im(\dot{\phi}_0)) = k - d = dim(\dot{S}(\theta_0)^\perp)$ we obtain

$$im(\dot{\phi}_0) = \dot{S}(\theta_0)^\perp.$$

Consequently, the direction r has to belong to $im(\dot{\phi}_0)$, which means that there exists a $\nu \in U$ with $\dot{\phi}_0\nu = r$. Now define a path

$$\{\theta_\eta\} = \{\phi(\eta\nu) \ : \ \eta \in \mathbb{R}, \ |\eta| < \varepsilon\}$$

for sufficiently small $\varepsilon > 0$, which obviously passes through θ_0 because of $\phi(0) = \theta_0$. Then, we have

$$\begin{aligned} |\theta_\eta - \theta_0 - \eta r| &= |\phi(\eta\nu) - \phi(0) - \eta r| \\ &\le \left|\eta\dot{\phi}_0\nu - \eta r\right| + o(|\eta|) \\ &= o(|\eta|). \end{aligned}$$

2.4 Conclusion of Chapter 2

In this chapter, we have proposed efficient estimators for (semi)parametric models whenever the parameter of interest is constrained, by updating an estimator that is efficient within the model without constraints.

In Section 2.2, we considered the case where the Euclidean parameter of interest is determined by a lower dimensional parameter via a continuously differentiable function. An efficient estimator of this lower dimensional parameter can be defined in terms of the restriction function and the efficient estimator of the original estimator. It has also been shown that this estimator attains the efficient lower bound, which is obtained via the Hájek-LeCam Convolution Theorem for regular parametric models and without projection techniques. In the case where the parameter of interest has to satisfy a functional equality constraint, we have proposed a construction of an efficient estimator of the parameter of interest without reparametrization in Section 2.3. This construction is based on the original parameter and the function defining the constraint. Unlike in the first case, we have derived the efficient lower bound by a projection technique here.

Part II

Application

Chapter 3

Semiparametric Copula-Based Score Level Fusion

3.1 Chapter Introduction

PURPOSE. This chapter presents the use of copula models to handle dependence between matchers in score level fusion. The use of copula models is aimed at improving on the fusion method that assumes independence between matchers.

CONTENTS. Section 3.2 explains a mathematical framework for building a semiparametric likelihood ratio-based fusion and for estimating the underlying parameters. The corresponding semiparametric model is made by modeling the marginal individual likelihood ratios nonparametrically and the dependence between them by parametric copulas. Some applications in real biometric scenarios are given briefly by computing the individual likelihood ratios via the PAV algorithm, by modeling dependence via some well-known parametric families of copulas, and by demonstrating how the best copula pair for genuine and impostor scores is obtained. The special case when the dependence between matchers for genuine and impostor scores is modeled by Gaussian copulas with the same correlation matrices and the individual likelihood ratios are computed by kernel density estimation (KDE), is presented in Section 3.3 and evaluated for standard biometric verification scenarios. A more general method, which is also based on copula models, is provided in Section3.4 for standard biometric verification scenarios as well by setting the false acceptance rate in advance. The Jeffreys' credible interval for comparing the proposed method to the simple fusion method, which assumes independence between matchers, is also provided in Section 3.4.

PUBLICATIONS. The manuscript presented in Section 3.3 has been published in [50], and Section 3.4 has been accepted for publication at BTAS 2016.

NOTES. The reader may focus on the following subsections:

(1) 3.2.4 explains how the LR is computed via copula models;
(2) 3.2.5.2 presents estimation of parameters for computing the LR;
(3) 3.2.6 demonstrates the choice of the best copula pair in some scenarios;
(4) 3.4.3 discusses Jefreys' method for comparing two fusion strategies.

Subsection 3.4.4 gives the mathematical background that has already been explained in Subsections 3.2.4 and 3.2.5.2. We also note that the manuscript presented in Section 3.3 was published before the manuscripts from Sections 3.2 and 3.4 had been written. Therefore, the individual likelihood ratios in Section 3.3 were computed by the KDE method instead of the PAV algorithm as used in Sections 3.2 and 3.4.

3.2 Semiparametric Likelihood-ratio-based Score Level Fusion via Parametric Copulas

3.2.1 Abstract

We present a mathematical framework for modelling dependence between matchers 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 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. We also compare the proposed method to Gaussian mixture model (GMM) and linear logistic (Logit) methods, which are also designed to handle dependence.

3.2.2 Introduction

In a biometric verification system, biometric samples (images of faces, fingerprints, voices, gaits, etc.) of people are compared and classifiers (matchers) indicate the level of similarity between any pair of samples by a 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 [3]. A common performance measure for this scenario is the *cost of log likelihood ratio* that can be decomposed into discrimination and calibration performance [8].

When our biometric system has two or more classifiers, one has to transform these multiple scores to a new score (a scalar) as a *fused 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 [15] and (2) it reflects evidential value in forensic individualization [3]. By assuming independency between classifiers, the fused LR is only 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 advantageous:

1. The fused score is an LR.
2. It can deal with dependent scores.

This paper uses the copula concept to handle dependence between classifiers. Although the copula model is already used in [50–53] 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 3.2.3, this paper will explain how a copula model splits the LR computation for two or more classifiers into a product of the individual likelihood ratio and a correction factor in Section 3.2.4. Section 3.2.5 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 3.2.6. Finally, our conclusions are presented in Section

3.2.7.

3.2.3 Score Level Fusion

There are three categories in score level fusion: transformation-based [13], classifier-based [14], and density-based (also called *likelihood-ratio-based*) [16]. The transformation-based fusion is done by mapping all components of the vector of comparison scores to a comparable 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 forensic scenario since 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 [15]. Moreover, the fused score, as an LR value, can be used for forensic evidence evaluation [3] in forensic individualization as a multiplicative factor for the information before analyzing the evidence (*prior odds*) to get the new information after taking the evidence into account (*posterior odds*) via Bayesian framework

$$O_{\text{posterior}} = \text{LR} \times O_{\text{prior}}. \tag{3.2.1}$$

Since these distributions are unknown in practice then the performance does depends on the accuracy of the LR computation.

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 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.). The parametric model is usually used because of its simplicity and nonparametric model is chosen because of its flexibility. However, the main problem in using parametric models is the difficulty in choosing the appropriate model whereas nonparametric estimators have sensitivity in the choice of the bandwidth or other smoothing parameters, especially for our multivariate case. A common parametric model to compute the density 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 [54,55], the same problem in choosing an appropriate model will also appear. On the other hand, a nonparametric approach called *Pool Adjacent Violators* (PAV) method, which seems promising because of its optimality in transforming score to its LR value [8], is only applicable for 1-dimensional score which means that it cannot be used to compute the LR for fusion.

Many studies of score level fusion assume independency between classifiers; see [55–57]. However, the independency assumption is not realistic since the scores are obtained from the same sample. To incorporate the dependency 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.

3.2.4 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

$$\text{LR}(s_1, \cdots, s_d) = \frac{f_{\text{gen}}(s_1, \cdots, s_d)}{f_{\text{imp}}(s_1, \cdots, s_d)}. \tag{3.2.2}$$

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

A d-variate copula is a distribution function on the unit cube $[0, 1]^d$, of which the marginals are uniformly distributed. Sklar [18] shows the existence of copula for any multivariate distribution functions f.

Theorem 3.2.1 (Sklar (1959)). *Let $d \geq 2$, and suppose H is a distribution function on $\mathbb{R}^d$ with one dimensional continuous marginal distribution functions $F_1, \cdots, F_d$. Then there is a unique copula C so that*

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)) \tag{3.2.3}$$

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

By taking the d-th derivative of (3.2.3), we will get the joint density function

$$\begin{aligned} h(x_1,\dots,x_d) = c(F_1(x_1),\dots,F_d(x_d)) \\ \times \prod_{i=1}^{d} f_i(x_i) \end{aligned} \tag{3.2.4}$$

where c is the copula density and f_i is the i-th marginal density for every $i = 1, \cdots, d$. This implies that (3.2.2) can be written as

$$\begin{aligned} \mathrm{LR}(s_1,\cdots,s_d) = \frac{c_{\mathrm{gen}}(F_{\mathrm{gen},1}(s_1),\cdots,F_{\mathrm{gen},d}(s_{\mathrm{d}}))}{c_{\mathrm{imp}}(F_{\mathrm{imp},1}(s_1),\cdots,F_{\mathrm{imp},d}(s_{\mathrm{d}}))} \\ \times \prod_{i=1}^{d} \frac{f_{\mathrm{gen},i}(s_i)}{f_{\mathrm{imp},i}(s_i)} \end{aligned} \tag{3.2.5}$$

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 (3.2.5), which is the product of the individual likelihood ratios

$$\mathrm{PLR}(\mathbf{s}) = \prod_{i=1}^{d} \frac{f_{\mathrm{gen},i}(s_i)}{f_{\mathrm{imp},i}(s_i)} = \prod_{i=1}^{d} \mathrm{LR}_i(s_i), \tag{3.2.6}$$

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

$$\mathrm{CF}(\mathbf{s})^{(C_{\mathrm{gen}},C_{\mathrm{imp}})} = \frac{c_{\mathrm{gen}}(F_{\mathrm{gen},1}(s_1),\cdots,F_{\mathrm{gen},d}(s_{\mathrm{d}}))}{c_{\mathrm{imp}}(F_{\mathrm{imp},1}(s_1),\cdots,F_{\mathrm{imp},d}(s_{\mathrm{d}}))} \tag{3.2.7}$$

will be called the *correction factor* where the superscript $(C_{\mathrm{gen}}, C_{\mathrm{imp}})$ means that the CF is modelled by copulas C_{gen} and C_{imp} for genuine and impostor scores, respectively. We call (3.2.7) as correction factor because it corrects the likelihood ratio computation under independence assumption.

3.2.5 Semiparametric Model for Likelihood Ratio Computation

The LR as defined in (3.2.5) 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 that 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 (3.2.5).

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

$$\mathbf{W}_1, \ldots, \mathbf{W}_{n_{\text{gen}}} \tag{3.2.8}$$

and

$$\mathbf{B}_1, \ldots, \mathbf{B}_{n_{\text{imp}}} \tag{3.2.9}$$

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 variable of genuine and impostor scores are continuous.

3.2.5.1 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 [4] for a brief explanation of these methods. In this paper, we choose the PAV method because of its optimality [57].

For every $k = 1, \cdots, d$, PAV sorts and assigns a posterior probability of 1 and 0 to the k-th component of genuine and impostor scores, respectively. It then finds the non-monotonic adjacent group of probabilities and replaces it with 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 (3.2.8) and (3.2.9) 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 (3.2.8) and (3.2.9) can be computed according to the Bayesian formula by

$$\widehat{\mathrm{LR}}_k(\cdot) = \frac{P(H_1|(\cdot))}{1 - P(H_1|(\cdot))} \times \frac{n_{\mathrm{imp}}}{n_{\mathrm{gen}}} \tag{3.2.10}$$

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

3.2.5.2 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 $\theta = \begin{pmatrix}\theta_{\mathrm{gen}} \\ \theta_{\mathrm{imp}}\end{pmatrix}$. Thus our correction factor model is defined by

$$\mathcal{CF} = \{\mathrm{CF}_{\theta,F}^{(C_{\mathrm{gen}},C_{\mathrm{imp}})} \ : \ \theta \in \Theta, \ F \in \mathcal{F}\} \tag{3.2.11}$$

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

$$F = (F_{\mathrm{gen},1}, \cdots, F_{\mathrm{gen},d}, F_{\mathrm{imp},1}, \cdots, F_{\mathrm{imp},d}). \tag{3.2.12}$$

Note that if the marginal distributions $F_{\mathrm{gen},k}$ and $F_{\mathrm{imp},k}$ for $k = 1, \ldots, d$ are known then the log-likelihood of the combined samples (3.2.8) and (3.2.9) can be written as

$$L = \sum_{i=1}^{n_{\mathrm{gen}}} \log c_{\theta_{\mathrm{gen}}}(\mathbf{U}_{\mathrm{gen},i}) + \sum_{j=1}^{n_{\mathrm{imp}}} \log c_{\theta_{\mathrm{imp}}}(\mathbf{U}_{\mathrm{imp},j}) \tag{3.2.13}$$

where

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

and

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

for $i = 1, \ldots, n_{\text{gen}}$ and $j = 1, \ldots, n_{\text{imp}}$. Differentiating (3.2.13) with respect to θ gives

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

and

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

As a consequence, if $F_{\text{gen},k}$ and $F_{\text{imp},k}$ are replaced by their *modified* 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 estimator* (PMLE) of θ_{gen} and θ_{imp}, respectively, as studied in [58] and extended in [19]. Our modified empirical distribution function based on a sample $X_1, \ldots, X_n$ is defined by

$$\hat{F}_n(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}, \quad \forall x \in \mathbb{R}. \tag{3.2.14}$$

Under some regularity conditions, we can derive the convergence of

$$\hat{\theta}_n = \begin{pmatrix} \hat{\theta}_{\text{gen},n_{\text{gen}}} \\ \hat{\theta}_{\text{imp},n_{\text{imp}}} \end{pmatrix} \tag{3.2.15}$$

in the following theorem.

Theorem 3.2.2. *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 conditions C2-C4 and assumption A1-A5 explained in Section 3 of Chen and Fan [19] then*

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \to \mathcal{N}(0, \Sigma) \tag{3.2.16}$$

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

Proof. Proof is given in the Appendix 3.2.8. □

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.

3.2.5.3 Choosing the best copula pair

Note that the LR at score $\mathbf{s} = (s_1, \cdots, s_{\mathrm{d}})$ under correction factor model (3.2.11) can be computed by a *rule of thumb*:

$$\mathrm{LR}^{(C_{\mathrm{gen}}, C_{\mathrm{imp}})}(\mathbf{s}) = \prod_{k=1}^{d} \widehat{\mathrm{LR}}_k(s_k) \times \mathrm{CF}_{\hat{\theta}_n, \hat{F}_n}^{(C_{\mathrm{gen}}, C_{\mathrm{imp}})}(\mathbf{s}). \tag{3.2.17}$$

Here $\widehat{\mathrm{LR}}_k(\cdot)$ and $\hat{\theta}_n$ are given by (3.2.10) and (3.2.15), respectively, while $\hat{F}_n$ is the modified empirical version of (3.2.12), 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 3.2.2 is still valid although the copula pair $(C_{\mathrm{gen}}, C_{\mathrm{imp}})$ misspecified the true pair. It means that we will still get a reasonable estimator of the dependence parameter whenever a copula pair is chosen.

As noted in Section 3.2.2, 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 [59] will only give the copula pair that is closest to the pair $(c_{\mathrm{gen}}, c_{\mathrm{imp}})$, but whose ratio is not necessarily closest to the ratio $c_{\mathrm{gen}}/c_{\mathrm{imp}}$, 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

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

as given in (3.2.17). 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 independence copula in $\mathcal{C}$ to guarantee that the chosen fused classifier performs at least as good as fusion under independence assumption.

3.2.6 Applications

We will present how our correction factor works and improves the PLR fusion in some practical scenarios: maximizing TMR at certain FMR, minimizing half total error rate (HTER), and minimizing discrimination loss. The first two scenarios are usually used in a standard biometric verification while the last one is used 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 [52, 53].

Suppose that we have genuine and impostor scores as given in (3.2.8) and (3.2.9), 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

Table 3.1: 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

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 [55] for the Logit method while the parameters in the GMM method are fitted by the algorithm proposed in [60], 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, 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 3.1.

3.2.6.1 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* (FMR(Δ)) with threshold Δ, while the probability of rejecting a genuine score is called the *False Non-Match*

Rate (FNMR(Δ)). The complement of the FNMR(Δ) is called the *True Match Rate* (TMR(Δ)), 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 TMR(Δ) $= 1-$FNMR(Δ). The most common method to see the performance of a biometric person verification system is by plotting the relation between FMR(Δ) and TMR(Δ) for all $\Delta \in (-\infty, \infty)$, which is known as *Receiver Operating Characteristic* (ROC) [1].

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

$$W_1, \cdots, W_{n_{\text{gen}}} \tag{3.2.18}$$

$$B_1, \cdots, B_{n_{\text{imp}}} \tag{3.2.19}$$

be 1-dimensional genuine and impostor scores, respectively. In our case, these are the fused scores of the testing set. According to [2], the TMR$_\alpha$ 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 (3.2.18) and (3.2.19), respectively while $\hat{Q}_{\hat{F}^{-}_{\text{imp}}}$ is the empirical quantile function with respect to $\hat{F}^{-}_{\text{imp}}$. Slightly different from (3.2.14), the left-continuous empirical distribution function based on a sample $X_1, \cdots, X_n$ is defined by

$$\hat{F}^{-}_{n}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i < x\}}, \quad \forall x \in \mathbb{R} \tag{3.2.20}$$

and its corresponding quantile function is defined by

$$\hat{Q}_{\hat{F}^{-}_{n}}(p) = \sup\{y \ : \ \hat{F}^{-}_{n}(y) \leq p\}, \quad \forall p \in [0,1]. \tag{3.2.21}$$

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 [61] and Face-3D [62, 63] data to simulate fingerprint and face authentication, respectively.

- 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

Table 3.2: 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}

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 [62, 63] 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 3.2 and Table 3.3, 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 3.1, 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.

