Asymptotic optimal classification of three populations by ML-estimators

I. Steinke
Department of Mathematics, University of Rostock, Germany

Abstract: Let π_1 and π_2 be independent populations distributed according to Q_{ϑ_1} and Q_{ϑ_2} with unknown parameters ϑ_1, $\vartheta_2 \in \mathcal{H}$ and let π_3 be a third population whose distribution Q_{ϑ_3} coincides with that of either π_1 or π_2. Assume there are given samples of size $n_1 = n_2$ and n_3, it is to be decided whether $Q_{\vartheta_3} = Q_{\vartheta_1}$ or $Q_{\vartheta_3} = Q_{\vartheta_2}$.

The decision problem is described by a sequence of localized experiments. The corresponding lower Hajek-LeCam-bound is established and there is constructed a sequence of decision rules based on ML-estimators which attains this bound and is therefore considered to be asymptotic optimal.

AMS Subject Classification: 62F05, 62F12

Keywords: Classification of 3 populations, Hajek-LeCam bound, weakly convergent experiments

1 Introduction

Let $\{Q_\vartheta, \vartheta \in \mathcal{H}\}$ be a family of probability measures defined on a measurable space $(\mathcal{X}, \mathfrak{A})$ and π_1, π_2 be independent populations distributed according to Q_{ϑ_1} and Q_{ϑ_2}, respectively, where the parameters ϑ_1, $\vartheta_2 \in \mathcal{H}$ are unknown. π_3 is a third population that is distributed according to either Q_{ϑ_1} or Q_{ϑ_2}. Taking samples from each population one has to decide whether π_3 has distribution Q_{ϑ_1} or Q_{ϑ_2}.

To identify the correct distribution it is natural to apply the following classificatory procedure: By means of samples of π_1, π_2 and π_3 the parameters ϑ_1, ϑ_2 and ϑ_3 are estimated. Let $\widehat{\vartheta}_i$ be an estimator for ϑ_i, $i = 1, 2, 3$. Then one decides for the population π_1 if $\widehat{\vartheta}_3$ lies closer to $\widehat{\vartheta}_1$ than to $\widehat{\vartheta}_2$ and, conversely, it is decided for population π_2 if $\widehat{\vartheta}_3$ is closer to $\widehat{\vartheta}_2$ than to $\widehat{\vartheta}_1$.

It is clear that in this approach, "good" estimators will lead to "good" classification rules. As maximum likelihood estimators (MLE) are efficient under some regularity conditions one conjectures that classification rules based on ML-estimators are best in some sense. It is the motivation of this paper to show that this natural classificatory procedure is under certain conditions an optimal one. It turns out that the necessary conditions for this optimality essentially coincide with that needed for asymptotic normality of ML-estimators. Using methods from asymptotic decision theory for sequences of localized models an asymptotic lower bound for the risk of sequences of classification rules is established. Then optimality of the classification rule based on maximum likelihood estimators is shown to be asymptotically optimal in the sense that the sequence of risks tends to the universal lower bound.

2 Notations and results

Because of asymptotic normality of maximum likelihood estimators we first consider the case that the underlying distributions are univariate normal, i.e. $\mathcal{H} = I\!R_1$ and $Q_h = N(h, \sigma^2)$, where $N(h, \sigma^2)$ represents the univariate normal distribution with mean h and variance σ^2. Let π_i, $i = 1, 2, 3$, be distributed according to $N(h_i, \sigma^2)$, where $h_3 = h_1$ or $h_3 = h_2$, and denote the sample from π_i, $i = 1, 2$, by $X_{i,1}, \dots, X_{i,m}$. Let furthermore $Y_1, \dots, Y_n$ be the sample from π_3, i.e. $\mathcal{L}(Y_i) = N(h_\tau, \sigma^2)$ and $\tau = 1$ or $\tau = 2$. For fixed σ^2 the arithmetic means $\overline{X}_{i,m} = \sum_{j=1}^m X_{i,j}/m$, $\overline{Y}_n = \sum_{j=1}^n Y_j/n$ are sufficient statistics for the distributions $\mathcal{L}(X_{i,1}, \dots, X_{i,m})$ and $\mathcal{L}(Y_1, \dots, Y_n)$, respectively, with $\mathcal{L}(\overline{X}_{i,m}) = N(h_i, \sigma^2/m)$, $\mathcal{L}(\overline{Y}_n) = N(h_\tau, \sigma^2/n)$. Therefore the decision problem may be reduced to the case of sample sizes 1 but, if $m \neq n$, unequal variances for the reference populations π_1, π_2 on one side and π_3 on the other side.

Hence we may assume in the following that there is given a sample X_1, X_2, Y of independent random variables such that $\mathcal{L}(X_i) = N(h_i, \sigma^2)$, $\mathcal{L}(Y) = N(h_\tau, \sigma_0^2)$, $i, \tau \in \{1, 2\}$, where the h_i are unknown and one has to decide whether the mean of Y corresponds to the mean of X_1 or X_2.

To perform the statistical procedure of classification a classification function is applied, as was proposed by Kudo (1959), in analogy to that of randomized test functions in hypothesis testing. Let $\mathfrak{B}_k$ denote the σ-algebra of Borel sets of $I\!R_k$.

