Far East Journal of Mathematical Sciences (FJMS)
Volume 79, Number 2, 2013, Pages 219-239
Published Online: September 2013
Available online at http://pphmj.com/journals/fjms.htm
Published by Pushpa Publishing House, Allahabad, INDIA

PARAMETRIC ESTIMATES OF J-DIVERGENCE FOR CREDIT SCORING MODELS

Martin Řezáč

Department of Mathematics and Statistics
Masaryk University
Kotlářská 2, 611 37 Brno, Czech Republic
e-mail: mrezac@math.muni.cz

Abstract

J-divergence is widely used to describe the difference between two probability distributions. It is also called the Information value for the purpose of scoring models, e.g., credit scoring models used to predict a probability of clients default. The aim of the paper is to show formulas for frequently occurred situations when scores are normally, beta and gamma distributed. A simulation study showing performance of parametric estimates and an empirical estimate of J-divergence is included.

1. Introduction

J-divergence is widely used to describe the difference between two probability distributions. It is also called the Information value for the purpose of scoring models, e.g., credit scoring models used to predict a probability of clients default.

Credit scoring models are used for the majority of credit decisions in the financial sector. Methodology of credit scoring models and some measures of

2010 Mathematics Subject Classification: Primary 62P20; Secondary 91B82, 91G70.
Keywords and phrases: J-divergence, credit scoring, quality indexes.
Received June 20, 2013

their quality were discussed in works like Hand and Henley [5] or Thomas [14] and books like Anderson [1], Siddiqi [13] and Thomas [15]. Further remarks connected to credit scoring issues can be found there as well.

The paper deals with the J-divergence, which is one of the widely used indexes (beside Gini index and K-S statistic, see [18] for more details) for assessing credit scoring models. Commonly it is computed by discretization of data into bins using deciles with requirement on the nonzero number of cases for all bins. As an alternative method to the empirical estimates one can use the kernel smoothing theory, which allows to estimate unknown densities and consequently, using some numerical method for integration, to estimate value of the J-divergence. See Koláček and Řezáč [8] for more details. Another alternative is the empirical estimates with supervised interval selection (ESIS) proposed and discussed in Řezáč [11].

The main aim of the paper is to provide formulas for normally, beta and gamma distributed scores, which allow the parametric estimating of the J-divergence. Furthermore, they allow to assess the quality of nonparametric estimators of the J-divergence via MSE, bias and so forth. So far the formulas could not be found in the literature, especially not in the context of scoring models.

The choice of mentioned distribution types is not random. When we used real credit scoring data (the data set is described in Řezáč [11]) and made common transformation, we can get these three types of distributions.

Consider logistic regression model: $\ln\left(\frac{p}{1-p}\right) = b_0 + \sum_{i=1}^{k} b_i \cdot x_i$, where p is the modeled outcome (probability of default), b_i are coefficients and x_i are regressors. Then we use the following markings:

- score $= b_0 + \sum_{i=1}^{k} b_i \cdot x_i$,
- pgood $= p = \frac{1}{1 + e^{-b_0 - \sum_{i=1}^{k} b_i \cdot x_i}} = \frac{1}{1 + e^{-score}}$,

- $\text{odds} = \dfrac{p}{1-p} = e^{b_0 + \sum_{i=1}^{k} b_i \cdot x_i} = e^{score}$.

Consider *score* to be normally distributed random variable, which is quite common consideration. Then it can be shown that *pgood* is beta distributed and *odds* follows log-normal distribution. It is well known (see [6]) that both log-normal and gamma distributions can be used quite effectively to analyze positively skewed data sets. Furthermore, these two distributions are often interchangeable (see Wiens [17]). Nevertheless, Kundu and Manglick [9] proposed a test to discriminate between log-normal and gamma distribution. The test is based on ratio of likelihood functions (RML) of log-normal and gamma distributions. The natural logarithm of RML, denoted as *T*, plays the role of test statistic and the procedure of the test is quite simple. The log-normal distribution is recommended if $T > 0$, otherwise one should prefer the gamma distribution. Since $T = -4.193610^5$ in case of our real credit scoring data, we prefer the gamma distribution.

Tables 1 to 3 show results of goodness of fit tests. Figures 1 to 3 show histograms and fitted density functions for scores of good and bad clients. All of them were generated by SAS using Univariate procedure.

