

On the Relation between Generalized Entropy and the Bayes Decision Error

M. Horáček¹

¹Centre of Biomedical Informatics, Institute of Computer Science, Academy of Sciences of the Czech Republic

Supervisor: Prof. RNDr. Jana Zvárová, DrSc.

Summary

We deal with the relation between the generalized entropy (f -entropy, a family of functions that include several biodiversity measures) of a discrete random variable and the minimal probability of error (Bayes error) when the value of this random variable is estimated. Namely the tightness of their relation is studied.

Morales and Vajda [1] recently introduced a measure called average inaccuracy that aims to quantify the tightness of the relation between the posterior Bayes error and the power entropies. It is defined as a standardized average difference between the upper and the lower bound for the posterior Bayes error under given entropy. Their concept can be generalized to any strictly concave f -entropy and used to evaluate its relation to the Bayes probability of error. However, due to a complex form of the formula of the average inaccuracy, it is difficult to compare the average inaccuracies of most f -entropies analytically.

We propose a smooth approximation of the lower bound for the posterior Bayes error under given f -entropy that simplifies the formula of the average inaccuracy. We show that under this approximation, the quadratic entropy has the tightest relation to the posterior Bayes error among f -entropies.

The quadratic entropy has the tightest relation to the posterior Bayes error (in the sense described in this paper) than the Shannon's entropy and other functions that belong to the family of f -entropies, like Emlen's index, Ferreri's index and Good's index.

Keywords: generalized entropy, f -entropy, Bayes error, average inaccuracy, power entropy, quadratic entropy, Shannon's entropy, Emlen's index, Ferreri's index, Good's index

1. Introduction

Let X and Y be discrete random variables with values in finite sets $X = (x_1, \dots, x_r)$ and $Y = (y_1, \dots, y_s)$. Let Z be a random variable that is absolutely continuous with respect to the Lebesgue measure with probability distribution P_z on some set $Z \subseteq \mathbf{R}$. We denote $P(x) = P(X = x)$, the joint distribution of X and Y as $p(x, y) = P(X = x, Y = y)$, the conditional distribution of X under given $Z = z$ as $p(x|z) = P(X = x|Z = z)$ etc., i.e. the arguments in the probability distribution p denote the space and random variable which they refer to.

The uncertainty associated with a random variable X is traditionally measured by the Shannon's entropy

$$H(X) = - \sum_{x \in X} p(x) \ln p(x)$$

with convention $0 \ln 0 = 0$. It is intuitively related to the minimal probability of error (the prior Bayes error)

$$e(X) = 1 - \max_{x \in X} p(x)$$

when the value of X is estimated. The conditional Shannon's entropy

$$\begin{aligned} H(X | Z) &= \int_Z H(X | Z = z) dP_z(z) = \\ &= - \int_Z \sum_{x \in X} p(x | z) \ln p(x | z) dP_z(z) \end{aligned}$$

and the average minimal posterior probability of error (the posterior Bayes error)

$$e_B(X | Z) = 1 - \int_Z \max_{x \in X} p(x | z) dP_z(z)$$

are related in a similar manner. Fano [2] and Kovalevsky [3] were among the first who studied the relationship between the entropy and the Bayes error. They derived attainable bounds for Shannon's entropy $H(X)$ under given prior Bayes error $e(X)$. Since these bounds are continuous and strictly increasing functions of the Bayes error e on the interval $[0, (r-1)/r]$, the bounds for the prior Bayes error under given entropy are well defined. Fano's and Kovalevsky's findings were later broaden to more general classes of entropies by various authors, namely to all Schur-concave entropies by Vajda and Vašek [4]. Vajda and Vašek also retrieved corresponding bounds for conditional entropies and the posterior Bayes error.

Recently, Morales and Vajda [1] studied tightness of the relationship between the prior and posterior Bayes errors and a class of power entropies. They numerically compared tightness of the relationship between the power entropies and the Bayes error and showed that the quadratic entropy seems to give at average the most accurate estimate of the posterior Bayes error.