Definition 2.1 *A* ***classification function*** *q (for the classification problem discussed above) is a two-dimensional vector function, $q : I\!R_3 \to I\!R_2$, $q = (q_1, q_2)$, satisfying the following conditions:*

1. *q is $(I\!R_3, \mathfrak{B}_3)$-$(I\!R_2, \mathfrak{B}_2)$-measurable,*
2. *$q_1 + q_2 = 1$ and $q_i \geq 0$, i=1,2.*

$q_i(x_1, x_2, y)$ is the probability to decide for population π_i if x_1, x_2, and y are observed from π_1, π_2, and π_3, respectively. From the decision theoretical point of view the classification problem may be stated in the following way:

Let $P_\theta = N(h_1, \sigma^2) \times N(h_2, \sigma^2) \times N(h_\tau, \sigma_0^2)$, $\theta = (h_1, h_2, \tau) \in I\!R_2 \times \{1, 2\}$, denote the common multivariate distribution of X_1, X_2, and Y. Then

$$E = (I\!R_3, \mathfrak{B}_3, P_\theta, \theta \in I\!R_3 \times \{1, 2\})$$

is called the corresponding **experiment** of the decision problem and $I\!D = \{1, 2\}$ is said to be the **decision space**. Let $\mathcal{C}(E, I\!D)$ denote the set of all classification functions for this decision problem. If we introduce the loss function, $\theta = (h_1, h_2, \tau) \in I\!R_2 \times I\!D$,

$$L(\theta, d) = L(h_1, h_2, \tau, d) = \begin{cases} 1 \ , if \ \tau \neq d \\ 0 \ , if \ \tau = d \end{cases} , \tag{1}$$

we get for the risk function of the classification rule q

$$R(\theta, q) = R(h_1, h_2, \tau, q) = \int \sum_{i=1}^{2} L(\theta, i)\, q_i dP_\theta = 1 - \int q_\tau \, dP_\theta \ . \tag{2}$$

By $\overline{q} = (\overline{q}_{n,1}, \overline{q}_{n,2})$,

$$\bar{q}_1(x_1, x_2, y) = \begin{cases} 1 \ , if \ |x_1 - y| < |x_2 - y| \ , \\ \frac{1}{2} \ , if \ |x_1 - y| = |x_2 - y| \ , \\ 0 \ , if \ |x_1 - y| > |x_2 - y| \ , \end{cases} \tag{3}$$

$\overline{q}_{n,2} = 1 - \overline{q}_{n,1}$, is denoted the so-called "next-neighbour-rule". Note that x_1, x_2, and y represent in this case the estimators for the means h_1, h_2, and h_3, respectively, of the populations. Kudo (1959) proved that $\overline{q}$ is the uniformly best classification function within a certain class of classification functions:

Theorem 2.2 *(Kudo, 1959)*
Let $\mathcal{C}$ be the class of classification functions q satisfying the following conditions

1. $q(x_1 + b, x_2 + b, y + b) = q(x_1, x_2, y)$ *for all* $x_1, x_2, y \in \mathbb{R}_1$,
2. $R(h_1, h_2, 1, q) = R(h_1, h_2, 2, q)$ *for all* $h_1, h_2 \in \mathbb{R}_1$,
3. $R(h_1, h_2, \tau, q)$ *depends only through* $|h_1 - h_2|$ *on* h_1 *and* h_2.

Then the "next-neighbour-rule" $\overline{q}$ belongs to $\mathcal{C}$ and is uniformly the best in $\mathcal{C}$, that means

$$R(h_1, h_2, \tau, \bar{q}) \leq R(h_1, h_2, \tau, q) \ \ \textit{for all } h_1, h_2 \in \mathbb{R}_1, \ \tau \in \{1, 2\} \tag{4}$$

Similar conditions to that of Theorem 2.2 can be expressed by transformation groups. A classification function is said to satisfy the invariance condition (T), (S), or (P), respectively, iff

$$\begin{aligned} \textbf{(T)} \quad q_i(x_1, x_2, y) &= q_i(x_1 + b, x_2 + b, y + b) \ \text{ for all } x_1, x_2, y, b \in \mathbb{R}_1, \\ \textbf{(S)} \quad q_i(x_1, x_2, y) &= q_i(-x_1, -x_2, -y) \ \text{ for all } x_1, x_2, y \in \mathbb{R}_1, \\ \textbf{(P)} \quad q_1(x_1, x_2, y) &= q_2(x_2, x_1, y) \ \text{ for all } x_1, x_2, y \in \mathbb{R}_1. \end{aligned}$$

Lemma 2.3 *If $q \in \mathcal{C}(E, \mathbb{D})$ satisfies $(T), (S)$, and (P) then q satisfies the conditions 1.- 3. of Theorem 2.2 and (4) holds true.*

The proof of Lemma 2.3 will be given in section 3.

The uniform optimality of $\overline{q}$ with respect to the parameter space holds only in the class $\mathcal{C}$. But one can ask whether $\overline{q}$ satisfies a weaker minimax optimality in $\mathcal{C}(E, \mathbb{D})$, the family of all classification rules on E. Let $c > \Delta$ be positive real numbers and

$$\Theta_\Delta \ := \ \{(h_1, h_2, \tau) \in \mathbb{R}_2 \times \mathbb{D} : \ |h_1 - h_2| \geq \Delta\} \tag{5}$$

$$K_c \ := \ \{(h_1, h_2, \tau) \in \mathbb{R}_2 \times \mathbb{D} : \ |h_1| \leq c, |h_2| \leq c\} \tag{6}$$

Theorem 2.4 *Given the decision problem described above, it holds*

$$\inf_{q \in \mathcal{C}(E,I\!D)} \sup_{\theta \in \Theta_\Delta} R(\theta, q) = \sup_{\theta \in \Theta_\Delta} R(\theta, \overline{q}) = \lim_{c \uparrow \infty} \inf_{q \in \mathcal{C}(E,I\!D)} \sup_{\theta \in \Theta_\Delta \cap K_c} R(\theta, q) \ . \tag{7}$$

It should be noted that Theorem 2.4 may be derived from the generalized Theorem of Hunt and Stein (e.g. see Strasser (1985), chapter 48). But in order to avoid several additional sophisticated notions from the theory of amenable groups, we give a direct proof in section 3.

