Volume 13, No. 1 (2004), pp. Allerton Press, Inc.

MATHEMATICAL METHODS OF STATISTICS

DISCRIMINANT ANALYSIS OF STATIONARY FINITE MARKOV CHAINS

Yu. Kharin and A. Kostevich

National Research Centre for Applied Problems of Mathematics and Informatics
Belarussian State University, 4 Fr. Scoriny av., 220050 Minsk, Belarus
E-mail: kharin@bsu.by

Decision rules for classification of realizations of stationary finite Markov chains are constructed and their performance is evaluated for different levels of prior information: known parameters, unknown parameters, missing values. In the asymptotic setting of increasing number of observations and contiguous classes, asymptotic expansions of the misclassification probability are constructed and analyzed.

Key words: discriminant analysis, Markov chain, contiguous classes, risk, asymptotic expansion, missing values.

2000 Mathematics Subject Classification: Primary 62M02, Secondary 62H30.

Introduction

Let a discrete time series be observed which may be described by one of L finite homogeneous Markov chains ($2 \leq L < \infty$). These Markov chains specify L classes of the observed time series. The classes are assumed to differ in parameters of the Markov chains, i.e., in matrices of one-step transition probabilities. We consider the problem of classification of the observed time series into one of these classes.

This problem is very topical in applications to medical diagnostics, classification of DNA sequences [1, 21], technical diagnostics (e.g., faulty link detection in communication networks [9]), sequential detection of an abrupt change in the Markov chain distribution [16], intrusion detection in computer networks [8], etc.) In practice, this classification problem is often accompanied by some prior uncertainty: unknown parameters, missing values, etc. [6, 12].

The classification problem under consideration includes the construction of an optimal in some sense decision rule (DR) for classification of the observed time

This research was supported by the ISTC grant No. B-705.

series into one of L classes, and also the evaluation of the DR performance. The performance of a DR is usually described by the misclassification probability, which is also called the risk of classification [10, 15].

If the parameters of the classes are known, the Bayesian decision rule (BDR), which minimizes the misclassification probability, can be easily constructed (see, e.g., [9, 11]) using the traditional technique of discriminant analysis [10, 15]. But the evaluation of the exact misclassification probability of the BDR is a serious problem, that is why the asymptotic analysis of the misclassification probability for the BDR is needed. There are two main approaches in asymptotic analysis of the misclassification probability. According to the first approach, the parameters of the classes are fixed and the rate of convergence of the misclassification probability to zero is investigated using the large deviations technique as the length of the observed time series goes to infinity [11, 17, 19]. According to the second approach, the classes are assumed to be contiguous [5, 13] (or "close" [4]) as the length of the observed time series goes to infinity. Then the limiting value of the misclassification probability is sought, which is not equal to zero because of contiguity of the classes. The second approach is more general and seems to be more adequate for practice as the hardness of discrimination between classes is adapted to the size of experimental data [4, 5, 13]. The contiguous classes approach has not been applied in discriminant analysis of Markov chains before. The case of unknown parameters of the classes as well as the case of missing values in discrimination of Markov chains have not been investigated so far.

In the paper we construct the decision rules for classification of stationary finite Markov chains for three levels of prior uncertainty: known parameters, unknown parameters, missing values. For these three cases we construct and analyze asymptotic expansions of the misclassification probability using the contiguous classes approach.

1. Mathematical Model

Let a sequence of discrete random variables $\{X_t\}$, $X_t \in \mathcal{A} = \{1, 2, \ldots, N\}$, $t = 1, 2, \ldots$, be observed; it belongs to one of L classes Ω_1, $\Omega_2, \ldots, \Omega_L$ with prior probabilities $q_1, q_2, \ldots, q_L \in (0, 1)$ ($L \geq 2$, $q_1 + \ldots + q_L = 1$). A sequence of class Ω_l is a homogeneous finite Markov chain specified by the vector of initial probabilities $\pi^{(l)}$ and the matrix of one-step transition probabilities $P^{(l)}$:

$$\begin{aligned} &\pi^{(l)} = (\pi_i^{(l)}): \quad \pi_i^{(l)} = \Pr\{X_1 = i \mid \Omega_l\}, \\ &P^{(l)} = (p_{ij}^{(l)}): \quad p_{ij}^{(l)} = \Pr\{X_t = j \mid X_{t-1} = i, \Omega_l\}, \qquad i, j \in \mathcal{A}, \end{aligned} \tag{1}$$

where $l \in \{1, \ldots, L\}$. The Markov chains of classes $\{\Omega_l\}$ are assumed to be stationary and ergodic; the vector $\pi^{(l)}$ is the stationary distribution for the Markov chain of class Ω_l with $\pi_i^{(l)} > 0$, $i \in \mathcal{A}$. Excluding the singularities, we will assume that all one-step transitions have nonzero probabilities:

$$p_{ij}^{(l)} > 0, \qquad i, j \in \mathcal{A}, \quad l \in \{1, \ldots, L\}. \tag{2}$$

We suppose that the classes $\{\Omega_l\}$ differ in the matrices of one-step transition probabilities $\{P^{(l)}\}$.

Let a realization of length n from the class Ω_ν be observed:

$$X = (x_1, x_2, \dots, x_n), \qquad x_t \in \mathcal{A}, \quad t \in \{1, \dots, n\}, \tag{3}$$

where $\nu \in \{1, 2, \dots, L\}$ is an unobservable random classification indicator. The probability distribution of the random variable ν is determined by the prior probabilities $\Pr\{\nu = l\} = q_l$, $l \in \{1, \dots, L\}$.

We consider the problem of finding a decision rule d for classification of the observed realization X into one of the classes $\{\Omega_l\}$, $d = d(X)$, $X \in \mathcal{A}^n$, $d \in \{1, 2, \dots, L\}$. The performance of a DR $d(\cdot)$ is described by the misclassification probability:

$$r = \Pr\{d(X) \neq \nu\}. \tag{4}$$

2. Bayesian Decision Rule and its Performance

A decision rule $d_{BDR}(\cdot)$ that minimizes the classification risk $r = r(d(\cdot))$ (in our case, the misclassification probability (4)) for known values of the parameters is called the Bayesian decision rule (BDR) [10, 15]. We will construct the BDR for the model (1), (3) and find the asymptotic value of the misclassification probability (4).

Define statistical estimators of the $(L \times L)$-matrix of bivariate probabilities $\Pi = (\Pi_{ij})$, $\Pi_{ij} = \Pr\{x_t = i, x_{t+1} = j\}$, $i, j \in \mathcal{A}$, calculated from the realization (3):

$$\widehat{\Pi} = (\widehat{\Pi}_{ij}): \quad \widehat{\Pi}_{ij} = \frac{n_{ij}}{n}, \quad n_{ij} = \sum_{t=1}^{n-1} \boldsymbol{I}\{x_t = i, x_{t+1} = j\}, \quad i, j \in \mathcal{A},$$

where $\boldsymbol{I}\{A\}$ is the indicator function of the event A. Because of the norming condition for the probabilities $\{\Pi_{ij}\}$ we consider only $\{\Pi_{ij}, (i,j) \in \mathcal{A}_\Pi\}$ as unknown parameters to be estimated, where $\mathcal{A}_\Pi = \{\mathcal{A}^2 \setminus \{(N,N)\}\}$; $\Pi_{NN} = 1 - \sum_{(i,j) \in \mathcal{A}_\Pi} \Pi_{ij}$.

Theorem 1. *The BDR for classification of the Markov chains for the model* (1), (3) *is*:

$$d_{BDR}(X) = \arg \max_{1 \le l \le L} \left(\frac{1}{n} \log q_l + \frac{1}{n} \log \pi_{x_1}^{(l)} + \sum_{i,j \in \mathcal{A}} \widehat{\Pi}_{ij} \log p_{ij}^{(l)} \right), \quad X \in \mathcal{A}^n. \tag{5}$$

Proof. Using the log-likelihood function of the parameters $(\pi^{(l)}, P^{(l)})$ for the realization X in the BDR for discrete distributions [10] we obtain (5). □

Corollary 1. *In the case of two classes* $(L = 2)$ *the BDR* (5) *is*:

$$d_{BDR}(X) = \mathbf{1}(\Lambda(X)) + 1, \qquad X \in \mathcal{A}^n, \tag{6}$$

$$\Lambda(X) = \Lambda^*(X) + \frac{1}{n} \log \frac{q_2}{q_1} + \frac{1}{n} \log \frac{\pi_{x_1}^{(2)}}{\pi_{x_1}^{(1)}}, \qquad \Lambda^*(X) = \sum_{i,j \in \mathcal{A}} \widehat{\Pi}_{ij} \log \frac{p_{ij}^{(2)}}{p_{ij}^{(1)}}, \tag{7}$$

where $\Lambda(X)$ *is the discriminant function based on the log-likelihood functions of the parameters of the classes* Ω_1, Ω_2; $\mathbf{1}(x) = \boldsymbol{I}\{x > 0\}$ *is the Heaviside function.*

Now we explore the misclassification probability (4) for the case of two classes ($L = 2$). Define the contiguous classes asymptotics [4] for the model (1):

$$p_{ij}^{(2)} = p_{ij}^{(1)}(1 + b_{ij}\varepsilon), \qquad \varepsilon \to 0, \quad p_{ij}^{(2)} \to p_{ij}^{(1)}, \qquad i, j \in \mathcal{A}, \tag{8}$$

where $\{b_{ij}\}$ are some constant weight coefficients $\big(\sum_{j\in\mathcal{A}} p_{ij}^{(1)} b_{ij} = 0,\ i, j \in \mathcal{A}\big)$, ε is the "contiguity" parameter.

