MINIMAX NONPARAMETRIC PREDICTION UNDER RANDOM SAMPLE SIZE

MACIEJ WILCZYŃSKI*

Received May 31, 2000

ABSTRACT. Let U_0 be a random vector taking its values in a measurable space and having an unknown distribution P. Let $U_1, U_2, \ldots, U_N$ and $V_1, V_2, \ldots, V_m$ be independent simple random samples from P of a random size N and a fixed size m, respectively. Further, let $z_1, z_2, \ldots, z_k$ be real valued bounded functions defined on the same space. Assuming that only the first sample is observed, we find a minimax predictor $\boldsymbol{d^0}(N, U_1, \ldots, U_N)$ of the vector $\boldsymbol{Y^m} = \sum_{j=1}^m (z_1(V_j), z_2(V_j), \ldots, z_k(V_j))^T$ with respect to a quadratic error loss function.

1. Introduction

Let U_0 be a random vector taking its values in a measurable space $(\mathcal{Y}, \mathcal{B})$ whose unknown distribution P is assumed to be an element of the set

$$\mathcal{P} = \{ \text{ all probability measures on } (\mathcal{Y}, \mathcal{B}) \ \}.$$

Let $U_1, U_2, \ldots, U_N$ and $V_1, V_2, \ldots, V_m$ be independent, simple random samples from P of a random size N and a fixed size m, respectively. We assume that N is an ancillary statistics, i.e. a random variable, which takes values in a set $\{0, 1, 2, \ldots\}$, and whose known distribution does not depend on P. Further, let $\boldsymbol{z} = (z_1, z_2, \ldots, z_k)^T$ be a measurable, bounded function on the space $(\mathcal{Y}, \mathcal{B})$ with values in $(R^k, \mathcal{B}_{R^k})$. In the paper we consider the problem of predicting the value of a k-dimensional random vector $\boldsymbol{Y^m} = \sum_{j=1}^m \boldsymbol{z}(V_j)$ from the data $\boldsymbol{U^N} = (U_1, \ldots, U_N)$. Assuming that the loss function has the form

$$L(\boldsymbol{d}, \boldsymbol{Y^m}) = (\boldsymbol{d} - \boldsymbol{Y^m})^{\boldsymbol{T}} \boldsymbol{C} (\boldsymbol{d} - \boldsymbol{Y^m}), \tag{1}$$

where $\boldsymbol{C} = [c_{ij}]$ is nonnegative definite, symmetric $k \times k$ matrix, we find a minimax solution of the above problem of prediction. As we show, the minimax predictor $\boldsymbol{d^0}(N, \boldsymbol{U^N})$ of $\boldsymbol{Y^m}$ is an affine (inhomogeneous linear) function of the random vector $\boldsymbol{X^N} = \sum_{j=1}^N \boldsymbol{z}(U_j)$.

Using this result we find, for each $n \geq 1$, the predictor $\boldsymbol{d^1}(n, \boldsymbol{U^n})$ which is minimax when the value of $\boldsymbol{Y^m}$ is predicted from the sample $U_1, U_2, \ldots, U_n$ of a fixed size n. Then we show that the decision rule $\boldsymbol{d^1}(N, \boldsymbol{U^N})$ is not minimax when the sample size N is random and takes at least two different values with positive probabilities. This is an ancillarity paradox, because $\boldsymbol{d^1}(N, U^N)$ seems to be the best candidate for a minimax predictor of $\boldsymbol{Y^m}$ when the sample size N is random.

The first example of such an ancillarity paradox was given by Brown [3]. He showed that in the multiple linear regression the admissibility of the ordinary estimator of the constant

---

1991 *Mathematics Subject Classification.* 62C20; Secondary 62G05.

Key words and phrases. Minimax prediction, Dirichlet process, nonparametric prediction, ancillarity paradox.

Research supported by the grant KBN 341459.

This paper is in final form and no version of it will be submitted for publication elsewhere.

term depends on the distribution of the design matrix, which is an ancillary statistics. Next example of this paradox was presented by Kun He [6] who considered estimation of the multinomial probabilities $\boldsymbol{p} = (p_1, p_1, \dots, p_k)^T$ with respect to the loss (1), in which $\boldsymbol{C}$ was the identity matrix. He proved that the estimator of $\boldsymbol{p}$, which is minimax when the sample size is fixed, is neither minimax nor admissible when the sample size is random. Analogous results were presented by Amrhein [1] who studied minimax estimation of the multivariate hypergeometric proportion $p_i = M_i/M, i = 1, \dots, r$ with respect to the same loss as in Kun He.

In the last part of the paper we find minimax predictor of $\boldsymbol{Y^m}$ when the distribution of the size N of the observed sample is unknown.