Now we will consider the situation of more general underlying distributions. Then, in general, it is only possible to provide asymptotic results. Let $\{Q_\vartheta, \vartheta \in \mathcal{H}\}$, $\mathcal{H} \subseteq I\!R_1$, be a family of probability distributions on a measurable space $(\mathcal{X}, \mathfrak{A})$ and π_i, $i = 1, \ldots, 3$, independent populations distributed according to Q_{ϑ_i} where ϑ_i belongs to $\mathcal{H}$ and $\vartheta_3 = \vartheta_1$ or $\vartheta_3 = \vartheta_2$. Assume that there are samples $X_{n,1}^{(i)}, \ldots, X_{n,n}^{(i)}$ from π_i, $i = 1, 2$, of size n and a sample $Y_{n,1}, \ldots, Y_{n,k_n}$ from π_3 of size k_n. Then it is to be decided whether π_3 should be assigned to either π_1 or π_2. In addition, we demand that

$$\lim_{n \to \infty} \frac{k_n}{n} = \alpha > 0 \ . \tag{8}$$

For a suffiently large number of observations and a reasonable classificatory procedure it may be assumed that the corresponding error probabilities will tend to zero. Therefore, in order to measure the power of a sequence of classification rules, a localized parametrization is introduced. Let ϑ_0 belong to the interior kernel $\overset{\circ}{\mathcal{H}}$ of $\mathcal{H}$ and put $\mathcal{H}_n(\vartheta_0) = \{h \in I\!R_1 | \vartheta_0 + \frac{h}{\sqrt{n}} \in \mathcal{H}\}$, $\theta = (h_1, h_2, \tau) \in \mathcal{H}_n(\vartheta_0) \times \mathcal{H}_n(\vartheta_0) \times I\!D$ and $P_{n,\theta} = Q^n_{\vartheta_0 + h_1/\sqrt{n}} \times Q^n_{\vartheta_0 + h_2/\sqrt{n}} \times Q^{k_n}_{\vartheta_0 + h_\tau/\sqrt{n}}$. Then the experiment

$$E_n = (\Omega^{2n+k_n}, \mathfrak{A}^{2n+k_n}, P_{n,\theta}, \theta \in \mathcal{H}_n^2(\vartheta_0) \times I\!D)$$

describes the distribution of the underlying decision problem. Here Q^n_ϑ, Ω^n, and $\mathfrak{A}^n$ denote the n-fold direct product of the probability measures Q_ϑ, sample sets Ω, and σ-algebras $\mathfrak{A}$, respectively. In analogy to Definition 2.1 a map $q_n : \Omega^{2n+k_n} \to I\!R_2$, $q_n = (q_{n,1}, q_{n,2})$, is said to be a **classification function of the experiment $\boldsymbol{E_n}$** if q_n is $\mathfrak{A}^{2n+k_n}$- $\mathfrak{B}_2$-measurable and $q_{n,i} \geq 0, i = 1, 2$, as well as $q_{n,1} + q_{n,2} = 1$. Then $\mathcal{C}(E_n, I\!D)$ denotes the set of all classification functions for the experiment E_n. Using the loss function given in (1) the risk function is defined to be

$$R_n(\theta, q_n) = R_n(h_1, h_2, \tau, q_n) = 1 - \int q_{n,\tau} dP_{n,\theta} \ . \tag{9}$$

Impose the following conditions on the family of probability distributions $\{Q_\vartheta, \vartheta \in \Theta\}$:

(A1) $\{Q_\vartheta, \vartheta \in \mathcal{H}\}$ is dominated by a σ-finite measure μ and there exist a version f_ϑ of the density $dQ_\vartheta / d\mu$ such that $f_\vartheta > 0$ $\mu - a.s.$ and $\partial f_\vartheta / \partial \vartheta$ exists μ-a.s in an neighbourhood $U(\vartheta_0)$ of $\vartheta_0 \in \mathcal{H}$.

(A2) For all $\vartheta' \in U(\vartheta_0)$ holds

$$0 < I(\vartheta') := \int \left(\frac{\partial}{\partial \vartheta} \ln f_\vartheta \Big|_{\vartheta=\vartheta'} \right)^2 f_{\vartheta'} d\mu < \infty$$

and $I(\vartheta)$ is continuous in ϑ in $U(\vartheta_0)$.

$I(\vartheta)$ is called the **Fisher-Information** of $\{Q_\vartheta, \vartheta \in \mathcal{H}\}$ in ϑ. If **(A1)** and **(A2)** are fulfilled we define

$$\begin{aligned} Z_n \quad := \quad & \Big(\frac{1}{\sqrt{n I(\vartheta_0)}} \sum_{i=1}^{n} \frac{\partial}{\partial \vartheta} \ln f_\vartheta \Big|_{\vartheta=\vartheta_0} (X_{n,i}^{(1)}), \frac{1}{\sqrt{n} I(\vartheta_0)} \sum_{i=1}^{n} \frac{\partial}{\partial \vartheta} \ln f_\vartheta \Big|_{\vartheta=\vartheta_0} (X_{n,i}^{(2)}), \\ & \frac{1}{\sqrt{n \alpha I(\vartheta_0)}} \sum_{i=1}^{k_n} \frac{\partial}{\partial \vartheta} \ln f_\vartheta \Big|_{\vartheta=\vartheta_0} (Y_{n,i}) \Big) \end{aligned} \tag{10}$$