We introduce the following auxiliary variables:

$$a_l = (-1)^l \sum_{i\in\mathcal{A}} \pi_i^{(l)} \sum_{j\in\mathcal{A}} p_{ij}^{(l)} \log \frac{p_{ij}^{(2)}}{p_{ij}^{(1)}} > 0, \tag{9}$$

$$s_{ijuv}^{(l)} = \pi_i^{(l)} p_{ij}^{(l)} (\delta_{iu}\delta_{jv} - \pi_u^{(l)} p_{uv}^{(l)}) + p_{ij}^{(l)} p_{uv}^{(l)} (\pi_i^{(l)} c_{ju}^{(l)} + \pi_u^{(l)} c_{vi}^{(l)}), \tag{10}$$

$$c_{ju}^{(l)} = \sum_{k=0}^{\infty} (p_{ju}^{(l)}(k) - \pi_u^{(l)}) < \infty,$$

$$\sigma_{ijuv}^{(l)} = \frac{\delta_{iu}}{\pi_i^{(l)}} (\delta_{jv} p_{ij}^{(l)} - p_{ij}^{(l)} p_{uv}^{(l)}), \qquad i, j, u, v \in \mathcal{A}, \quad l \in \{1, 2\}, \tag{11}$$

where a_l is the weighted sum of the Kullback–Leibler information [2] for discrimination between $P^{(1)}$ and $P^{(2)}$ (the multiplier $(-1)^l$ ensures that $a_l > 0$); $\{s_{ijuv}^{(l)}\}$ and $\{\sigma_{ijuv}^{(l)}\}$ are some covariances, which will be explained in the proofs of Theorems 2 and 3; $p_{ju}^{(l)}(k) = ((P^{(l)})^k)_{ju}$ is the probability of the k-step transition from the state j to the state u of the Markov chain of class Ω_l; the series for $c_{ju}^{(l)}$ converges at an exponential rate and can be easily computed; δ_{ij} is the Kronecker delta.

The following lemma concerns the behavior of the auxiliary variables (9)–(11) and the stationary distribution in the contiguous classes asymptotics (8).

Lemma 1. *Under the assumption of contiguous classes* (8) *of stationary Markov chains the following expansions hold*:

$$\pi_j^{(2)} = \pi_j^{(1)} \left(1 + \varepsilon h_j + O\left(\varepsilon^2\right)\right); \quad \sigma_{ijuv}^{(2)} = \sigma_{ijuv}^{(1)} + O\left(\varepsilon\right), \quad s_{ijuv}^{(2)} \to s_{ijuv}^{(1)},$$

$$a_l = \frac{\varepsilon^2}{2} \sum_{i,j\in\mathcal{A}} b_{ij}^2 \pi_i^{(l)} p_{ij}^{(1)} + O\left(\varepsilon^3\right), \qquad l \in \{1, 2\}, \quad \frac{a_2}{a_1} \to 1,$$

where $|h_j| < +\infty$, $i, j, u, v \in \mathcal{A}$.

Proof. The first statement is based on the well-known result on error in solution of a system of linear algebraic equations under matrix distortions [7]. The other statements follow from the Taylor formula. □

Now we evaluate the misclassification probability (4) in the contiguous classes asymptotics (8) with "contiguity" parameter $\varepsilon = O\left(n^{-1/2}\right)$. Denote

$$\mu = \sum_{i,j\in\mathcal{A}} b_{ij}^2 \pi_i^{(1)} p_{ij}^{(1)}, \quad V = \sum_{(i,j),(u,v)\in\mathcal{A}_\Pi} (b_{ij} - b_{NN}) s_{ijuv}^{(1)} (b_{uv} - b_{NN}) > 0, \tag{12}$$

$$\Delta_l = \Delta + (-1)^l \frac{2\log(q_2/q_1)}{c\sqrt{V}}, \quad \Delta = \frac{c\mu}{\sqrt{V}} > 0, \qquad l \in \{1, 2\}, \tag{13}$$

where the covariances $\{s_{ijuv}^{(1)}\}$ are defined in (10)

Theorem 2. *For increasing number of observations and two contiguous classes* (8),

$$n \to \infty, \quad \varepsilon = \frac{c}{\sqrt{n}} \to 0, \quad 0 < c < \infty,$$

the misclassification probability (4) *of the BDR* (6) *has the limit*

$$r_0 \to \tilde{r}_0 = q_1 \Phi\Big(-\frac{\Delta_1}{2}\Big) + q_2 \Phi\Big(-\frac{\Delta_2}{2}\Big),$$

where $\Phi(\cdot)$ *is the standard normal distribution function,* Δ_1, Δ_2 *are defined in* (13).

Proof. According to (6) the conditional misclassification probabilities are:

$$\begin{aligned} r_1 &= \Pr\{d(X) \neq \nu \mid \nu = 1\} = 1 - \Pr\{\Lambda(X) < 0 \mid \nu = 1\}, \\ r_2 &= \Pr\{d(X) \neq \nu \mid \nu = 2\} = \Pr\{\Lambda(X) < 0 \mid \nu = 2\}. \end{aligned}$$

Let us find the probability distribution of $\Lambda(X)$ defined by (7).

Consider first the summand $\Lambda^*(X)$ of $\Lambda(X)$. From (7) we see that $\Lambda^*(X)$ is a linear combination of the random variables $\{\widehat{\Pi}_{ij}\}$. It is known [2] that if the observation X belongs to the class Ω_l then the statistics $\xi_{ij}^{(l)} = \sqrt{n}(\widehat{\Pi}_{ij} - \Pi_{ij}^{(l)})$ have the asymptotically normal probability distribution with zero means and covariances $\mathrm{Cov}\{\xi_{ij}^{(l)}, \xi_{uv}^{(l)}\} = s_{ijuv}^{(l)}$ defined by (10), where $\Pi_{ij}^{(l)} = \pi_i^{(l)} p_{ij}^{(l)}$. Therefore the conditional distribution of $\Lambda^*(X)$ is also asymptotically normal. The asymptotic mean of $\Lambda^*(X)$ obtains as a linear combination of the means of $\{\widehat{\Pi}_{ij}\}$ and is equal to $(-1)^l a_l$. The asymptotic variance of $\Lambda^*(X)$ is a quadratic form of $\{s_{ijuv}^{(l)}\}$, and taking into account the norming condition for $\{\Pi_{ij}\}$ the asymptotic variance is

$$\sigma_l^2 = \sum_{(i,j),(u,v) \in \mathcal{A}_\Pi} \log \frac{p_{ij}^{(2)}}{p_{ij}^{(1)}} \frac{p_{NN}^{(1)}}{p_{NN}^{(2)}} \log \frac{p_{uv}^{(2)}}{p_{uv}^{(1)}} \frac{p_{NN}^{(1)}}{p_{NN}^{(2)}} s_{ijuv}^{(l)} > 0.$$

Note that $\sigma_l^2 > 0$ because the covariance matrices $\{s_{ijuv}^{(l)}, (i,j), (u,v) \in \mathcal{A}_\Pi\}$ are nonsingular [2] and $P^{(1)} \neq P^{(2)}$. Under the contiguous classes asymptotics (8) σ_l^2 can be presented as

$$\sigma_l^2 = \varepsilon^2 \sum_{(i,j),(u,v) \in \mathcal{A}_\Pi} (b_{ij} - b_{NN}) s_{ijuv}^{(l)} (b_{uv} - b_{NN}) + O\left(\varepsilon^3\right). \tag{14}$$

Consider the last summand of $\Lambda(X)$ in (7): $\zeta = n^{-1} \log(\pi_{x_1}^{(2)} / \pi_{x_1}^{(1)})$. Now we get

$$\begin{aligned} \Pr\{\Lambda(X) < 0 \mid \nu = l\} &= \Pr\left\{\Lambda^*(X) + \frac{1}{n} \log \frac{q_2}{q_1} + \zeta < 0 \mid \nu = l\right\} \\ &= \Pr\left\{\sqrt{n} \frac{\Lambda^*(X) - (-1)^l a_l}{\sigma_l} + \frac{\sqrt{n}\zeta}{\sigma_l} < -\sqrt{n} \frac{(-1)^l a_l}{\sigma_l} - \frac{1}{\sqrt{n}\sigma_l} \log \frac{q_2}{q_1} \mid \nu = l\right\}. \end{aligned}$$