We deal with the tightness of relationship between the Bayes errors and the family of f -entropies, introduced by Zvárová [7], which include many of the functions used for measuring biological diversity with applications in biology and medicine,

i.e. in the cancer research (Maley et al. [5]), or in the research of relations among human populations (Chak-rabarty [6]). Some of the power entropies are also included in the f -entropies family. We propose an approximation of the lower bound of posterior Bayes error under given f -entropy that simplifies the formula for average inaccuracy. Then we show analytically that the quadratic entropy has the tightest relation to the prior Bayes error among f -entropies under our approximation.

2. The f -entropy

A generalization of entropy, called f -entropy, was suggested by Zvárová [7]. Here we will briefly recall its derivation.

Assume that $p(x)$ is absolutely continuous with respect to $q(x)$, i.e. $q(x) = 0 \Rightarrow p(x) = 0$, $x \in \chi$ and the convention $0f(0/0) = 0$ is used. The information divergence

$$D(p(x) \| q(x)) = \sum_{x \in X} p(x) \ln(p(x)/q(x))$$

of two discrete distributions $p(x)$ and $q(x)$ defined on the space χ was proposed by Kullback and Liebler [8].

The entropy

$$H(X) = - \sum_{x \in X} p(x) \ln p(x)$$

and the mutual information

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \ln \frac{p(x, y)}{p(x)p(y)}$$

were introduced by Shannon [9]. They satisfy equalities

$$I(X; Y) = D(p(x, y) \| p(x)p(y))$$

$$H(X) = I(X; X) \quad (1)$$

and inequality

$$0 \leq I(X; Y) \leq H(X)$$

A more general class of divergences, called f -divergences, was proposed by Csiszár [10]. The f -divergence of $q(x)$ from $p(x)$ is

$$D_f(p(x) \| q(x)) = \sum_{x \in X} q(x) f\left(\frac{p(x)}{q(x)}\right)$$

for any function f that is convex on the interval $[0, \infty)$, strictly convex in $f(1)$ and $f(1) = 0$. Following Zvárová [7], we can replace the information divergence in (1) by the f -divergence and define the generalized mutual information by

$$I_f(X; Y) = D_f(p(x, y) \| p(x)p(y)).$$

Let f be a function that satisfies the conditions required when defining f -divergences and also let $(f(x) - f(0))/x$ be a concave function on $[1, \infty)$. The f -entropy is then defined by the formula

$$H_f(X) = I_f(X, X) = \sum_{x \in X} p^2(x) f(1/p(x)) + f(0) \left(1 - \sum_{x \in X} p^2(x)\right).$$

Zvárová [7] showed that the f -entropies share many properties with the Shannon's entropy, e.g. that under given $r = \text{card}(\chi)$ they attain their minimal possible value (zero) when X is a constant with probability equal to one, attain their maximal possible value when X is uniformly distributed on χ and they satisfy the inequality

$$0 \leq I_f(X; Y) \leq H_f(X).$$

Further information about f -entropies, their characteristics and usage can be found in the paper of Zvárová, Vajda [11] and in the work Horáček [12].

3. The average inaccuracy

Let $r = \text{card}(\chi) \geq 2$ and denote $\Delta_r = \{(p(x_1), \dots, p(x_r)) : p(x_i) \geq 0, \sum_{i=1}^r p(x_i) = 1\}$. The bounds of f -entropy under given prior Bayes error are given by

$$H_f^+(e) = \max_{\{X: e(X)=e\}} H_f(X)$$

respectively by

$$H_f^-(e) = \min_{\{X: e(X)=e\}} H_f(X),$$

i.e. the maximum (minimum) is taken with respect to any discrete distribution $p(x)$ of random variable X on χ that satisfies $e(X) = 1 - \max_{x \in \chi} p(x) = e$. Since any f -entropy is a Schur-concave function on Δ_r (see e.g. Horáček [12]), the bounds satisfy