For any $n \in I\!N$ let P_n and Q_n probability measures defined on the same measurable spaces $(\Omega_n, \mathfrak{A}_n)$. Q_n is said to be **contiguous** w.r.t. P_n ($Q_n \triangleleft P_n$) iff for any sequence $(A_n)_{n \in I\!N}$, $A_n \in \mathfrak{A}_n$, holds:

$$\text{if} \quad P_n(A_n) \xrightarrow[n \to \infty]{} 0 \ , \quad \text{then} \quad Q_n(A_n) \xrightarrow[n \to \infty]{} 0 \ .$$

If $X_n : (\Omega_n, \mathfrak{A}_n) \to (I\!R_1, \mathfrak{B}_1)$ is a sequence of random variables and $Q_n \triangleleft P_n$ then $X_n \xrightarrow{P_n} 0$, i.e. for any $\varepsilon > 0$ $\lim_{n \to \infty} P_n(|X_n| > \varepsilon) = 0$, implies $X_n \xrightarrow{Q_n} 0$.

Let $P_{n,0} = Q_{\vartheta_0}^{2n+k_n}$, $\Theta = I\!R_2 \times I\!D$, $\Theta_n = \mathcal{H}_n(\vartheta_0) \times I\!D$

$$P_\theta = N(h_1, I^{-1}(\vartheta_0)) \times N(h_2, I^{-1}(\vartheta_0)) \times N(h_\tau, \tfrac{1}{\alpha} I^{-1}(\vartheta_0)) \ , \tag{11}$$

$\theta = (h_1, h_2, \tau) \in \Theta$, and

$$E = (I\!R_3, \mathfrak{B}_3, P_\theta, \theta \in \Theta) \ .$$

Then under **(A1)** and **(A2)** the asymptotic decision theory of LeCam provides the following result:

Theorem 2.5 *Under **(A1)** and **(A2)**, for any sequence $(q_n)_{n \in I\!N}$ of classification functions and any subset H_0, it holds*

$$\liminf_{n \to \infty} \sup_{\theta \in \Theta_n \mathcal{H}_0} R_n(\theta, q_n) \geq \inf_{q \in \mathcal{C}(E, I\!D)} \sup_{\theta \in \mathcal{H}_0} R(\theta, q) \ , \tag{12}$$

where $R(\theta, q)$ and $R_n(\theta, q)$ are defined as in (2) and (9). Furthermore, for any convergent sequence $(\theta_n)_{n \in I\!N}$, $\theta_n = (h_{n,1}, h_{n,2}, \tau) \in \Theta_n$, $h_{n,1} \to h_1$, $h_{n,2} \to h_2$, $\theta = (h_1, h_2, \tau) \in I\!R_2 \times I\!D$,

$$\mathcal{L}(Z_n | P_{n,\theta_n}) \Longrightarrow P_\theta \quad \textit{and} \quad P_{n,\theta_n} \triangleleft P_{n,0} \tag{13}$$

Here "$\Rightarrow$" denotes the weak convergence of the corresponding probability measures.

For a proof of Theorem 2.5 we refer to Strasser (1985), chapter 10 and chapter 13, dealing with asymptotic decision theory and asymptotic normality of sequences of experiments , respectively, in a more general way. Or see Witting (1985), Theorem 1.182, to convince yourself that **(A1)** and **(A2)** are sufficient for the L_2-differentiablility of the family $\{Q_\vartheta, \vartheta \in \mathcal{H}\}$ and Rüschendorf (1987), chapter 4, Theorem 4.2, for the asymptotic normality of the sequence of experiments E_n and (13), where the different sample sizes have to be taken into account. Finally Strasser (1985), Theorem 62.5, gives (12).

Note that $\mathcal{H}_0$ is any subset of Θ and that the lower bound for a sequence of decision rules depends on the choice of $\mathcal{H}_0$. In order to construct classification rules which attain the lower bound in (12), a sequence of statistics is used whose the distributions converge weakly under $P_{n,\theta}$ to the normal distributions given in (11). Note that Z_n fulfils this condition but it still depends on ϑ_0 which is assumed to be unknown.

(12) provides a lower bound for a sequence of classification functions for the experiments E_n. A sequence $(q_n)_{n\in\mathbb{N}}$, $q_n \in \mathcal{C}(E_n, \mathbb{D})$, of classification rules is now called to be **optimal** w.r.t. $\mathcal{H}_0$ iff

$$\lim_{n\to\infty} \sup_{\theta\in\Theta_n\cap\mathcal{H}_0} R_n(\theta, q_n) = \inf_{q\in\mathcal{C}(E,D)} \sup_{\theta\in\Theta\cap\mathcal{H}_0} R(\theta, q)$$

attains this lower bound.

Let $\widehat{\theta}_{n,1}, \widehat{\theta}_{n,2}$, and $\widehat{\theta}_{n,3}$ be a sequences of consistent ML-estimators, based on $Q^n_{\vartheta_0}, Q^n_{\vartheta_0}$, and $Q^{k_n}_{\vartheta_0}$, respectively, for $\vartheta_0 \in \mathcal{H}$. By

$$\begin{aligned} \overline{q}_n(X^{(1)}_{n,1}, \dots, X^{(2)}_{n,1}, \dots, Y_{n,k_n}) &= \overline{q}(\widehat{\theta}_{n,1}, \widehat{\theta}_{n,2}, \widehat{\theta}_{n,3}) \\ &= \overline{q}(\sqrt{n}(\widehat{\theta}_{n,1} - \vartheta_0), \sqrt{n}(\widehat{\theta}_{n,2} - \vartheta_0), \sqrt{n}(\widehat{\theta}_{n,3} - \vartheta_0)) \end{aligned} \tag{14}$$

is given a sequence of decision rules for our localized problem. Let Θ_Δ and K_c be defined as in (5) and (6).