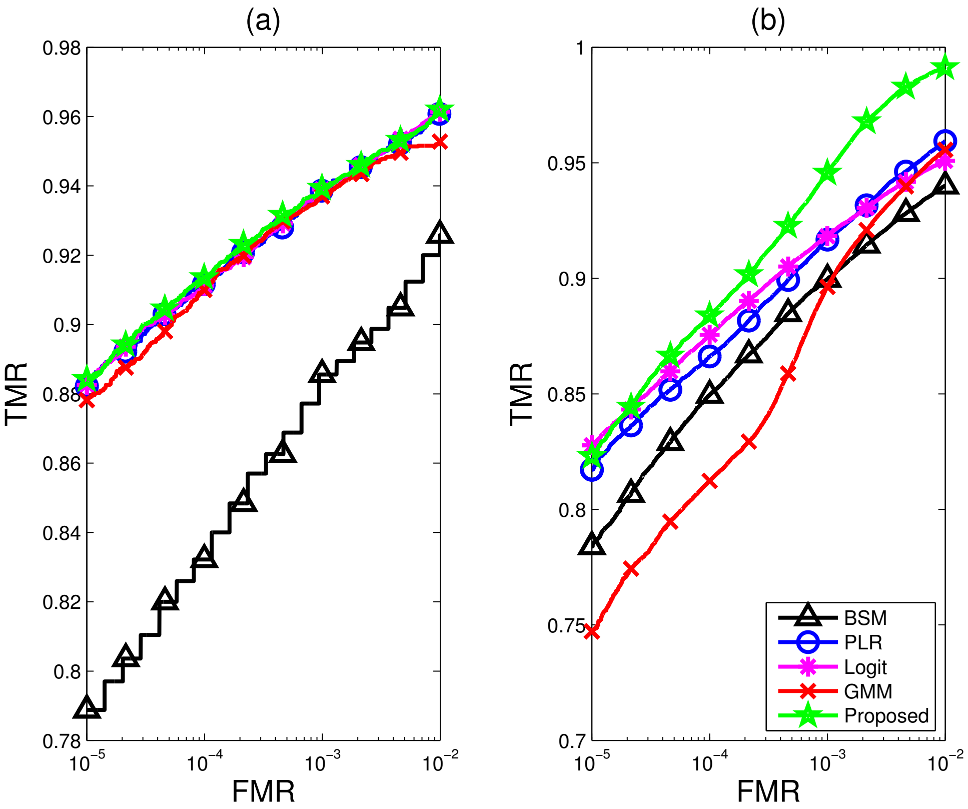


Figure 3.1: ROC curves of different fusion strategies on (a)NIST-finger (b)face-3D

Table 3.3: 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**

3.2.6.2 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 [17,56]. However, as pointed out in [64], 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 [17] 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 [65].

Table 3.4: 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, Best pair	**0.034** {Gu,fCl}	**0.047** {Gu,GC}	0.035 {t,t}	**0.131** {Gu,Cl}

- 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 [65].
- 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 [65].
- BSS3: This database is composed of two-dimensional scores from a wearable accelerometer and a gyroscope sensor [66].

Results: The HTERs of different fusion strategies are reported in Table 3.4. We do not present the performance of other fusion strategies used in [17] 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 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.

3.2.6.3 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 [3].

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 [8] 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 [67]. This measure is also used in forensic face scenarios in [12]. Note that the scores are interpretable as likelihood ratios when computing this measure. Given 1-dimensional genuine scores (3.2.18), which correspond to the hypothesis of the prosecution, and impostor scores (3.2.19), which correspond to the hypothesis of the defense, the cost of log likelihood ratio C_{llr} is defined by

$$\begin{aligned} C_{\text{llr}} = \frac{1}{2n_{\text{gen}}} \sum_{i=1}^{n_{\text{gen}}} \log_2 \left(1 + \frac{1}{W_i}\right) \\ + \frac{1}{2n_{\text{imp}}} \sum_{j=1}^{n_{\text{imp}}} \log_2 \left(1 + B_j\right). \end{aligned} \tag{3.2.22}$$

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 (3.2.22), 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*

$$C_{\text{llr}}^{\text{cal}} = C_{\text{llr}} - C_{\text{llr}}^{\text{min}}. \tag{3.2.23}$$

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 [64].

- Face-3D: The same database as used for the standard verification in Section 3.2.6.1.

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 3.5. 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 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 3.2.6.1, its performance is also worse than the best single classifier. It 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 3.2.6.1 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 our all fusion strategies (including our method) is pretty high on the Face-3D database as seen in Figure 3.2. In order to reduce this calibration loss, we proposed in our previous work [53] a method called *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 [53].

3.2.7 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

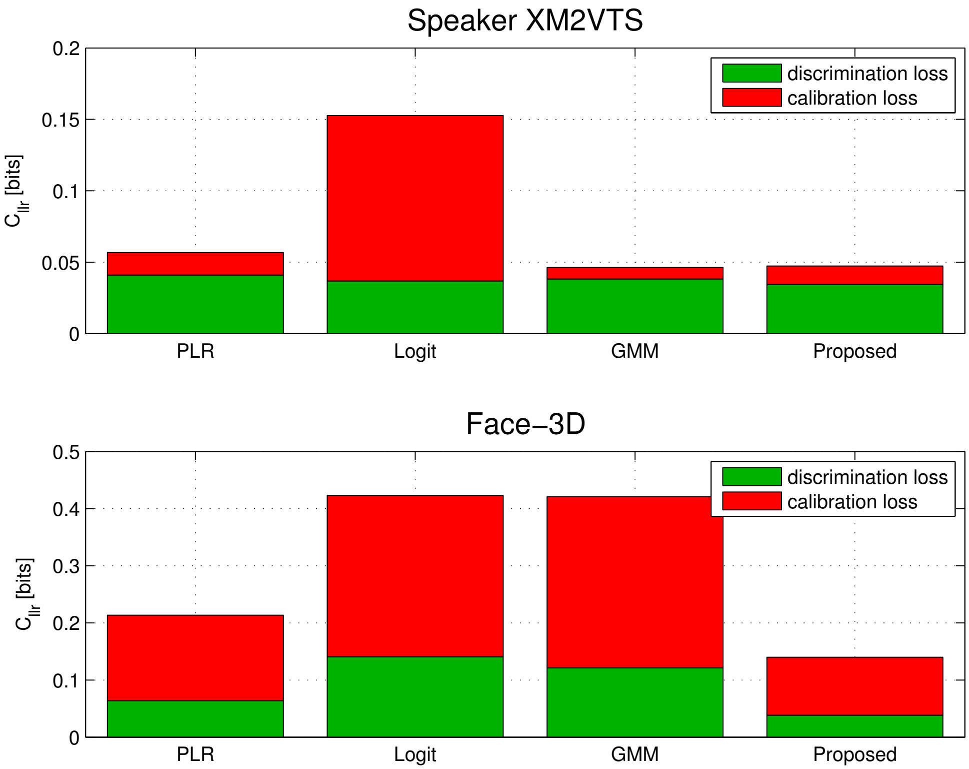


Figure 3.2: The discrimination and calibration loss of different fusion strategies on the XM2VTS and Face-3D databases

Table 3.5: 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}	

is chosen. 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.

3.2.8 Appendix: Proof of Theorem 3.2.2

According to Proposition 2 of Chen and Fan [19], 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

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

□

3.3 Semiparametric Score Level Fusion: Gaussian Copula Approach

3.3.1 Abstract

Score level fusion is an appealing method for combining multi-algorithms, multi-representations, and multi-modality biometrics due to its simplicity. Often, scores are assumed to be independent, but even for dependent scores, according to the Neyman-Pearson lemma, the likelihood ratio is the optimal score level fusion if the underlying distributions are known. However, in reality, the distributions have to be estimated. The common approaches are using parametric and nonparametric models. The disadvantage of the parametric method is that sometimes it is very difficult to choose the appropriate underlying distribution, while the nonparametric method is computationally expensive when the dimensionality increases. Therefore, it is natural to relax the distributional assumption and make the computation cheaper using a semiparametric approach.

In this paper, we will discuss the semiparametric score level fusion using Gaussian copula. The theory how this method improves the recognition performance of the individual systems is presented and the performance using synthetic data will be shown. We also apply our fusion method to some public biometric databases (NIST and XM2VTS) and compare the thus obtained recognition performance with that of several common score level fusion rules such as sum, weighted sum, logistic regression, and Gaussian Mixture Model.

3.3.2 Introduction

Multi-biometric system or biometric fusion is a combination of several biometric systems or algorithms in order to enhance the performance of the individual system or algorithm. In general, it can be characterized into six categories [68]: multi-sensor, multi-algorithm, multi-instance, multi-sample, multi-modal and hybrid. Several studies [68–71] show that combining information from multiple traits or algorithms can provide better performance. For example, Lu et al. [69] combining three different feature extractions (Principle Component Analysis, Independent Component Analysis and Linear Discriminant Analysis) which is related to the multi-algorithm biometric fusion. In the fingerprint biometric field, Prabhakar and Jain [72] use the left and right index fingers to verify an individual's identity which is an example of the multi-instance biometric fusion.

Biometric fusion can be done at the sensor, feature, match score, rank and decision levels either for verification or identification. In this paper, we will focus on the match score level for person verification. This means that scores from multiple biometric matchers for every pair of two subjects (user and enrollment) are transformed to a new score (a scalar) as a combined score. Once the new score has been generated, one has to decide whether the user and enrollment are from the same person or not. To do this, a threshold has to be set such that a score greater than or equal to the threshold is recognized as *genuine score* which means that the user and enrollment are the same subject while a score less than the threshold will lead to the conclusion that the user and enrollment are different people which will be called by *impostor score.* This threshold is determined using a set which is called the *training set* and is evaluated using a disjoint set which is called the *testing set*

There are three categories in biometric fusion: transformation-based [13], classifier-based [14], and density-based. The last category would be optimal if the underlying densities were known. However, in practice, such densities have to be estimated from the training set so that the performance relies on how well these two densities are estimated. The parametric models suffers from the limitation in choosing the appropriate parametric model to the data. The most successful parametric approach is the Gaussian Mixture Model (GMM) [16]. However, the number of the mixture components which is the most important part in estimating GMM is very hard to be determined. The author in his paper used GMM fitting algorithm proposed in [60] that automatically estimates the number of the mixture components using an EM algorithm and the minimum message length criterion. However, the computational cost is time consuming when the sample size is big or the the number of mixture components increases. On the other hands, the nonparametric models have a problem in choosing bandwidth and computational cost when working in the multidimensional space.

This paper focuses on the fusion strategy for dependent matchers. Using synthetic data, we will show that our approach is robust in handling the dependent classifiers even with an extremely high dependence structure. We will also apply our method on the public databases NIST-BSSR1 and XM2VTS. The rest of this paper is organized as follows. In Section 3.3.3, we will review the theory of Gaussian copula, why it is suitable to be chosen and how to do Gaussian copula based fusion. Some experimental results on the synthetic data are presented in Section 3.3.4 to show the robustness of our method in handling the dependence issues and the results on the public database will be provided to show the applicability of our method in the real world. Finally, this paper will be closed by our conclusions.

3.3.3 Gaussian Copula Fusion

3.3.3.1 Likelihood ratio based fusion

Suppose we have d matchers and let $\mathbf{X} = (X_1, \cdots, X_d)$ denote the d components of the matching(similarity or distance) scores where X_i is the random variable corresponding to the i-th match score where $\mathbf{X}$ takes its values in $\Omega \subset \mathbb{R}^d$. The decision function is a map $\psi : \mathbb{R}^d \mapsto \{0, 1\}$ where 0 and 1 corresponds to negative and positive decisions which are denoted by H_0 and H_1, respectively. A system can make two types of error(false): accepting an impostor score or rejecting a genuine score. The probability of accepting impostor score $P(\psi(\mathbf{X}) = 1|H_0)$ is called by *False Acceptance Rate (FAR)* while the probability of rejecting genuine score $P(\psi(\mathbf{X}) = 0|H_1)$ is called by *False Rejection Rate (FRR).* From the definition of FRR, it can be understood that the probability of accepting genuine score that will be called by *True Positive Rate (TPR)* is TPR = 1 − FFR. In application, the FAR has to be set very small since the cost of accepting an impostor may be much more expensive than the cost of rejecting a genuine user. For example, in security, allowing a forbidden person to access a secret place is much more dangerous that rejecting a "nice" person to access it. Therefore, for every given FAR, our fusion has to maximize the TPR.

Neyman and Pearson established the most powerful test based on the likelihood ratio [15]. Let f_{gen} and f_{imp} be the density of genuine and impostor scores, respectively. The likelihood ratio at a point $\mathbf{x} = (x_1, \cdots, x_d)$ is defined by

$$\text{LR}(\mathbf{x}) = \frac{f_{\text{gen}}(\mathbf{x})}{f_{\text{imp}}(\mathbf{x})}. \tag{3.3.1}$$

According to the Neyman-Pearson theorem, in order to get the maximum TPR for every fixed FAR, say α, we have to decide

$$\psi(X) = 1 \iff \text{LR}(\mathbf{x}) \geq \eta \tag{3.3.2}$$

where η is implicitly defined by

$$P(\text{LR}(\mathbf{X}) \geq \eta) = \alpha. \tag{3.3.3}$$

As a consequence, the optimal performance can be reached by defining the fused score as the likelihood ratio of the vector consisting of all matching scores.

3.3.3.2 Gaussian copula

Computing (3.3.1) means that the estimation of f_{gen} and f_{imp} is a must. Let H be any distribution function on $\mathbb{R}^d$ with density h. A classical result of Sklar [18] shows that H can be uniquely factorized into its univariate marginal distributions and a distribution function on the unit cube $[0,1]^d$ in $\mathbb{R}^d$ with uniform marginal distributions which is called by *copula:*

Theorem 3.3.1 (Sklar (1959)). *Let $d \geq 2$ and suppose H is a distribution function on $\mathbb{R}^d$ with one dimensional continuous marginal distribution functions $F_1, \cdots, F_d$. Then there is a unique copula C so that*

$$H(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)) \;\; \forall (x_1, \dots, x_d) \in \mathbb{R}^d. \tag{3.3.4}$$

This paper assumes that C is determined by a multivariate normal distribution with standard normal marginals and correlation matrix R. Note that this assumption is more flexible than assuming H to be multivariate normally distributed. The main difference is that each marginal of the multivariate normal has to be normally distributed while each marginal of a Gaussian copula can be any continuous distribution function. In section 3.3.4, we will see that our generated data follow a Gaussian copula distribution with normal and weibull marginal.

The key concept of the Gaussian copula is the assumption of the existence of a componentwise transformation $\tau : \mathbb{R}^d \mapsto \mathbb{R}^d$ such that $\tau(\mathbf{X}) \sim N(0, R)$. Here, each component τ_i of τ is a monotone continuous function. One can show that

$$\tau_i(x_i) = \Phi^{-1}(H_i(x_i)) \tag{3.3.5}$$

for $i = 1, \dots, d$ where Φ and H_i denote the standard normal distribution function and the marginal distribution of the $i-$th component.

This means that (3.3.4) can be rewritten as

$$H(x_1, \dots, x_d) = \Phi_R(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)), \tag{3.3.6}$$

where $u_i = F(x_i)$, Φ the one-dimensional standard normal distribution function, and Φ_R the d-dimensional standard normal distribution function with correlation matrix R. Consequently, the density function of H is

$$h(x_1, \dots, x_d) = \frac{1}{|R|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{u}^T (R^{-1} - I)\mathbf{u}\right) \prod_{i=1}^{d} f_i(x_i), \tag{3.3.7}$$

with $\mathbf{u} = (\Phi^{-1}(F_1(x_1)), \cdots, \Phi^{-1}(F_d(x_d)))^T$.

3.3.3.3 Gaussian copula based fusion

Our fused score using the Gaussian copula approach is defined by (3.3.1) with the numerator f_{imp} and the denominator f_{gen} as in (3.3.7), i.e.,

$$\text{LR}(x_1, \ldots, x_d) = \frac{|R_{\text{imp}}|^{1/2} \times \exp\left(-\frac{1}{2}\mathbf{u}_{\text{gen}}{}^T(R_{\text{gen}}^{-1} - I)\mathbf{u}_{\text{gen}}\right) \times \prod_{i=1}^d f_{gen,i}(x_i)}{|R_{\text{gen}}|^{1/2} \times \exp\left(-\frac{1}{2}\mathbf{u}_{\text{imp}}{}^T(R_{\text{imp}}^{-1} - I)\mathbf{u}_{\text{imp}}\right) \times \prod_{i=1}^d f_{imp,i}(x_i)}. \tag{3.3.8}$$

Here, R_{gen} and R_{imp} denote the correlation matrices of transformed genuine and impostor scores, respectively, $\mathbf{u}_{\text{gen}}$ and $\mathbf{u}_{\text{imp}}$ are given by

$$\mathbf{u}_{\text{gen}} = (\Phi^{-1}(F_{gen,1}(x_1)), \cdots, \Phi^{-1}(F_{gen,d}(x_d)))^T$$

and

$$\mathbf{u}_{\text{imp}} = (\Phi^{-1}(F_{imp,1}(x_1)), \cdots, \Phi^{-1}(F_{imp,d}(x_d)))^T,$$

respectively. To obtain the LR value as given by (3.3.8), we need to estimate the correlation matrices $R_{\text{gen}}(R_{\text{imp}})$, the marginal densities $f_{gen,i}(f_{imp,i})$ and marginal distribution functions $F_{gen,i}(F_{imp,i})$ using a training set. Given a training set, we can extract to the genuine and impostor scores. Note that the scores often are dependent within the group of genuine scores, within the group of impostor scores, and between these two groups. However, we shall proceed as if all scores are independent. The resulting estimators are still reliable because most scores will be independent.

Let $\mathbf{W}_1, \ldots, \mathbf{W}_{n_{\text{gen}}}$ and $\mathbf{B}_1, \ldots, \mathbf{B}_{n_{\text{imp}}}$ be the two samples representing the genuine and impostor scores, respectively.

Matchers dependence As stated above, some genuine and impostor scores are dependent. However, we are interested in the correlation matrices of the match scores, which we will assume to be the same, $R_{\text{gen}} = R_{\text{imp}} = R$. We shall estimate R using the combined sample, i.e.,

$$(\mathbf{X}_1, \ldots, \mathbf{X}_n) = (\mathbf{W}_1, \ldots, \mathbf{W}_{n_{\text{gen}}}, \mathbf{B}_1, \ldots, \mathbf{B}_{n_{\text{imp}}})$$

with $n = n_{\text{gen}} + n_{\text{imp}}$. Our experiments show that such restriction will improve the performance of the fused score. It is reasonable since we are estimating the matchers dependence not only the genuine or impostor scores dependence.