2. Minimax estimate.

Before stating the main result we will introduce the following notation: We denote by $\boldsymbol{Z}$, $\boldsymbol{p}$ and $R_1(P)$ the random vector $\boldsymbol{z}(U_0)$, its expected value and the sum of the variances of its components weighted by the matrix $\boldsymbol{C}$, i.e. we put

$$\begin{aligned} \boldsymbol{Z} &= \boldsymbol{z}(U_0) \\ \boldsymbol{p} &= E_P \boldsymbol{Z}, \\ R_1(P) &= E_P \boldsymbol{(Z - p)^T\, C\, (Z - p)}. \end{aligned} \tag{2}$$

Now, let (P_j) be any sequence of probability measures on $(\mathcal{Y}, \mathcal{B})$ such that

$$\lim_{j\to\infty} R_1(P_j) = \sup_{P\in\mathcal{P}} R_1(P) \tag{3}$$

and let $(\boldsymbol{p_j})$, where

$$\boldsymbol{p_j} = E_{P_j} \boldsymbol{Z}, \tag{4}$$

be the corresponding sequence of points from the convex k–dimensional cube

$$\mathcal{M} = [-M, M]^k,$$

where

$$M^2 \stackrel{def}{=} \sup_{y\in\mathcal{Y}} \boldsymbol{z}(y)^T \boldsymbol{z}(y).$$

Because of the boundedness of $\boldsymbol{z}$, the number M is finite and $\mathcal{M}$ is compact in R^k. Therefore, the sequence $(\boldsymbol{p_j})$ has a cluster point, which will be denoted throughout by $\boldsymbol{p_0}$.

Suppose now that the following condition is satisfied

$$m[f(0) - \sum_{n=1}^{\infty} \frac{f(n)}{n}] \le 1, \tag{5}$$

where $f(n)$, $n \ge 0$, denotes the probability that the random variable N takes the value n. Then, by the same arguments as in Kun He [6], there exists a positive real number A_1 which satisfies the following equation

$$\sum_{n=0}^{\infty} \frac{(n+A)^2 + nm - mA^2}{(n+A)^2} f(n) = 0. \tag{6}$$

Since the above series is a decreasing function of the variable $A > 0$ this number is unique. Moreover, $A_1 < \infty \iff m > 1$.

Now, let the number A_0 be defined by

$$A_0 = \begin{cases} A_1, & if \quad m[f(0) - \sum_{n=1}^{\infty} \frac{f(n)}{n}] \leq 1, \\ 0, & if \quad m[f(0) - \sum_{n=1}^{\infty} \frac{f(n)}{n}] > 1. \end{cases} \tag{7}$$

Then the following theorem, which is the main result of the paper, holds.

Theorem 1. *If $m > 1$, then*

$$\boldsymbol{d^0}(N, \boldsymbol{U^N}) = \begin{cases} m\dfrac{\boldsymbol{X^N} + A_0\boldsymbol{p_0}}{N + A_0}, & if \quad N > 0, \\ m\boldsymbol{p_0}, & if \quad N = 0 \end{cases} \tag{8}$$

is the minimax predictor of the unobservable vector $\boldsymbol{Y^m}$ and its minimax risk equals

$$\sup_{P\in\mathcal{P}} R(\boldsymbol{d^0}, P) = m^2 \left[f(0) + \sum_{n=1}^{\infty} \frac{A_0^2 f(n)}{(n + A_0)^2} \right] \sup_{P\in\mathcal{P}} R_1(P).$$

If $m = 1$, then

$$\boldsymbol{d^0}(N, \boldsymbol{U^N}) = \boldsymbol{p_0} \tag{9}$$

is the minimax predictor of $\boldsymbol{Y^1}$ and

$$\sup_{P\in\mathcal{P}} R(\boldsymbol{d^0}, P) = \sup_{P\in\mathcal{P}} R_1(P).$$

3. Proof of the main result

Let $\mathcal{D}$ stand for the class of all predictors $\boldsymbol{d}$ of the unobservable vector $\boldsymbol{Y^m}$. For a predictor $\boldsymbol{d} = \boldsymbol{d}(N, \boldsymbol{U^N}) \in \mathcal{D}$ we denote by $R(\boldsymbol{d}, P)$ the risk function for $\boldsymbol{d}$, i.e. we put

$$R(\boldsymbol{d}, P) = E_P L(\boldsymbol{d}, \boldsymbol{Y^m}) = E_P \left(\boldsymbol{d}(N, \boldsymbol{U^N}) - \boldsymbol{Y^m} \right)^T \boldsymbol{C} \left(\boldsymbol{d}(N, \boldsymbol{U^N}) - \boldsymbol{Y^m} \right).$$

Since the vectors $\boldsymbol{U^N}$ and $\boldsymbol{Y^m}$ are independent and since $\boldsymbol{z}(V_1), \ldots, \boldsymbol{z}(V_m)$ are i.i.d. random vectors with the expected values equal to $\boldsymbol{p}$,

$$E_P \boldsymbol{Y^m} = E_P \sum_{j=1}^{m} \boldsymbol{z}(V_j) = m\boldsymbol{p} \tag{10}$$

and

$$R(\boldsymbol{d}, P) = E_P(\boldsymbol{d} - m\boldsymbol{p})^T \boldsymbol{C}(\boldsymbol{d} - m\boldsymbol{p}) + E_P(\boldsymbol{Y^m} - m\boldsymbol{p})^T \boldsymbol{C}(\boldsymbol{Y^m} - m\boldsymbol{p}).$$