Theorem 2.6 *Let the family of probability measures $\{Q_\vartheta, \vartheta \in \mathcal{H}\}$ fulfil the following conditions:* ***(A1)***, ***(A2)***, *and there is a neighbourhood $U(\vartheta_0)$ of ϑ_0, such that*

(A3) $\frac{\partial^2}{\partial\vartheta^2} f_\vartheta$ *exists and is continuous in ϑ, $\vartheta_0 \in U(\vartheta_0)$,*

(A4) $\int \frac{\partial}{\partial\vartheta} f_\vartheta d\mu = \int \frac{\partial^2}{\partial\vartheta^2} f_\vartheta d\mu = 0 \ \forall\vartheta \in U(\vartheta_0)$

(A5) *For any $\vartheta \in U(\vartheta_0)$ there is a $\delta_\vartheta > 0$ with $U_{\delta_\vartheta}(\vartheta) \subset U(\vartheta_0)$ and a measurable function $M_\vartheta(z)$ with $\int |M_\vartheta(X_1)|\, dQ_\vartheta < \infty$, such that*

$$\left| \frac{\partial}{\partial\vartheta} \ln f_\vartheta(z) \Big|_{\vartheta=\vartheta'} \right| \leq M_\vartheta(z) \ \text{ for all } \vartheta' \in U_{\delta_\vartheta}(\vartheta)$$

Then for consistent ML-estimators $\widehat{\theta}_{n,i}$ for ϑ_0 and $\overline{q}_n$ from (14) holds for any $c > \Delta$

$$\lim_{n\to\infty} \sup_{\theta\in\Theta_n\cap K_c\cap\Theta_\Delta} |R_n(\theta, \overline{q}_n) - R(\theta, \overline{q})| = 0 \qquad \text{and} \tag{15}$$

$$\lim_{c\to\infty} \lim_{n\to\infty} \sup_{\theta\in\Theta_\Delta\cap K_c\cap\Theta_n} R_n(\theta, \overline{q}_n) = \lim_{c\to\infty} \inf_{q\in\mathcal{C}(E,D)} \sup_{\theta\in\Theta_\Delta\cap K_c} R(\theta, q) = \sup_{\theta\in\Theta_\Delta} R(\theta, \overline{q}) \tag{16}$$

(15) means that the risk function $R_n(\theta, \overline{q}_n)$ converge uniformly on the compact parameter set $K_c \cap \Theta_\Delta$ to the risk function of the next-neighbour-rule for normal distributed populations where c may be chosen arbitrarily large while (refmain2) states the convergence of

the maximal risks of $\overline{q}_n$ to the minimax risk of the normal distributed decision problem if c tends to infinity.

The conditions **(A1)** - **(A5)** correspond to that of Witting, Noelle (1970) and guarantee that any consistent sequence of ML-estimators is even asymptotically normal distributed. In any special situation the conditions of Theorem 2.6 have to be established. Especially, it is easy to show that these conditions are met by an one-dimensional exponential family with natural parametrization, i.e. if $\{Q_\vartheta, \vartheta \in \mathcal{H}\}$ is dominated and there is a dominating measure μ such that its density w.r.t. μ may be represented in the form

$$\frac{dQ_\vartheta}{d\mu}(x) = \exp\{x\vartheta - K(\vartheta)\} \ .$$

3 Proofs of the auxiliary results

Proof of Lemma 2.3: First it is shown that the following relations hold true:
If $q \in \mathcal{C}(E, I\!D)$ satisfies

1. (T) $\Longrightarrow R(h_1, h_2, \tau, q) = R(h_1 + b, h_2 + b, \tau, q)$ for all $h_1, h_2, b \in I\!R_1, \tau \in I\!D$.
2. (S) $\Longrightarrow R(h_1, h_2, \tau, q) = R(-h_1, -h_2, \tau, q)$ for all $h_1, h_2 \in I\!R_1$, $\tau \in I\!D$.
3. (P) $\Longrightarrow R(h_1, h_2, \tau, q) = R(h_2, h_1, 3 - \tau, q)$ for all $h_1, h_2 \in I\!R_1, \tau \in I\!D$.

For $\theta = (h_1, h_2, \tau) \in I\!R_2 \times I\!D$, let

$$N_\theta = N(h_1, \sigma^2) \times N(h_2, \sigma^2) \times N(h_\tau, \sigma_0^2) \ ,$$