Klaasen and Wellner [73] give an explicit formula to obtain an optimal estimator for the correlation matrix R via normal rank correlation by taking $\hat{R} = \left(\hat{\rho}_{rs}^{(n)}\right)$ where

$$\hat{\rho}_{rs}^{(n)} = \frac{\frac{1}{n}\sum_{j=1}^{n} \Phi^{-1}\left(\frac{n}{n+1}\mathbb{F}_r^{(n)}(X_{rj})\right) \Phi^{-1}\left(\frac{n}{n+1}\mathbb{F}_s^{(n)}(X_{sj})\right)}{\frac{1}{n}\sum_{j=1}^{n}\left[\Phi^{-1}\left(\frac{j}{n+1}\right)\right]^2} \tag{3.3.9}$$

where Φ denotes the one-dimensional standard normal distribution function while $\mathbb{F}_r^{(n)}$ and $\mathbb{F}_s^{(n)}$ are the marginal empirical distributions of F_r and F_s, respectively, is an efficient estimator for ρ_{rs} for every $1 \leq r < s \leq d$.

Marginal density estimation To estimate the marginal density functions, we use the kernel bandwidth optimization as studied by Shimazaki and Shinomoto [74]. This method has two different kinds of choosing the optimal bandwidth. The first bandwidth choice is similar with the regular bandwidth selection but it performs much faster than the built-in *ksdensity* matlab. The second one is a local bandwidth optimization. This approach works very well in handling the data that have "spikes".

Marginal distribution function estimation The empirical distribution function is an optimal estimator for the marginal distribution function and very easy to be implemented and very fast to be computed (see Figure 3.3 for an example in biometric). The empirical distribution function, $\hat{F}$, is the distribution function that puts mass $1/n$ at each data point x_i where n is the number of the observation. In this paper, since we need to compute the quantile of the standard normal, then to avoid singularity, we prefer to put mass $1/(n+1)$. Explicitly, the empirical distribution function of genuine and impostor scores are given by

$$\hat{F}_{\text{gen}}(x) = \frac{1}{n_{\text{gen}}+1}\sum_{i=1}^{n_{\text{gen}}}(1)_{[W_i \leq x]} \text{ and } \hat{F}_{\text{imp}}(x) = \frac{1}{n_{\text{imp}}+1}\sum_{i=1}^{n_{\text{imp}}}(1)_{[B_i \leq x]}. \tag{3.3.10}$$

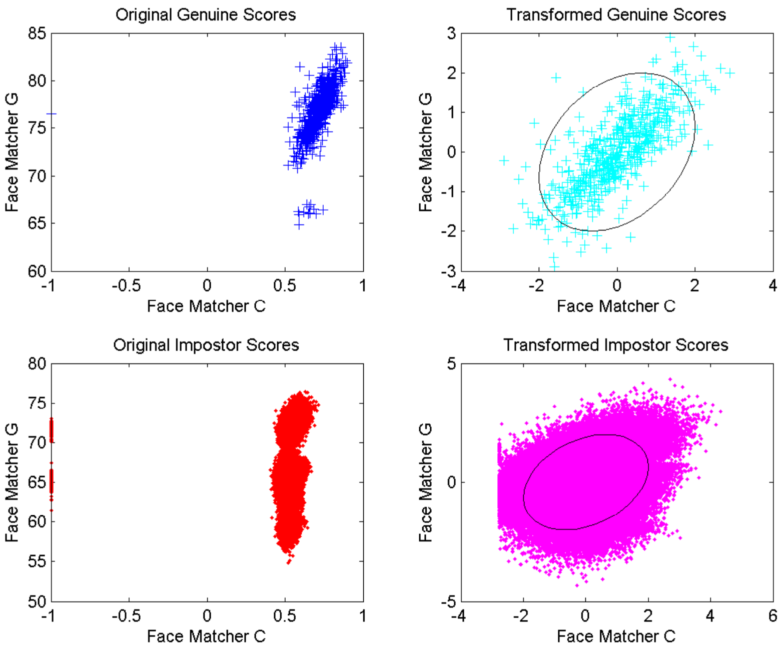


Figure 3.3: Two face matchers scores from NIST-Multimodal

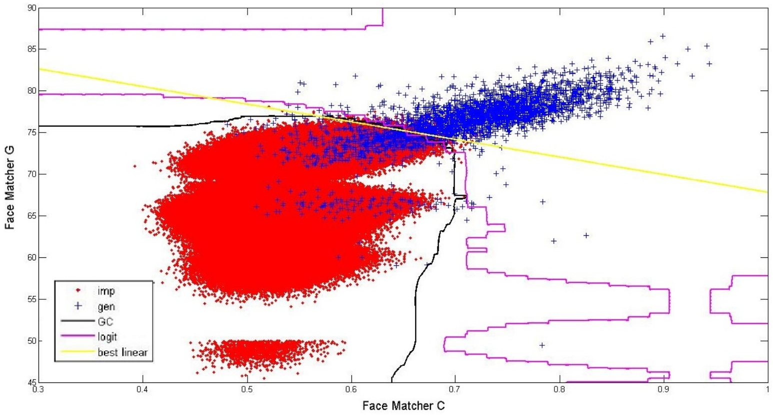


Figure 3.4: Boundary decisions at 0.01% FAR

3.3.4 Experimental Results

To study the robustness of our method in fusing biometrics scores related to the classifiers dependence, genuine and impostor scores are generated that follow three different distribution functions and have three different dependence levels. Here, we assume that there are 1000 subjects with 2 biometric specimens for each subject, one is put as user and the other for enrollment. We also assume that we have 2 different biometric systems. Therefore, the size of genuine and impostor scores are 2×1000 and 2×9999000, respectively which we will use as training data. The testing data are obtained in the same way. The parameters for generating the data are:

- multivariate normal scores with correlations $0.99, 0.5$ and 0.1 with genuine means $[1,3]^T$, $[5,3]^T$ and $[5,3]^T$, respectively. All impostor means are set to be $[0,0]^T$.
- Gaussian copula with correlation value $0.9, 0.5$ and 0.1. The genuine and impostor marginals of the first matcher are set to follow weibull distribution with the shape parameters 3 and 1, respectively, and the common scale parameter 4. For the second matcher, the genuine and impostor marginals follow normal distribution with parameter $(5,1)$ and $(1,0)$, respectively.

Once all data have been generated, for every pair of training and testing set, the exact likelihood ratio is computed which is called by true fusion. The next step is performing the sum rule with min-max and z-norm normalization and also the weighted sum using Fisher criterion [75]. Subsequently, we pick the best results. For the Logit fusion, we use nonlinear logistic regression as given by W. Chen and Y. Chen [76]. The performance of several methods compared with the true fusion is provided in Table 3.6. The bold value is the best non-true fusion which indicated the TPR (%) at 0.01% FAR. We can see that our method is the most robust approach especially for the data with high dependence.

Table 3.6: Influence of Dependence in Biometric Fusion

Methods	**High**			**Moderate**			**Low**		
	MV	GC	Gu	MV	GC	Gu	MV	GC	Gu
True Fusion	90.70	93.20	99.90	91.00	90.70	97.40	96.90	90.90	84.70
Best Linear	89.80	90.40	94.00	**91.00**	90.20	90.90	**96.90**	89.90	83.50
Logistic Regression	00.10	88.20	87.60	90.60	90.50	87.40	**96.90**	**90.80**	82.80
Gaussian Copula	**90.10**	**92.80**	**99.70**	89.80	**90.70**	**93.50**	96.50	90.60	**84.70**

*MV: Multivariate Normal, GC: Gaussian Copula, Gu: Gumbel Copula.

We will also apply our method on the public databases: NIST-BSSR1 [61] and XM2VTS [64]. The NIST-BSSR1 database has three different set:

- NIST-Multimodal: Two fingerprints and Two face matchers applied to 517 subjects,
- NIST-Face: Two face matchers applied to 3000 subjects,
- NIST-Finger: Two fingerprints applied to 6000 subjects.

For every experiment, each set is split up randomly into two subsets, one is used for training and the other is used for testing. Then the naive sum rule with min-max normalization, naive sum with Z-normalization, weighted sum with Fisher criterion, nonlinear logistic regression, and our method are performed and the TPR at 0.01% is computed for every fusion strategy. This procedure is repeated 20 times and the average of all TPR at 0.01% for each fusion strategy is provided in the Table 3.7. We do not include the Gaussian Mixture Model (GMM) fusion strategy because the computation is very time consuming when it is done on a normal computer. However, we also provide the result of the GMM strategy as reported in [16] and we compare the 95% Confidence Interval on increase in TPR at 0.01% as given by Table 3.8. We can see that our approach outperforms all other fusion strategies (the bold value is the best one) even with GMM fusion which is computationally expensive. Also for the XM2VTS database that contains match scores from five face matchers and three speech matchers applied to 295 subjects with the partition of the training and testing set have been defined in [64], our method is the highest among all reported TPR at 0.01% FAR.

Table 3.7: TPR (%) values for different methods at 0.01% FAR on the public databases

Method	**NIST Multi modal**	**NIST Face**	**NIST Finger print**	**XM2VTS**
Naive Sum min-max	97.97	76.47	91.33	97.50
Naive Sum Z-norm	97.87	76.48	91.33	97.50
Weighted Sum	97.97	76.48	91.40	97.50
Logistic Regression	98.74	76.48	91.46	98.50
Gaussian Mixture Model [16]	99.10	**77.20**	91.40	98.70
This paper	**99.48**	**77.21**	**91.60**	**99.00**

Table 3.8: Comparison with LR fusion using Gaussian Mixture Model on the NIST-BSSR1 database

Database	**Mean TPR (%) at 0.01% FAR**			**95% Confidence Interval on increase in TPR (%) at 0.01% FAR**	
	BSM	GMM	GC	GMM	GC
NIST-Multimodal	85.30	99.10	**99.48**	[13.50,14.00]	**[13.51,14.84]**
NIST-Face	71.20	**77.20**	**77.21**	[4.70, 7.30]	[4.69, 7.32]
NIST-Fingerprint	83.50	91.40	**91.60**	[7.60, 8.20]	**[7.63, 8.57]**

*BSM: Best Single Matcher, GMM: Gaussian Mixture Model, GC: Gaussian Copula (used in this paper).

3.3.5 Conclusion

The Gaussian copula is a semiparametric model which is easy to be implemented, cheap in computation, and able to handle the dependence structure that usually appears in multi-algorithm fusion. Using several synthetic data, we have shown that our approach performs very well in dependent classifiers fusion even for extreme dependence structures when the performance of other approaches drops dramatically. We also see that our method works well when it is applied on the NIST-BSSR1 database (see Figure 3.4 for the comparison of the boundary decision with another approaches on this database) and even on the XM2VTS it reaches the highest TPR at 0.01% FAR among all reported results. However, it has limitations in estimating the tail density because estimation is based on the kernel density method. Our experiments show that although our approach works well at 0.01% FAR, it is sometimes much worse than individual classifiers at 0.001% FAR.

3.4 Fixed FAR Correction Factor of Score Level Fusion

3.4.1 Abstract

In biometric score level fusion, the scores are often assumed to be independent to simplify the fusion algorithm. In some cases, the "average" performance under this independence assumption is surprisingly successful, even competing with a fusion that incorporates dependence. We present two main contributions in score level fusion: (i) proposing a new method of measuring the performance of a fusion strategy at fixed FAR via Jeffreys credible interval analysis and (ii) subsequently providing a method to improve the fusion strategy under the independence assumption by taking the dependence into account via parametric copulas, which we call fixed FAR fusion. Using synthetic data, we will show that one should take the dependence into account even for scores with a low dependence level. Finally, we test our method on some public databases (FVC2002, NIST-face, and Face3D), compare it to Gaussian mixture model and linear logistic methods, which are also designed to handle dependence, and notice its significance improvement with respect to our evaluation method.

3.4.2 Introduction

In a score based biometric person verification system, a *threshold* has to be set to decide whether a matching score between two biometric samples (query and template) is a *genuine* or an *impostor* score. A genuine score leads to the conclusion that the query and template originate from the same person while an impostor score means that the query and template stem from different people. We will assume that the matching score is a similarity score. Note that once the threshold is set, the 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 Acceptance Rate (FAR)*, while the probability of rejecting a genuine score is called the *False Rejection Rate (FRR)*. The complement of the FRR is called the *True Positive Rate (TPR)*, 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 TPR $= 1 -$ FRR. The most common method to evaluate a biometric person verification system is by plotting the relation between FAR and TPR, which is known as *Receiver Operating Characteristics* (ROC).

When there are two or more matchers, one has to transform these multiple scores to a new score (a scalar) as a fused score, which is called score level fusion. There are three categories in score level fusion. The most commonly used one is the transformation-based one which is done by mapping all components of the vector of matching scores to a comparable domain and applying some simple rules such as sum, mean, max, med, etc. [13]. However, this approach relies heavily on the niceness of the training set used for the transformation. For example if one wants to normalize each component of the vector of matching scores to the unit interval [0,1] (which is called minmax normalization), then the maximum and the minimum of all scores have to be determined.

Unfortunately, when the maximum and minimum scores have to be estimated from the training set that has outlier(s), the estimation will be very bad. The second approach is classifier-based fusion which is done by stacking all components of the vector of matching scores and applying a classifier to separate the genuine and impostor scores [14]. The last approach is based on estimation of the densities of the genuine and impostor scores [16]. According to [15] this approach, which is also known as likelihood ratio based, would be optimal if the underlying densities were known. However, in practice, such densities have to be estimated from data so that the performance relies on how well the two densities are estimated.

In this paper, we will focus on score level fusion for dependent matchers. The likelihood ratio based fusion automatically incorporates the dependence between matchers. However, this approach needs to estimate two density functions, which is a challenging task. While the choice of an appropriate parametric model is sometimes difficult, nonparametric estimators suffer from the difficulty that they are sensitive to the choice of the bandwidth or of other smoothing parameters. To simplify, many researchers assume that all genuine and impostor scores are independent so that the likelihood ratio is only the product of the individual likelihood ratios of the matchers (henceforth called PLR fusion); see [55–57]. However, the independence assumption is not realistic since the scores are obtained from the same sample. A study of incorporating dependence instead of using PLR fusion is presented in [77] where the authors investigate the effect of considering correlation and compare their method to PLR fusion by computing the difference between the areas their respective ROCs. However, in practice the FAR has to be set in advance. For example, in a security application, the FAR is set to be very small and usually less than 0.1% or even 0.01%. Since area under ROC does not always reflect the performance at small FAR, we will compare the performance between dependent and PLR fusion at specific FAR.

This paper has two main contributions: proposing an evaluation of biometric fusion at fixed FAR and proposing a method to improve PLR fusion. In Section 3.4.3, we present our method to evaluate biometric fusion at fixed FAR. Instead of using parametric or nonparametric models, we propose a semiparametric approach, which will be called *fixed FAR fusion*, by modeling the marginal densities nonparametrically and the dependence between them by parametric copulas as explained in Section 3.4.4. We will see the gain of considering dependence using synthetic data and subsequently compare our method to GMM [16] and Logit [55] fusions, which are also intended to deal with matcher dependence, on some real databases (FVC2002, NIST-face, Face3D) in Section

3.4.5. Although also vector machine (SVM) fusion can handle dependence, we do not include it because it is a classifier tool so that we cannot set the FAR value beforehand (the FAR value of SVM fusion is automatically determined by the classifier). Finally, our conclusions are presented in Section 3.4.6.

3.4.3 Performance of biometric fusion at fixed FAR

Suppose we have d matchers. In biometric fusion, one has to find a function $\psi : \mathbb{R}^d \to \mathbb{R}$, which will be called a *fusion.* Let

$$\mathbf{W}_1, \ldots, \mathbf{W}_{n_{\text{gen}}} \tag{3.4.1}$$

and

$$\mathbf{B}_1, \ldots, \mathbf{B}_{n_{\text{imp}}} \tag{3.4.2}$$

be i.i.d copies of the d-dimensional random variable of genuine scores $\mathbf{S}_{\text{gen}}$ and impostor scores $\mathbf{S}_{\text{imp}}$, respectively. In this section, we will present how to measure the performance of a fusion at fixed FAR.

Let α be a fixed FAR. The exact TPR is

$$\text{TPR} = P(\psi(\mathbf{S}_{\text{gen}}) \geq \tau) \tag{3.4.3}$$

where the threshold τ is explicitly determined via relation

$$P(\psi(\mathbf{S}_{\text{imp}}) \geq \tau) = \alpha. \tag{3.4.4}$$

This means that all fused scores greater than or equal to τ will be recognized as genuine scores. In practice, we do not know the distribution functions of $\mathbf{S}_{\text{gen}}$ and $\mathbf{S}_{\text{imp}}$. However, we can compute the empirical value of TPR based on (3.4.1) and (3.4.2) by

$$\widehat{\text{TPR}} = \widehat{F}^{\psi}_{\text{gen}}(\hat{\tau}). \tag{3.4.5}$$

where

$$\hat{\tau} = \inf\{x \ : \ \widehat{F}^{\psi}_{\text{imp}}(x) \geq 1 - \alpha\}. \tag{3.4.6}$$

Here, $\widehat{F}^{\psi}_{\text{gen}}$ and $\widehat{F}^{\psi}_{\text{imp}}$ are *modified* empirical distribution functions based on the two samples

$$\psi(\mathbf{W}_1), \ldots, \psi(\mathbf{W}_{n_{\text{gen}}})$$

and

$$\psi(\mathbf{B}_1), \ldots, \psi(\mathbf{B}_{n_{\text{imp}}}),$$

respectively. Our modified empirical distribution function based on a sample $X_1, \ldots, X_n$ is defined by

$$\hat{F}(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}, \quad \forall x \in \mathbb{R}. \tag{3.4.7}$$

The $\widehat{\mathrm{TPR}}$ is only an estimated rate, which may be viewed as the probability of a Bernoulli experiment [63]. With n_{gen} genuine scores $\widehat{\mathrm{TPR}}$ has a binomial distribution with success probability TPR, which may be approximated by $\mathrm{Bin}(n_{\mathrm{gen}}, \widehat{\mathrm{TPR}})$. We employ Jeffreys method to construct a credible interval (CI) from this. It is one of the more trusted ways to obtain a CI here [75,78]. In conclusion, for a given significance level $0 < \varepsilon << 1$, we will have the $100(1-\varepsilon)\%$ Jeffreys CI $[L, U]$ where

$$L = B(\varepsilon/2; \beta_1, \beta_2) \tag{3.4.8}$$

and

$$U = B(1-\varepsilon/2; \beta_1, \beta_2) \tag{3.4.9}$$

with

$$\beta_1 = n_{\mathrm{gen}}\widehat{\mathrm{TPR}} + \frac{1}{2} \text{ and } \beta_2 = n_{\mathrm{gen}}(1 - \widehat{\mathrm{TPR}}) + \frac{1}{2}.$$

Here, $B(\varepsilon; p_1, p_2)$ denotes the ε quantile of a $\mathrm{Beta}(p_1, p_2)$ distribution. This means that it is approximately $100(1-\varepsilon)\%$ certain that the true TPR is in-between L and U.