Moreover,

$$E_P(\boldsymbol{Y^m} - m\boldsymbol{p})^T \boldsymbol{C}(\boldsymbol{Y^m} - m\boldsymbol{p}) = mE_P \boldsymbol{(Z - p)^T \, C \, (Z - p)} = mR_1(P), \tag{11}$$

which implies that the risk for any predictor $\boldsymbol{d}(N, \boldsymbol{U^N}) \in \mathcal{D}$ can be rewritten as

$$R(\boldsymbol{d}, P) = E_P(\boldsymbol{d} - m\boldsymbol{p})^T \boldsymbol{C}(\boldsymbol{d} - m\boldsymbol{p}) + mR_1(P). \tag{12}$$

According to the definition of minimaxity, to prove that the predictor $\boldsymbol{d^0}(N, \boldsymbol{U^N})$ defined in Theorem 1 is minimax it is necessary to show that

$$\sup_{P\in\mathcal{P}} R(\boldsymbol{d^0}, P) = \inf_{d\in\mathcal{D}} \sup_{P\in\mathcal{P}} R(\boldsymbol{d}, P). \tag{13}$$

To prove this result for $m > 1$ we use the method which is analogous to that proposed in Wilczynski [7]. First we show that $\boldsymbol{d^0}$ is minimax if the class of predictors is restricted to a subset $\mathcal{D}_0 \subset \mathcal{D}$ which consists of all predictors $\boldsymbol{d^a}$, with $\boldsymbol{a} \in \mathcal{M}$, of the form

$$\boldsymbol{d^a}(N, \boldsymbol{U^N}) = \begin{cases} m\dfrac{\boldsymbol{X^N} + A_0\boldsymbol{a}}{N + A_0}, & if \quad N > 0, \\ m\boldsymbol{a}, & if \quad N = 0. \end{cases} \tag{14}$$

Next we calculate the upper bound for the risk $R(\boldsymbol{d^0}, P)$ of $\boldsymbol{d^0} = \boldsymbol{d^{p_0}}$ and then, via nonparametric Bayes approach, we construct a sequence of priors on $\mathcal{P}$ for which the corresponding sequence of Bayes risks converges to this upper bound. From this we deduce minimaxity of $\boldsymbol{d^0}$ when $m > 1$. Then, using a different approach, we prove minimaxity of $\boldsymbol{d^0}$ for $m = 1$.

We begin the whole proof from the first case in which $m > 1$ and the condition (5) holds, which implies that $A_0 \in (0, \infty)$. For simplicity we denote the risk function of a predictor $\boldsymbol{d^a} \in \mathcal{D}_0$ by $R(\boldsymbol{a}, P)$. Since the number A_0 satisfies the equation (6), we obtain, by (12) and (14),

$$\begin{aligned} R(\boldsymbol{a}, P) &= mR_1(P) + m^2 \sum_{n=0}^{\infty} \frac{nR_1(P) + A_0^2(\boldsymbol{a} - \boldsymbol{p})^{\boldsymbol{T}}\boldsymbol{C}(\boldsymbol{a} - \boldsymbol{p})}{(n + A_0)^2} f(n) \\ &= m \sum_{n=0}^{\infty} \frac{[(n + A_0)^2 + mn]R_1(P) + mA_0^2(\boldsymbol{a} - \boldsymbol{p})^{\boldsymbol{T}}\boldsymbol{C}(\boldsymbol{a} - \boldsymbol{p})}{(n + A_0)^2} f(n) \\ &= m^2 \left[R_1(P) + (\boldsymbol{a} - \boldsymbol{p})^{\boldsymbol{T}}\boldsymbol{C}(\boldsymbol{a} - \boldsymbol{p})\right] \sum_{n=0}^{\infty} \frac{A_0^2 f(n)}{(n + A_0)^2}, \\ &= m^2 \left[R_1(P) + (\boldsymbol{a} - \boldsymbol{p})^{\boldsymbol{T}}\boldsymbol{C}(\boldsymbol{a} - \boldsymbol{p})\right] \left[f(0) + \sum_{n=1}^{\infty} \frac{A_0^2 f(n)}{(n + A_0)^2}\right], \\ &= m^2 \left[E_P\boldsymbol{Z^T C Z} - 2\boldsymbol{a^T C} E_P \boldsymbol{Z} + \boldsymbol{a^T C a}\right] \left[f(0) + \sum_{n=1}^{\infty} \frac{A_0^2 f(n)}{(n + A_0)^2}\right]. \end{aligned} \tag{15}$$

This results from the equalities (cf. (10) and (11))

$$E_P\boldsymbol{X^n} = n\boldsymbol{p} \quad \text{and} \quad E_P(\boldsymbol{X^n} - n\boldsymbol{p})^T\boldsymbol{C}(\boldsymbol{X^n} - n\boldsymbol{p}) = nR_1(P), \quad n \geq 1$$

and from the boundedness of the random vector $\boldsymbol{Z}$, which implies that the function $R_1(P)$ can be rewritten as

$$R_1(P) = E_P(\boldsymbol{Z} - \boldsymbol{p})^{\boldsymbol{T}}\, \boldsymbol{C}\, (\boldsymbol{Z} - \boldsymbol{p}) = E_P\boldsymbol{Z^T C Z} - \boldsymbol{p^T C p}.$$