We see from Lemma 1 and (14) that $\zeta = O_P(\varepsilon/n)$,

$$\sqrt{n}\frac{a_l}{\sigma_l} = \sqrt{n}\frac{\frac{1}{2}\varepsilon^2 \sum_{i,j\in\mathcal{A}} b_{ij}^2 \pi_i^{(l)} p_{ij}^{(1)} + O\left(\varepsilon^3\right)}{\sqrt{\varepsilon^2 \sum_{(i,j),(u,v)\in\mathcal{A}_\Pi} (b_{ij} - b_{NN}) s_{ijuv}^{(l)} (b_{uv} - b_{NN}) + O\left(\varepsilon^3\right)}},$$

$$\frac{\log(q_2/q_1)}{\sqrt{n}\sigma_l} = \frac{\log(q_2/q_1)}{\sqrt{n}\sqrt{\varepsilon^2 \sum_{(i,j),(u,v)\in\mathcal{A}_\Pi} (b_{ij} - b_{NN}) s_{ijuv}^{(l)} (b_{uv} - b_{NN}) + O\left(\varepsilon^3\right)}},$$

$$-\sqrt{n}\frac{(-1)^l a_l}{\sigma_l} - \frac{\log(q_2/q_1)}{\sqrt{n}\sigma_l} \to -(-1)^l \frac{\Delta}{2} - \frac{\log(q_2/q_1)}{c\sqrt{V}},$$

and $\sqrt{n}\zeta/\sigma_l = O_P(n^{-1/2}) \to 0$ in probability. Using the well-known result [18] (see Theorem 15) on convergence in distribution for the sum of $\sqrt{n}\zeta/\sigma_l$, which converges to 0 in probability, and $\Lambda^*(X)$, which has an asymptotically normal distribution, we get

$$r_1 = 1 - \Pr\{\Lambda(X) < 0 \mid \nu = 1\} \to 1 - \Phi\left(\frac{\Delta}{2} - \frac{\log(q_2/q_1)}{c\sqrt{V}}\right) = \Phi\left(-\frac{\Delta_1}{2}\right),$$

$$r_2 = \Pr\{\Lambda(X) < 0 \mid \nu = 2\} \to \Phi\left(-\frac{\Delta}{2} - \frac{\log(q_2/q_1)}{c\sqrt{V}}\right) = \Phi\left(-\frac{\Delta_2}{2}\right),$$

and $r_0 = q_1 r_1 + q_2 r_2 \to \tilde{r}_0$. □

Corollary 2. *If the classes are equiprobable* ($q_1 = q_2 = \frac{1}{2}$) *then the limiting value of the risk is* $\tilde{r}_0 = \Phi(-\Delta/2)$.

Remark 1. Taking into account the proof of Theorem 2, in the asymptotics (8) the BDR (5) is equivalent to the decision rule

$$d(X) = \arg\max_{1\le l\le L}\left(\frac{1}{n}\log q_l + \sum_{i,j\in\mathcal{A}} \widehat{\Pi}_{ij} \log p_{ij}^{(l)}\right), \qquad X \in \mathcal{A}^n.$$

3. The Case of Unknown Parameters

3.1. Plug-in DR and its risk. If the parameters of the classes (1) are unknown then a classified "training sample" is assumed to be observed:

$$\mathbb{X} = \{X^{(1)}, X^{(2)}, \dots, X^{(L)}\}, \tag{15}$$

$$X^{(l)} = (x_1^{(l)}, x_2^{(l)}, \dots, x_{n_l}^{(l)}), \qquad x_t^{(l)} \in \mathcal{A}, \quad t \in \{1, \dots, n_l\},$$

where $X^{(l)}$ is a realization of length n_l of the Markov chain from the class Ω_l, $l \in \{1, \dots, L\}$. It is assumed that X and $X^{(1)}, \dots, X^{(L)}$ are jointly independent.

The ML-estimators of the unknown matrices of one-step transition probabilities $P^{(l)}$ can be calculated from the "training sample":

$$\widehat{P}^{(l)} = (\hat{p}_{ij}^{(l)}) : \quad \hat{p}_{ij}^{(l)} = \frac{n_{ij}^{(l)}}{n_{i\cdot}^{(l)}}, \qquad i, j \in \mathcal{A}, \quad l \in \{1, \dots, L\},$$

$$n_{ij}^{(l)} = \sum_{t=1}^{n_l - 1} \boldsymbol{I}\{x_t^{(l)} = i, x_{t+1}^{(l)} = j\}, \qquad n_{i\cdot}^{(l)} = \sum_{j\in\mathcal{A}} n_{ij}^{(l)}.$$

Because of the norming condition for the probabilities $\{p_{ij}^{(l)}\}$ we consider only $\{p_{ij}^{(l)}, (i,j) \in \mathcal{A}_P\}$ as unknown parameters to be estimated, where $\mathcal{A}_P = \{(i,j) : i \in \mathcal{A}, j \in \mathcal{A} \setminus \{N\}\}$; $p_{iN}^{(l)} = 1 - \sum_{j=1}^{N-1} p_{ij}^{(l)}$.

The plug-in Bayesian decision rule (PBDR) is obtained from the BDR (5) if the unknown parameters $\{P^{(l)}\}$ are replaced by their ML-estimators $\{\widehat{P}^{(l)}\}$:

$$d_{PBDR}(X, \mathbb{X}) = \arg \max_{1 \le l \le L} \left(\frac{1}{n} \log q_l + \frac{1}{n} \log \hat{\pi}_{x_1}^{(l)} + \sum_{i,j \in \mathcal{A}} \widehat{\Pi}_{ij} \log \hat{p}_{ij}^{(l)} \right),$$

where $\hat{\pi}_i^{(l)} = n_{i\cdot}^{(l)}/n$. In case of two classes the PBDR can be represented as

$$d_{PBDR}(X, \mathbb{X}) = \mathbf{1}\left(\widehat{\Lambda}(X, \mathbb{X})\right) + 1, \tag{16}$$

$$\widehat{\Lambda}(X, \mathbb{X}) = \widehat{\Lambda}^*(X, \mathbb{X}) + \frac{1}{n} \log \frac{q_2}{q_1} + \frac{1}{n} \log \frac{\hat{\pi}_{x_1}^{(2)}}{\hat{\pi}_{x_1}^{(1)}}, \quad \widehat{\Lambda}^*(X, \mathbb{X}) = \sum_{i,j \in \mathcal{A}} \widehat{\Pi}_{ij} \log \frac{\hat{p}_{ij}^{(2)}}{\hat{p}_{ij}^{(1)}}.$$

Theorem 3. *For increasing number of observations n, n_1, n_2 and two contiguous classes* (8):

$$n, n_l \to \infty, \quad n_l/n = \tilde{\lambda}_l > 0, \quad l = 1,2; \qquad \varepsilon = cn^{-1/2} \to 0, \quad 0 < c < \infty, \tag{17}$$

the misclassification probability (4) *of the PBDR* (16) *has the limit*:

$$\begin{aligned} r \to \tilde{r} &= q_1 \Phi\left(-\frac{\widetilde{\Delta}_1}{2}\right) + q_2 \Phi\left(-\frac{\widetilde{\Delta}_2}{2}\right), \\ \widetilde{\Delta}_l &= \frac{c\mu}{\sqrt{V + \widetilde{V}_l}} + (-1)^l \frac{2\log(q_2/q_1)}{c\sqrt{V + \widetilde{V}_l}}, \\ \widetilde{V}_l &= \frac{1}{\tilde{\lambda}_{3-l}} \sum_{(i,j),(u,v) \in \mathcal{A}_P} \pi_i^{(1)} \pi_u^{(1)} (b_{ij} - b_{iN})(b_{uv} - b_{uN}) \sigma_{ijuv}^{(1)} > 0, \end{aligned} \tag{18}$$

where μ, V are defined in (12), *the covariances $\{\sigma_{ijuv}^{(1)}\}$ are defined in* (11).

Proof. Consider the event $Z = \{n_{ij}^{(l)} \neq 0, l \in \{1,2\}, i,j \in \mathcal{A}\}$. The PBDR is defined only if the event Z occurs. Put $d_{PBDR}(X, \mathbb{X}) = 0$ if the complement $\bar{Z}$ of the event Z occurs. Consider the conditional misclassification probabilities $r_l = \Pr\{d_{PBDR}(X, \mathbb{X}) \neq \nu \mid \nu = l\}$, $l \in \{1,2\}$:

$$r_l = \Pr\{\{d_{PBDR}(X, \mathbb{X}) \neq \nu\} \cap Z \mid \nu = l\} + \Pr\{\{d_{PBDR}(X, \mathbb{X}) \neq \nu\} \cap \bar{Z} \mid \nu = l\}.$$

Since the Markov chains are stationary (with $\pi_i^{(l)} > 0$, $i \in \mathcal{A}$) and recurrent, taking into account assumption (2), in the asymptotics of increasing number of observations we get $\Pr\{\bar{Z}\} \to 0$ and $\Pr\{\{d_{PBDR}(X, \mathbb{X}) \neq \nu\} \cap \bar{Z} \mid \nu = l\} \to 0$.

Consider now the summand $\Pr\{\{d_{PBDR}(X,\mathbb{X}) \neq \nu\} \cap Z \mid \nu = l\}$ and find the probability distribution of the statistic $\widehat{\Lambda}(X,\mathbb{X})$.