3.4.4 Fixed FAR correction factor

According to the Neyman-Pearson lemma [15], the optimal fusion is the likelihood-ratio-based method, i.e., by taking $\psi = \mathrm{LR}$ where

$$\mathrm{LR}(\mathbf{s}) = \frac{f_{\mathrm{gen}}(\mathbf{s})}{f_{\mathrm{imp}}(\mathbf{s})} \tag{3.4.10}$$

where f_{gen} and f_{imp} are the densities of genuine and impostor scores, respectively, which are unknown in practice. Therefore, we have to estimate the LR from data.

3.4.4.1 Correction factor

A copula is a distribution function on the unit cube $[0,1]^d$, $d \geq 2$, of which the marginals are uniformly distributed. Susyanto et al. [50] use a specific copula called Gaussian copula to handle dependence between classifiers in biometric fusion. However, since the Gaussian copula is appropriate for only a limited number of biometric data sets, we will use a family of well-known parametric copulas from the collection of elliptic and Archimedean copulas.

For any continuous multivariate distribution function there exists a copula function [18].

Theorem 3.4.1 (Sklar (1959)). *Let $d \geq 2$, and suppose H is a distribution function on $\mathbb{R}^d$ with one dimensional continuous marginal distribution functions $F_1, \cdots, F_d$. Then there is a unique copula function $C : [0,1]^d \to [0,1]$, so that*

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)) \tag{3.4.11}$$

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

The joint density function can be computed by taking the d-th derivative of (3.4.11):

$$\begin{aligned} h(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \\ \times \prod_{i=1}^{d} f_i(x_i) \end{aligned} \tag{3.4.12}$$

where c is the copula density and f_i is the i-th marginal density for every $i = 1, \cdots, d$. Note that according to (3.4.12), we can estimate separately the dependence structure represented by the copula density c and the individual densities f_i in order to get the joint density h. If C_α is determined by a finite dimensional Euclidean parameter α then it is called a parametric copula. In this case, we can estimate the dependence parameter α based on i.i.d. observations

$$\mathbf{X}_1, \ldots, \mathbf{X}_n$$

with

$$\mathbf{X}_i = (X_{1i}, \ldots, X_{di}) \quad \forall i = 1, \ldots, n$$

by the pseudo-maximum likelihood estimator (PMLE). Mathematically, the

PMLE of α has to maximize

$$\frac{1}{n}\sum_{i=1}^{n}\log c_\alpha\left(\hat{F}_1(X_{1i}),\ldots,\hat{F}_d(X_{di})\right) \tag{3.4.13}$$

where $\hat{F}_j$ is the modified empirical distribution function as defined in (3.4.7) based on $X_{j1},\ldots,X_{jn}$ for $1 \leq j \leq d$ and c_α is the copula density.

Let C_{gen} and C_{imp} be the copula corresponding to genuine and impostor scores with copula densities c_{gen} and c_{imp}, respectively. In view of (3.4.10) and (3.4.12), the likelihood ratio at score $\mathbf{s} = (s_1,\cdots,s_{\text{d}})$ can be written as

$$\text{LR}(\mathbf{s}) = \text{PLR}(\mathbf{s}) \times \text{CF}(\mathbf{s})$$

where

$$\text{PLR}(\mathbf{s}) = \prod_{i=1}^{d}\text{LR}_i(s_i) \tag{3.4.14}$$

is the product of the individual likelihood ratios and

$$\text{CF}(\mathbf{s}) = \frac{c_{\text{gen}}(F_{\text{gen},1}(s_1),\cdots,F_{\text{gen},d}(s_{\text{d}}))}{c_{\text{imp}}(F_{\text{imp},1}(s_1),\cdots,F_{\text{imp},d}(s_{\text{d}}))} \tag{3.4.15}$$

is the copula density ratio that will be called the *correction factor*. Here, $F_{\text{gen},i}$ and $F_{\text{imp},i}$ denote the distribution functions of genuine and impostor scores, respectively.

Note that for every i-th component of score $\mathbf{s} = (s_1,\cdots,s_{\text{d}})$, the posterior probability $P(H_1|s_i)$ can be estimated optimally by the Pool-Adjacent-Violators (PAV) algorithm as shown in [57] where H_1 correspond to a genuine user. Therefore, from the Bayesian relation

$$\frac{P(H_1|s_i)}{P(H_0|s_i)} = \frac{P(s_i|H_1)}{P(s_i|H_0)} \times \frac{P(H_1)}{P(H_0)}$$

where H_0 corresponds to an impostor user, we can estimate LR_i optimally by

$$\widehat{\text{LR}}_i = \frac{P(H_1|s_i)}{1-P(H_1|s_i)} \times \frac{n_{\text{imp}}}{n_{\text{gen}}} \tag{3.4.16}$$

as used in [8] for calibrating scores in the field of speaker recognition. Therefore, we only need to estimate the correction factor CF.

3.4.4.2 Fixed FAR fusion

Estimating CF can be done by estimating c_{gen} and c_{imp} separately. Of course we will not estimate these copula densities nonparametrically since it will lead to the same problems as when estimating the original density functions directly. We will approximate CF by the following parametric copulas: Gaussian copula (GC), Student's t (t), Frank (Fr), Clayton (Cl), and Gumbel (Gu). We also include the independence copula (ind) to guarantee that our fusion is better than the PLR method. Readers interested in copulas are referred to [79] for a more detailed explanation. To have more dependence models and because the Clayton and Gumbel copulas are not symmetric, their flipped forms (flipped Clayton (fCl) and flipped Gumbel (fGu)) will be included as well (if U has copula C then $1 - U$ has copula flipped C). Therefore, the copulas c_{gen} and c_{imp} are chosen from the copula family

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

Note that the best copula pair must have the best performance among other pairs in the sense that it has the highest TPR at fixed FAR. Applying a goodness-of-fit test as provided in [59] 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 $c_{\text{gen}}/c_{\text{imp}}$. Therefore, we propose to choose the best copula pair directly by maximizing the empirical TPR at the given FAR$= \alpha$ as explained in Section 3.4.3. Given a fixed FAR $= \alpha$, a set $\mathcal{C}$ of n_c candidate copulas and a training set, our fixed FAR fusion is very simple. The first step is computing PLR by the PAV algorithm and multiplying it by each of all copula pairs $\hat{c}_{\text{gen}}/\hat{c}_{\text{imp}}$ in which the dependence parameters have been estimated by the PMLEs as defined in (3.4.13). Of the $n_c \times n_c$ resulting different combined scores we choose the one that maximizes the TPR.

3.4.5 Experimental Results

To study the performance of our fixed FAR fusion in improving the simple PLR method we apply it to synthetic and real databases, which are split up into training and testing sets. Given a training set, we will compute the product of the individual likelihood ratios and select the best copula pair. The corresponding testing set is used for evaluation only. We compare our fixed FAR fusion to the linear Logit fusion explained in [55] and the GMM fusion proposed in [16] at FAR$= 0.01\%$ for all experiments. The Jeffreys CIs of all fusions are computed at significance level 0.01 and the improvement of fusion ψ compared to PLR fusion in TPR at 0.01% FAR is defined by $[L_\psi - U, U_\psi - L]$ where $[L_\psi, U_\psi]$ and $[L, U]$ are the 99% Jeffreys CIs of fusion ψ and PLR fusion, respectively, as explained in Section 3.4.3.

Given genuine and impostor scores

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

and

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

in the training set, our procedure to choose the best copula pair is simple. We randomize the genuine (impostor) scores and take two disjoint subsets with size

$$n_b = \min\{10,000; \lfloor n_{\text{gen}}/2 \rfloor\}$$

and

$$n_w = \min\{10,000; \lfloor n_{\text{imp}}/2 \rfloor\}.$$

This re-sampling method is aimed at increasing the computation speed because it will be repeated 100 times to see the consistency. Once the product of the individual likelihood ratios is computed, it is multiplied by the 64 copula pair estimates $\hat{c}_{\text{gen}}/\hat{c}_{\text{imp}}$. After all 64 combined scores are obtained using the first subset, the empirical TPR at 0.01% FAR is then computed. The final TPR for each copula pair is the average over all 100 experiments. The best copula pair is the pair having the highest average of the TPR values. If there are several pairs having the same averages, we choose the pair with the smallest variance. If there is still more than one pair having the smallest means and variances then we choose one of them at random.

3.4.5.1 Synthetic Data

To get synthetic data that behave like real data, we take two algorithms presented in [63]. The first algorithm measures the similarity of the left half of the face between two images and the second one the similarity of the right half. The density and distribution functions of the genuine and impostor scores for each algorithm are estimated by a mixture of logconcave densities [80]. We choose this estimation method because it is more general than a Gaussian mixture and more robust for handling skewness. To obtain scores with *explicit* dependence that can be represented by a copula C, we generate random samples of the copula C and apply the inverse transform technique, using the estimates of the two marginal distribution functions. In this way the generated scores have as marginal distribution functions these estimates of the distribution functions of data generated by the two algorithms. Recall that if F is a continuous distribution function then U is uniformly distributed if and only if $F^{-1}(U)$ has distribution function F.

In our experiment, we generate 10,000 genuine and 1,000,000 impostor scores in the way as explained above. The dependence is made by putting 4 different copula pairs

$$\{(\mathrm{GC}, \mathrm{GC}), (\mathrm{t}, \mathrm{fCl}), (\mathrm{fGu}, \mathrm{GC}), (\mathrm{Cl}, \mathrm{Gu})\}$$

completed with 9 dependence level pairs obtained from the cross pairs

$$\{\mathrm{low}, \mathrm{moderate}, \mathrm{high}\}.$$

In order to know the effect of dependence in biometric fusion, the low, moderate, and high dependence levels are set to have correlation values 0.1, 0.5, and 0.9 for Gaussian and Student's t copulas while for other copulas we put parameters 1, 10, and 50. Student's t copula has 3 degrees of freedom for all experiments.

By following our procedure, we get that the best copula pair is the true one for every experiment. Then, the fixed FAR fusion is compared to the PLR fusion to see the gain of considering dependence in biometric fusion. Figure 3.5 shows the improvement by the fixed FAR fusion compared to the PLR fusion. We can see that we really have to take the dependence into account when the dependence between the impostor scores is higher than between the genuine ones. Moreover, the dependence between classifiers should be taken into account even for low levels of dependence.

3.4.5.2 FVC2002-DB1 database

This data set [81] consists of 100 fingers with 8 impressions per finger. We will use the same experimental set up as used in [77] by putting the first two impressions as templates and the remaining ones as queries. Two 600 × 100 scores matrices are obtained by matching each query to both the templates using a minutiae matcher [82]. The purpose of this experiment is to see the improvement in using our fixed FAR method for multi-instances scenarios. To have a big enough testing set so that the CIs are not too large, we did 1,000 experiments. In every experiment, we randomized the 100 subjects, and took 70 subjects for training and the remaining 30 for testing. Our fixed FAR

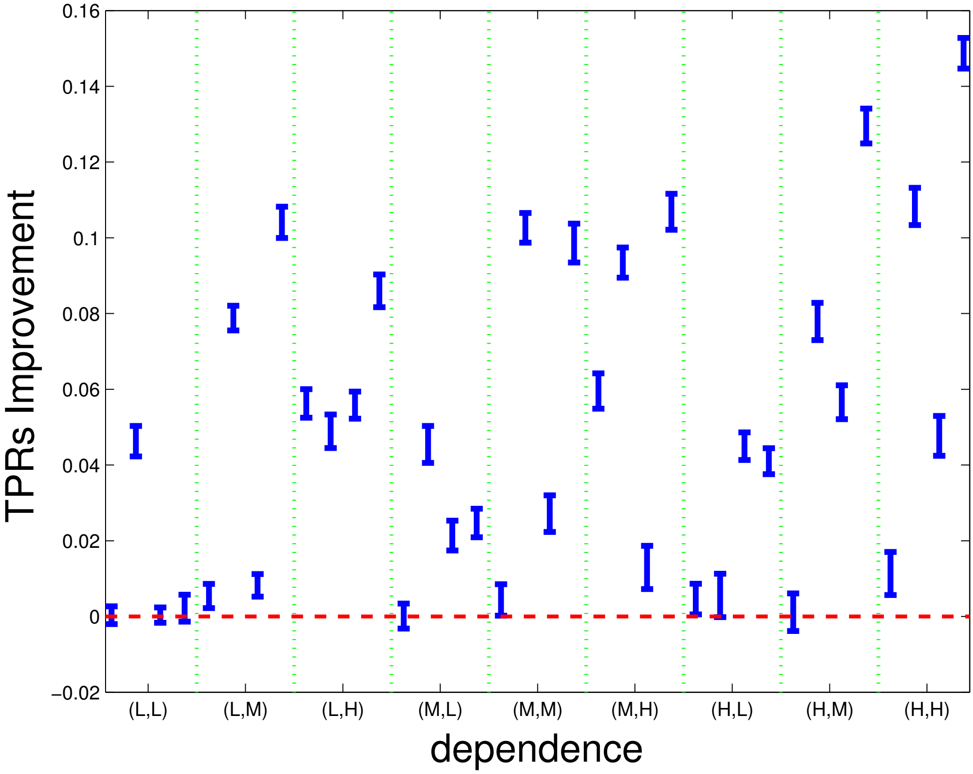


Figure 3.5: Gain of considering dependence between classifiers. The blue thick lines are the 99% Jeffreys CI of fixed FAR fusion compared to PLR fusion. The blue thick lines that do not intersect the red dashed line, mean that the gain of considering dependence is significant. On the x-axis the databases are indicated in 9 groups of 4, each group having the same dependence level pair for each of the 4 chosen copula pairs. Database (L,L) has low and low dependence levels for genuine and impostor scores, (L,M) low and moderate, (L,H) low and high, etc.

Table 3.9: Performances at 0.01% FAR on FVC2002-DB1.

Methods	TPR	99% Jeffreys CI compared to PLR in TPR at 0.01% FAR
BSM	77.5%	N/A
PLR	81.8%	N/A
Logit	81.9%	[−0.4%, 0.6%]
GMM	83.6%	[1.3%, 2.3%]
FFF	**83.9%**	[1.7%, 2.6%]

BSM: Best Single Matcher, GMM: Gaussian Mixture Model, Logit: Logistic Regression, PLR: Product of Likelihood Ratios, FFF: our fixed FAR fusion. The bold number is the best one and the underlined number is the worst one.

and benchmark fusion methods were trained on the first subset and evaluated on the second subset. As a result, each fusion method has 180 genuine and 5,220 impostor scores for every experiment. The average TPR is computed by pooling all genuine scores from the 1,000 experiments in one set and all impostor scores in the other set [1]. Therefore, we have 180,000 genuine and 5,220,000 impostor scores in total.

For every experiment, we train our fixed FAR fusion method by following the procedure explained at the beginning of this section and the pair (ind,fCl) is obtained as the best copula pair. The difference of the area under ROC of our fixed FAR and the PLR fusion is around 0.1%, which is relatively small. At first sight it is consistent with the results in [77] , which claims that considering dependence will not improve the PLR fusion significantly. However, if we highlight the TPR at FAR= 0.01% (see Figure 3.6), we can see that the improvement is significant. Detailed TPR values for our fixed FAR and benchmark fusions are provided in Table 3.9. On this database, our fixed FAR fusion is slightly better than the GMM fusion and both of them improve the PLR fusion at significance level 0.01. On the other hand, the Logit and PLR fusions have almost the same performances.

3.4.5.3 NIST-face database

The NIST-face BSSR1 database is published by the National Institute of Standards and Technology [61]. The data contain similarity scores from two face algorithms run on images from 3,000 subjects with each subject having two probe images and one gallery image. To evaluate the performance of our

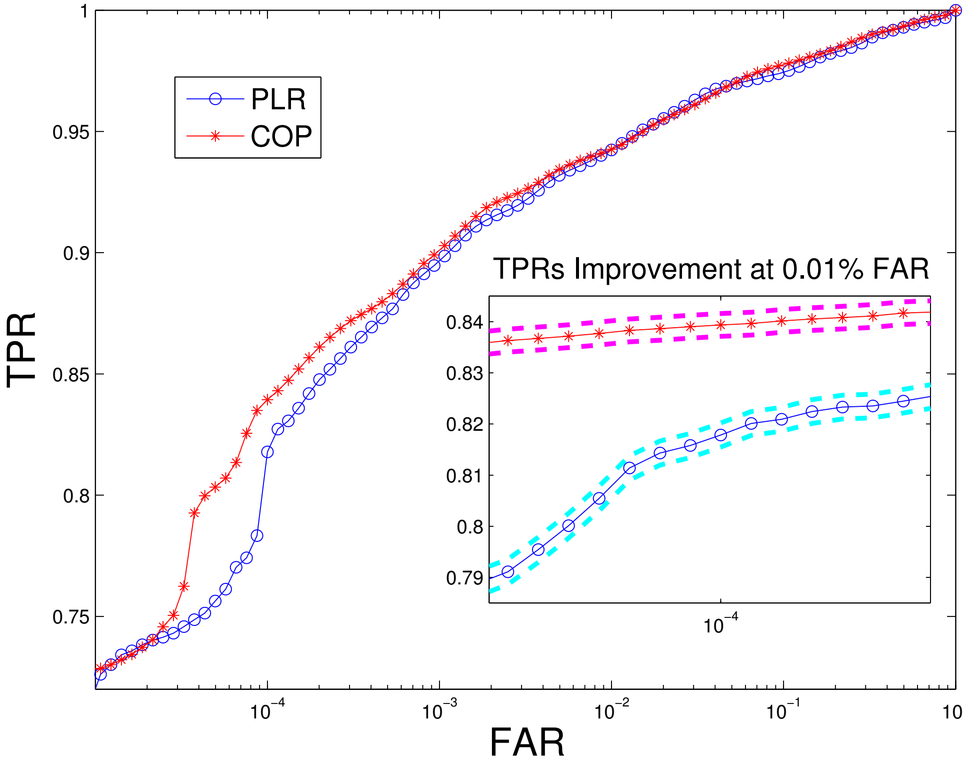


Figure 3.6: Comparison between the PLR and our fixed FAR fusion methods on FVC2002-DB1 database. The small box contains the highlighted performance at around 0.01% FAR. The dashed lines are the 99% Jeffrey CIs.