$\theta' = (h_2, h_1, \tau)$, $\overline{\theta} = (-h_1, -h_2, \tau)$, and for any $b \in I\!R_1$: $\theta_b = (h_1 + b, h_2 + b, h_\tau + b)$. Then

$$\begin{aligned}
(1)\ R(h_1 + b, h_2 + b, \tau, q) &= 1 - \iiint q_\tau(x_1, x_2, y) N_{\theta_b}(dx_1, dx_2, dy) \\
&= 1 - \iiint q_\tau(x_1 + b, x_2 + b, y + b) N_\theta(dx_1, dx_2, dy) \\
&= R(h_1, h_2, \tau, q) \\
(2) \qquad R(-h_1, -h_2, \tau, q) &= 1 - \iiint q_\tau(x_1, x_2, y) N_{\overline{\theta}}(dx_1, dx_2, dy) \\
&= 1 - \iiint q_\tau(-x_1, -x_2, -y) N_\theta(dx_1, dx_2, dy) = R(h_1, h_2, \tau, q) \\
(3) \quad R(h_2, h_1, 3 - \tau, q) &= 1 - \iiint q_{3-\tau}(x_1, x_2, y) N_{\theta'}(dx_1, dx_2, dy) \\
&= 1 - \iiint q_\tau(x_2, x_1, y) N_\theta(dx_2, dx_1, dy) = R(h_1, h_2, \tau, q)
\end{aligned}$$

It follows: if $q \in \mathcal{C}(E, I\!D)$ satisfies (T), (S) and (P) then holds for all $h_1, h_2 \in I\!R_1$, $\tau \in I\!D$,

$$\begin{aligned}
R(h_1, h_2, 1, q) &\overset{(3)}{=} R(h_2, h_1, 2, q) \overset{(2)}{=} R(-h_2, -h_1, 2, q) \overset{(1)}{=} R(h_1, h_2, 2, q) \\
R(h_1, h_2, \tau, q) &\overset{(1)}{=} R(h_1 - h_2, 0, \tau, q) \overset{(2)}{=} R(h_2 - h_1, 0, \tau, q)
\end{aligned}$$

Hence $R(h_1, h_2, \tau, q) = R(|h_1 - h_2|, 0, \tau, q)$. $\square$

Proof of Theorem 2.4: The proof is carried out into 3 steps.
(1) Let first $\tilde{q}$ denote a classification function that is invariant w.r.t. translations, i.e. it satisfies (T). We define by

$$\widehat{q}_1(x_1, x_2, y) = \frac{\tilde{q}_1(x_1, x_2, y) + \tilde{q}_1(-x_1, -x_2, -y) + \tilde{q}_2(x_2, x_1, y) + \tilde{q}_2(-x_2, -x_1, -y)}{4} \ ,$$

$\widehat{q}_2 = 1 - \widehat{q}_1$, a classification function that fulfils all assumptions of Lemma 2.3. Therefore

$$\sup_{\theta \in \Theta_\Delta} R(\theta, \overline{q}) \leq \sup_{\theta \in \Theta_\Delta} R(\theta, \widehat{q}) \leq \sup_{\theta \in \Theta_\Delta} R(\theta, \tilde{q})$$

Hence $\overline{q}$ is minimax-optimal within all classification functions satisfying (T).
(2) Let now q be any classifcation function. Then $(q_n)_{n \in I\!N}$,

$$q_{n,i}(x_1, x_2, y) = \frac{1}{2n} \int_{-n}^{n} q_i(x_1 + t, x_2 + t, y + t) dt \ ,$$

is a sequence of classification functions. Let $b \in I\!R_1$ be fixed.

$$\begin{aligned}
&|q_{n,i}(x_1 + b, x_2 + b, y + b) - q_{n,i}(x_1, x_2, y)| \\
&= \ |\frac{1}{2n} \int_{-n}^{n} (q_{n,i}(x_1 + t + b, x_2 + t + b, y + t + b) - q_{n,i}(x_1 + t, x_2 + t, y + t))dt| \\
&= \ \frac{1}{2n} \ | \int_{n}^{n+b} q_{n,i}(x_1 + t, x_2 + t, y + t)dt - \int_{-n}^{-n+b} q_{n,i}(x_1 + t, x_2 + t, y + t)dt| \\
&= \ \frac{|b|}{n} \longrightarrow 0 \quad , \ n \to \infty \ .
\end{aligned} \tag{17}$$

By Witting (1985), Corollary 2.15, the class of test functions which corresponds to the class of classification functions is weakly sequential compact, i.e. for any sequence of classification functions (p_n) there is a subsequence (p_{n_k}) and a clasification function p_0 such that

$$\lim_{k \to \infty} \int p_{n_k} d\mu \longrightarrow \int p_0 d\mu \tag{18}$$

for any finite measure μ on $(I\!R_3, \mathfrak{B}_3)$ dominated by the Lebesgue-measure λ_3. Let for any $\underline{h} = (h_1, h_2, h_3) \in I\!R_3$

$$N_{\underline{h}} = N(h_1, \sigma^2) \times N(h_2, \sigma^2) \times N(h_3, \sigma_0^2) \ .$$

Then there is a classification function $\check{q}$ and a subsequence (q_{n_k}) of (q_n) such that

$$\lim_{k \to \infty} \iiint q_{n_k,i}(x_1, x_2, y) N_{\underline{h}}(dx_1, dx_2, dy) = \iiint \check{q}_i(x_1, x_2, y) N_{\underline{h}}(dx_1, dx_2, dy)$$

for all $h_1, h_2, h_3 \in I\!R_1$. Therefore, by (17), (18) and Lebegues's Theorem

$$\begin{aligned}
&| \iiint (\check{q}_i(x_1 + b, x_2 + b, y + b) - \check{q}_i(x_1, x_2, y)) N_{\underline{h}}(dx_1, dx_2, dy)| \\
&= \ \lim_{k \to \infty} | \iiint \left[q_{n_k,i}(x_1 + b, x_2 + b, y + b) - q_{n_k,i}(x_1, x_2, y) \right] N_{\underline{h}}(dx_1, dx_2, dy)| = 0 \ .
\end{aligned}$$