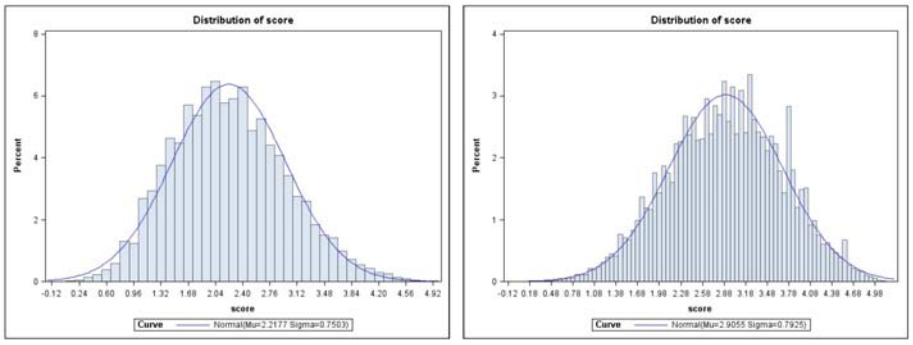
Table 1. Goodness of fit tests for score: (a) for bad, (b) for good clients

(a)

Goodness of fit tests for normal distribution				
Test	Statistic		*p*-value	
Kolmogorov-Smirnov	D	0.0548	Pr > D	> 0.150
Cramer-von Mises	W-Sq	0.0552	Pr > W-Sq	> 0.250
Anderson-Darling	A-Sq	0.3353	Pr > A-Sq	> 0.250

(b)

Goodness of fit tests for normal distribution				
Test	Statistic		*p*-value	
Kolmogorov-Smirnov	D	0.0220	Pr > D	> 0.150
Cramer-von Mises	W-Sq	0.1015	Pr > W-Sq	0.109
Anderson-Darling	A-Sq	0.6267	Pr > A-Sq	0.103



(a) (b)

Figure 1. Fitted distribution of score: (a) for bad, (b) for good clients.

It can be seen in Table 1 that *p*-values of all tests of normality were greater than 0.1. Also histograms and fitted densities show good match (see Figure 1). Hence both scores for bad and good clients could be considered as normally distributed.

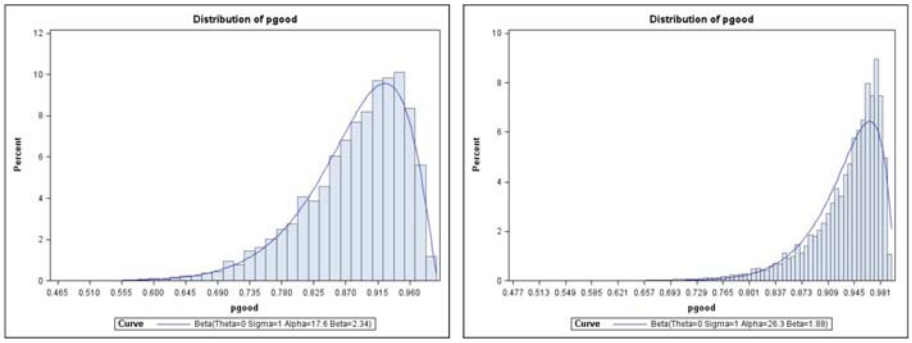
Table 2. Goodness of fit tests for pgood: (a) for bad, (b) for good clients

(a)

Goodness of fit tests for beta distribution				
Test	Statistic		*p*-value	
Kolmogorov-Smirnov	D	0.0737	Pr > D	> 0.250
Cramer-von Mises	W-Sq	0.1301	Pr > W-Sq	> 0.250
Anderson-Darling	A-Sq	1.0226	Pr > A-Sq	> 0.250

(b)

Goodness of fit tests for beta distribution				
Test	Statistic		*p*-value	
Kolmogorov-Smirnov	D	0.0475	Pr > D	0.036
Cramer-von Mises	W-Sq	0.5568	Pr > W-Sq	0.031
Anderson-Darling	A-Sq	4.6499	Pr > A-Sq	0.005



(a) (b)

Figure 2. Fitted distribution of pgood: (a) for bad, (b) for good clients.