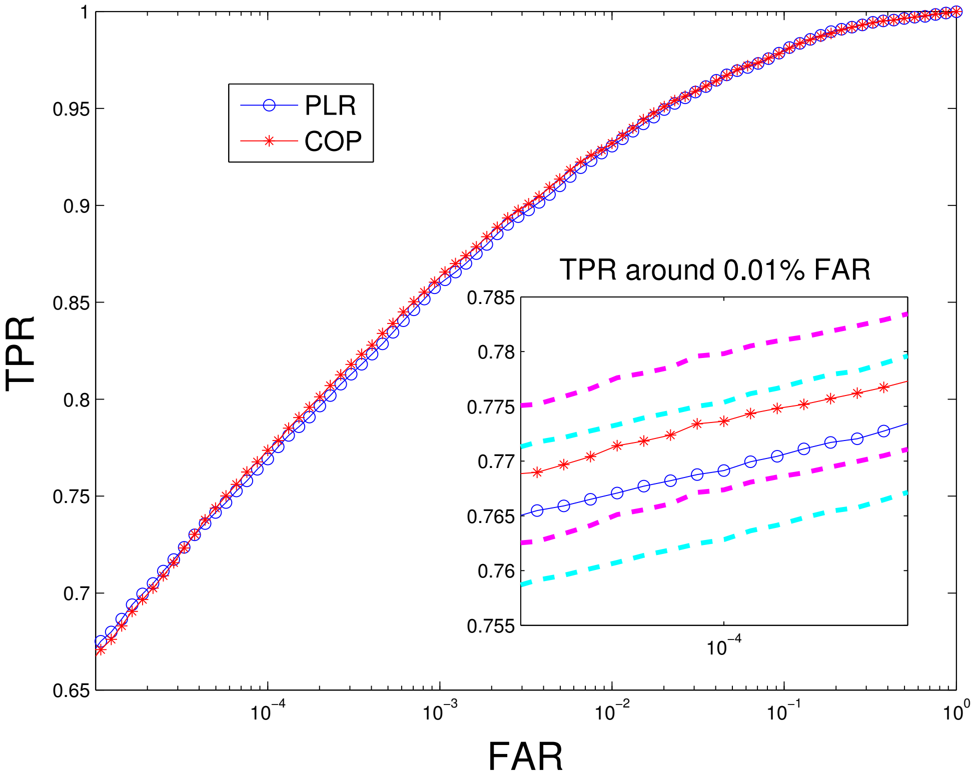


Figure 3.7: Comparison between the PLR and our fixed FAR fusion methods on NIST-face database. The small box contains the highlighted performance at around 0.01% FAR. The dashed lines are the 99% Jeffrey CIs.

benchmark fusion strategies, we randomize the subjects and split the set into two disjoint sets with size 1,500 each. Each fusion strategy is trained on the first subset and evaluated on the second subset. This procedure is repeated 10 times. Then, we collect all genuine scores from all 10 experiments in one set and all impostor scores in another set resulting in 30,000 genuine and 44,970,000 impostor scores.

Figure 3.7 shows that the ROC of our fixed FAR fusion method almost coincides with the ROC of the PLR fusion. Although our fixed FAR fusion has the highest TPR, we should not conclude that it is the best one because all 99% Jeffreys CIs are overlapping (see Table 3.10). This means that on this database, the simple PLR fusion method is comparable to other fusion methods that take dependence into account.

Table 3.10: Performances at 0.01% FAR on NIST-face database.

Methods	TPR	99% Jeffreys CI compared to PLR in TPR at 0.01% FAR
BSM	71.2%	N/A
PLR	76.9%	N/A
Logit	76.1%	[−2.0%, 0.5%]
GMM	76.8%	[−1.4%, 1.1%]
FFF	77.4%	[−0.8%, 1.7%]

BSM: Best Single Matcher, GMM: Gaussian Mixture Model, Logit: Logistic Regression, PLR: Product of Likelihood Ratios, FFF: our fixed FAR fusion.

3.4.5.4 Face3D database

This database is used in [62, 63] for 3D face recognition. It is quite realistic for biometric verification because both the training and the testing set contain very different images (taken with different cameras, backgrounds, poses, expressions, illuminations and time). In his papers, the author proposes 60 different classifiers by measuring the similarity of different regions. In our experiment, we only take 5 regions out of these 60: similarity of the full face, the left half, the right half, the bottom part, and the upper part. The results of these 5 algorithms are rather correlated, of course. This choice is made to see the performance of our benchmark methods in handling the dependence between classifiers. By following our procedure, we get as the best copula pair (ind,Fr).

Figure 3.8 shows clearly that considering dependence can improve the performance significantly. We can see that our fixed FAR fusion method is the only fusion strategy that can handle the dependence on this database as given in Table 3.11. While our fixed FAR fusion performs very well in handling the dependence, the GMM fusion is even worse than the best single matcher. This happens because the estimated number of components in the GMM is equal to the the maximum value (20) of it when being estimated by the minimum message length criterion as proposed in [60]. It means that the number of components may be more than 20. However, if we increase the number of components then the estimator becomes less reliable.

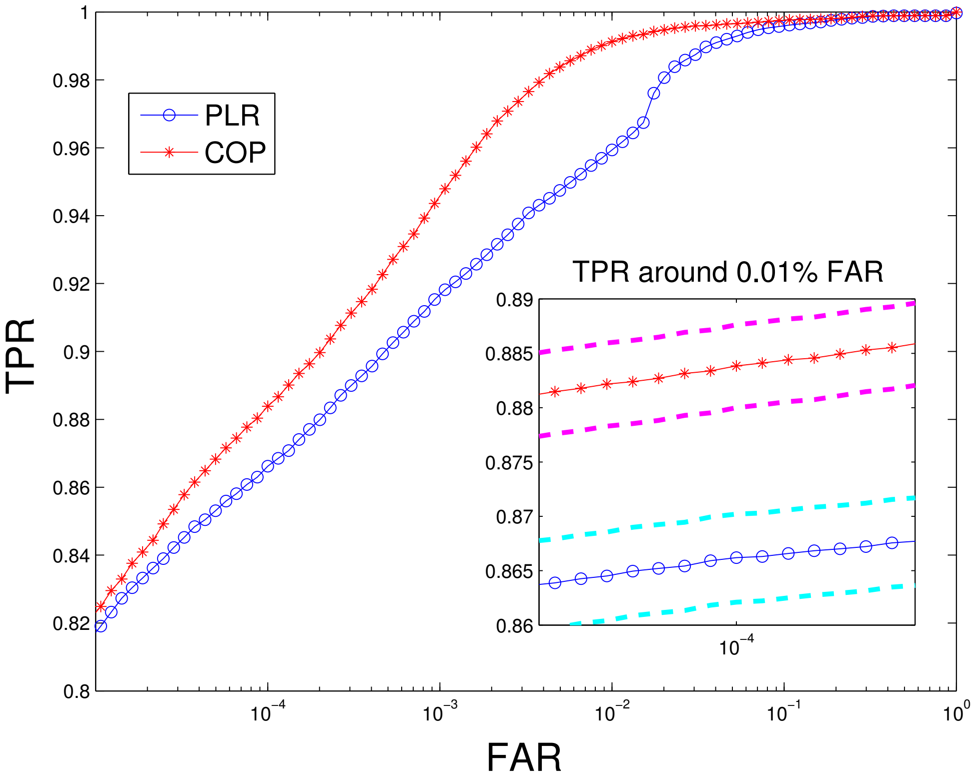


Figure 3.8: Comparison between the PLR and our fixed FAR fusion methods on Face3D database. The small box contains the highlighted performance at around 0.01% FAR. The dashed lines are the 99% Jeffrey CIs.

Table 3.11: Performances at 0.01% FAR on Face3D database.

Methods	TPR	99% Jeffreys CI compared to PLR in TPR at 0.01% FAR
BSM	84.9%	N/A
PLR	86.6%	N/A
Logit	87.6%	[0.1%, 1.7%]
GMM	<u>81.2%</u>	[−6.3%, −4.5%]
FFF	**88.4%**	[1.0%, 2.6%]

BSM: Best Single Matcher, GMM: Gaussian Mixture Model, Logit: Logistic Regression, PLR: Product of Likelihood Ratios, FFF: our fixed FAR fusion. The bold number is the best one and the underlined number is the worst one.

3.4.6 Conclusion

We have proposed and used an alternative method for evaluating the performance of biometric fusion methods at fixed FAR using Jeffreys credible intervals. We have also proposed a fixed FAR fusion method to improve via parametric copulas the PLR fusion strategy. From a simulation study with synthetic data, we have concluded that it is always useful to take the dependence into account even for low dependence levels. It has also been shown that our fixed FAR fusion method is the best method on real databases compared to the GMM and Logit fusion methods, which are also designed to handle dependence. Instead of a rule of thumb to always take dependence in biometric fusion into account, we propose to always check whether our fixed FAR method improves on the PLR fusion method by a simple test as follows: define relevant training and testing sets, follow our procedure in choosing the best copula pair on the training set, and finally check the significance improvement using our evaluation method on the testing set. We can see from the FVC2002-DB1 database that the existing rule of thumb concludes the unimportance in considering dependence. However, when the FAR value is fixed (0.01%), we get a significant improvement of around 82% to 84% (around 2%). Although it is a relatively small improvement, our fixed FAR fusion method reduces the number of people that have to be checked manually from 18 to 16 for every 100 people. This means that if the manual checking needs 10 minutes per person then we save 20 minutes for every 100 people.

3.5 Conclusion of Chapter 3

In this chapter we presented a mathematical framework for semiparametric LR-based biometric fusion via parametric copula families. The convergence of the parameters determining the LR computation was discussed and a detailed procedure to train the proposed method was also demonstrated. Experimental results in all sections of this chapter have shown that our method outperforms the other LR-based fusion methods (the GMM and Logit methods) and of course the PLR method especially for the standard biometric verification scenario.

Chapter 4

Fusion in Forensic Face Recognition

4.1 Chapter Introduction

PURPOSE. This chapter presents the results of a study into the effect of incorporating dependence between matchers in score level fusion for forensic face recognition. Of course, it is hoped that taking dependence into account yields a better performance than simple fusion under the independence assumption between matchers.

CONTENTS. Section 4.2 introduces a new method of score level fusion for forensic face recognition based on the PAV algorithm and copula models. The detailed procedure of the proposed method is given in Subsection 4.2.4 and some experiments on synthetic and real databases are presented in Subsection 4.2.5.

PUBLICATIONS. The manuscript presented in Section 4.2 has been published in [53].

NOTES. The reader might focus on the following subsections:

(1) 4.2.4 explains how our proposed method, which we call the two-step calibration method, can be used in combining two or more dependent face recognition systems in order to get better performance than with the simple fusion method under the independence assumption;
(2) 4.2.5 provides experimental results on synthetic and real databases.

Subsection 4.2.3 gives some performance measures that have been discussed in Chapter 1, Subsection 1.1.1.

4.2 Two-step Calibration Method for Multi-algorithm Score-based Face Recognition Systems by Minimizing Discrimination Loss

4.2.1 Abstract

We propose a new method for combining multi-algorithm score-based face recognition systems, which we call the two-step calibration method. Typically, algorithms for face recognition systems produce dependent scores. The two-step method is based on parametric copulas to handle this dependence. Its goal is to minimize discrimination loss. For synthetic and real databases (NIST-face and Face3D) we will show that our method is accurate and reliable using the cost of log likelihood ratio and the information-theoretical empirical cross-entropy (ECE).

4.2.2 Introduction

The likelihood ratio (LR) approach of evidence evaluation is increasingly accepted in forensic science [3]. The LR of evidence e is defined as the ratio between the probability of the evidence given *prosecution* and *defense* hypotheses, i.e.,

$$\mathrm{LR}(e) = \frac{P(e|H_\mathrm{p})}{P(e|H_\mathrm{d})} \tag{4.2.1}$$

where H_p and H_d are two mutually exclusive hypotheses respectively supporting whether or not the suspect is the donor of the biometric trace. This quantitative value is computed by a forensic scientist and can be used to support the fact finder (judge/jury) in court to make an objective decision. The Bayesian framework explains elegantly how the LR supports the decision via relation

$$\frac{P(H_\mathrm{p}|e)}{P(H_\mathrm{d}|e)} = \frac{P(e|H_\mathrm{p})}{P(e|H_\mathrm{d})} \times \frac{P(H_\mathrm{p})}{P(H_\mathrm{d})}. \tag{4.2.2}$$

This means that the LR can be interpreted as a multiplicative factor for the information before analyzing the evidence (*prior odds*) to get the new information after taking the evidence into account (*posterior odds*).

In this paper we are studying *multi-algorithm score-based face recognition systems*, in which two or more different algorithms compute a similarity score for any pair of face images. This means that the evidence e is a vector of scores in which every score describes the similarity of the image found at the crime scene and an image of the suspect. It is intuitively understandable that combining several algorithms might be advantageous. For instance, every individual algorithm can be selected for its good performance under a specific condition, such as varying pose, illumination, or robustness. Therefore, an appropriate combination is hoped to integrate the complementary information of the individual algorithms. Indeed, several studies [62, 63, 69] show that a multi-algorithm method might enhance the recognition performance.

Several methods of deriving the LR from a biometric comparison score, which is also called *calibration*, have been proposed and evaluated for single-algorithm face recognition systems; see [4, 5] for a survey of these methods. In contrast, to the best of our knowledge, there is no method of combining two or more face recognition systems for forensic evidence evaluation. In this paper, we propose such a method, which we will call the two-step calibration method, for calibrating multi-algorithm face recognition systems via parametric copulas. We will compare our method to the *linear* logistic regression (Logit) method, which is commonly used in the field of speaker recognition [54, 55], and also to the Gaussian Mixture Model (GMM) [16] and the simple Product of Likelihood Ratios (PLR) [56], which are originally proposed in biometric fusion for person authentication. We also show through simulation and real data that the logistic regression method used in the field of speaker recognition [54, 55] is not recommendable for use in forensic face scenarios.

The rest of this paper is organized as follows. Section 4.2.3 reviews the cost of log likelihood ratio and the ECE plot, which measure the accuracy and reliability of calibration methods. Our two-step calibration method is presented in Section 4.2.4. Section 4.2.5 demonstrates the excellent performance of our method for both synthetic and real databases. Finally, our conclusions are presented in Section 4.2.6.

4.2.3 Performance Evaluation of Likelihood Ratio Computation

There are two types of measures for the reliability of calibration methods: *application-dependent* [6,7] and *application-independent* [8–11] measures. Since forensic scientists do not have access to the prior odds, we will focus on application-independent ones.

4.2.3.1 Cost of log likelihood ratio

The cost of log likelihood ratio (C_{llr}) is introduced by Brümmer and du Preez [8] in the field of speaker recognition, is based on a generalization of cost evaluation metrics, and is used in forensic face scenarios in [12]. This measure may be interpreted as a summary of a LR computation [67]. Note that a face recognition system does not necessarily produce a similarity score as an LR value. Thus, a calibration is needed to make this *original score* interpretable as an accepted measure of strength of evidence in court by mapping it into LR value, which we also call *LR score.* A score is called *genuine* if it is associated to 2 images of the same person, and is called an *impostor* score if it involves 2 images of two different persons. Let $\mathcal{M}$ denote a method to calibrate original scores into LR values. Given a set of scores, let $\mathcal{LR}_{\text{p}}$ denote the set of N_{gen} genuine $\mathcal{M}$-calibrated scores, which correspond to the hypothesis of the prosecution, and $\mathcal{LR}_{\text{d}}$ the set of N_{imp} impostor $\mathcal{M}$-calibrated scores, which correspond to the hypothesis of the defense. The cost of log likelihood ratio C_{llr} is defined by

$$\begin{aligned} C_{\text{llr}} = {} & \frac{1}{2N_{\text{gen}}} \sum_{\text{LR}\in\mathcal{LR}_{\text{p}}} \log_2\left(1+\frac{1}{\text{LR}}\right) \\ & + \frac{1}{2N_{\text{imp}}} \sum_{\text{LR}\in\mathcal{LR}_{\text{d}}} \log_2\left(1+\text{LR}\right). \end{aligned} \tag{4.2.3}$$

To explain the name of this metric we note that LR in formula (4.2.3) may be rewritten in terms of the logarithm of LR. Interestingly, this metric can be decomposed into a *discrimination* and *calibration* form via relation

$$C_{\text{llr}} = C_{\text{llr}}^{\text{min}} + C_{\text{llr}}^{\text{cal}}. \tag{4.2.4}$$

Here, $C_{\text{llr}}^{\text{min}}$ and $C_{\text{llr}}^{\text{cal}}$ denote the discrimination and calibration loss, respectively. Discrimination loss is the opposite of discrimination power (the ability of the system to distinguish between genuine and impostor scores). The smaller the value of this quantity, the higher the discrimination power. The $C_{\text{llr}}^{\text{min}}$ is defined as the minimum C_{llr} value under evaluation by preserving the discrimination power which is attained by the Pool-Adjacent-Violators (PAV) algorithm as proved in [8]. Therefore, the $C_{\text{llr}}^{\text{min}}$ is computed by plugging the $\mathcal{M}$-calibrated scores after PAV transformation into (4.2.3). On the other hand, calibration loss indicates the calibration performance on a separate evaluation set.