Put $\mathcal{Q} = \{N(h_1, \sigma^2) \times N(h_2, \sigma^2) \times N(h_3, \sigma_0^2), h_1, h_2, h_3 \in I\!R_1\}$. Because the family $\mathcal{Q}$ of distributions is complete, i.e. $\int h(x)Q(dx) = 0$ for any $Q \in \mathcal{Q}$ implies $h = 0$ $\mathcal{Q}$-a.s., we have a $\mathcal{Q}$-almost-sure invariance: for all $b \in I\!R_1$

$$\check{q}(x_1, x_2, y) = \check{q}(x_1 + b, x_2 + b, y + b) \quad \mathcal{Q} - a.s. \ .$$

According to Witting(1985), Theorem 3.109, for $\mathcal{Q}$ is closed w.r.t. translations and $\mathcal{Q}$ is equivalent to the Lebegue-measure λ_3 there is a classification function $\tilde{q}_0$ such that

$$\tilde{q}_0(x_1, x_2, y) = \check{q}(x_1, x_2, y) \quad \mathcal{Q} - a.s.$$

Consequently, for any $\theta = (h_1, h_2, \tau) \in \Theta_\Delta$

$$\begin{aligned} R(h_1, h_2, \tau, \tilde{q}_0) &= \lim_{k\to\infty} R(h_1, h_2, \tau, q_{n_k}) \\ &= 1 - \frac{1}{2n_k} \iiint \int_{-n_k}^{n_k} q_\tau(x_1 + t, x_2 + t, y + t)\, dt\, N_\theta(dx_1, dx_2, dy) \\ &= \frac{1}{2n_k} \int_{n_k}^{n_k} R(h_1 + t, h_2 + t, \tau) dt \le \sup_{\theta\in\Theta_\Delta} R(\theta, q) \end{aligned}$$

Hence for any $q \in \mathcal{C}(E, D)$ there is a classification function $\tilde{q}_0$ that is invariant w.r.t. translations such that

$$\sup_{\theta\in\Theta_\Delta} R(\theta, q) \ge \sup_{\theta\in\Theta_\Delta} R(\theta, \tilde{q}_0)$$

and by Lemma 2.3 we get the left equality in (7).
(3) For the second equality first note that "$\ge$" holds trivially. Now we assume that there is a sequence $(c_n)_{n\in I\!N}$, $c_n \in I\!R_1$, $c_n \uparrow \infty$, such that

$$\lim_{n\to\infty} \inf_{q\in\mathcal{C}(E,I\!D)} \sup_{\theta\in\Theta_\Delta\cap K_{c_n}} R(\theta, q) < \inf_{q\in\mathcal{C}(E,I\!D)} \sup_{\theta\in\Theta_\Delta} R(\theta, q) =: A \tag{19}$$

i.e. there is a sequence of classification functions $(q_n)_{n\in I\!N}$ and a positive real number ε so that

$$\sup_{\theta\in\Theta_\Delta\cap K_{c_n}} R(\theta, q_n) \le A - \varepsilon \ .$$

Because $C(E, I\!D)$ is weakly sequential compact there is a classification function $\tilde{q}$ and a subsequence (q_{n_k}) such that

$$R(\theta, \tilde{q}) = \lim_{k\to\infty} R(\theta, q_{n_k}) \quad \text{for all } \theta \in \Theta_\Delta \ .$$

And therefore for all $\theta \in \Theta_\Delta$, $c_n \uparrow \infty$,

$$R(\theta, \tilde{q}) \le \limsup_{n\to\infty} \sup_{\theta\in\Theta_\Delta\cap K_{c_n}} R(\theta, q_n) \le A - \varepsilon$$

and, in particular, $\sup\{R(\theta, \tilde{q}) : \theta \in \Theta_\Delta\} < A$ which contradicts the definition of A in (19). $\square$

Proof of Theorem 2.6: According to Witting, Nölle (1970) the conditions **(A1)** to **(A5)** imply for any consistent sequence of maximum liklihood estimators

$$\begin{aligned} \mathcal{L}(\sqrt{n}(\widehat{\theta}_{n,i} - \vartheta_0)|Q^n_{\vartheta_0}) &\Longrightarrow N(0, I^{-1}(\vartheta_0)) \quad \text{and} \\ \sqrt{n}(\hat{\theta}_{n,i} - \vartheta_0) - Z_{n,i} &\xrightarrow{P_{n,0}} 0 \ , i = 1, 2. \end{aligned}$$

Similarily, for $\widehat{\theta}_{n,3}$ which is based on a sample of size k_n

$$\mathcal{L}(\sqrt{\frac{n}{k_n}}\sqrt{k_n}(\widehat{\theta}_{n,3} - \vartheta_0)|Q^{k_n}_{\vartheta_0}) \Longrightarrow N(0, \frac{1}{\alpha} I^{-1}(\vartheta_0)) \quad \text{and}$$

$$\begin{aligned} \sqrt{n}(\widehat{\theta}_{n,3} - \vartheta_0) - Z_{n,3} &= \sqrt{\frac{n}{k_n}}\Big(\sqrt{k_n}(\widehat{\theta}_{n,3} - \vartheta_0) - \frac{k_n}{\alpha n}\frac{1}{\sqrt{k_n} I(\vartheta_0)} \sum_{i=1}^{k_n} \frac{\partial}{\partial \vartheta} \ln f_\vartheta \Big|_{\vartheta = \vartheta_0} (Y_{n,i}) \\ &\xrightarrow{Q^{k_n}_{\vartheta_0}} 0 \end{aligned}$$