Obviously, to prove that the decision rule $\boldsymbol{d^0}(N, \boldsymbol{U^N})$, defined by (8), is minimax in $\mathcal{D}_0$ it suffices to show that

$$\sup_{P\in\mathcal{P}} R(\boldsymbol{p_0}, P) = \inf_{a\in\mathcal{M}} \sup_{P\in\mathcal{P}} R(\boldsymbol{a}, P).$$

This can easily be deduced from the paper of Wilczyński [7] in which it is proved, using minmax Nikaido Theorem (cf. Aubin [2]), that the function $R(\boldsymbol{a}, P)$ (multiplied by some constant) satisfies the following condition

$$\sup_{P\in\mathcal{P}} R(\boldsymbol{p_0}, P) = \inf_{a\in\mathcal{M}} \sup_{P\in\mathcal{P}} R(\boldsymbol{a}, P) = \sup_{P\in\mathcal{P}} \inf_{a\in\mathcal{M}} R(\boldsymbol{a}, P). \tag{16}$$

This implies that the predictor $\boldsymbol{d^0}(N, \boldsymbol{U^N})$ is minimax in $\mathcal{D}_0$ and its minimax risk equals

$$\inf_{a\in\mathcal{M}} \sup_{P\in\mathcal{P}} R(\boldsymbol{a}, P) = m^2 \left[f(0) + \sum_{n=1}^{\infty} \frac{A_0^2 f(n)}{(n + A_0)^2}\right] \sup_{P\in\mathcal{P}} R_1(P) \tag{17}$$

because, for a fixed distribution $P \in \mathcal{P}$, the convex function $R(\boldsymbol{a}, P)$ of the variable $\boldsymbol{a}$ attains its global minimum over $\mathcal{M}$ at the point $\boldsymbol{a}(P) = \boldsymbol{p}$.

To show that $\boldsymbol{d^0}(N, \boldsymbol{U^N})$ is minimax in $\mathcal{D}$ we make use of the nonparametric Bayes approach proposed in Ferguson [5]. The structure of the arguments will be analogous to those appearing in Wilczyński [7].

Let Π_j, $j \geq 1$, be a Dirichlet prior process on $(\mathcal{Y}, \mathcal{B})$ with a parameter $\beta_j = A_0 P_j$, where (P_j) is a sequence defined by (3). Note first that, by (12), the Bayes predictor of $\boldsymbol{Y^m}$ is equal to the Bayes estimator of the parameter $m\boldsymbol{p}$. Therefore, by Ferguson [5] example b, the Π_j Bayes predictor of $\boldsymbol{Y^m}$ is given by

$$m\left[\frac{A_0 E_{P_j}\boldsymbol{Z}}{n + A_0} + \frac{n}{n + A_0}\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{z}(U_j)\right] = m\frac{\boldsymbol{X^n} + A_0\boldsymbol{p}_j}{n + A_0} = \boldsymbol{d^{p_j}}(n, \boldsymbol{U^n}),$$

whenever $N = n \geq 0$. This implies that $\boldsymbol{d^{p_j}}(N, \boldsymbol{U^N})$ is the Π_j Bayes predictor of $\boldsymbol{Y^m}$. Moreover, the Bayes risk $\rho(j)$ for this decision rule is given by

$$\rho(j) \stackrel{def}{=} E_{\Pi_j} R(\boldsymbol{d^{p_j}}, P) = E_{\Pi_j} R(\boldsymbol{p_j}, P) = m^2\left[f(0) + \sum_{n=1}^{\infty}\frac{A_0^2 f(n)}{(n + A_0)^2}\right] R_1(P_j),$$

because, by (15), the risk $R(\boldsymbol{p_j}, P)$ of the predictor $\boldsymbol{d^{p_j}}(N, \boldsymbol{U^N})$ equals

$$R(\boldsymbol{p_j}, P) = m^2\left[f(0) + \sum_{n=1}^{\infty}\frac{A_0^2 f(n)}{(n + A_0)^2}\right][E_P \boldsymbol{Z^T C Z} - 2\boldsymbol{p_j^T C} E_P \boldsymbol{Z} + \boldsymbol{p_j^T C p_j}]$$

and (cf. Ferguson [5] Theorem 3)

$$E_{\Pi_j}[E_P \boldsymbol{Z^T C Z}] = E_{P_j}\boldsymbol{Z^T C Z} \quad \text{and} \quad E_{\Pi_j}[E_P \boldsymbol{Z}] = E_{P_j}\boldsymbol{Z} = \boldsymbol{p_j}. \tag{18}$$

Since $\lim_{j \longrightarrow \infty} R_1(P_j) = \sup_{P \in \mathcal{P}} R_1(P)$, the Bayes risk $\rho(j)$ converges to

$$m^2\left[f(0) + \sum_{n=1}^{\infty}\frac{A_0^2 f(n)}{(n + A_0)^2}\right]\sup_{P \in \mathcal{P}} R_1(P)$$

which, by (17), is the upper bound for the risk of $\boldsymbol{d^0}(N, \boldsymbol{U^N})$. This implies that $\boldsymbol{d^0}(N, \boldsymbol{U^N})$ is minimax (see Ferguson [4], Theorem 2, p.91), when (5) holds and $m > 1$.