4.2.3.2 ECE plot

The Empirical Cross Entropy (ECE) plot is an application-independent method of measuring the reliability of calibration with an information theoretical interpretation [10]. The ECE function is defined as a function of the log prior odds by

$$\begin{aligned}\mathrm{ECE}(\mathrm{lp}) &= \frac{1}{2N_{\mathrm{gen}}}\sum_{\mathrm{LR}\in\mathcal{LR}_{\mathrm{p}}}\log_2\left(1+\frac{1}{\mathrm{LR}\times e^{\mathrm{lp}}}\right)\\ &+\frac{1}{2N_{\mathrm{imp}}}\sum_{\mathrm{LR}\in\mathcal{LR}_{\mathrm{d}}}\log_2\left(1+\mathrm{LR}\times e^{\mathrm{lp}}\right)\end{aligned} \tag{4.2.5}$$

for every lp $\in(-\infty,\infty)$. Clearly $C_{\mathrm{llr}}=\mathrm{ECE}(0)$ holds, which shows that the ECE generalizes the cost of log likelihood ratio.

Figure 4.1 is an example of the ECE plot of a system. The solid red curve represents the performance of the calibration, the dashed blue curve is the minimum ECE value under evaluation by preserving the discrimination power which is attained by PAV transformation, and the dashed black curve is the entropy of the neutral system without considering the evidence, i.e., all LR values equal to 1. The difference between the solid red and dashed blue curves is the calibration loss. Since the ECE value can be interpreted as the average information loss by taking the system into account, we can see that the system will lose more information than the neutral system for log prior odds greater than 2. Therefore, forensic scientists should provide the usual LR and also explain to the fact finder that he should not use the forensic system if his log prior odds are greater than 2.

4.2.4 Evidential Value Computation of Multi-algorithm Systems by Minimizing Discrimination Loss

This section explains how to get calibrated scores for d-algorithm face recognition systems, i.e., computing the LR at evidence $e=(s_1,\cdots,s_{\mathrm{d}})$. Of course, if the joint density functions of the evidence under both hypotheses, which will be denoted by f_{gen} and f_{imp} for genuine and impostor scores, respectively, are known then the exact LR can be easily obtained. However, in practice,

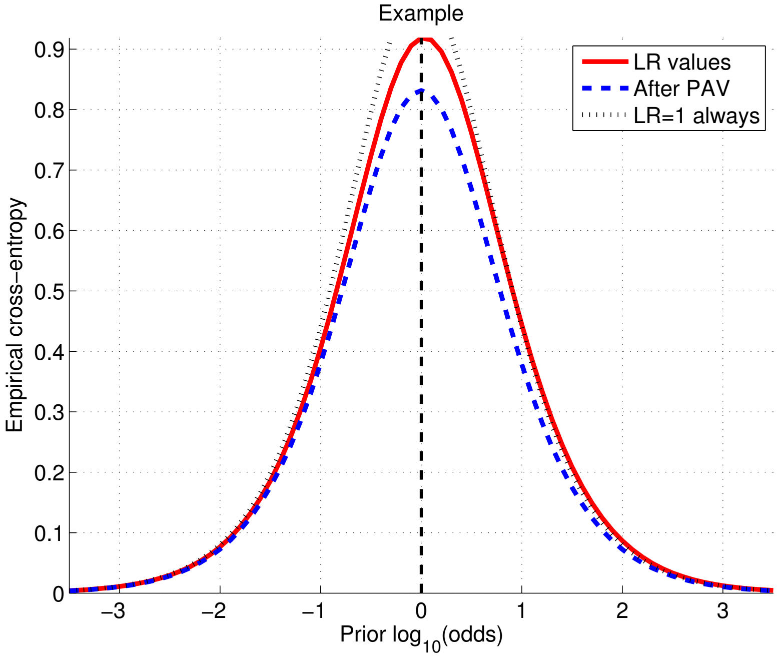


Figure 4.1: Example of ECE plot

these density functions have to be estimated from data. This classical problem in statistics can be solved by parametric (e.g., normal distribution, Weibull distribution) and nonparametric (e.g., histogram, kernel density estimation) models. However, the choice of an appropriate parametric model is sometimes difficult while nonparametric estimators suffer from the difficulty that they are sensitive to the choice of the bandwidth or of other smoothing parameters, especially for our multivariate case. Therefore, it is natural to approach our estimation problem semiparametrically, modelling the marginal densities nonparametrically and the dependence between them by parametric copulas.

4.2.4.1 Dependence through Copula

Mathematically, a copula is a distribution function on the unit cube $[0,1]^d$, $d \geq 2$, of which the marginals are uniformly distributed. In practice, it is widely used to describe the dependence of random variables; see e.g. [83, 84] for application in econometrics and finance. In biometric fusion, Susyanto et al. [50] use a specific copula called Gaussian copula to handle the dependence between classifiers. A classical result of Sklar [18] relates any continuous multivariate distribution function to a copula.

Theorem 4.2.1 (Sklar (1959)). *Let $d \geq 2$, and suppose H is a distribution function on $\mathbb{R}^d$ with one dimensional continuous marginal distribution functions $F_1, \cdots, F_d$. Then there is a unique copula C so that*

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)) \tag{4.2.6}$$

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

The joint density function can be computed by taking the d-th derivative of (4.2.6):

$$\begin{aligned} h(x_1, \ldots, x_d) = c(&F_1(x_1), \ldots, F_d(x_d)) \\ &\times \prod_{i=1}^{d} f_i(x_i) \end{aligned} \tag{4.2.7}$$

where c is the copula density and f_i is the i-th marginal density for every $i = 1, \cdots, d$. We can see that the density h is a product of the copula density depending only on the marginal distributions $F_1, \cdots, F_d$ and its marginal densities. It means that we can estimate separately the dependence structure represented by the copula density c and the individual densities f_i in order to get the joint density h. If C_α is determined by a finite dimensional Euclidean parameter α then it is called parametric copula. In this case, we can estimate the dependence parameter α based on i.i.d. observations

$$\mathbf{X}_1, \ldots, \mathbf{X}_n$$

with

$$\mathbf{X}_i = (X_{1i}, \ldots, X_{di}) \quad \forall i = 1, \ldots, n$$

by the pseudo-maximum likelihood estimator (PMLE). Mathematically, the PMLE of α has to maximize

$$\frac{1}{n} \sum_{i=1}^{n} \log c_\alpha \left(\hat{F}_1(X_{1i}), \ldots, \hat{F}_d(X_{di}) \right) \tag{4.2.8}$$

where

$$\hat{F}_j(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbf{1}_{\{X_{ji} \leq x\}}, \quad \forall 1 \leq j \leq d$$

is a *modified* empirical distribution function and c_α is the copula density.

Let C_{gen} and C_{imp} be the copula corresponding to genuine and impostor scores with copula densities c_{gen} and c_{imp}, respectively. In view of (4.2.1) and (4.2.7), the likelihood ratio at $e = (s_1, \cdots, s_{\text{d}})$ can be written as

$$\begin{aligned}\text{LR}(e) = {} & \frac{c_{\text{gen}}(F_{\text{gen},1}(s_1), \cdots, F_{\text{gen},d}(s_{\text{d}}))}{c_{\text{imp}}(F_{\text{imp},1}(s_1), \cdots, F_{\text{imp},d}(s_{\text{d}}))} \\ & \times \prod_{i=1}^{d} \text{LR}_i(s_i) \end{aligned} \tag{4.2.9}$$

where $F_{\text{gen},i}$ and $F_{\text{imp},i}$ denote the distribution functions of genuine and impostor scores, respectively. The i-th individual LR can be computed for each $i = 1, \ldots, d$ by the PAV algorithm, which is optimal for calibrating 1-dimensional scores. Therefore, we only need to estimate the first factor at the right hand side of (4.2.9): the copula part.

4.2.4.2 Two-step Calibration Methods

As noted before, the density functions f_{gen} and f_{imp} have to be estimated, which implies that the copula densities c_{gen} and c_{imp} must be estimated as well. Estimating copula density functions nonparametrically will lead to the same problems as when estimating the original density functions directly. Therefore, we will approximate the copula part of (4.2.9) by some well-known parametric copulas. We will use the following copulas: Gaussian copula (GC), Student's t (t), Frank (Fr), Clayton (Cl), and Gumbel (Gu). We also include the independent (ind) to guarantee that our combined system is better than the simple product of likelihood ratios. Readers interested in copulas are referred to [79] for a more detailed explanation. To have more dependence models and because the Clayton and Gumbel copulas are not symmetric, their flipped forms (flipped Clayton (fCl) and flipped Gumbel (fGu)) will be included as well (if U has copula C then $1-U$ has copula flipped C). Therefore, the copulas c_{gen} and c_{imp} are chosen from the copula family

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

We can choose the best copula for c_{gen} and c_{imp} from the family $\mathcal{C}$ for f_{gen} and f_{imp} by a goodness-of-fit test as provided in [59]. However, this method only guarantees that the selected copulas are closest to c_{gen} and c_{imp}, but not necessarily good enough to model $c_{\text{gen}}/c_{\text{imp}}$. Therefore, we propose to choose the best copula pair directly by minimizing the discrimination loss as explained in Section 4.2.3.1. Since the estimated copula pair is not the true copula and only minimizing the discrimination loss among other pairs, the combined scores can be poorly calibrated. To solve this problem, we apply the PAV algorithm once combined scores have been obtained via the best copula pair.

Given a set $\mathcal{C}$ of n_c candidate copulas and a training set, our two-step calibration method is very simple. The first step is computing the product of the individual likelihood ratios by the PAV algorithm and multiplying this product by each of all copula pairs $\hat{c}_{\text{gen}}/\hat{c}_{\text{imp}}$ in which the dependence parameters have been estimated by the PMLEs as defined in (4.2.8). Of the $n_c \times n_c$ resulting different combined scores we choose the one that minimizes the discrimination loss. The second step is transforming the combined scores by the PAV algorithm so that the final scores have high discrimination power and are also well-calibrated.

4.2.5 Experimental Results

To study the performance of our two-step calibration method we apply it to synthetic and real databases, which are split up into training and testing sets. Given a training set, we will compute the product of the individual likelihood ratios, select the best copula pair, and calibrate the combined scores. The corresponding testing set is used for evaluation only. The ECE plot is chosen for evaluation because it is more general than the cost of log likelihood ratio. On the real databases, beside plotting the ECE curves, we also highlight the discrimination loss and the ECE values for log prior odds -2, 0, and 2; see Table 4.1. We compare our two-step method to the Logit method studied in [55] and the GMM method where the number of the mixture components is automatically estimated by the minimum message length criterion as proposed in [60]. For all experiments, the maximum value of the number of the mixture components is 20.

Given genuine and impostor scores

$$W_1, \ldots, W_{n_{\text{gen}}}$$

and

$$B_1, \ldots, B_{n_{\text{imp}}}$$

in the training set, our procedure to choose the best copula pair is simple. We randomize the genuine (impostor) scores and take two disjoint subsets with size

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

and

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

This re-sampling method is aimed at increasing the computation speed because it will be repeated 100 times to see the consistency. Once the product of the individual likelihood ratios is computed, it is multiplied by the 64 copula pair estimates $\hat{c}_{\text{gen}}/\hat{c}_{\text{imp}}$. After all 64 combined scores are obtained using the first subset, the discrimination loss is then computed. The final discrimination loss for each copula pair is the average over all 100 experiments. The best copula pair is the pair having the smallest average of the discrimination loss values. If there are several pairs having the same averages, we choose the pair with the smallest variance. If there is still more than one pair having the smallest means and variances then we choose one of them at random.

4.2.5.1 Synthetic data

To get synthetic data that behave like real data, we take two algorithms presented in [63]. The first algorithm measures the similarity of the left half of the face between two images and the second one the similarity of the right half. The density and distribution functions of the genuine and impostor scores for each algorithm are estimated by kernel density estimation. To obtain scores with *explicit* dependence that can be represented by a copula C, we generate random samples of the copula C and apply the inverse transform technique, using the estimates of the two marginal distribution functions. In this way the generated scores have as marginal distribution functions these estimates of the distribution functions of data generated by the two algorithms. Recall that if F is a continuous distribution function then U is uniformly distributed if and only if $F^{-1}(U)$ has distribution function F.

In our experiment, we generate 10,000 genuine and 1,000,000 impostor scores in the way as explained above. The dependence is made by putting 4 different copula pairs

$$\{(\text{GC}, \text{GC}), (\text{t}, \text{fCl}), (\text{fGu}, \text{GC}), (\text{Cl}, \text{Gu})\}$$

completed with 9 dependence level pairs obtained from the cross pairs

$$\{\text{low}, \text{moderate}, \text{high}\}.$$

The low, moderate, and high dependence levels are set to have correlation

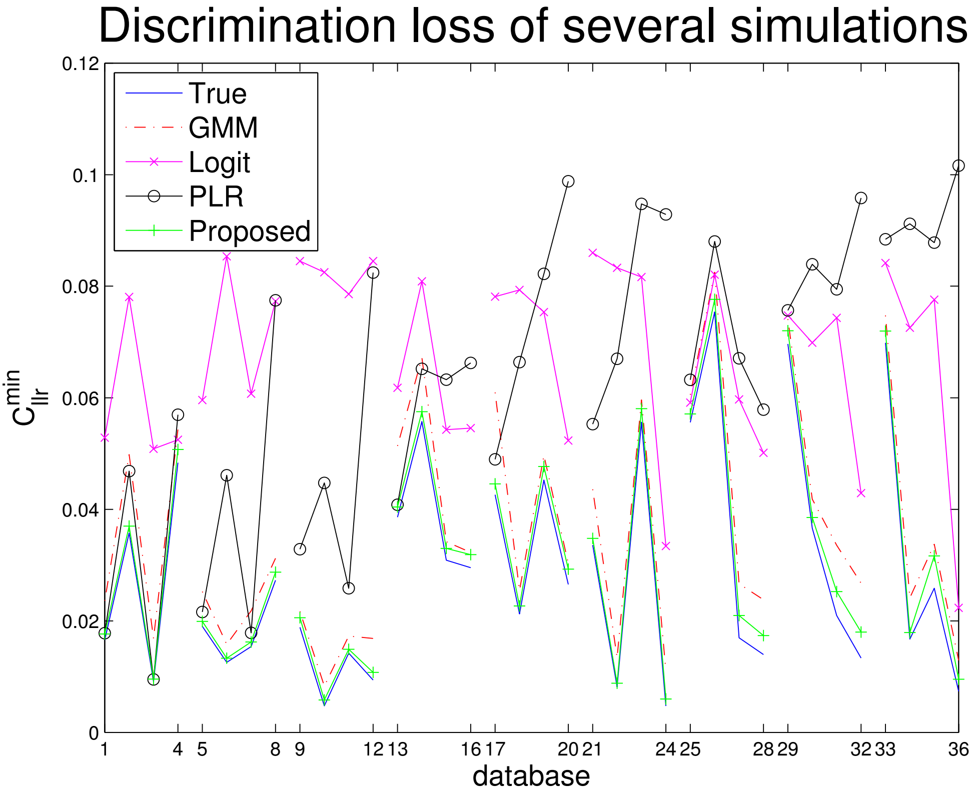


Figure 4.2: Performance on synthetic data. On the x-axis the databases are indicated in 9 groups of 4, each group having the same dependence level pair for each of the 4 chosen copula pairs. Database 1-4 has low and low dependence levels for genuine and impostor scores, 5-8 low and moderate, 9-12 low and high, 13-16 moderate and low, etc.

values 0.1, 0.5, and 0.9 for Gaussian and Student's *t* copulas while for other copulas we put 1, 10, and 50. Student's *t* copula has 3 degrees of freedom for all experiments.

Figure 4.2 is the plot of discrimination loss of our 36 simulated databases. The true LR can be computed exactly because the underlying distributions are known. We can see that our two-step method outperforms the others. As expected, the PLR method performs poorly when the dependence between algorithms is moderate or high because much information will lose by assuming the independence between algorithms. The GMM method is the second best method but the computation time is much longer than for the two-step method. We can also see that the logistic regression method, which is commonly used in the field of speaker recognition, has the worst performance among all methods.

4.2.5.2 NIST-face BSSR1 database

The NIST-face BSSR1 database is published by National Institute of Standards and Technology [61]. The data contain similarity scores from two face systems run on images from 3000 subjects with each subject having two probe images and one gallery image. We only take 2992 subjects because the scores of the other 8 subjects on the first system are always -1. It is reported that these images are not accepted by the facial recognition system and are therefore excluded. To evaluate the performance of our benchmark calibration methods, we randomize the subjects and split the set into two disjoint sets with size 1496. We follow the procedure explained at the beginning of this section and the pair (ind,GC) is obtained as the best copula pair. We repeat this experiment 20 times and all of them give almost the same result. Therefore, we decided to show only one of these results.

The GMM, PLR, and our two-step method have almost the same performance as seen from the ECE plot in Figure 4.3. Although the Logit method performs reasonably well for small values of the log prior odds, it has the highest calibration loss among all methods and it is even dangerous for use in forensic face scenarios for large values of the log prior odds (greater than 2). By only considering the discrimination loss, Table 4.1 of the NIST-face part tells us that the GMM is the best calibration method. However, the ECE values show that the two-step method is actually the best one.