A similar situation was also in testing the beta distribution for *pgood* for bad clients (*p*-values are greater than 0.25). Slightly worse results we had for *pgood* for good clients. Nevertheless, with significance level 0.03 or less and using Kolmogorov-Smirnov or Cramer-von Mises test we could accept hypothesis that *pgood* for good clients was beta distributed (see Table 2 and Figure 2).

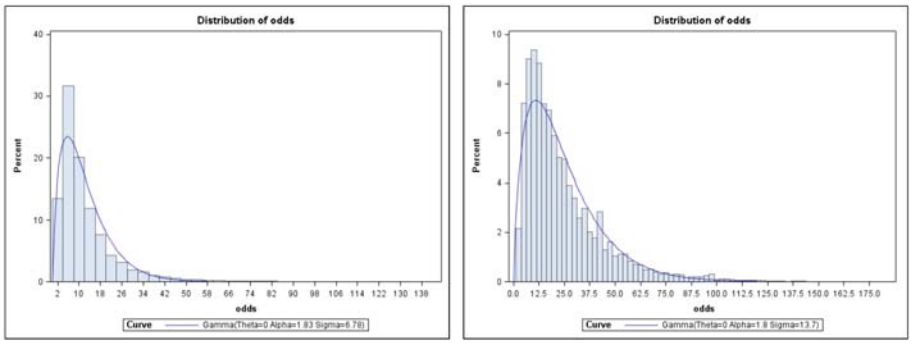
Table 3. Goodness of fit tests for odds: (a) for bad, (b) for good clients

(a)

Goodness of fit tests for gamma distribution				
Test	Statistic		*p*-value	
Kolmogorov-Smirnov	D	0.0761	Pr > D	> 0.250
Cramer-von Mises	W-Sq	0.1387	Pr > W-Sq	> 0.250
Anderson-Darling	A-Sq	1.3259	Pr > A-Sq	0.228

(b)

Goodness of fit Tests for gamma distribution				
Test	Statistic		*p*-value	
Kolmogorov-Smirnov	D	0.0346	Pr > D	0.234
Cramer-von Mises	W-Sq	0.3111	Pr > W-Sq	0.129
Anderson-Darling	A-Sq	3.2605	Pr > A-Sq	0.021



(a) (b)

Figure 3. Fitted distribution of odds: (a) for bad, (b) for good clients.

In case of *odds*, we can see in Table 3 and Figure 3 that *p*-values are high enough and histograms and estimated densities are matched sufficiently to state that *odds* follow the gamma distribution.

But let us go back to the merits of the paper. First of all we give the definition of J-divergence.

Definition 1.1. Jeffreys divergence (called J-divergence) of random variables X and Y with probability density functions $f(x)$ and $g(x)$ is a symmetrized Kullback-Leibler divergence given by

$$D_J(X, Y) = D_{KL}(X : Y) + D_{KL}(Y : X)$$

$$= \int_{-\infty}^{\infty} (f(x) - g(x)) \ln\left(\frac{f(x)}{g(x)}\right) dx, \tag{1.1}$$

where the Kullback-Leibler divergence $D_{KL}(X : Y)$ is given by

$$D_{KL}(X : Y) = \int_{-\infty}^{\infty} (f(x)) \ln\left(\frac{f(x)}{g(x)}\right) dx. \tag{1.2}$$

The connection to more general class of divergences, like general Alpha-divergences, Alpha-Rényi or Jensen-Shannon divergence, can be found in the literature, e.g., Cichocki and Amari [2]. Important properties like convexity or strict positivity of D_J are direct consequences of Alpha-divergence properties, which could be found there as well.

2. Normally Distributed Scores

Let X and Y be normally distributed random variables with density functions $f_0(x)$ and $f_1(x)$ given by

$$f_0(x) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}}, \quad x \in (-\infty, \infty), \tag{2.1}$$

$$f_1(x) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}, \quad x \in (-\infty, \infty), \tag{2.2}$$

where $\mu_0, \mu_1, \sigma_0 > 0$ and $\sigma_1 > 0$ are real constants.