Consider first the summand $\widehat{\Lambda}^*(X,\mathbb{X})$ of $\widehat{\Lambda}(X,\mathbb{X})$. Suppose the realization X belongs to the class Ω_l. It is seen from (16) that $\widehat{\Lambda}^*(X,\mathbb{X})$ is a function of the estimators of one-step transition probabilities and the estimators of the bivariate probabilities: $\widehat{\Lambda}^*(X,\mathbb{X}) = f(\widehat{P}^{(1)}, \widehat{P}^{(2)}, \widehat{\Pi})$. It is known [2] that the statistics $\theta_{ij}^{(l)} = \sqrt{n_l}(\hat{p}_{ij}^{(l)} - p_{ij}^{(l)}) = \sqrt{\tilde{\lambda}_l n}(\hat{p}_{ij}^{(l)} - p_{ij}^{(l)})$ are asymptotically normal with zero means and covariances $\operatorname{Cov}\{\theta_{ij}^{(l)}, \theta_{uv}^{(l)}\} = \sigma_{ijuv}^{(l)}$ defined by (11), and the statistics $\xi_{ij}^{(l)} = \sqrt{n}\left(\widehat{\Pi}_{ij} - \Pi_{ij}^{(l)}\right)$ are asymptotically normal with zero means and covariances $\operatorname{Cov}\{\xi_{ij}^{(l)}, \xi_{uv}^{(l)}\} = s_{ijuv}^{(l)}$ defined by (10), $\Pi_{ij}^{(l)} = \pi_i^{(l)} p_{ij}^{(l)}$, $i,j,u,v \in \mathcal{A}$. Furthermore, independence of X, $X^{(1)}$, $X^{(2)}$ implies independence of the statistics $\{\xi_{ij}^{(l)}\}$, $\{\theta_{ij}^{(1)}\}$, $\{\theta_{ij}^{(2)}\}$. By the Anderson theorem [20] on functional transformations of asymptotically normal random variables, it follows that $\widehat{\Lambda}^*(X,\mathbb{X})$ has asymptotically normal distribution

$$\mathcal{L}\left\{\sqrt{n}\frac{\widehat{\Lambda}^*(X,\mathbb{X}) - (-1)^l a_l}{\widetilde{\sigma}_l} \mid \nu = l\right\} \to \mathcal{N}(0,1)$$

with the mean of the form $f\big(\mathsf{E}\{\widehat{P}^{(1)}\}, \mathsf{E}\{\widehat{P}^{(2)}\}, \mathsf{E}\{\widehat{\Pi}\}\big)$, which is equal to $(-1)^l a_l$ (see (9)), and the variance $\widetilde{\sigma}_l^2$ being the following quadratic form of the covariance matrices $(\sigma_{ijuv}^{(l)}/\tilde{\lambda}_l)$, $(s_{ijuv}^{(l)})$ and a vector of partial derivatives of f [20]:

$$\begin{aligned}
\widetilde{\sigma}_l^2 = &\sum_{(i,j),(u,v)\in\mathcal{A}_P} \frac{\partial f}{\partial \hat{p}_{ij}^{(1)}} \frac{\partial f}{\partial \hat{p}_{uv}^{(1)}} \frac{\sigma_{ijuv}^{(1)}}{\tilde{\lambda}_1} + \sum_{(i,j),(u,v)\in\mathcal{A}_P} \frac{\partial f}{\partial \hat{p}_{ij}^{(2)}} \frac{\partial f}{\partial \hat{p}_{uv}^{(2)}} \frac{\sigma_{ijuv}^{(2)}}{\tilde{\lambda}_2} \\
&+ \sum_{(i,j),(u,v)\in\mathcal{A}_\Pi} \frac{\partial f}{\partial \widehat{\Pi}_{ij}} \frac{\partial f}{\partial \widehat{\Pi}_{uv}} s_{ijuv}^{(l)} \\
= &\sum_{(i,j),(u,v)\in\mathcal{A}_P} \left(\frac{\pi_i^{(l)} p_{ij}^{(l)}}{p_{ij}^{(1)}} - \frac{\pi_i^{(l)} p_{iN}^{(l)}}{p_{iN}^{(1)}}\right)\left(\frac{\pi_u^{(l)} p_{uv}^{(l)}}{p_{uv}^{(1)}} - \frac{\pi_u^{(l)} p_{uN}^{(l)}}{p_{uN}^{(1)}}\right)\frac{\sigma_{ijuv}^{(1)}}{\tilde{\lambda}_1} \\
&+ \sum_{(i,j),(u,v)\in\mathcal{A}_P} \left(\frac{\pi_i^{(l)} p_{ij}^{(l)}}{p_{ij}^{(2)}} - \frac{\pi_i^{(l)} p_{iN}^{(l)}}{p_{iN}^{(2)}}\right)\left(\frac{\pi_u^{(l)} p_{uv}^{(l)}}{p_{uv}^{(2)}} - \frac{\pi_u^{(l)} p_{uN}^{(l)}}{p_{uN}^{(2)}}\right)\frac{\sigma_{ijuv}^{(2)}}{\tilde{\lambda}_2} \\
&+ \sum_{(i,j),(u,v)\in\mathcal{A}_\Pi} \log\frac{p_{ij}^{(2)}}{p_{ij}^{(1)}}\frac{p_{NN}^{(1)}}{p_{NN}^{(2)}} \log\frac{p_{uv}^{(2)}}{p_{uv}^{(1)}}\frac{p_{NN}^{(1)}}{p_{NN}^{(2)}} s_{ijuv}^{(l)}.
\end{aligned}$$

We have $\widetilde{\sigma}_l^2 > 0$ because $\nabla f(P^{(1)}, P^{(2)}, \Pi) \neq 0$ and the covariance matrices $\{s_{ijuv}^{(l)}$, (i,j), $(u,v) \in \mathcal{A}_\Pi\}$, $\{\sigma_{ijuv}^{(l)}, (i,j), (u,v) \in \mathcal{A}_P\}$ are nonsingular, $l \in \{1,2\}$.

Under the contiguous classes asymptotics (8) $\widetilde{\sigma}_l^2$ can be represented as

$$\begin{aligned}
(19) \qquad \widetilde{\sigma}_l^2 = &\frac{\varepsilon^2}{\tilde{\lambda}_{3-l}} \sum_{(i,j),(u,v)\in\mathcal{A}_P} \pi_i^{(l)} \pi_u^{(l)} (b_{ij} - b_{iN})(b_{uv} - b_{uN})\sigma_{ijuv}^{(l)} \\
&+ \varepsilon^2 \sum_{(i,j),(u,v)\in\mathcal{A}_\Pi} (b_{ij} - b_{NN})(b_{uv} - b_{NN}) s_{ijuv}^{(l)} + O\left(\varepsilon^3\right).
\end{aligned}$$

Using Lemma 1, (17), and (19) we get

$$\sqrt{n}\frac{a_l}{\widetilde{\sigma}_l} \to \frac{c\mu}{2\sqrt{V+\widetilde{V}_l}}, \qquad \frac{\log q_2/q_1}{\sqrt{n}\widetilde{\sigma}_l} \to \frac{\log(q_2/q_1)}{c\sqrt{V+\widetilde{V}_l}}.$$

Consider the summand $\hat{\zeta} = n^{-1}\log(\hat{\pi}^{(2)}_{x_1}/\hat{\pi}^{(1)}_{x_1})$ of $\widehat{\Lambda}(X,\mathbb{X})$. The estimators of the stationary distributions are consistent ($\hat{\pi}^{(l)}_i \to \pi^{(l)}_i > 0$ in probability, $i \in \mathcal{A}$) and $\hat{\zeta}$ is analyzed in the same way as in the proof of Theorem 2. The rest of the proof follows the lines of the proof of Theorem 2. □

Remark 2. If n_1, n_2 increase faster than n, so that $\tilde{\lambda}_1, \tilde{\lambda}_2 \to \infty$, then $\widetilde{V}_1, \widetilde{V}_2 \to 0$ and the limiting value of the PBDR risk converges to the limiting value of the Bayesian risk: $\tilde{r} \to \tilde{r}_0$.

3.2. Asymptotic expansion of the risk of PBDR. Now we shall investigate convergence of the PBDR risk to the BDR risk as $n_1, n_2 \to \infty$ for a fixed n in case of two classes ($L = 2$). Consider the difficult for discrimination case, where the classes are equiprobable ($q_1 = q_2 = \frac{1}{2}$).