$$H_f^+(e) = H(1-e, e/(r-1), \dots, e/(r-1)) =$$

$$= (1-e)^2 f\left(\frac{1}{1-e}\right) + \frac{e^2}{r-1} f\left(\frac{r-1}{e}\right) + ef(0) \left(2 - e \frac{r}{r-1}\right)$$

when $e \in [0, (r-1)/r]$ and

$$H_f^-(e) = H(1-e, \dots, 1-e, 1-(1-e)k, 0, \dots, 0) =$$

$$= k(1-e)^2 f\left(\frac{1}{1-e}\right) + [1-k(1-e)^2] f\left(\frac{1}{1-k(1-e)}\right) +$$

$$+ f(0) [k(1-e)^2 + 1 - (1-k(1-e))^2]$$

when $e \in [(k-1)/k, k/(k+1)]$, $k \in \{1, \dots, r-1\}$.

Since the function $(f(x) - f(0))/x$ is concave on the interval $[1, \infty)$ and therefore also the function $x^2[f(x^{-1}) - f(0)]$ is concave on the interval $[0, 1]$, it is not difficult to show that the upper bound $H_f^+(e)$ is concave on the interval $[0, (r-1)/r]$ and the lower bound $H_f^-(e)$ is piecewise concave on each of the intervals $[(k-1)/k, k/(k+1)]$ where $k \in \{1, \dots, r-1\}$. Moreover, the convexity of f implies that the piecewise linear function

$$\Phi(t) = H_f\left(\frac{k-1}{k}\right) + \frac{t - \frac{k-1}{k}}{\frac{k}{k+1} - \frac{k-1}{k}} \cdot$$

$$\left[h_f\left(\frac{k}{k+1}\right) - h_f\left(\frac{k-1}{k}\right) \right]$$

when $t \in [(k-1)/k, k/(k+1)]$, $k \in \{1, \dots, r-1\}$ is convex on the interval $[0, (r-1)/r]$.

The upper bound for f -entropy under given posterior Bayes error e_B is defined by

$$H_f^+(e_B) = \max_{\{(X,Z): e_B(X|Z)=e_B\}} H_f(X|Z),$$

i.e. the maximum is taken with respect to any discrete random variable X defined on χ and random variable Y that satisfy $e_B(X|Z) = e_B$. The lower bound is defined accordingly. Since the bounds for f -entropy under prior Bayes error satisfy the concavity and convexity rules mentioned above, the proof of the theorem 4.1 in Morales and Vajda [1] can be followed to show that

$$\begin{aligned} H_f^+(e_B) &= (1-e_B)^2 f\left(\frac{1}{1-e_B}\right) + \\ &+ \frac{e_B^2}{r-1} f\left(\frac{r-1}{e_B}\right) + \\ &+ e_B f(0) \left(2 - e_B \frac{r}{r-1}\right) \end{aligned}$$

when $e \in [0, (r-1)/r]$ and that the lower bound is a convex, piecewise linear function with values

$$H_f^-(e_B) = (1-e_B) f\left(\frac{1}{1-e_B}\right) + f(0) e_B$$

when $e_B = \frac{k}{k+1}$, $k \in \{0, \dots, r-1\}$,

i.e. $H_f^-(e_B) = \Phi(e_B)$.

Since these bounds are continuous strictly increasing functions on the interval $[0, (r-1)/r]$, they implicitly set the bounds for the posterior Bayes error under given entropy. Following Morales and Vajda, we denote these bounds

$$\begin{aligned} e_{B,f}^+(H) &= \max_{H_f^-(e_B) \leq H} e_B \\ e_{B,f}^-(H) &= \max_{H_f^+(e_B) \leq H} e_B. \end{aligned}$$

Let us also denote

$$a_{r,f} = H_f\left(\frac{1}{r}, \dots, \frac{1}{r}\right)$$

and recall that the power entropies are defined by the formula

$$H_\alpha(X) = \frac{1}{\alpha-1} \left(1 - \sum_{x \in X} p^\alpha(x)\right)$$

when $\alpha > 0$.