Lemma 2.1. *Consider X, Y to be random variables. Let* $f_0(x)$ *and* $f_1(x)$ *be density functions given by* (2.1) *and* (2.2), *respectively. Then it holds*

$$D_{KL}(X : Y) = \ln\left(\frac{\sigma_1}{\sigma_0}\right) + \frac{(\mu_1 - \mu_0)^2 + (\sigma_0^2 - \sigma_1^2)}{2\sigma_1^2}. \tag{2.3}$$

Proof.

$$\begin{aligned} D_{KL}(X : Y) &= \int_{-\infty}^{\infty} (f_0(x)) \ln\left(\frac{f_0(x)}{f_1(x)}\right) dx \\ &= \ln\left(\frac{\sigma_1}{\sigma_0}\right) + \frac{\sigma_0^2 - \sigma_1^2}{2\sigma_1^2\sigma_0^2} \int_{-\infty}^{\infty} x^2 \cdot f_0(x) dx \\ &\quad + \frac{2(\mu_0\sigma_1^2 - \mu_1\sigma_0^2)}{2\sigma_1^2\sigma_0^2} \int_{-\infty}^{\infty} x \cdot f_0(x) dx \\ &\quad + \frac{\mu_1^2\sigma_0^2 - \mu_0^2\sigma_1^2}{2\sigma_1^2\sigma_0^2} \int_{-\infty}^{\infty} f_0(x) dx \\ &= \ln\left(\frac{\sigma_1}{\sigma_0}\right) + \frac{(\mu_1 - \mu_0)^2 + (\sigma_0^2 - \sigma_1^2)}{2\sigma_1^2}. \end{aligned}$$

□

Using this lemma, we can get a formula for the J-divergence. It is provided in the following Theorem 2.1.

Theorem 2.1. *Consider X, Y to be random variables. Let* $f_0(x)$ *and* $f_1(x)$ *be density functions given by* (2.1) *and* (2.2), *respectively. Then it holds*

$$D_J(X, Y) = \frac{(\mu_1 - \mu_0)^2 \cdot (\sigma_1^2 + \sigma_0^2) + (\sigma_1^2 - \sigma_0^2)^2}{2\sigma_0^2\sigma_1^2}. \tag{2.4}$$

Proof. Following Definition 1.1 and the previous Lemma 2.1, we have

$$\begin{aligned} D_J(X, Y) &= D_{KL}(X : Y) + D_{KL}(Y : X) \\ &= \ln\left(\frac{\sigma_1}{\sigma_0}\right) + \frac{(\mu_1 - \mu_0)^2 + (\sigma_0^2 - \sigma_1^2)}{2\sigma_1^2} \end{aligned}$$

$$+ \ln\left(\frac{\sigma_0}{\sigma_1}\right) + \frac{(\mu_0 - \mu_1)^2 + (\sigma_1^2 - \sigma_0^2)}{2\sigma_0^2}$$

$$= \frac{(\mu_1 - \mu_0)^2 \cdot (\sigma_1^2 + \sigma_0^2) + (\sigma_1^2 - \sigma_0^2)^2}{2\sigma_0^2\sigma_1^2}. \qquad \square$$

For parametric estimation of D_J, one can use (2.4) with MLE estimates of the parameters. Consider $X_1^j, ..., X_{n_j}^j \sim N(\mu_j, \sigma_j)$, $j = 0, 1$ is a random sample of scores. MLE estimates (see [6]) are given by

$$\hat{\mu}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_i, \tag{2.5}$$

$$\hat{\sigma}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} (x_i - \hat{\mu}_j)^2. \tag{2.6}$$

3. Beta Distributed Scores

Let X and Y be beta distributed random variables with density functions $f_0(x)$ and $f_1(x)$ given by

$$\begin{aligned} f_0(x) &= \frac{1}{B(\alpha_0, \beta_0)} \cdot x^{\alpha_0 - 1} \cdot (1 - x)^{\beta_0 - 1}, \quad x \in (0, 1), \\ f_0(x) &= 0, \quad x \in (-\infty, 0] \cup [1, \infty), \end{aligned} \tag{3.1}$$

$$\begin{aligned} f_1(x) &= \frac{1}{B(\alpha_1, \beta_1)} \cdot x^{\alpha_1 - 1} \cdot (1 - x)^{\beta_1 - 1}, \quad x \in (0, 1), \\ f_1(x) &= 0, \quad x \in (-\infty, 0] \cup [1, \infty), \end{aligned} \tag{3.2}$$

where $\alpha_0 > 0$, $\alpha_1 > 0$, $\beta_0 > 0$ and $\beta_1 > 0$ are real constants.