Let us introduce some notation: $\widetilde{\Lambda}(X) = q_2 L_2(X) - q_1 L_1(X)$ is the discriminant function based on the likelihood functions of the classes; $L_l(X)$ is the likelihood function of the parameters of the class Ω_l for the realization X:

$$L_l(X) = \pi^{(l)}_{x_1} \prod_{i,j\in\mathcal{A}} (p^{(l)}_{ij})^{n_{ij}(X)},$$

where $n_{ij}(X)$ is the bivariate frequency calculated from the realization X, $i, j \in \mathcal{A}$. Taking into account the expressions for the conditional misclassification probabilities:

$$\begin{aligned} r_1 &= \Pr\{d_{BDR}(X) = 2 \mid \nu = 1\} = \sum_{X\in\mathcal{A}^n} L_1(X)\mathbf{1}\left(\Lambda(X)\right), \\ r_2 &= 1 - \Pr\{d_{BDR}(X) = 2 \mid \nu = 2\} = 1 - \sum_{X\in\mathcal{A}^n} L_2(X)\mathbf{1}\left(\Lambda(X)\right), \end{aligned}$$

we get the exact value of the BDR risk in the form:

$$r_0 = q_1 r_1 + q_2 r_2 = q_2 - \sum_{X\in\mathcal{A}^n} \widetilde{\Lambda}(X)\mathbf{1}\left(\Lambda(X)\right) = q_2 - \sum_{X\in\mathcal{A}^n, \Lambda(X)\geq 0} \widetilde{\Lambda}(X). \tag{20}$$

Averaging the risk $r_{PBDR}(\mathbb{X})$ of the PBDR $d_{PBDR}(\cdot,\mathbb{X})$ over the "training sample" (15) we obtain the exact value of the (unconditional) risk of the PBDR:

$$r = q_2 - \sum_{X\in\mathcal{A}^n} \widetilde{\Lambda}(X)\mathsf{E}\left\{\mathbf{1}(\widehat{\Lambda}(X,\mathbb{X}))\right\}. \tag{21}$$

One can see that the discriminant functions $\widehat{\Lambda}(X,\mathbb{X})$ and $\Lambda(X)$ depend on the one-step transition probabilities and the stationary distributions. Let us represent the discriminant functions as functions of the bivariate probabilities only, which are determined by the one-step transition probabilities and the stationary distributions:

$$
\begin{aligned}
\widehat{\Lambda}(X,\mathbb{X}) &= \sum_{i,j\in\mathcal{A}} \frac{n_{ij}(X)}{n}\log\frac{\widehat{\Pi}^{(2)}_{ij}}{\widehat{\Pi}^{(1)}_{ij}} - \sum_{i\in\mathcal{A}} \frac{(n_{i\cdot}(X)-\delta_{ix_1})}{n}\log\frac{\widehat{\Pi}^{(2)}_{i\cdot}}{\widehat{\Pi}^{(1)}_{i\cdot}},\\
\Lambda(X) &= \sum_{i,j\in\mathcal{A}} \frac{n_{ij}(X)}{n}\log\frac{\Pi^{(2)}_{ij}}{\Pi^{(1)}_{ij}} - \sum_{i\in\mathcal{A}} \frac{(n_{i\cdot}(X)-\delta_{ix_1})}{n}\log\frac{\Pi^{(2)}_{i\cdot}}{\Pi^{(1)}_{i\cdot}},
\end{aligned}
$$

$$
\widehat{\Pi}^{(l)} = \left(\widehat{\Pi}^{(l)}_{ij}\right): \quad \widehat{\Pi}^{(l)}_{ij} = \frac{n^{(l)}_{ij}}{n_l}, \qquad \Pi^{(l)} = \left(\Pi^{(l)}_{ij}\right): \quad \Pi^{(l)}_{ij} = \pi^{(l)}_i p^{(l)}_{ij}, \qquad i,j\in\mathcal{A},
$$

where $\widehat{\Pi}^{(l)}_{i\cdot} = \sum_{j\in\mathcal{A}}\widehat{\Pi}^{(l)}_{ij}$, $\Pi^{(l)}_{i\cdot} = \sum_{j\in\mathcal{A}}\Pi^{(l)}_{ij}$, $l\in\{1,2\}$.

Introduce the notation:

$$
\begin{aligned}
B_n(X) &= \sum_{i,j\in\mathcal{A}} \frac{n_{ij}(X)}{n} b_{ij} + \sum_{i\in\mathcal{A}} \frac{\delta_{ix_1}}{n} h_i,\\
D_n(X) &= \frac{\lambda_1+\lambda_2}{\lambda_1\lambda_2} \sum_{(i,j),(u,v)\in\mathcal{A}_\Pi} g^{(1)}_{ij}(X) g^{(1)}_{uv}(X) s^{(1)}_{ijuv},\\
g^{(l)}_{ij}(X) &= \frac{1}{n}\left(\frac{n_{ij}(X)}{\Pi^{(l)}_{ij}} - \frac{n_{NN}(X)}{\Pi^{(l)}_{NN}}\right) - \frac{1}{n}\left(\frac{n_{i\cdot}(X)}{\Pi^{(l)}_{i\cdot}} - \frac{n_{N\cdot}(X)}{\Pi^{(l)}_{N\cdot}}\right),\\
\mathcal{A}^{(l)} &= \left\{X\in\mathcal{A}^n : g^{(l)}(X) = 0\right\}, \quad g^{(l)}(X) = \left(g^{(l)}_{ij}(X)\right), \quad (i,j)\in\mathcal{A}_\Pi,
\end{aligned}
$$

where $\{h_i\}$ are as in Lemma 1, $\lambda_1,\lambda_2>0$; the function $g^{(l)}_{ij}(X)$ is the partial derivative of $\widehat{\Lambda}(X,\mathbb{X})$ with respect to $\widehat{\Pi}^{(l)}_{ij}$ at $\widehat{\Pi}^{(1)} = \Pi^{(1)}$, $\widehat{\Pi}^{(2)} = \Pi^{(2)}$, $l\in\{1,2\}$.

Theorem 4. *Assume that the classes are equiprobable* ($q_1 = q_2 = \frac{1}{2}$), $n_* = \min\{n_1,n_2\}$, *n is fixed. Under the contiguous classes asymptotics* (8) *and increasing lengths of realizations from the "training sample"*:

$$
(22) \qquad n_*\to\infty, \quad n_l/n_* = \lambda_l > 0, \quad l\in\{1,2\}, \quad \varepsilon = cn_*^{-1/2}\to 0, \quad 0<c<\infty,
$$

the following expansion for the increment of the PBDR risk holds:

$$
(23) \qquad r = r_0 + \frac{\tilde{c}}{\sqrt{n_*}} + o\!\left(\frac{1}{\sqrt{n_*}}\right),
$$

$$
\tilde{c} = \frac{cn}{2}\sum_{X\in\mathcal{A}^n\setminus\mathcal{A}^{(1)}} L_1(X) B_n(X)\Phi\left(-|\Delta(X)|\right) > 0, \quad \Delta(X) = \frac{cB_n(X)}{\sqrt{D_n(X)}}.
$$

Proof. The discriminant function $\widehat{\Lambda}(X,\mathbb{X})$ is a function of the statistics $\widehat{\Pi}^{(1)}$, $\widehat{\Pi}^{(2)}$: $\widehat{\Lambda} = g(\widehat{\Pi}^{(1)},\widehat{\Pi}^{(2)})$. The statistics $\{\sqrt{n}\,(\widehat{\Pi}^{(l)}_{ij} - \Pi^{(l)}_{ij})\}$ are asymptotically normally distributed (see the proof of Theorem 3). By the Anderson theorem [20] on functional transformations of asymptotically normal random variables, it follows that the discriminant function $\widehat{\Lambda}(X, \mathbb{X})$ has the asymptotically normal distribution

$$\mathcal{L}\left\{\sqrt{n_*}\frac{\widehat{\Lambda}(X, \mathbb{X}) - \Lambda(X)}{\sigma(X)}\right\} \to \mathcal{N}(0, 1),$$

where the mean obtains as $g(\mathsf{E}\{\widehat{\Pi}^{(1)}\}, \mathsf{E}\{\widehat{\Pi}^{(2)}\})$ and is equal to $\Lambda(X)$ and the variance $\sigma^2(X)$ is given by the following quadratic form of the covariance matrices $(s_{ijuv}^{(l)}/\lambda_l)$ and the vector of partial derivatives $g^{(l)}(X)$ [20]:

$$\sigma^2(X) = \sum_{(i,j),(u,v)\in\mathcal{A}_\Pi} g_{ij}^{(1)}(X)g_{uv}^{(1)}(X)\frac{s_{ijuv}^{(1)}}{\lambda_1} + \sum_{(i,j),(u,v)\in\mathcal{A}_\Pi} g_{ij}^{(2)}(X)g_{uv}^{(2)}(X)\frac{s_{ijuv}^{(2)}}{\lambda_2}.$$

Under the contiguous classes asymptotics (8) we get $\Pi_{ij}^{(2)} = \Pi_{ij}^{(1)}(1 + \varepsilon(b_{ij} + h_i)) + O\left(\varepsilon^2\right)$. Construct the Taylor expansion of $\Lambda(X)$ and $\sigma^2(X)$.