4.2.5.3 Face-3D database

This database is used in [62,63] for 3D face recognition. It is quite realistic for the forensic face problem, because both the training and the testing set contain very different images (taken with different cameras, backgrounds, poses, expressions, illuminations and time). In his papers, the author proposes 60 classifiers operating on 30 different facial regions with 2 different image registration methods. In our experiment, we only take 3 classifiers out of these 60: similarity of the full face, the left and the right half. The results of these 3 classifiers are rather correlated, of course. This choice is made to see the performance of our benchmark methods in handling the dependence among classifiers. Although they are not different algorithms, we use these 3 clas-

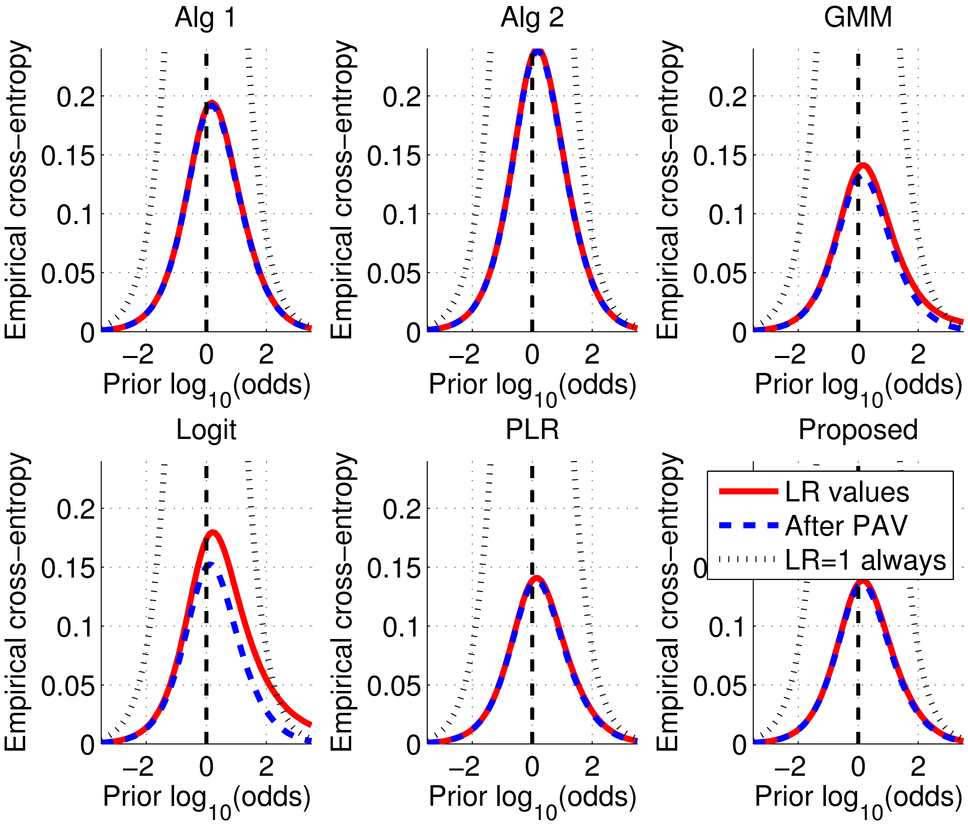


Figure 4.3: ECE plot of NIST-face

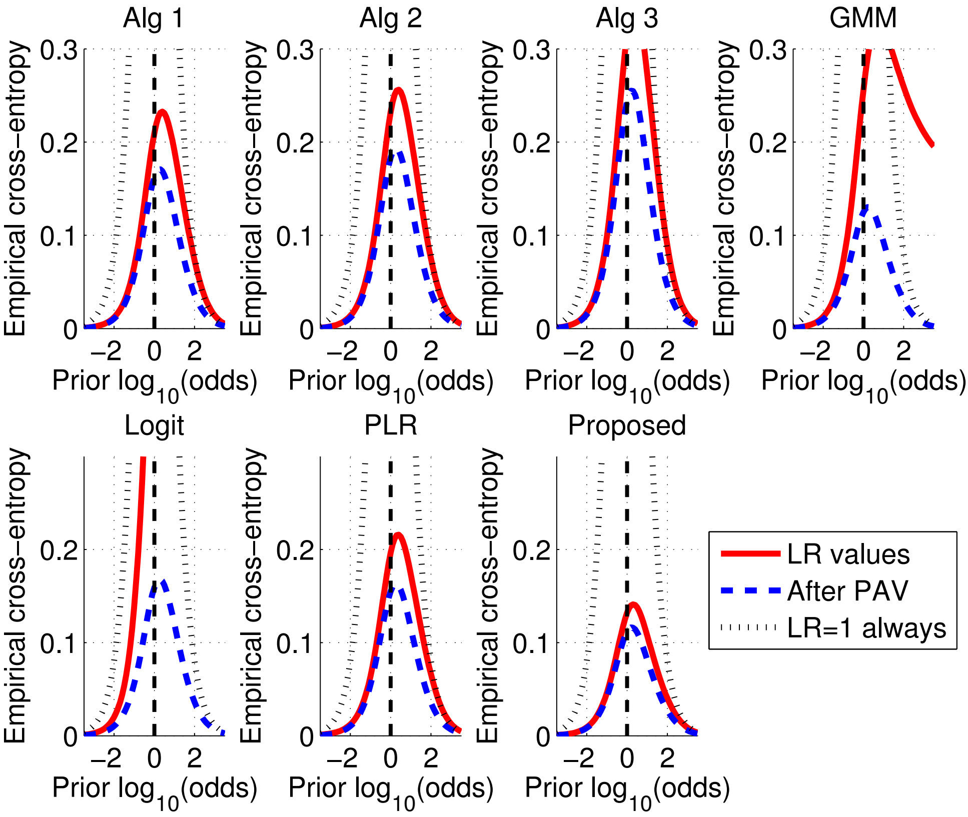


Figure 4.4: ECE plot of Face3D

sifiers to have different types of data and see the performance our two-step method for multi-classifiers scenario as well. By following our procedure, we get as the best copula pair (ind,t). The performances on this database are provided by Figure 4.4 for the ECE plot and Table 4.1 for the discrimination loss and some values of the ECE.

We can see that our two-step method is the best one using all evaluation metrics ($C_{\text{llr}}^{\text{min}}$, C_{llr}, ECE at small prior odds, and ECE at high prior odds). As before, the Logit method performs poorly on this database since the underlying distributions are not gaussian. Surprisingly, the GMM method also has poor performance on this database; it is even worse than the simple PLR. This may be because the number of the mixture components is more than the the maximum value that we set. However, if we increase the number of components then we will have a problem with the limitation of the sample size.

Methods	NIST-Face				Face 3D			
	$C_{\text{llr}}^{\text{min}}$	ECE			$C_{\text{llr}}^{\text{min}}$	ECE		
		-2	0	2		-2	0	2
BSS	0.187	0.153	0.189	0.037	0.162	0.012	0.210	0.073
GMM	**0.131**	0.123	0.139	0.033	0.125	0.013	0.256	0.260
Logit	0.150	0.138	0.174	0.055	0.159	0.020	0.598	0.828
PLR	0.137	0.123	0.139	0.028	0.155	0.012	0.197	0.066
Proposed	0.134	**0.120**	**0.136**	**0.027**	**0.112**	**0.009**	**0.132**	**0.040**

Table 4.1: Discrimination loss and ECE of different methods on the real databases. BSS: Best Single System, GMM: Gaussian Mixture Model, Logit: Logistic Regression, PLR: Product of Likelihood Ratios. The bold number is the best one in every column.

4.2.6 Conclusion

We propose a two-step calibration method to compute the likelihood ratio of multi-algorithm score-based face recognition systems in forensic evidence evaluation. The first step of the two-step method is computing the product of the individual likelihood ratios multiplied by the density ratio of the best copula pair determined by minimizing discrimination loss. The simple second step is applying the PAV algorithm in order to get well-calibrated scores. Using several synthetic data sets, we have shown that our approach performs very well in handling all dependence levels (low, moderate, and high). We also see that our two-step method on the real databases NIST-face BSSR1 and Face3D. We conclude that the GMM method, which works quite well in biometric fusion for person authentication, can somehow perform poorly in forensic face scenarios. We also recommend to avoid the logistic method, which is commonly used in the field of speaker recognition, to compute the likelihood ratio in forensic face recognition because it has high discrimination loss and sometimes it is much worse than the neutral system.

4.3 Conclusion of Chapter 4

In this chapter, we studied the use of copula models in forensic face scenarios and proposed the two-step calibration method. Using some well-known parametric copula families, it was demonstrated how to choose the best copula pairs for genuine and impostor scores on some synthetic and public databases. Finally, we also noticed that the GMM and Logit method may somehow perform poorly in forensic face scenarios.

Chapter 5

Conclusion

This chapter concludes this PhD thesis and presents recommendations for future research. The thesis contains contributions to Statistical Theory and Biometric Applications. The Statistical Theory part of this thesis was inspired by the dependence between two comparison scores that involve at least one common person. This dependence was modelled by a Gaussian copula with constraints on the covariance matrix. We studied efficient estimation in this model and generalized this to efficient estimation in quite general semiparametric models with constraints on the parameters.

Our LR-based score level fusion was proposed in the Biometric Application part to handle dependence between matchers. The results from the Statistical Theory part were not used in the Biometric Application part because they did not give a significant improvement and sometimes they even degraded the performance of our fusion strategy. In the following, we highlight the main contributions of this thesis by reviewing the research questions posed in Chapter 1 and explaining how the thesis answers these questions.

5.1 Answers to the research questions

5.1.1 Statistical Theory

Consider a quite arbitrary (semi)parametric model with a Euclidean parameter of interest and assume that an asymptotically (semi)parametrically efficient estimator of it is given.

- If the parameter of interest is known to lie on a general surface (image of a continuously differentiable vector valued function), what is the lower bound on the performance of estimators under this restriction and how can an efficient estimator be constructed?

A semiparametric submodel is defined in which the parameter of interest is the lower dimensional parameter determining the general surface. The semiparametric lower bound for estimators of it is obtained via the Hájek-LeCam Convolution Theorem for regular parametric models. Furthermore, the efficient score function for the underlying parameter is determined by the efficient score function for the original parameter and the Jacobian of the function defining the general surface, via a simple chain rule for score functions. An efficient estimator for the underlying parameter is constructed in terms of the efficient estimator for the original estimator, the Jacobian of the function defining the general surface, and a consistent estimator of the optimal lower bound. This consistent estimator is based on empirical characteristic functions and a sample splitting technique. Finally, some simple examples are given in location-scale, Gaussian copula, and semiparametric regression models, and in parametric models under linear restrictions.

- If the parameter of interest belongs to the zero set of a continuously differentiable function (for which it might be impossible to parametrize it as the image of a continuously differentiable vector valued function), what is the lower bound on the performance of estimators under this restriction and how can an efficient estimator be constructed?

A semiparametric submodel is defined in which the parameter of interest is restricted by an functional equality constraint. The efficient influence function for the constrained parameter is obtained by a projection technique and an updated estimator is proposed for the constrained parameter in terms of the efficient estimator of the parameter without restrictions and the function defining the equality constraint. An efficient estimator for the constrained parameter itself is then obtained by finding the closest point in the zero set of the function defining the constraint to the updated estimator. Finally, some simple examples are given in location-scale, Gaussian copula, semiparametric regression, and parametric models.

5.1.2 Biometric Application

Suppose we have score-based multibiometric matchers, in which two or more different matchers compute a similarity score for any pair of two biometric samples.

- How can copula models handle dependence between matchers? How do we estimate the dependence parameters from training data? What are the performances of handling dependence compared to the simple independence assumption between matchers in applications?

A what we call semiparametric LR-based score level fusion strategy is proposed. The LR of the joint matchers is computed by splitting the marginal likelihood ratios and the dependence between matchers via the copula concept. As a result, the LR can be computed by multiplying the product of the individual likelihood ratios and the copula density ratio called Correction Factor. A semiparametric model to compute the Correction Factor is proposed by computing the individual likelihood ratios nonparametrically via the PAV algorithm and by modelling dependence between matchers by parametric copulas. It is then discussed how the estimator for the Correction Factor parameter can be obtained and how it behaves like, asymptotically. Finally, some applications simulating real biometric scenarios are presented and it is demonstrated how our LR-based fusion is applied to them including the method to choose the best copula pairs.

- How can copula models be used in standard biometric verification? How can we compare copula-based biometric fusion to the simple independence assumption between matchers?

In line with the semiparametric LR-based score level fusion strategy, a method is proposed, called fixed FAR fusion, by maximizing the true positive rate (TPR) at fixed false acceptance rate (FAR) semiparametrically. After computation of the individual likelihood ratios via the PAV algorithm, the best Correction Factor modelled by some well-known parametric copulas is determined by optimization of the TPR at fixed FAR, which is set beforehand, using a resampling method. Our fixed FAR fusion is then compared to simple fusion under the independence assumption between matchers by use of Jeffreys' method. Our fixed FAR fusion is also compared to other LR-based methods (GMM and Logit) on synthetic and real databases.

- How can copula models be used in forensic applications for combining multi-algorithm face recognition systems, which are usually dependent?

Following the semiparametric LR-based score level fusion strategy, a method called *two-step calibration* is proposed by choosing from a family of some well-known parametric copulas the copula pair that gives the smallest discrimination loss after the product of the individual likelihood ratios has been computed via the PAV algorithm. Once the best copula pair has been chosen, the training data are fused via this copula pair and the fused data are used to train the PAV algorithm in order to make the fused scores well calibrated. Some experimental results on real databases show that the two-step method outperforms the PLR, GMM, and Logit methods with respect to the cost of loglikelihood and the ECE plot.

5.2 Final remarks

There are two types of dependence in score level fusion: dependence between scores (that involve at least one common person) produced by one matcher and dependence between matchers. The first type of dependence may be modelled by a Gaussian copula under constraints, which is a special case of a general semiparametric model with constrained parameters. Although such semiparametric models can be applied to some common problems in statistics, the obtained results do not help in improving the accuracy of estimates of the LR and they might even degrade the performance. Causes might be that most scores are independent and that the assumption of the observations following a Gaussian copula distribution is rather strong and hard to be satisfied for real biometric data. On the other hand, the dependence between matchers is always relevant and should be taken care of in order to improve the PLR method as can be seen from the results in Chapter 3. It is also emphasized that the best copula pair may be different for every performance measure.

5.3 Recommendations for future research

5.3.1 Statistical Theory

In Section 2.3 semiparametric estimation of Euclidean parameters under equality constraints is discussed. In future research more general constraints can be studied by considering both equality and inequality constraints as already studied in parametric models.

5.3.2 Biometric Application

Chapters 3 and 4 discussed the importance of incorporating dependence between matchers in score level fusion. A semiparametric LR-based method was proposed where the individual likelihood ratios are computed by the PAV algorithm and the Correction Factor is approximated by a copula density ratio. Future work may investigate the following problems.

- The accuracy of the PAV algorithm can be very poor for small sample sizes. A more robust method that can handle small sample size problems for estimating 1-dimensional likelihood ratios, may potentially improve the semiparametric LR-based method.
- The parametric copula families that are used in Chapter 3, model the dependencies between all matchers the same. More precisely, if there are three matchers then the dependence of all three pairs of two matchers is assumed to be the same. In practice, the dependence between the first and second matcher may be different from the dependence between the first and the third one or between the second and third one. Therefore, by allowing each pair of two matchers to have different dependence may also lead to improvement.

Bibliography

[1] T. Fawcett, “An introduction to roc analysis,” *Pattern Recogn. Lett.*, vol. 27, pp. 861–874, June 2006.

[2] F. Hsieh and B. W. Turnbull, “Nonparametric and semiparametric estimation of the receiver operating characteristic curve,” *Ann. Statist.*, vol. 24, pp. 25–40, 02 1996.

[3] O. E. Facey and R. J. Davis, “Re: Expressing evaluative opinions; a position statement,” *Science & Justice*, vol. 51, no. 4, pp. 212 –, 2011.

[4] T. Ali, L. J. Spreeuwers, and R. N. J. Veldhuis, “Forensic face recognition: A survey,” in *Face Recognition: Methods, Applications and Technology* (A. Quaglia and C. M. Epifano, eds.), Computer Science, Technology and Applications, p. 9, Nova Publishers, 2012.

[5] T. Ali, L. J. Spreeuwers, R. N. J. Veldhuis, and D. Meuwly, “Effect of calibration data on forensic likelihood ratio from a face recognition system,” in *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems, BTAS 2013, Washington, DC, U.S.A.*, IEEE explore digital library, (United States), pp. 1–8, IEEE, September 2013.

[6] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, “The NIST speaker recognition evaluation overview, methodology, systems, results, perspective,” *Speech Communication*, vol. 31, no. 23, pp. 225 – 254, 2000.

[7] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The det curve in assessment of detection task performance,” pp. 1895–1898, 1997.

[8] N. Brümmer and J. du Preez., “Application-independent evaluation of speaker detection,” *Computer Speech & Language*, vol. 20, no. 2, pp. 230–275, 2006.

[9] D. A. van Leeuwen and N. Brümmer, “Speaker classification i,” ch. An Introduction to Application-Independent Evaluation of Speaker Recognition Systems, pp. 330–353, Berlin, Heidelberg: Springer-Verlag, 2007.

[10] D. Ramos and J. Gonzalez-Rodriguez, “Reliable support: Measuring calibration of likelihood ratios,” *Forensic Science International*, vol. 230, no. 13, pp. 156 – 169, 2013. EAFS 2012 6th European Academy of Forensic Science Conference The Hague, 20-24 August 2012.

[11] D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, and C. Aitken, “Information-theoretical assessment of the performance of likelihood ratio computation methods,” *Journal of Forensic Sciences*, vol. 58, no. 6, pp. 1503–1518, 2013.

[12] M. I. Mandasari, M. Günther, R. Wallace, R. Saeidi, S. Marcel, and D. A. van Leeuwen, “Score calibration in face recognition,” *IET Biometrics*, vol. 3, no. 4, pp. 246–256, 2014.

[13] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, “On combining classifiers,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226–239, Mar 1998.

[14] Y. Ma, B. Cukic, and H. Singh, “A classification approach to multibiometric score fusion,” in *Proceedings of the 5th International Conference on Audio- and Video-Based Biometric Person Authentication*, AVBPA’05, (Berlin, Heidelberg), pp. 484–493, Springer-Verlag, 2005.

[15] J. Neyman and E. S. Pearson, “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*, vol. 231, no. 694-706, pp. 289–337, 1933.

[16] K. Nandakumar, Y. Chen, S. Dass, and A. Jain, “Likelihood ratio-based biometric score fusion,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, pp. 342–347, Feb 2008.

[17] Y. Makihara, D. Muramatsu, H. Iwama, T. T. Ngo, Y. Yagi, and M. A. Hossain, “Score-level fusion by generalized delaunay triangulation,” in *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pp. 1–8, Sept 2014.