Now we consider the second case in which $m > 1$ and the condition (5) is not satisfied. Then $A_0 = 0$ and, as it is easy to calculate, the risk function for the predictor $\boldsymbol{d^0}$ is given by

$$R(\boldsymbol{d^0}, P) = mR_1(P) + m^2\left[(\boldsymbol{p_0} - \boldsymbol{p})^{\boldsymbol{T}}\boldsymbol{C}(\boldsymbol{p_0} - \boldsymbol{p})f(0) + R_1(P)\sum_{n=1}^{\infty}\frac{f(n)}{n}\right].$$

Since $m[f(0) - \sum_{n=1}^{\infty}\frac{f(n)}{n}] > 1$, this risk satisfies the inequality

$$R(\boldsymbol{d^0}, P) \leq m^2[(\boldsymbol{p_0} - \boldsymbol{p})^{\boldsymbol{T}}\boldsymbol{C}(\boldsymbol{p_0} - \boldsymbol{p}) + R_1(P)]f(0),$$

which immediately implies that the upper bound for the risk of $\boldsymbol{d^0}$ is given by

$$\sup_{P \in \mathcal{P}} R(\boldsymbol{d^0}, P) \leq m^2 f(0) \sup_{P \in \mathcal{P}} R_1(P), \tag{19}$$

because, by (16) and (15),

$$\begin{aligned} &\sup_{P\in\mathcal{P}} \left[(\boldsymbol{p_0}-\boldsymbol{p})^{\boldsymbol{T}}\boldsymbol{C}(\boldsymbol{p_0}-\boldsymbol{p}) + R_1(P)\right] = \\ (20) \qquad &= \inf_{a\in\mathcal{M}} \sup_{P\in\mathcal{P}} \left[(\boldsymbol{a}-\boldsymbol{p})^{\boldsymbol{T}}\boldsymbol{C}(\boldsymbol{a}-\boldsymbol{p}) + R_1(P)\right] \\ &= \sup_{P\in\mathcal{P}} \inf_{a\in\mathcal{M}} \left[(\boldsymbol{a}-\boldsymbol{p})^{\boldsymbol{T}}\boldsymbol{C}(\boldsymbol{a}-\boldsymbol{p}) + R_1(P)\right] = \sup_{P\in\mathcal{P}} R_1(P). \end{aligned}$$

As before, to prove minimaxity of $\boldsymbol{d^0}$ we construct a sequence of priors on $\mathcal{P}$ for which the corresponding sequence of Bayes risk converges to this upper bound. From this we deduce minimaxity of $\boldsymbol{d^0}(N, Y^N)$.

Let Π_j, $j \geq 1$, be a Dirichlet prior process on $(\mathcal{Y}, \mathcal{B})$ with a parameter $\alpha_j = A_j P_j$, where (A_j) is a sequence of positive real numbers, which converges to 0 and (P_j) is a sequence of probability measures on $(\mathcal{Y}, \mathcal{B})$ defined by (3). Then, as in the previous case, the Π_j Bayes predictor of $\boldsymbol{Y^m}$ is given by

$$\boldsymbol{d^j}(N, \boldsymbol{U^N}) = \begin{cases} m\dfrac{\boldsymbol{X^N} + A_j\boldsymbol{p_j}}{N + A_j}, & if \quad N > 0, \\ m\boldsymbol{p_j}, & if \quad N = 0, \end{cases}$$

where $\boldsymbol{p_j}$ is defined by (4). Furthermore, the risk function $R(\boldsymbol{d^j}, P)$ equals, by (15),

$$R(\boldsymbol{d^j}, P) \quad = \quad mR_1(P) + m^2 \sum_{n=0}^{\infty} \frac{nR_1(P) + A_j^2(\boldsymbol{p_j}-\boldsymbol{p})^{\boldsymbol{T}}\boldsymbol{C}(\boldsymbol{p_j}-\boldsymbol{p})}{(n+A_j)^2} f(n).$$

To calculate the Bayes risk $\rho(j)$ for this decision rule we note that, by Theorem 4 of Ferguson [5],

$$\begin{aligned} E_{\Pi_j}\boldsymbol{p^T C p} &= E_{\Pi_j}(E_P\boldsymbol{Z})^{\boldsymbol{T}}\boldsymbol{C}(E_P\boldsymbol{Z}) = \frac{E_{P_j}\boldsymbol{Z^T C Z} + A_j(E_{P_j}\boldsymbol{Z})^{\boldsymbol{T}}\boldsymbol{C}(E_{P_j}\boldsymbol{Z})}{A_j+1} \\ &= \frac{E_{P_j}\boldsymbol{Z^T C Z} + A_j\boldsymbol{p_j^T C p_j}}{A_j+1} = \frac{R_1(P_j)}{A_j+1} + \boldsymbol{p_j^T C p_j}. \end{aligned}$$