Lemma 3.1. *For real constants* $\alpha_0 > 0$, *and* $\beta_0 > 0$ *holds*

$$\int_0^1 \frac{1}{B(\alpha_0, \beta_0)} x^{\alpha_0 - 1} \cdot (1 - x)^{\beta_0 - 1} \cdot \ln(x)\, dx = \psi(\alpha_0) - \psi(\alpha_0 + \beta_0), \tag{3.3}$$

where $\psi(x)$ *is the digamma function given by* $\psi(x) = \dfrac{\Gamma(x)'}{\Gamma(x)}$.

Proof. Using the formula

$$\int_0^1 x^{\mu-1}\cdot(1-x^r)^{\nu-1}\cdot\ln(x)\,dx=\frac{1}{r^2}B\left(\frac{\mu}{r},\,\nu\right)\left[\psi\left(\frac{\mu}{r}\right)-\psi\left(\frac{\mu}{r}+\nu\right)\right],$$

which holds for all real constants $\mu>0$, $\nu>0$ and $r>0$ (see [3, p. 570, eq. 4.253.1]), the statement of the lemma follows considering $\mu=\alpha_0$, $r=1$ and $\nu=\beta_0$. □

More details about the digamma function can be found for instance in Gradshtein and Ryzhik [3] or Medina and Moll [10]. Using the previous lemma, we can get a formula for D_{KL}.

Lemma 3.2. *Consider X, Y to be random variables. Let* $f_0(x)$ *and* $f_1(x)$ *be density functions given by* (3.1) *and* (3.2), *respectively. Then it holds*

$$D_{KL}(X:Y)=\ln\left(\frac{B(\alpha_1,\beta_1)}{B(\alpha_0,\beta_0)}\right)+(\alpha_0-\alpha_1)\psi(\alpha_0)+(\beta_0-\beta_1)\psi(\beta_0)$$

$$+(\alpha_1-\alpha_0+\beta_1-\beta_0)\psi(\alpha_0+\beta_0). \tag{3.4}$$

Proof. Using the previous Lemma 3.1 and the fact that $f_0(1-x)=\frac{1}{B(\beta_0,\,\alpha_0)}\cdot x^{\beta_0-1}\cdot(1-x)^{\alpha_0-1}$, we have

$$D_{KL}(X:Y)=\int_0^1 f_0(x)\cdot\ln\left(\frac{f_0(x)}{f_1(x)}\right)dx$$

$$=\int_0^1 f_0(x)\cdot\ln\left(\frac{\frac{1}{B(\alpha_0,\,\beta_0)}\cdot x^{\alpha_0-1}\cdot(1-x)^{\beta_0-1}}{\frac{1}{B(\alpha_1,\,\beta_1)}\cdot x^{\alpha_1-1}\cdot(1-x)^{\beta_1-1}}\right)dx$$

$$=\int_0^1 f_0(x)\cdot\left[\ln\frac{B(\alpha_1,\,\beta_1)}{B(\alpha_0,\,\beta_0)}+\ln x^{\alpha_0-\alpha_1}+\ln(1-x)^{\beta_0-\beta_1}\right]dx$$

$$= \ln\left(\frac{B(\alpha_1, \beta_1)}{B(\alpha_0, \beta_0)}\right) + (\alpha_0 - \alpha_1)\psi(\alpha_0) + (\beta_0 - \beta_1)\psi(\beta_0)$$

$$+ (\alpha_1 - \alpha_0 + \beta_1 - \beta_0)\psi(\alpha_0 + \beta_0). \quad \square$$

The formula for J-divergence is the direct consequence of the previous lemma.

Theorem 3.1. *Consider X, Y to be random variables. Let* $f_0(x)$ *and* $f_1(x)$ *be density functions given by* (3.1) *and* (3.2), *respectively. Then it holds*

$$D_J(X, Y) = (\alpha_1 - \alpha_0) \cdot (\psi(\alpha_1) - \psi(\alpha_0)) + (\beta_1 - \beta_0) \cdot (\psi(\beta_1) - \psi(\beta_0))$$

$$+ (\alpha_1 - \alpha_0 + \beta_1 - \beta_0) \cdot (\psi(\alpha_0 + \beta_0) - \psi(\alpha_1 + \beta_1)). \quad (3.5)$$