[18] M. Sklar, *Fonctions de Répartition À N Dimensions Et Leurs Marges.* Université Paris 8, 1959.

[19] Y. F. Xiaohong Chen, "Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection," *The Canadian Journal of Statistics*, vol. 33, no. 3, pp. 389–414, 2005.

[20] C. A. J. Klaassen and N. Susyanto, "Semiparametrically Efficient Estimation of Constrained Euclidean Parameters," *ArXiv e-prints*, Aug. 2015.

[21] C. A. J. Klaassen and N. Susyanto, "Semiparametrically Efficient Estimation of Euclidean Parameters under Equality Constraints," *ArXiv e-prints*, June 2016.

[22] C. van Eeden, *Restricted Parameter Space Estimation Problems: Admissibility and Minimaxity Properties.* Lecture Notes in Statistics, Springer New York, 2006.

[23] P. Bickel, C. Klaassen, Y. Ritov, and J. Wellner, *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins series in the mathematical sciences, Springer New York, 1998.

[24] C. A. J. Klaassen, "Consistent estimation of the influence function of locally asymptotically linear estimators," *The Annals of Statistics*, vol. 15, no. 4, pp. 1548–1562, 1987.

[25] P. H. D. Charles W. Cobb, "A theory of production," *The American Economic Review*, vol. 18, no. 1, pp. 139–165, 1928.

[26] R. Stone, *The Measurement of Consumers' Expenditure and Behaviour in the United Kingdom, 1920-1938.* No. no. 1 in Studies in the national income and expenditure of the United Kingdom, University Press, 1954.

[27] H. Nyquist, "Restricted estimation of generalized linear models," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 40, no. 1, pp. 133–141, 1991.

[28] J. M. G. T. Dong K. Kim, "The restricted em algorithm for maximum likelihood estimation under linear restrictions on the parameters," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 708–716, 1995.

[29] Z. Sidak, P. Sen, and J. Hajek, *Theory of Rank Tests.* Probability and Mathematical Statistics, Elsevier Science, 1999.

[30] R. A. Khan, “A remark on estimating the mean of a normal distribution with known coefficient of variation,” *Statistics*, vol. 49, no. 3, pp. 705–710, 2015.

[31] J. D. H. Leon Jay Gleser, “Estimating the mean of a normal distribution with known coefficient of variation,” *Journal of the American Statistical Association*, vol. 71, no. 356, pp. 977–981, 1976.

[32] R. A. Khan, “A note on estimating the mean of a normal distribution with known coefficient of variation,” *Journal of the American Statistical Association*, vol. 63, no. 323, pp. 1039–1041, 1968.

[33] C. A. Klaassen and J. A. Wellner, “Efficient estimation in the bivariate normal copula model: normal margins are least favourable,” *Bernoulli*, vol. 3, pp. 55–77, 02 1997.

[34] P. D. Hoff, X. Niu, and J. A. Wellner, “Information bounds for gaussian copulas,” *Bernoulli*, vol. 20, pp. 604–622, 05 2014.

[35] J. Segers, R. van den Akker, and B. J. M. Werker, “Semiparametric gaussian copula models: Geometry and efficient rank-based estimation,” *Ann. Statist.*, vol. 42, pp. 1911–1940, 10 2014.

[36] A. Schick, “On efficient estimation in regression models,” *Ann. Statist.*, vol. 21, pp. 1486–1521, 09 1993.

[37] G. Cheng, H. H. Zhang, and Z. Shang, “Sparse and efficient estimation for partial spline models with increasing dimension,” *Annals of the Institute of Statistical Mathematics*, vol. 67, no. 1, pp. 93–127, 2015.

[38] B. Y. Levit, “On the efficiency of a class of non-parametric estimates,” *Theory of Probability & Its Applications*, vol. 20, no. 4, pp. 723–740, 1976.

[39] Y. A. Koshevnik and B. Y. Levit, “On a non-parametric analogue of the information matrix,” *Theory of Probability & Its Applications*, vol. 21, no. 4, pp. 738–753, 1977.

[40] S. J. Haberman, “Adjustment by minimum discriminant information,” *Ann. Statist.*, vol. 12, pp. 971–988, 09 1984.

[41] A. Sheehy, “Kullbackleibler constrained estimation of probability measures,” 1988.

[42] U. U. Müller and W. Wefelmeyer, “Estimators for models with constraints involving unknown parameters,” vol. 11, no. 2, pp. 221–235, 2002.

[43] M. Broniatowski and A. Keziou, "Divergences and duality for estimation and test under moment condition models," *Journal of Statistical Planning and Inference*, vol. 142, no. 9, pp. 2554 – 2573, 2012.

[44] J. Aitchison and S. D. Silvey, "Maximum-likelihood estimation of parameters subject to restraints," *Ann. Math. Statist.*, vol. 29, pp. 813–828, 09 1958.

[45] M. Jamshidian, "On algorithms for restricted maximum likelihood estimation," *Computational Statistics & Data Analysis*, vol. 45, no. 2, pp. 137 – 157, 2004.

[46] J. D. Gorman and A. O. Hero, "Lower bounds for parametric estimation with constraints," *IEEE Transactions on Information Theory*, vol. 36, pp. 1285–1301, Nov 1990.

[47] T. L. Marzetta, "A simple derivation of the constrained multiple parameter cramer-rao bound," *IEEE Transactions on Signal Processing*, vol. 41, pp. 2247–2249, Jun 1993.

[48] P. Stoica and B. C. Ng, "On the cramer-rao bound under parametric constraints," *IEEE Signal Processing Letters*, vol. 5, pp. 177–179, July 1998.

[49] A. W. v. d. Vaart, *Asymptotic statistics.* Cambridge series in statistical and probabilistic mathematics, Cambridge (UK), New York (N.Y.): Cambridge University Press, 1998. Autre tirage : 2000 (dition broche), 2005, 2006, 2007.

[50] N. Susyanto, C. A. J. Klaassen, R. N. J. Veldhuis, and L. J. Spreeuwers, "Semiparametric score level fusion: Gaussian copula approach," in *Proceedings of the 36th WIC Symposium on Information Theory in the Benelux, Brussels*, (Brussels), pp. 26–33, Université Libre de Bruxelles, May 2015.

[51] S. G. Iyengar, P. K. Varshney, and T. Damarla, "A parametric copula-based framework for hypothesis testing using heterogeneous data," *IEEE Transactions on Signal Processing*, vol. 59, pp. 2308–2319, May 2011.

[52] N. Susyanto, R. N. J. Veldhuis, L. J. Spreeuwers, and C. A. J. Klaassen, "Fixed FAR correction factor of score level fusion," 2016. Accepted for publication at The 8th IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS2016).

[53] N. Susyanto, R. N. J. Veldhuis, L. J. Spreeuwers, and C. A. J. Klaassen, "Two-step calibration method for multi-algorithm score-based face recognition systems by minimizing discrimination loss," in *The 9th IAPR International Conference on Biometrics (ICB 2016), 13-16 June, Halmstad*, June 2016.

[54] G. S. Morrison, "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*, vol. 53, no. 2, pp. 242 – 256, 2011.

[55] G. S. Morrison, "Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio," *Australian Journal of Forensic Sciences*, vol. 45, no. 2, pp. 173–197, 2013.

[56] Q. Tao and R. N. J. Veldhuis, "Robust biometric score fusion by naive likelihood ratio via receiver operating characteristics," *IEEE Transactions on Information Forensics and Security*, vol. 8, pp. 305–313, February 2013.

[57] B. Zadrozny and C. Elkan, "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), pp. 694–699, ACM, 2002.

[58] L.-P. R. C. Genest, K. Ghoudi, "A semiparametric estimation procedure of dependence parameters in multivariate families of distributions," *Biometrika*, vol. 82, no. 3, pp. 543–552, 1995.

[59] J.-D. Fermanian, "Goodness-of-fit tests for copulas," *Journal of Multivariate Analysis*, vol. 95, no. 1, pp. 119 – 152, 2005.

[60] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 381–396, March 2002.

[61] National Institute of Standards and Technology, "Nist biometric scores set - release 1," 2004. Available at http://www.itl.nist.gov/iad/894.03/biometricscores.

[62] L. Spreeuwers, "Fast and accurate 3D face recognition," *International Journal of Computer Vision*, vol. 93, no. 3, pp. 389–414, 2011.

[63] L. Spreeuwers, “Breaking the 99% barrier: optimisation of three-dimensional face recognition,” *Biometrics, IET*, vol. 4, no. 3, pp. 169–178, 2015.

[64] N. Poh and S. Bengio, “Database, protocols and tools for evaluating score-level fusion algorithms in biometric authentication,” *Pattern Recogn.*, vol. 39, pp. 223–233, Feb. 2006.

[65] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, “The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition,” *Information Forensics and Security, IEEE Transactions on*, vol. 7, pp. 1511–1521, Oct 2012.

[66] N. Trung, Y. Makihara, H. Nagahara, R. Sagawa, Y. Mukaigawa, and Y. Yagi, “Performance evaluation of gait recognition using the largest inertial sensor-based gait database,” in *The 5th IAPR International Conference on Biometrics (ICB 2012), Mar 29 - Apr 1, New Delhi*, Mar 2012.

[67] N. Brümmer and E. de Villiers, “The BOSARIS toolkit: Theory, algorithms and code for surviving the new DCF,” *CoRR*, vol. abs/1304.2865, 2013.

[68] A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Multibiometrics (International Series on Biometrics).* Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[69] X. Lu, Y. Wang, and A. Jain, “Combining classifiers for face recognition,” in *Multimedia and Expo, 2003. ICME ’03. Proceedings. 2003 International Conference on*, vol. 3, pp. III–13–16 vol.3, July 2003.

[70] A. Ross, A. Jain, and J. Reisman, “A hybrid fingerprint matcher,” in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 3, pp. 795–798 vol.3, 2002.

[71] B. Ulery, A. Hicklin, C. Watson, W. Fellner, P. Hallinan, and C. M. Gutierrez, “Studies of biometric fusion nistir 7346,” 2006.

[72] S. Prabhakar and A. K. Jain, “Decision-level fusion in fingerprint verification,” *Pattern Recognition*, vol. 35, no. 4, pp. 861 – 874, 2002.

[73] C. A. Klaassen and J. A. Wellner, “Efficient estimation in the bivariate normal copula model: normal margins are least favourable,” *Bernoulli*, vol. 3, pp. 55–77, 02 1997.

[74] H. Shimazaki and S. Shinomoto, “Kernel bandwidth optimization in spike rate estimation,” *J. Comput. Neurosci.*, vol. 29, pp. 171–182, Aug. 2010.

[75] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics).* Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[76] W. Chen, Y. Chen, Y. Mao, and B. Guo, “Density-based logistic regression,” in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’13, (New York, NY, USA), pp. 140–148, ACM, 2013.

[77] K. Nandakumar, A. Ross, and A. K. Jain, “Biometric fusion: Does modeling correlation really matter?,” in *Biometrics: Theory, Applications, and Systems, 2009. BTAS ’09. IEEE 3rd International Conference on*, pp. 1–6, Sept 2009.

[78] L. D. Brown, T. T. Cai, and A. DasGupta, “Interval estimation for a binomial proportion,” *Statist. Sci.*, vol. 16, pp. 101–133, 05 2001.

[79] H. Joe, *Multivariate Models and Multivariate Dependence Concepts.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis, 1997.

[80] G. T. Chang and G. Walther, “Clustering with mixtures of log-concave distributions,” *Computational Statistics & Data Analysis*, vol. 51, no. 12, pp. 6242 – 6251, 2007.

[81] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain, “Fvc2002: Second fingerprint verification competition,” in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 3, pp. 811–814 vol.3, 2002.

[82] J. Abraham, J. Gao, and P. Kwan, *Fingerprint Matching Using A Hybrid Shape and Orientation Descriptor.* INTECH Open Access Publisher, 2011.

[83] P. Embrechts, “Copulas: A personal view,” *Journal of Risk and Insurance*, vol. 76, no. 3, pp. 639–650, 2009.

[84] Y. Fan, , and A. J. Patton, “Copulas in econometrics,” *Annual Review of Economics*, vol. 6, pp. 179–200, 2014.

Acknowledgements

This thesis would not have been completed without the supports of all people during my PhD journey. I would like to use this opportunity to thank:

- Prof. Chris Klaassen for his kindness, dedication and attention to detail have been a great inspiration to me. He always gave me guidance, support, and motivation in tough times of my PhD journey.

- Prof. Raymond Veldhuis and Dr. Luuk Spreeuwers also for their kindness, dedication and attention during my stay at the University of Twente. They always helped me to find the possible ”bridges” whenever I saw a ”biometric island” from my mathematical world.

- Members of my graduation committee for reading and reviewing my work.

- Other members of the Forensic Face Recognition project: Chris Zeinstra and Arnout Ruifrok.

- M. Kranenburg, E. Wallet, and M.E.M. Onderwater for making my research at our institute became easier.

- B.F.J. Scholten-Koop, S.E. Engbers, and G.J. Laanstra for answering my all questions during my stay at the University of Twente.

- My colleagues at the Korteweg-de Vries Institute UvA and at the SCS group UT.

- Department of Mathematics, Universitas Gadjah Mada, Yogyakarta, for

permitting and supporting me to take this PhD.

- Last but not least, terima kasih untuk keluarga di Nyamplung dan Traji yang selalu memberikan do'a dan semangat selama kami tinggal di Belanda. Khususnya untuk Ibukku tercinta Siti Fauzanah, terima kasih atas inspirasi, motivasi, dan do'a yang tak henti-hentinya engkau berikan kepada kami.

Hengelo, August 2016

Summary

Semiparametric Copula Models for Biometric Score Level Fusion

In biometric recognition, biometric samples (images of faces, fingerprints, voices, gaits, etc.) of people are compared and matchers (classifiers) indicate the level of similarity between any pair of samples by a 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. If we model the joint distribution of all scores by a (semiparametric) Gaussian copula model, the resulting correlation matrix will be structured. It has many zeros and many correlations have a common value. Estimation of these parameters is a problem in constrained semiparametric estimation, a topic that we study in quite some generality in the Statistical Theory part of this thesis. The Biometric Application part of it focuses on score level fusion and models the dependence between classifiers also by semiparametric copula models.

The Statistical part of this thesis studies semiparametric estimation of constrained Euclidean parameters. When the Euclidean parameter is known to lie on a general surface (image of a continuously differentiable vector valued function), the lower bound for estimators for the underlying parameter is obtained via the Hájek-LeCam Convolution Theorem for regular parametric models and subsequently an efficient estimator attaining this bound is constructed in terms of the original estimator and the function defining the surface.

A projection technique is used for computing the lower bound for the estimators when the Euclidean parameter belongs to the zero set of a continuously differentiable function (for which it might be impossible to parametrize it as the image of a continuously differentiable vector valued function) and an efficient estimator under this constraint is also provided in terms of an efficient estimator within the unconstrained model and the function defining the constraint.

The Biometric part proposes a semiparametric likelihood ratio-based score level fusion strategy by modelling the marginal individual likelihood ratios nonparametrically and the dependence between them by parametric copulas. The dependence parameter is estimated by pseudo-likelihood estimation and its convergence is discussed. A detailed procedure to train the proposed method is provided and applications on real data for the biometric standard verifications and in forensic scenarios are also demonstrated.

Samenvatting

Semiparametrische Copula Modellen voor Biometrische Score Level Fusion

Bij biometrische herkenning worden biometrische *sample*s (vingerafdrukken, opnamen van gezichten, stemmen, manieren van lopen, etc.) vergeleken en geven vergelijkers (matchers) door middel van een score de mate van gelijkenis aan tussen paren van *samples*.

Als twee *samples* van eenzelfde persoon worden vergeleken, wordt een *genuine score* verkregen. Bij het vergelijken van *samples* van verschillende personen heet de resulterende score een *impostor score*. Als we de simultane verdeling van alle scores modelleren met een (semiparametrisch) *Gaussian copula* model, dan is de resulterende correlatiematrix gestructureerd. Deze heeft veel nullen en veel correlaties hebben dezelfde waarde.

Het schatten van deze waarden is een probleem in het semiparametrisch schatten met voorwaarden op de parameters. Dit is een onderwerp dat we in zijn volle algemeenheid bestuderen in het Statistische Theorie-deel van het proefschrift. Het Biometrische Toepassingen-deel richt zich op het samenvoegen van scores en modelleert de onderlinge afhankelijkheid van vergelijkers ook door middel van semiparametrische *copula* modellen.

Het statistische deel van dit proefschrift bestudeert het semiparametrisch schatten van ingeperkte Euclidische parameters. Wanneer bekend is dat de Euclidische parameter op een algemeen oppervlak (het beeld van een continu-differentieerbare vectorwaardige functie) ligt, krijgen we de ondergrens voor schatters van de onderliggende parameter via de Hájek-LeCam Convolutie

Stelling voor reguliere parametrische modellen. Vervolgens wordt een efficiënte schatter geconstrueerd die deze ondergrens bereikt en die gedefinieerd is in termen van een efficiënte schatter binnen het niet-ingeperkte model en van de functie die het oppervlak definieert. Een projectietechniek wordt gebruikt om de ondergrens te bepalen voor schatters, wanneer de Euclidische parameter behoort tot de nulpuntenverzameling van een continu-differentieerbare functie (waarvoor het onmogelijk kan zijn om die te herparametriseren als beeld van een continu-differentieerbare vectorwaardige functie), en ook wordt een efficiënte schatter onder zo'n inperking gegeven in termen van een efficiënte schatter binnen het niet-ingeperkte model en van de functie die de inperking definieert.

Het Biometrische deel stelt een semiparametrische op *likelihood ratio* gebaseerde samenvoegingsstrategie voor scores voor die de individuële *likelihood ratio* niet-parametrisch modelleert en hun onderlinge afhankelijkheden via parametrische koppelingsfuncties. De afhankelijkheidsparameters worden met *pseudo-likelihood estimation* geschat en hun convergentie wordt besproken. Een gedetailleerde procedure wordt beschreven om de voorgestelde methode te trainen en toepassingen op echte gegevens worden gepresenteerd voor standaard biometrische verificatie en voor forensische situaties.

ISBN: 978-94-6295-513-4