i.e. for $T_n = (T_{n,1}, T_{n,2}, T_{n,3}) = (\sqrt{n}(\widehat{\theta}_{n,1} - \vartheta_0), \sqrt{n}(\widehat{\theta}_{n,2} - \vartheta_0), \sqrt{n}(\widehat{\theta}_{n,3} - \vartheta_0))$

$$Z_n - T_n \xrightarrow{P_{n,0}} 0 \ .$$

Let $(\theta_n)_{n \in I\!N}$, $\theta_n = (h_{n,1}, h_{n,2}, \tau)$ be any convergent sequence, $(h_{n,1}, h_{n,2}, \tau) \to (h_1, h_2, \tau) =: \theta \in I\!R_2 \times I\!D$ and $P_\theta = N(h_1, I^{-1}(\vartheta_0)) \times N(h_2, I^{-1}(\vartheta_0)) \times N(h_\tau, \frac{1}{\alpha} I^{-1}(\vartheta_0))$ then the contiguity of P_{n,θ_n} w.r.t. $P_{n,0}$ (according to (13)) implies

$$Z_n - T_n \xrightarrow{P_{n,\theta_n}} 0 \ . \tag{20}$$

Hence, by $\mathcal{L}(Z_n|\, P_{n,0}) \Rightarrow P_\theta$, (20) and Slutsky's Theorem $\mathcal{L}(T_n|P_{n,\theta_n}) \Rightarrow P_\theta$. Let $\overline{q}_i$ be defined as in (3) and $D_{\overline{q}_i}$ denote the set of all discontinuity points of $\overline{q}_i$, then holds

$$P_\theta(D_{\overline{q}_i}) = 0 \quad \text{for all } \theta \in \Theta \ . \tag{21}$$

According to Billingsley, that implies the weak convergence $\mathcal{L}(\overline{q}_i|\, P^{T_n}_{n,\theta_n}) \Rightarrow \mathcal{L}(\overline{q}_i|P_\theta)$ and because of the boundedness of $\overline{q}_i$, $i = 1, 2$,

$$\int \overline{q}_i(T_{n,1}, T_{n,2}, T_{n,3}) dP_{n,\theta_n} \longrightarrow \int \overline{q}_i dP_\theta \tag{22}$$

for any $\theta_n \to \theta$.

Let $\varphi^i(\theta) = \int \overline{q}_i dP_\theta$ and $\varphi^i_n(\theta) = \int q_i(\widehat{\theta}_{n,1}, \ldots, \widehat{\theta}_{n,3}) dP_{n,\theta_n}$ for a compact parameter set $K \subseteq \Theta$. For $\{P_\theta, \theta \in \Theta\}$ is as a subfamily of normal multivariate distributions it is weakly continuous, i.e. for any $\theta_n \to \theta$ holds $P_{\theta_n} \Rightarrow P_\theta$, and by (21) and Billingsley $\varphi^i(\theta)$ is continuous. Standard arguments from analysis now say that, if $\varphi^i(\theta)$ is a continuous, real-valued function, defined on a compact set K, and if for a sequence of real-valued functions φ^i_n, defined on the same set K, holds $\varphi^i_n(\theta_n) \to \varphi^i(\theta)$ for any $\theta_n \to \theta$ then (φ^i_n) converges even uniformly on K. Therfore by (22)

$$\lim_{n\to\infty} \sup_{\theta \in K} |\varphi^i_n(\theta) - \varphi^i(\theta)| = 0 \ ,$$

and, hence for $K_n = \Theta_n \cap K_c \cap \Theta_\Delta \uparrow K = K_c \cap \Theta_\Delta$

$$\lim_{n\to\infty} \sup_{\theta \in \Theta_n \cap K_c \cap \Theta_\Delta} |R_n(\theta, \overline{q}_n) - R(\theta, \overline{q})| = 0 \tag{23}$$

for any $c > \Delta$. Hence for $c \uparrow \infty$

$$\begin{aligned}\lim_{c\uparrow\infty}\lim_{n\to\infty}\sup_{\theta\in\Theta_\Delta\cap\Theta_n\cap K_c} R_n(\theta,\overline{q}_n) &= \lim_{c\uparrow\infty}\sup_{\theta\in\Theta_\Delta} R(\theta,\overline{q}) = \sup_{\theta\in\Theta_\Delta} R(\theta,\overline{q}) \\ &= \lim_{c\uparrow\infty}\inf_{q\in\mathcal{C}(E,I\!D)}\sup_{\theta\in\Theta_\Delta\cap\Theta_n\cap K_c} R(\theta,q)\end{aligned}$$

according to Theorem 2.4.

□

References

Billingsley, P. (1969): Convergence of Probability Measures. Wiley, New York.

Kudo, A. (1959), *The Classificatory Problem Viewed as a Two-decision Problem.* Memoirs of the Faculty of Science, Ser. A, Vol.4,No.1, Kyushu University.

LeCam, L. (1986), *Asymptotic Methods in Statistical Decision Theory.* Springer, Berlin, Heidelberg, New York.

Neuhaus, G. (1989), *Einige Kapitel der finiten und asymptotischen Entscheidungstheorie von LeCam.* Hamburg.

Rüschendorf, L. (1987), *Asymptotische Statistik.* Skripten zur Mathematischen Statistik Nr. 13, Münster.

Strasser, H. (1985), *Mathematical Theory of Statistics.* de Gruyter, Berlin.

Witting, H. (1985), *Mathematische Statistik I*, Teubner, Stuttgart.

Witting, H., Nölle, G. (1970), *Angewandte Mathematische Statistik.* BSG B.G. Teubner Verlagsgesellschaft.