(The entropy H_1 is defined as a limit and is equal to the Shannon's entropy.) Note that the power entropies are f -entropies when $\alpha = (0, 2]$, namely the quadratic entropy $H_2(X)$ is f -entropy with $f(x) = x-1$.

Morales and Vajda introduced a measure of tightness of the relationship between the power entropies and the posterior Bayes error, called the average inaccuracy. It is defined by the formula

$$\begin{aligned} AI_r(e_B | H_\alpha) &= \frac{1}{a_{r,\alpha}} \cdot \\ &\cdot \int_0^{a_{r,\alpha}} [e_{B,\alpha}^+(H) - e_{B,\alpha}^-(H)] dH. \end{aligned}$$

This measure can be straightforwardly generalized to any strictly concave f -entropy by replacing $\alpha_{r,\alpha}$ by $\alpha_{r,f}$ and $e_{B,\alpha}$ by $e_{B,f}$ in the formula.

Morales and Vajda stated, on the basis of numerical computations with various values of the parameter α , that the quadratic entropy seems to have the lowest average inaccuracy when estimating posterior Bayes error among power entropies. However, due to the form of the lower bound for the posterior Bayes error, it is difficult to verify this postulate analytically. We suggest to approximate the piecewise linear function $H_f^-(e_B)$ by a function

$$\tilde{H}_f^-(e_B) = (1-e_B) f\left(\frac{1}{1-e_B}\right) + f(0) e_B$$

that acquire the same values as the original function when $e_B = k/(k+1)$, $k \in \{0, \dots, r-1\}$. This alternative lower bound is also a convex function on the interval $[0, (r-1)/r]$, it acquires the same values as the original piecewise linear bound in $\{(k-1)/k, k \in \{1, \dots, r-1\}\}$ and therefore it cannot be higher than the original bound. Since $H_f^-(e_B)$ is again a continuous strictly increasing function, the inverse function $\tilde{e}_{B,f}^-(H)$ is well defined and we can approximate the average inaccuracy by

$$\tilde{AI}_r(e_B | H_f) = \frac{1}{a_{r,f}} \int_0^{a_{r,f}} [\tilde{e}_{B,f}^+(H) - \tilde{e}_{B,f}^-(H)] dH.$$

The approximated bound is smooth when the function f is smooth. An example of the bounds $H^+(e_B)$, $H^-(e_B)$ and $\tilde{H}^-(e_B)$ is displayed in the figure 1.

Using this approximation, the following theorem can be formulated.

Theorem 1 Under the approximation described above, the quadratic entropy $H_2 = 1 - \sum_{x \in X} p^2(x)$ has the tightest relation to posterior Bayes error, i.e.

$$\tilde{AI}_r(e_B | H_2) \leq \tilde{AI}_r(e_B | H_f)$$

for any f -entropy H_f and any given integer $r \geq 2$.

Proof. Recall that when H_f is a f -entropy, the function f is convex on the interval $[0, \infty)$ and strictly convex in $f(1)$ with $f(1) = 0$. Let us first consider only those functions f that satisfy $f(0)=0$. Then it follows that $f(x)$ has to be above zero for all $x > 1$. Since $H_f^+(0) = H_f^-(0)$ and also $H_f^+((r-1)/r) = H_f^-((r-1)/r)$, we can calculate the approximate average inaccuracy $\tilde{AI}_r(e_B | H_f)$, i.e. the relative area between bounds $\tilde{e}_{B,f}^+(H)$ and $\tilde{e}_{B,f}^-(H)$, by calculating the area between bounds $H_f^+(e_B)$ and $\tilde{H}_f^-(e_B)$ then applying the corresponding normalizing constant.