From this and (18) we conclude that

$$E_{\Pi_j}R_1(P) = E_{\Pi_j}\left(E_P\boldsymbol{Z^T C Z} - \boldsymbol{p^T C p}\right) = \frac{A_j}{A_j+1}R_1(P_j),$$

and

$$E_{\Pi_j}(\boldsymbol{p_j}-\boldsymbol{p})^{\boldsymbol{T}}\boldsymbol{C}(\boldsymbol{p_j}-\boldsymbol{p}) = E_{\Pi_j}\boldsymbol{p^T C p} - \boldsymbol{p_j^T C p_j} = \frac{R_1(P_j)}{A_j+1}.$$

Therefore,

$$\rho(j) = m\frac{A_j}{A_j+1}R_1(P_j) + m^2\frac{A_j}{A_j+1}R_1(P_j)\sum_{n=0}^{\infty}\frac{f(n)}{n+A_j}$$

and

$$\lim_{j\longrightarrow\infty}\rho(j) = m^2 f(0)\sup_{P\in\mathcal{P}} R_1(P),$$

because $A_j \longrightarrow 0$ and $R_1(P_j) \longrightarrow \sup_{P\in\mathcal{P}} R_1(P)$. Since $m^2 f(0)\sup_{P\in\mathcal{P}} R_1(P)$ is, by (19), the upper bound for the risk of $\boldsymbol{d^0}(N, \boldsymbol{U^N})$, this implies that $\boldsymbol{d^0}(N, \boldsymbol{U^N})$ is minimax when $m > 1$ and the condition (5) is not satisfied.

Now we consider the last case in which $m = 1$ and thus the predictor $\boldsymbol{d^0}(N, \boldsymbol{U^N})$ is defined by (9). Then, for any $\boldsymbol{d} \in \mathcal{D}$, we obtain, by (12) and (20),

$$\begin{aligned}\sup_{P\in\mathcal{P}} R(\boldsymbol{d}, P) &\geq m \sup_{P\in\mathcal{P}} R_1(P) = \sup_{P\in\mathcal{P}} R_1(P)\\ &= \sup_{P\in\mathcal{P}} \left[R_1(P) + \boldsymbol{(p_0 - p)^T C(p_0 - p)}\right] = \sup_{P\in\mathcal{P}} R(\boldsymbol{d^0}, P),\end{aligned}$$

which implies minimaxity of $\boldsymbol{d^0}(N, \boldsymbol{U^N})$ in that case. The proof of Theorem 1 is complete.

4. The failure of the minimax predictor for a fixed sample size

Suppose now that we want to predict $\boldsymbol{Y^m}$, $m > 1$, from the sample $U_1, U_2, \dots, U_n$ of a fixed size n. Then, from Theorem 1, the minimax predictor has the form

$$\boldsymbol{d^1}(n, \boldsymbol{U^n}) = m\frac{\boldsymbol{X^n} + A(n)\boldsymbol{p_0}}{n + A(n)} \tag{21}$$

where the positive real number $A(n)$ solves the equation (6) with $f(n) = 1$ and $f(j) = 0$, $j \neq n$, i.e.

$$A(n) = \frac{n + \sqrt{nm(n+m-1)}}{m-1}. \tag{22}$$

It seems that the predictor $\boldsymbol{d^1}(N, \boldsymbol{U^N})$ should be minimax when the sample size is random. Obviously, if N is not random and takes only one value, say n, then $A_0 = A(n)$ and $\boldsymbol{d^1}(N, \boldsymbol{U^N})$ has this optimal property. Otherwise, the following theorem holds:

Theorem 2. *If an ancillary statistics N is not concentrated on one point, then $\boldsymbol{d^1}(N, \boldsymbol{U^N})$ is not minimax.*

Proof. Assume that $\boldsymbol{d^1}$ is minimax, i.e. $\sup_{P\in\mathcal{P}} R(\boldsymbol{d^1}, P) = \sup_{P\in\mathcal{P}} R(\boldsymbol{d^2}, P)$, and, for fixed $0 < \alpha < 1$, consider the predictor $\boldsymbol{d^2} = \alpha\boldsymbol{d^0} + (1-\alpha)\boldsymbol{d^1}$. Then, as it is easy to calculate,

$$\begin{aligned}R(\boldsymbol{d^2}, P) &= \alpha R(\boldsymbol{d^0}, P) + (1-\alpha)R(\boldsymbol{d^1}, P) - \alpha(1-\alpha)E_P(\boldsymbol{d^1} - \boldsymbol{d^0})^T \boldsymbol{C}(\boldsymbol{d^1} - \boldsymbol{d^0})\\ &= \alpha R(\boldsymbol{d^0}, P) + (1-\alpha)R(\boldsymbol{d^1}, P)\\ &- \alpha(1-\alpha)\sum_{n=1}^{\infty} \frac{(A_0 - A(n))^2[nR_1(P) + n^2\boldsymbol{(p_0 - p)^T C(p_0 - p)}]}{(n + A(n))^2(n + A_0)^2} f(n).\end{aligned}$$