Consider first the case $X \in \mathcal{A}^n \setminus \mathcal{A}^{(1)} \cup \mathcal{A}^{(2)}$, i.e., $n_{ij}(X) \neq n_{i\cdot}(X)p_{ij}^{(l)}$, $l \in \{1, 2\}$. Then the Taylor expansion of $\Lambda(X)$ and $\sigma^2(X)$ for any fixed X gives

$$\Lambda(X) = \varepsilon B_n(X) + O\left(\varepsilon^2\right), \quad \sigma^2(X) = D_n(X) + O\left(\varepsilon\right),$$

where $B_n(X) = O\left(1\right)$, $D_n(X) = O\left(1\right)$. We have $D_n(X) > 0$ because $g^{(1)}(X) \neq 0$, and the covariance matrix $\{s_{ijuv}^{(1)}, (i, j), (u, v) \in \mathcal{A}_\Pi\}$ is nonsingular.

Consider now the case $X \in \mathcal{A}^{(1)}$, i.e., $n_{ij}(X) = n_{i\cdot}(X)p_{ij}^{(1)}$. Taking into account the norming condition $\sum_{j\in\mathcal{A}} p_{ij}^{(1)}b_{ij} = 0$, the Taylor expansion of $\Lambda(X)$ and $\sigma^2(X)$ gives

$$\begin{aligned}\Lambda(X) &= \varepsilon\sum_{i\in\mathcal{A}}\frac{\delta_{ix_1}}{n}h_i + O\left(\varepsilon^2\right),\\ \sigma^2(X) &= \varepsilon^2\sum_{(i,j),(u,v)\in\mathcal{A}_\Pi}\left(\frac{n_{i\cdot}(X)}{n\pi_i^{(1)}}b_{ij} + \frac{n_{N\cdot}(X)}{n\pi_N^{(1)}}b_{NN}\right)\\ &\qquad\times\left(\frac{n_{u\cdot}(X)}{n\pi_u^{(1)}}b_{uv} + \frac{n_{N\cdot}(X)}{n\pi_N^{(1)}}b_{NN}\right)\frac{s_{ijuv}^{(2)}}{\lambda_2} + O\left(\varepsilon^3\right).\end{aligned}$$

We have $\sigma^2(X) > 0$ because the covariance matrix $\{s_{ijuv}^{(2)}, (i, j), (u, v) \in \mathcal{A}_\Pi\}$ is nonsingular.

The case $X \in \mathcal{A}^{(2)}$ can be considered in the same way.

So, under the asymptotics (8), (22) we get

$$\sqrt{n_*}\frac{\Lambda(X)}{\sigma(X)} \to \Delta^*(X) = \begin{cases}\Delta(X), & \text{if } X \in \mathcal{A}^n \setminus \mathcal{A}^{(1)} \cup \mathcal{A}^{(2)},\\ \operatorname{sign}(\Lambda(X)) \cdot \infty, & \text{if } X \in \mathcal{A}^{(1)} \cup \mathcal{A}^{(2)}.\end{cases}$$

Then under the conditions of the theorem the expectation in (21) satisfies the asymptotics:

$$\begin{aligned}\mathsf{E}\left\{\mathbf{1}\left(\widehat{\Lambda}(X,\mathbb{X})\right)\right\} &= 1-\Pr\left\{\widehat{\Lambda}(X,\mathbb{X})<0\right\}\\ &= 1-\Pr\left\{\sqrt{n_*}\frac{\widehat{\Lambda}(X,\mathbb{X})-\Lambda(X)}{\sigma(X)}<-\sqrt{n_*}\frac{\Lambda(X)}{\sigma(X)}\right\}\to 1-\Phi\left(-\Delta^*(X)\right).\end{aligned}$$

Combining this with (21) and taking into account that n is fixed and $\Phi\left(\Delta^*(X)\right) = 1-\Phi\left(-\Delta^*(X)\right)$, we have:

$$\begin{aligned} r &= \frac{1}{2} - \sum_{X\in\mathcal{A}^n} \widetilde{\Lambda}(X)\Phi\left(\Delta^*(X)\right)+o\left(1\right)\\ &= \frac{1}{2} - \sum_{\substack{X\in\mathcal{A}^n,\\ \Lambda(X)\geq 0}} \widetilde{\Lambda}(X) + \sum_{\substack{X\in\mathcal{A}^n,\\ \Lambda(X)\geq 0}} \widetilde{\Lambda}(X) - \sum_{X\in\mathcal{A}^n} \widetilde{\Lambda}(X)\Phi\left(\Delta^*(X)\right)+o\left(1\right)\\ &= r_0 + \sum_{\substack{X\in\mathcal{A}^n,\\ \Lambda(X)\geq 0}} \widetilde{\Lambda}(X)\left(1-\Phi\left(\Delta^*(X)\right)\right) - \sum_{\substack{X\in\mathcal{A}^n,\\ \Lambda(X)< 0}} \widetilde{\Lambda}(X)\Phi\left(\Delta^*(X)\right)+o\left(1\right).\end{aligned}$$

Since $\operatorname{sign}(\widetilde{\Lambda}(X)) = \operatorname{sign}(\Lambda(X)) = \operatorname{sign}(\Delta^*(X))$ for any $X\in\mathcal{A}^n$, we get:

$$\begin{aligned}&\sum_{X\in\mathcal{A}^n,\,\Lambda(X)\geq 0} \widetilde{\Lambda}(X)\left(1-\Phi\left(\Delta^*(X)\right)\right)\\ &\qquad= \sum_{\substack{X\notin\mathcal{A}^{(1)}\cup\mathcal{A}^{(2)},\\ \Lambda(X)\geq 0}} |\widetilde{\Lambda}(X)|\Phi\left(-|\Delta(X)|\right) + \sum_{\substack{X\in\mathcal{A}^{(1)}\cup\mathcal{A}^{(2)},\\ \Lambda(X)\geq 0}} |\widetilde{\Lambda}(X)|\Phi\left(-|\Delta^*(X)|\right),\end{aligned}$$

$$\begin{aligned}&\sum_{X\in\mathcal{A}^n,\,\Lambda(X)< 0} \widetilde{\Lambda}(X)\Phi\left(\Delta^*(X)\right)\\ &\qquad= \sum_{\substack{X\notin\mathcal{A}^{(1)}\cup\mathcal{A}^{(2)},\\ \Lambda(X)< 0}} -|\widetilde{\Lambda}(X)|\Phi\left(-|\Delta(X)|\right) - \sum_{\substack{X\in\mathcal{A}^{(1)}\cup\mathcal{A}^{(2)},\\ \Lambda(X)< 0}} |\widetilde{\Lambda}(X)|\Phi\left(-|\Delta^*(X)|\right),\end{aligned}$$

and $\Phi\left(-|\Delta^*(X)|\right)=0$ for any $X\in\mathcal{A}^{(1)}\cup\mathcal{A}^{(2)}$. So, we obtain:

$$r = r_0 + \sum_{X\notin\mathcal{A}^{(1)}\cup\mathcal{A}^{(2)}} |\widetilde{\Lambda}(X)|\Phi\left(-|\Delta(X)|\right)+o\left(1\right). \tag{24}$$

Now we consider $|\widetilde{\Lambda}(X)| = \frac{1}{2}|L_2(X)-L_1(X)|$. Using the Taylor expansion $\Lambda(X) = \varepsilon B_n(X) + O\left(\varepsilon^2\right)$ for $X\notin\mathcal{A}^{(1)}\cup\mathcal{A}^{(2)}$ we obtain:

$$|\widetilde{\Lambda}(X)| = \frac{L_1(X)}{2}\left|\frac{L_2(X)}{L_1(X)}-1\right| = \frac{L_1(X)}{2}\left|e^{n\Lambda(X)}-1\right| = \frac{nL_1(X)}{2}\left|\varepsilon B_n(X)+O\left(\varepsilon^2\right)\right|.$$

Putting this expansion into (24), omitting the terms of order $O\left(\varepsilon^2\right)$, and substituting $\varepsilon = c/\sqrt{n_*}$ we obtain (23). □

4. Discriminant Analysis of Markov Chains with Missing Values

Let there be missing values in the realization $X = (x_1, \ldots, x_n)$ of the Markov chain under classification. We use the vector M of missing value indicators in order to indicate the location of missing values in the realization X:

$$M = (m_1, m_2, \ldots, m_n), \qquad m_t \in \{0,1\}, \quad t \in \{1, \ldots, n\}, \tag{25}$$

that is assumed to be known and fixed. Here $m_t = 0$ means that the observation x_t is missing, $m_t = 1$ means that the value x_t is registered $(m_1 \equiv m_n \equiv 1)$. The vector M determines the model of data registration, in some sense it determines the experimental design. There are two approaches to describe the missing-data mechanisms [14]: the probabilistic model of M assuming that $m_1, m_2, \ldots$ is a random sequence (e.g., Bernoulli trials, a Markov chain) and the deterministic model assuming that M is a nonrandom parameter of the data registration process. In this paper we follow the second approach.

4.1. The likelihood function for a Markov chain with missing values.

Let T be the number of fragments without missing values in the realization X, so T is equal to the number of series of ones in the vector M $(T \geq 2)$. Let us represent (X, M) in the following form:

$$X = (x_1, \ldots, x_n) = \left(X_{(1)} \vdots \overline{X}_{(1)} \vdots X_{(2)} \vdots \ldots \vdots \overline{X}_{(T-1)} \vdots X_{(T)}\right),$$
$$X_{(t)} = \left(x_{(t),1}, x_{(t),2}, \ldots, x_{(t),M_t^*}\right), \qquad t \in \{1, \ldots, T\},$$

where $X_{(t)}$ is the tth observed fragment of length M_t^* of the realization X that corresponds to the tth series of ones in M; $\overline{X}_{(s)}$ is the sth missing fragment of length $\overline{M}_s^*$ of the realization X that corresponds to the tth series of zeroes in M.