$$\begin{aligned}
\tilde{A}I_r(e_B | H_f) &= \\
&= \frac{1}{a_{r,f}} \int_0^{a_{r,f}} [\tilde{e}_{B,f}^+(H) - \tilde{e}_{B,f}^-(H)] dH = \\
&= \frac{1}{a_{r,f}} \int_0^{r-1} [H_f^+(e_B) - \tilde{H}_f^-(e_B)] de_B = \\
&= \frac{r}{f(r)} \int_0^{r-1} (1-e_B)^2 f\left(\frac{1}{1-e_B}\right) + \\
&\quad + \frac{e_B^2}{r-1} f\left(\frac{r-1}{e_B}\right) - (1-e_B) \cdot \\
&\quad \cdot f\left(\frac{1}{1-e_B}\right) de_B = \frac{r}{f(r)} \int_0^{r-1} (1-e_B)^2 \cdot \\
&\quad \cdot f\left(\frac{1}{1-e_B}\right) + \frac{e_B^2}{r-1} f\left(\frac{r-1}{e_B}\right) - \\
&\quad - (1-e_B) f\left(\frac{1}{1-e_B}\right) de_B = \frac{r}{f(r)} \cdot \\
&\quad \cdot \int_0^{r-1} e_B \left[f\left(\frac{r-1}{e_B}\right) \frac{e_B}{r-1} - f\left(\frac{1}{1-e_B}\right) - \right. \\
&\quad \left. - (1-e_B) \right] de_B
\end{aligned}$$

After transformation $z = (r/r-1)e_B$ we have

$$\begin{aligned}
\tilde{A}I_r(e_B | H_f) &= \frac{r}{f(r)} \int_0^1 \left(\frac{r-1}{r}\right)^2 z \cdot \\
&\quad \cdot \left[f\left(\frac{r}{z}\right) \frac{z}{r} - f\left(\frac{r}{r-zr+z}\right) \frac{r-zr+z}{r} \right] dz
\end{aligned}$$

(2).

Since f is convex, $z \in (0,1]$ and $r \geq 1$, we get following inequality

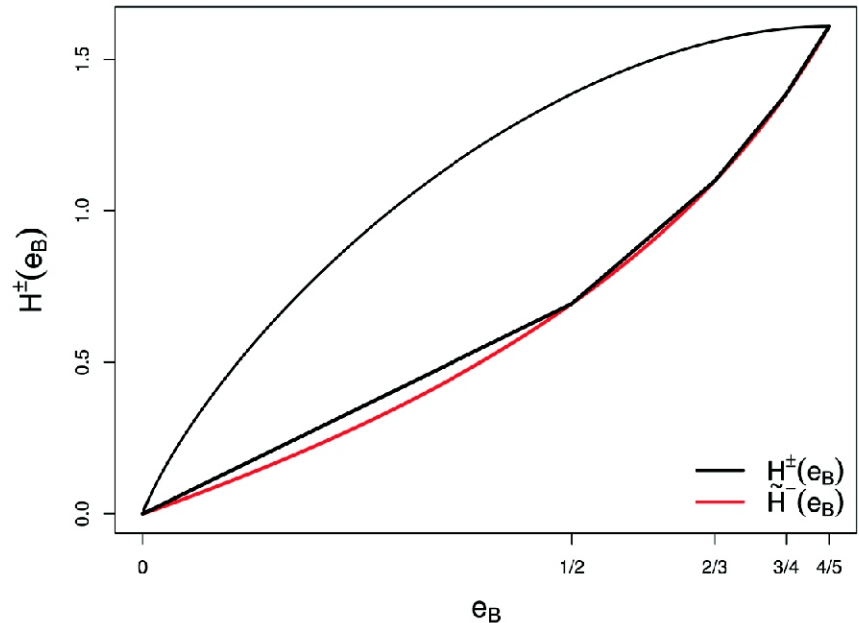


Fig. 1. The upper and lower bounds $H^\pm(e_B)$ and the approximate lower bound $\tilde{H}^-(e_B)$ for Shannon's entropy.