Since the ancillary statistics N is not concentrated on one point, the real number $\max_{n\geq1}(A_0 - A(n))^2 f(n)$ is greater than zero. Therefore, there must exist a sequence (P_j) of probability measures for which the following two conditions are satisfied

$$\lim_{j\to\infty} R(\boldsymbol{d^0}, P_j) = \lim_{j\to\infty} R(\boldsymbol{d^1}, P_j) = \sup_{P\in\mathcal{P}} R(\boldsymbol{d^0}, P)$$

and

$$\lim_{j\to\infty} R_1(P_j) = \lim_{j\to\infty} (\boldsymbol{p_0} - \boldsymbol{p_j})^T \boldsymbol{C}(\boldsymbol{p_0} - \boldsymbol{p_j}) = 0, \quad \text{with} \quad \boldsymbol{p_j} = E_{P_j}\boldsymbol{Z}.$$

Otherwise, minimaxity of $\boldsymbol{d^0}$ and $\boldsymbol{d^1}$ implies that

$$\sup_{P\in\mathcal{P}} R(\boldsymbol{d^2}, P) < \sup_{P\in\mathcal{P}} R(\boldsymbol{d^0}, P),$$

which is impossible. However, such a sequence (P_j) does not exist, because the risk function $R(\boldsymbol{d^0}, P)$ equals, by (15),

$$R(\boldsymbol{d^0}, P) = m^2\left[R_1(P) + \boldsymbol{(p_0 - p)^T C(p_0 - p)}\right]\left[f(0) + \sum_{n=1}^{\infty}\frac{A_0^2 f(n)}{(n + A_0)^2}\right],$$

and the above two conditions immediately imply that

$$\sup_{P\in\mathcal{P}} R(\boldsymbol{d^0},P) = \lim_{j\to\infty} R(\boldsymbol{d^0},P_j) =$$

$$= \lim_{j\to\infty} m^2 \left[R_1(P_j) + \boldsymbol{(p_0-p_j)^T C(p_0-p_j)}\right] \left[f(0) + \sum_{n=1}^{\infty} \frac{A_0^2 f(n)}{(n+A_0)^2}\right] = 0,$$

which is impossible. Therefore, $\boldsymbol{d^1}$ can't be minimax, which completes the proof of the theorem. □

5. The minimax predictor when the distribution of N is unknown.

We have derived minimax predictor of $\boldsymbol{Y^m}$ assuming that the distribution of the ancillary statistics N is known. Now we drop this assumption and prove the following theorem.

Theorem 3. *Suppose that the distribution of the ancillary statistics N is unknown. Then, for $m > 1$,*

$$\boldsymbol{d}^*(N,\boldsymbol{U^N}) = \begin{cases} \dfrac{\boldsymbol{X^N}}{N} & if \quad N>0 \\ \boldsymbol{p_0} & if \quad N=0 \end{cases} \tag{23}$$

is the minimax predictor of the unobservable vector $\boldsymbol{Y^m}$ and its minimax risk equals

$$\sup_{P\in\mathcal{P}} R(\boldsymbol{d}^*,P) = m^2 \sup_{P\in\mathcal{P}} R_1(P).$$

Proof. Obviously, the decision rule $\boldsymbol{d}^*$ coincide with the minimax predictor of $\boldsymbol{Y^m}$ derived under the assumptions that the distribution of N is known and the inequality (5) is not satisfied. Therefore, from (19), the risk of $\boldsymbol{d}^*$ is bounded from above by $m^2 \sup_{P\in\mathcal{P}} R_1(P)$, because $f(0) \le 1$. Moreover, this upper bound is the minimax risk of $\boldsymbol{d^0}$ when the ancillary statistic N is concentrated on the point zero. Therefore, this bound must be attained by any decision rule used to predict $\boldsymbol{Y^m}$ under the unknown distribution of N. This completes the proof. □

6. Examples

As an application of the results obtained in the paper we consider the following three examples in which we assume that $m > 1$. In each of them the minimax predictor $\boldsymbol{d^0}(N,\boldsymbol{U^N})$ is given by (8). The number A_0 is defined by (7), if the distribution of the ancillary statistics N is known, and is equal to zero otherwise. The form of the vector $\boldsymbol{p_0}$ will be found below.

Example 1. Suppose that the set $\mathcal{Y}$ is centrosymmetric about $\mathbf{0}$ and that, for each $y \in \mathcal{Y}$, $\boldsymbol{z}(y) = -\boldsymbol{z}(-y)$. Let (P_j) be a sequence for which (3) holds and let P_j^- denotes the distribution of the random vector $-U_0$, whenever U_0 is distributed according to P_j. Then the sequence $(P_j^{'})$, with $P_j^{'} = (1/2)(P_j + P_j^-)$, satisfies (3), because

$$\begin{aligned} R_1(P_j^{'}) &= E_{P_j^{'}}\boldsymbol{Z^T C Z} - (E_{P_j^{'}}\boldsymbol{Z})^T \boldsymbol{C}(E_{P_j^{'}}\boldsymbol{Z}) = E_{P_j}\boldsymbol{Z^T C Z} - 0 \\ &\ge E_{P_j}\boldsymbol{Z^T C Z} - (E_{P_j}\boldsymbol{Z})^T \boldsymbol{C}(E_{P_j}\boldsymbol{Z}) = R_1(P_j). \end{aligned}$$

Therefore, we may assume that $\boldsymbol{p_j} = E_{P_j^{'}}\boldsymbol{Z} = \mathbf{0}$, which implies that $\boldsymbol{p_0} = \mathbf{0}$.