Theorem 5. *The likelihood function of the Markov chain parameters (π, P) for the realization with missing values (X, M) is:*

$$L(\pi, P; X, M) = \pi_{x_{(1),1}} \left(\prod_{t=1}^{T} L_t(P, X_{(t)})\right) \left(\prod_{t=1}^{T-1} p_{x_{(t),M_t^*}, x_{(t+1),1}}(\overline{M}_t^* + 1)\right), \tag{26}$$

where $p_{ij}(k) = (P^k)_{ij}$ is the probability of the k-step transition from the state i to the state j; $L_s(P, X_{(s)})$ is the probability of the fragment $X_{(s)}$ given the fixed first state $x_{(s),1}$:

$$L_s = L_s(P; X_{(s)}) = \prod_{t=1}^{M_s^*-1} p_{x_{(s),t}, x_{(s),t+1}}. \tag{27}$$

Proof. One can see that the observed fragments of the realization (X, M) form the non-homogeneous Markov chain with the transition probabilities depending on

M [6]: the probability of the transition from x_t to x_{t+1} with $m_t = m_{t+1} = 1$ is equal to the probability of one-step transition $p_{x_t x_{t+1}}$; the probability of the transition from x_t to x_{t+k} with $m_t = m_{t+k} = 1$ and $k-1$ consecutive missing observations in between $(m_{t+1} = \ldots = m_{t+k-1} = 0)$ is equal to the probability of k-step transition $p_{x_t x_{t+k}}(k)$ calculated from the Kolmogorov-Chapman equation. □

One can see from (26) that the likelihood function is a complicated nonlinear function of the transition probabilities $\{p_{ij}\}$. Let us construct an approximation of the likelihood function (26).

Let $\overline{M}_-^* = \min_{1 \le t \le T-1} \overline{M}_t^*$ be the minimal length of the series of missing values in the realization X.

Theorem 6. *If the stationary Markov chain with parameters* (π, P) *is observed with missing values* (25)*, and there exists a positive integer* M_0 *such that*

$$\overline{M}_-^* \ge M_0, \qquad \rho = 1 - \min_{i,j \in \mathcal{A}} p_{ij}(M_0) < 1,$$

then the following multiplicative approximation of the likelihood function (26) *by likelihood functions for fragments without missing values* $\{L(\pi, P; X_{(t)})\}$ *is valid:*

$$L(\pi, P; X, M) = \prod_{t=1}^{T} L(\pi, P; X_{(t)}) + \delta(\pi, P; X, M), \tag{28}$$

$$\left| \frac{\delta(\pi, P; X, M)}{L(\pi, P; X, M)} \right| = O\left(T \rho^{\overline{M}_-^* / M_0} \right),$$

where $L(\pi, P; X_{(t)}) = \pi_{x_{(t),1}} L_t(P; X_{(t)})$, $L_t(\cdot;\cdot)$ *is defined in* (27).

Proof. Consider the case of one missing fragment $\big(X = (X_{(1)} \vdots \overline{X}_{(1)} \vdots X_{(2)})$, $T = 2\big)$ and evaluate the approximation accuracy $\delta(\pi, P; X, M)$:

$$\begin{aligned}
|\delta(\pi, P; X, M)| &= \left| L(\pi, P; X, M) - L(\pi, P; X_{(1)}) L(\pi, P; X_{(2)}) \right| \\
&= \left| \pi_{x_{(1),1}} L_1 L_2 p_{x_{(1),M_1^*}, x_{(2),1}}(\overline{M}_-^* + 1) - L(\pi, P; X_{(1)}) L(\pi, P; X_{(2)}) \right| \\
&= \left| \pi_{x_{(1),1}} L_1 L_2 \big(p_{x_{(1),M_1^*}, x_{(2),1}}(\overline{M}_-^* + 1) - \pi_{x_{(2),1}} + \pi_{x_{(2),1}} \big) - \pi_{x_{(1),1}} L_1 \pi_{x_{(2),1}} L_2 \right| \\
&= \pi_{x_{(1),1}} L_1 L_2 | p_{x_{(1),M_1^*}, x_{(2),1}}(\overline{M}_-^* + 1) - \pi_{x_{(2),1}} |.
\end{aligned}$$

Using the inequality $|p_{x_{(1),M_1^*}, x_{(2),1}}(\overline{M}_-^* + 1) - \pi_{x_{(2),1}}| \le c \rho^{[(\overline{M}_-^* + 1)/M_0] - 1}$ (see [3]), we get:

$$\left| \frac{\delta(\pi, P; X, M)}{L(\pi, P; X, M)} \right| \le \frac{c}{p_{x_{(1),M_1^*}, x_{(2),1}}(\overline{M}_-^* + 1)} \rho^{[(\overline{M}_-^* + 1)/M_0] - 1}.$$

The case $T > 2$ is considered in a similar way. □

Remark 3. By (2) one can take $M_0 = 1$ and $\rho = 1 - \min_{i,j \in \mathcal{A}} p_{ij}$, $\rho \in (0, 1)$.

We will use the asymptotics of increasing lengths of series of missing values: $\overline{M}_-^* \to \infty$. In practice, this asymptotics corresponds to "switches" of the observer for long time periods between registration of the realization X and registration of other realizations. Note that under the probabilistic approach to missing-data mechanism [14], the vector M can be generated as a realization of a binary Markov chain with "attraction": $\Pr\{m_{t+1} = 1 \mid m_t = 1\}$ and $\Pr\{m_{t+1} = 0 \mid m_t = 0\}$ are close to 1.

Corollary 3. *Under the assumptions of Theorem* 6 *and asymptotics of increasing number of series and increasing lengths of series of missing values,*

$$T \to \infty, \quad \overline{M}_-^* \to \infty, \quad T\rho^{\overline{M}_-^*} \to 0, \tag{29}$$

the following almost sure convergence holds:

$$\left| \frac{\delta(\pi, P; X, M)}{L(\pi, P; X, M)} \right| \to 0.$$

Thus under the conditions of Corollary 3 the fragments without missing values of the realization X may be interpreted as a set of independent "subrealizations" that are described by the same model of the Markov chain but without missing values.

Let us assume that the asymptotics (29) holds and so the approximation error in (28) may be neglected. Therefore we shall use the following multiplicative approximation:

$$L(\pi, P; X, M) = \prod_{t=1}^{T} L(\pi, P; X_{(t)}). \tag{30}$$

This enables us to use the results of Sections 2 and 3 of this paper in case of missing values.

4.2. A decision rule in case of known parameters of the classes. The approximation (30) under the assumption (29) enables us to generalize the results of Section 2 for the case of missing values in the realization X under classification.

Let $M^* = \sum_{t=1}^{T} M_t^*$ be the total number of registered observations in the realization with missing values (X, M).

Theorem 7. *Under the asymptotics*

$$M^*, T, \overline{M}_-^* \to \infty, \quad T\rho_l^{\overline{M}_-^*} \to 0, \qquad l \in \{1, \dots, L\}, \tag{31}$$

the BDR using the approximated likelihood functions for the model (1), (3), (25) *is*:

$$d(X) = \arg\max_{1 \le l \le L} \left(\frac{1}{M^*} \log q_l + \frac{1}{M^*} \sum_{i \in \mathcal{A}} \nu_i \log \pi_i^{(l)} + \sum_{i,j \in \mathcal{A}} \widehat{\Pi}_{ij} \log p_{ij}^{(l)} \right), \tag{32}$$

$$\widehat{\Pi}_{ij} = \frac{n_{ij}}{M^*}, \quad n_{ij} = \sum_{t=1}^{n-1} m_t m_{t+1} \cdot \boldsymbol{I}\{x_t = i, x_{t+1} = j\}, \quad \nu_i = \sum_{t=1}^{T} \boldsymbol{I}\{x_{(t),1} = i\},$$

where $i, j \in \mathcal{A}$, $\rho_l = 1 - \min_{i,j\in\mathcal{A}} p_{ij}^{(l)}$.

Proof. The proof follows the lines of the proof of Theorem 1 using the approximate likelihood function (30). □

Let us now find the misclassification probability (4) of the BDR (32) in the case of two classes ($L = 2$).

Theorem 8. *For* $L = 2$ *under the asymptotics* (31) *and the contiguous classes asymptotics* (8),

$$\varepsilon = \frac{c}{\sqrt{M^*}} \to 0, \quad T\varepsilon \to 0, \qquad 0 < c < \infty,$$

the misclassification probability (4) *of the BDR* (32) *has the limit*:

$$r_0 \to \tilde{r}_0 = q_1 \Phi\Big(-\frac{\Delta_1}{2}\Big) + q_2 \Phi\Big(-\frac{\Delta_2}{2}\Big),$$

where Δ_1, Δ_2 *are defined in* (13).