$$\frac{f(r/z) - f(1)}{r/z - 1} \geq \frac{f(r) - f(1)}{r - 1}$$

$$\frac{f(r/z)}{r/z - 1} \geq \frac{f(r)}{r - 1}$$

$$\frac{f(r/z)z}{f(r)} \geq \frac{r-z}{r-1}$$

(3).

Similarly the formula

$$\frac{f[r/(z-zr+r)] - f(1)}{r/(z-zr+r) - 1} \leq \frac{f(r) - f(1)}{r - 1}$$

can be converted into

$$\frac{f[r/(z-zr+r)](z-rz+r)}{f(r)} \leq z$$

(4).

After inserting inequalities (3) and (4) into the equality (2) we get

$$\begin{aligned}
\tilde{A}I_r(e_B | H_f) &\geq \int_0^1 \left(\frac{r-1}{r}\right)^2 z \left(\frac{r-z}{r-1} - z\right) dz \\
&= \int_0^1 \left(\frac{r-1}{r}\right) z(1-z) dz = \frac{1}{6} \frac{r-1}{r} = \tilde{A}I_r(e_B | H_2).
\end{aligned}$$

Now let us consider the situation when $f(0)=0$. Then $\tilde{A}I_r(e_B | H_f)$ satisfies

$$\begin{aligned}
\tilde{A}I_r(e_B | H_f) &= \\
&= \frac{1}{a_{r,f}} \int_0^1 \left(\frac{r-1}{r}\right)^2 z \left[f\left(\frac{r}{z}\right) \frac{z}{r} - f\left(\frac{r}{r-zr+z}\right) \frac{r-zr+z}{r} \right] dz + \\
&\quad + \frac{1}{a_{r,f}} f(0) \frac{r-1}{r} \tilde{A}I_r(e_B | H_2).
\end{aligned}$$

Here we use the inequalities (3) and (4) again (multiplied by $f(r)/r$ to avoid a potential division by zero or by a negative number) and we receive

$$\tilde{A}I_r(e_B | H_f) \geq \frac{1}{a_{r,f}}.$$

$$\cdot \left[\left(\frac{f(r)}{r} + f(0) \frac{r-1}{r} \right) \tilde{A}I_r(e_B | H_2) \right]$$

$$\tilde{A}I_r(e_B | H_f) = \frac{r}{f(r) + (r-1)f(0)}.$$

$$\cdot \left[\frac{f(r) + f(0)(r-1)}{r} \tilde{A}I_r(e_B | H_2) \right] =$$

$$= \tilde{A}I_r(e_B | H_2).$$

4. Conclusions

This result covers the power entropies H_α when $\alpha \in (0, 2]$ and also other f -entropies, for example Emlen's index [13]

$$H_E(p(x)) = \sum_x p(x) e^{-p(x)} - e^{-1},$$

where

$$f(x) = x \left(e^{-1/x} - e^{-1} \right) \chi_{(0, \infty)}(x),$$

Good's index [14]

$$H_{G(\alpha, \beta)}(p(x)) = \sum_x p(x)^\alpha (-\ln p(x))^\beta$$

when $\alpha \in [1/2, 1]$, $\beta = 1$, where

$$f(x) = x^{2-\alpha} (\ln x)^\beta I_{(1, \infty)}(x)$$

its generalization, the index of Sharma and Taneja [15]

$$H_{ST(s,t)}(p(x)) = (s-t)^{-1} \sum_{i \in M} (p(x)^t - p(x)^s)$$

when $t \in [0, 1]$, $s \in (\max\{t, 1-t\}, 3-t]$ (defined as a limit when $s=t$) for