Example 2. Suppose that $\boldsymbol{C} = [c_{ij}]$ is a diagonal matrix and that there exist two sequences $\{\overline{y}_j\}$ and $\{\overline{\overline{y}}_j\}$ of points from $\mathcal{Y}$ such that, for each $1 \le i \le k$,

$$\lim_{j\longrightarrow\infty} z_i(\overline{y}_j) = \inf_{y\in Y} z_i(y), \quad \lim_{j\longrightarrow\infty} z_i(\overline{\overline{y}}_j) = \sup_{y\in Y} z_i(y).$$

Let the distribution P_j of U_0, $j \geq 1$, be defined by:

$$P_j(U_0 = \overline{y}_j) = P_j(U_0 = \overline{\overline{y}}_j) = 0.5.$$

Then, as it is easy to verify, for each $1 \leq i \leq k$,

$$\begin{aligned} \sup_{P \in \mathcal{P}} \left[E_P \left(z_i(U_0)\right)^2 - \left(E_P z_i(U_0)\right)^2 \right] &= \lim_{j \longrightarrow \infty} \left[E_{P_j} \left(z_i(U_0)\right)^2 - \left(E_{P_j} z_i(U_0)\right)^2 \right] \\ &= \lim_{j \longrightarrow \infty} \frac{|z_i(\overline{y}_j) - z_i(\overline{\overline{y}}_j)|^2}{4}. \end{aligned}$$

This implies that (P_j) is a sequence of distributions defined in (3), because $\boldsymbol{C}$ is assumed to be a diagonal matrix and thus

$$R_1(P) = \sum_{i=1}^{k} c_{ii} \left[E_P \left(z_i(U_0)\right)^2 - \left(E_P z_i(U_0)\right)^2 \right].$$

Therefore, the coordinates of the point $\boldsymbol{p_0} = (p_{01}, p_{02}, \ldots, p_{0k})^T$ are given by

$$p_{0i} = \lim_{j \longrightarrow \infty} \frac{z_i(\overline{y}_j) + z_i(\overline{\overline{y}}_j)}{2} = \frac{\inf_{y \in \mathcal{Y}} z_i(y) + \sup_{y \in \mathcal{Y}} z_i(y)}{2}, \quad 1 \leq i \leq k.$$

Example 3. Let $\{A_i\}_{i=1}^k$ be a measurable partition of $\mathcal{Y}$, i.e. let $A_1, \ldots, A_k$ be measurable, pairwise disjoint, subsets of $\mathcal{Y}$ whose union equals $\mathcal{Y}$. Furthermore, let $z_i(y) = \mathbf{1}_{A_i}(y)$, $1 \leq i \leq k$, be the indicator function of the set A_i. Then the random vectors $\boldsymbol{Z} = \boldsymbol{z}(U_0)$, $\boldsymbol{X^n}$ and $\boldsymbol{Y^m}$ have $(1, \boldsymbol{p})$, $(n, \boldsymbol{p})$ and $(m, \boldsymbol{p})$ multinomial distributions in which the parameter $\boldsymbol{p} = E_P \boldsymbol{Z}$ takes its values in a simplex $S = \left\{ (s_1, s_2, \ldots, s_k) \ : \ \bigwedge_{1 \leq i \leq k} \ s_i \geq 0 \quad \text{and} \quad s_1 + s_2 + \ldots + s_k = 1 \right\}$. Furthermore, as it is easy to calculate, $R_1(P) = \boldsymbol{c^T p} - \boldsymbol{p^T C p}$, where $\boldsymbol{c} = (c_{11}, c_{22}, \ldots, c_{kk})^T$ stands for the diagonal of the matrix $\boldsymbol{C} = [c_{ij}]$. Therefore, the vector $\boldsymbol{p_0}$ satisfies the equation

$$\boldsymbol{c^T p_0} - \boldsymbol{p_0^T C p_0} = \max_{p \in S} \left[\boldsymbol{c^T p} - \boldsymbol{p^T C p} \right].$$

References

[1] Amrhein, P. (1995). "Minimax Estimation of Proportions Under Random Sample Size," *Journal of the American Statistical Association* 90, 1107-1111.

[2] Aubin, J.P. (1980). *Mathematical Methods of Game and Economic Theory.* North-Holland Publishing Company.

[3] Brown, L.D. (1990). "An Ancillarity Paradox Which Appears in Multiple Linear Regression" *Ann. Statist.*, 18

[4] Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach.* Academic, New York.

[5] Ferguson, T.S. (1973)."A Bayesian Analysis of Some Nonparametric Problems," *Ann. Statist.* 1, 209-230.

[6] He, K. (1990). "An Ancillarity Paradox in the Estimation of Multinomial Probabilities," *Journal of the American Statistical Association* 85, 824-829.

[7] Wilczyński, M. (1992). "Minimax Nonparametric Estimation," *Sankhyā* 54, Series A, Pt. 3, 425-431.

*Institute of Mathematics, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, PL-50-370 Wrocław, Poland

E-mail address: wilczyn@@im.pwr.wroc.pl