Proof. One can see that under the conditions of the theorem the approximation (30) of the likelihood function for the realization (X, M) is valid. Therefore the statistical estimators $\{\widehat{\Pi}_{ij}\}$ from “incomplete” data have the same asymptotic properties as statistical estimators from “full” data. The proof follows the lines of the proof of Theorem 2. □

4.3. A DECISION RULE IN CASE OF UNKNOWN PARAMETERS OF THE CLASSES. Consider the case where the parameters of the classes (1) are unknown and the “training sample” $\mathbb{X}$ is observed also with missing values:

$$\mathbb{X} = \big\{(X^{(1)}, M^{(1)}), (X^{(2)}, M^{(2)}), \ldots, (X^{(L)}, M^{(L)})\big\},$$

where for each lth realization $X^{(l)}$ of length n_l from the class Ω_l there is the corresponding vector of miss-indicators $M^{(l)} = \big(m_1^{(l)}, m_2^{(l)}, \ldots, m_{n_l}^{(l)}\big)$, $m_t^{(l)} \in \{0, 1\}$, $t \in \{1, \ldots, n_l\}$, $l \in \{1, \ldots, L\}$.

Let T_l be the number of fragments without missing values in the realization $X^{(l)}$ ($T_l \geq 2$). Let $X_{(t)}^{(l)}$ be the tth observed fragment of the realization $X^{(l)}$ that corresponds to the tth series of ones in $M^{(l)}$; let $M_{(l),t}^{*}$ denote the length of $X_{(t)}^{(l)}$; let $\overline{X}_{(s)}^{(l)}$ be the sth missing fragment of the realization $X^{(l)}$ that corresponds to the sth series of zeroes in $M^{(l)}$; let $\overline{M}_{(l),t}^{*}$ denote the length of $\overline{X}_{(t)}^{(l)}$, $t \in \{1, \ldots, T_l\}$, $s \in \{1, \ldots, T_l - 1\}$, $l \in \{1, \ldots, L\}$. Let $M_{(l)}^{*} = \sum_{t=1}^{T_l} M_{(l),t}^{*}$ be the number of registered observations in the realization $X^{(l)}$; let $\overline{M}_{(l),-}^{*} = \min_{1\leq t\leq T_l-1} \overline{M}_{(l),t}^{*}$ denote the minimal length of the fragment of missing values in the realization $X^{(l)}$, $l \in \{1, \ldots, L\}$.

The asymptotics

$$(33) \qquad T_l \to \infty, \quad \overline{M}_{(l),-}^{*} \to \infty, \quad T_l \rho_l^{\overline{M}_{(l),-}^{*}} \to 0, \qquad l \in \{1, \ldots, L\},$$

enables us to use the approximation (30) of the likelihood functions for all realizations from $\mathbb{X}$.

As in Section 3 we shall use the plug-in DR that is obtained from the BDR (32) if the unknown parameters $\{P^{(l)}\}$ are replaced by their estimators $\{\widehat{P}^{(l)}\}$:

$$(34)\ d(X,\mathbb{X}) = \arg\max_{1\le l\le L}\left(\frac{1}{M^*}\log q_l + \frac{1}{M^*}\sum_{i\in\mathcal{A}}\nu_i\log\hat{\pi}_i^{(l)} + \sum_{i,j\in\mathcal{A}}\widehat{\Pi}_{ij}\log\hat{p}_{ij}^{(l)}\right),$$

$$\widehat{\Pi}_{ij} = \frac{n_{ij}}{M^*},\quad n_{ij} = \sum_{t=1}^{n-1} m_t m_{t+1}\cdot \boldsymbol{I}\{x_t = i, x_{t+1} = j\},$$

$$\hat{p}_{ij}^{(l)} = \frac{n_{ij}^{(l)}}{n_{i\cdot}^{(l)}},\quad \hat{\pi}_i^{(l)} = \frac{n_{i\cdot}^{(l)}}{n_l},\quad n_{ij}^{(l)} = \sum_{t=1}^{n_l-1} m_t^{(l)} m_{t+1}^{(l)}\cdot \boldsymbol{I}\{x_t^{(l)} = i, x_{t+1}^{(l)} = j\},$$

where the bivariate frequencies $\{n_{ij}^{(l)}\}$ are calculated from the observed fragments of the realization $X^{(l)}$, $i,j\in\mathcal{A}$, $l\in\{1,\dots,L\}$.

Let us find now the misclassification probability (4) of the DR (34) in the case of two classes ($L=2$).

Theorem 9. *For $L=2$ under the asymptotics* (31), (33)*, and the contiguous classes asymptotics* (8)*,*

$$M_{(l)}^* \to \infty,\quad M_{(l)}^*/M^* = \tilde{\lambda}_l > 0,\quad \varepsilon = \frac{c}{\sqrt{M^*}} \to 0,\quad T\varepsilon \to 0,\quad 0 < c < \infty,$$

the misclassification probability (4) *of the DR* (34) *has the limit*:

$$r \to \tilde{r} = q_1\Phi\biggl(-\frac{\widetilde{\Delta}_1}{2}\biggr) + q_2\Phi\biggl(-\frac{\widetilde{\Delta}_2}{2}\biggr),$$

where $\widetilde{\Delta}_1$, $\widetilde{\Delta}_2$ are defined in (18)*.*

Proof. One can see that under the conditions of the theorem the approximation (30) of the likelihood functions for the realizations (X,M), $(X^{(l)},M^{(l)})$ is valid. Therefore the statistical estimators $\{\widehat{\Pi}_{ij}\}$ and $\{\hat{p}_{ij}^{(l)}\}$ from "incomplete" data have the same asymptotic properties as statistical estimators from "full" data. The proof follows the lines of the proof of Theorem 3. □

5. Conclusion

In this paper the classification statistical problem for stationary finite Markov chains is considered for different levels of prior information: the Bayesian decision rule and the plug-in Bayesian decision rule are constructed and their performance is evaluated in the case of two contiguous classes and increasing number of observations.

The obtained results are generalized to the case of missing values in realizations to be classified and in "training samples".

Acknowledgements

The authors would like to thank the anonymous referee, Prof. Dr. D. M. Chibisov, and Prof. Dr. H. G. Bock for their helpful comments.

References

[1] H. A. Amagor, *A Markov analysis of DNA squences*, J. Theoret. Biol., 104 (1983), 633–642.

[2] I. V. Basawa and B. L. C. Rao, *Statistical Inference for Stochastic Processes*, Academic Press, New York, 1980.

[3] P. Billingsley, *Statistical Methods in Markov Chains*, Ann. Math. Statist., 32 (1961), 12–40.

[4] A. A. Borovkov, *Mathematical Statistics*, Nauka, Moscow, 1984. (In Russian.)

[5] D. M. Chibisov, *A theorem on admissible tests and its application to an asymptotical problem of testing hypotheses*, Theor. Veroyat. Primen., 12 (1967), 96–111. (In Russian.)

[6] B. F. Cole, M. L. T. Lee, G. A. Whitmore, and A. M. Zaslavsky, *An empirical Bayes model for Markov-dependent binary sequences with randomly missing observations*, J. Amer. Statist. Assoc., 90 (1995), 1364–1372.

[7] R. A. Horn, C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1986.

[8] W.-H. Ju and Y. Vardi, *A hybrid high-order Markov chain model for computer intrusion detection*, J. Comput. Graphical Statist., 10 (2001), 277–295.

[9] D. Kazakos, *The Bhattacharyya distance and detection between Markov chains*, IEEE Trans. Inform. Theory, IT-24 (1978), 747–745.

[10] Yu. S. Kharin, *Robustness in Statistical Pattern Recognition*, Kluwer, Dordrecht–Boston–London, 1996.

[11] L. H. Koopmans, *Asymptotic rate of discrimination for Markov processes*, Ann. Math. Statist., 31 (1960), 982–994.

[12] M. T. L. Lee, *A two-scale Markov model. Statistical issues in drug testing and drug evaluation*, Comm. Statist. – Theory Methods, 23 (1994), 615–623.

[13] L. M. Le Cam and G. L. Yang, *Asymptotics in Statistics: Some Basic Concepts*, Springer, 1990.

[14] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, Wiley, 1987.

[15] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York, 1992.

[16] M. B. Malyutov and I. I. Tsitovich, *Second order optimal sequential discrimination between Markov chains*, Math. Methods Statist., 10 (2001), 446–464.

[17] A. V. Nagaev, *On the Precise Asymptotic Formula for the Bayes Risk Under Discriminating Between Two Markov Chains*, Preprint, 2000.

[18] V. Petrov, *Sums of Independent Random Variables*, Springer, New York, 1972.

[19] P. Scheffel and H. v. Weizsacker, *On risk rates and large deviations in finite Markov chain experiments*, Math. Methods Statist., 6 (1997), 293–312.

[20] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*, Wiley, New York, 1980.

[21] M. S. Waterman *et al.*, *Mathematical Methods for DNA Sequences*, CRC Press, Boca Raton, 1989.

[*Received March* 2003; *revised May* 2004]