$$f(x) = (s-t)^{-1} (x^{2-t} - x^{2-s}) \chi_{[1, \infty)}(x),$$

and the index of Ferreri [16]

$$H_{Fe(\alpha)}(p(x)) = (1+\alpha)\alpha^{-1} \ln(1+\alpha) -$$

$$-\alpha^{-1} \sum_x (1+\alpha p(x)) \ln(1+\alpha p(x))$$

$$\alpha > 0$$

where

$$f(x) = \left[\frac{x(1+\alpha)}{\alpha} \ln(1+\alpha) - \right.$$

$$\left. - \frac{x(x+\alpha)}{\alpha} \ln \frac{x+\alpha}{x} \right] I_{(0, \infty)}(x).$$

Acknowledgment

This work was supported by the project 1M06014 MŠMT ČR and SVV-2010-265513. The author thanks I. Vajda for introduction to the problem, and J. Zvárová and M. Studený for various useful suggestions during the preparation of the work.

References

- [1] Morales D., Vajda I.: Generalized Information Criteria for Optimal Bayes Decisions. Research report 2274. ÚTIA AV ČR, v.v.i.; 2010.
- [2] Fano R.M.: Transmission of Information: A Statistical Theory of Communications. New York: MIT Press; 1961.
- [3] Kovalovsky V.A.: The problem of character recognition from the point of view of mathematical statistics. Character Readers and Pattern Recognition. New York: Spartan Books; 1968;3-30.
- [4] Vajda I., Vašek K.: Majorization, concave entropies and comparison of experiments. Problems of Control and Information Theory. 1985;14:105-15.
- [5] Maley C.C., Galipeau P.C., Finley J.C., Wongsurawat V.J., Li X., Sanchez C.A., Paulson T.G., Blount P.L., Risques R.A., Rabinovitch P.S., Reid B.J.: Ecological Diversity Measures Predict Cancer Progression In Barrett's Esophagus. Gastroenterology. 2006;131:672-673.
- [6] Chakraborty R.: Mitochondrial DNA Polymorphism Reveals Hidden Heterogeneity within Some Asian Populations. American Journal of Human Genetic. 1990;47:87-94.
- [7] Zvárová J.: On Measures of Statistical Dependence. Časopis pro pěstování

matematiky. 1974;99:15-29.

- [8] Kullback S., Leibler R.A.: On Information and Sufficiency. Annals of Mathematical Statistics. 1951;22:7986.
- [9] Shannon C.E.: A mathematical theory of communication. Bell System Technical Journal. 1948;27:379-423,623-56.
- [10] Csizsár I.: Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. Publ. Math. Inst. Hungar. Acad. Sci. 1963;8:85-107.
- [11] Zvárová J., Vajda I.: On Genetic Information, Diversity and Distance. Methods of Inform. in Medicine. 2006;2:173-179.
- [12] Horáček M.: Míry biodiverzity a jejich aplikace. (Measures of biodiversity and their applications.) Master thesis. Prague, Charles university. Supervisor J. Zvárová. 2009.
- [13] Emlen J.M.: Ecology: An Evolutionary Approach. Reading: Addison-Wesley; 1973.
- [14] Good I.J.: The population frequencies of species and the estimation of population parameters. Biometrika. 1953;40:237-64.
- [15] Sharma B.D., Taneja I.J.: Entropy of type (;) and other generalized measures in information theory. Metrika. 1975;22:205-15.
- [16] Ferreri C.: Hypoentropy and related heterogeneity divergence measures. Statistica. 1980;40:55-118.

Contact

Mgr. Martin Horáček

Centre of Biomedical Informatics,
Institute of Computer Science,
Academy of Sciences of the Czech
Republic
Pod Vodárenskou věží 2
182 07 Prague 8
Czech Republic
e-mail: horacek@euromise.cz