Proof. Follows from Definition 1.1 and the previous Lemma 3.2:

$$D_J(X, Y) = D_{KL}(X : Y) + D_{KL}(Y : X)$$

$$= (\alpha_1 - \alpha_0) \cdot (\psi(\alpha_1) - \psi(\alpha_0)) + (\beta_1 - \beta_0) \cdot (\psi(\beta_1) - \psi(\beta_0))$$

$$+ (\alpha_1 - \alpha_0 + \beta_1 - \beta_0) \cdot (\psi(\alpha_0 + \beta_0) - \psi(\alpha_1 + \beta_1)). \quad \square$$

If we do not like to use the digamma function, we can use its approximation. Then the formula for the J-divergence can be expressed in the following way.

Corollary 3.1.1. *Following assumptions of Theorem* 3.1 *and approximation of digamma function* $\psi(x) \approx \ln(x - 0.5)$ (*can be found in* [7, *p*. 223]), *it holds*

$$D_J(X, Y) \approx \ln\left[\left(\frac{\alpha_1 - 0.5}{\alpha_0 - 0.5}\right)^{\alpha_1 - \alpha_0} \cdot \left(\frac{\beta_1 - 0.5}{\beta_0 - 0.5}\right)^{\beta_1 - \beta_0}\right.$$

$$\left. \cdot \left(\frac{\alpha_0 + \beta_0 - 0.5}{\alpha_1 + \beta_1 - 0.5}\right)^{\alpha_1 - \alpha_0 + \beta_1 - \beta_0}\right]. \quad (3.6)$$

When the beta distribution is considered, one can use (3.5) or (3.6) for parametric estimation of D_J with moment or MLE estimates of the parameters. Consider $X_1^j, ..., X_{n_j}^j \sim Beta(\alpha_j, \beta_j), \quad j = 0, 1$ is a random sample of scores. MLE estimates of α_j and β_j (see [7]) can be computed iteratively by

$$\begin{bmatrix} \alpha_j^{(k+1)} \\ \beta_j^{(k+1)} \end{bmatrix} = \begin{bmatrix} \alpha_j^{(k)} \\ \beta_j^{(k)} \end{bmatrix} - g(\alpha_j^{(k)}, \beta_j^{(k)})^T \cdot [h(\alpha_j^{(k)}, \beta_j^{(k)})]^{-1}, \tag{3.7}$$

where

$$g(\alpha, \beta) = \begin{bmatrix} n_j \cdot (\psi(\alpha + \beta) - \psi(\alpha)) + \sum_{i=1}^{n} \ln x_i^j \\ n_j \cdot (\psi(\alpha + \beta) - \psi(\alpha)) + \sum_{i=1}^{n} \ln(1 - x_i^j) \end{bmatrix},$$

$$h(\alpha, \beta) = \begin{bmatrix} n_j \cdot (\psi_1(\alpha + \beta) - \psi_1(\alpha)) & n_j \cdot \psi_1(\alpha + \beta) \\ n_j \cdot \psi_1(\alpha + \beta) & n_j \cdot (\psi_1(\alpha + \beta) - \psi_1(\beta)) \end{bmatrix}$$

and $\psi_1(x)$ is the trigamma function defined as $\psi_1(x) = \dfrac{\partial \psi(x)}{\partial x} = \dfrac{\partial^2 \ln \Gamma(x)}{\partial^2 x}$.

As an initial approximation one can take the moment estimates (see [7] for more details).

4. Gamma Distributed Scores

Let X and Y be gamma distributed random variables with density functions $f_0(x)$ and $f_1(x)$ given by

$$\begin{aligned} f_0(x) &= \frac{\lambda_0^{\alpha_0}}{\Gamma(\alpha_0)} x^{\alpha_0 - 1} e^{-\lambda_0 x}, & x \in (0, \infty), \\ f_0(x) &= 0, & x \in (-\infty, 0], \end{aligned} \tag{4.1}$$

$$f_1(x) = \frac{\lambda_1^{\alpha_1}}{\Gamma(\alpha_1)} x^{\alpha_1 - 1} e^{-\lambda_1 x}, \qquad x \in (0, \infty),$$
$$f_1(x) = 0, \qquad x \in (-\infty, 0], \tag{4.2}$$

where $\alpha_0 > 0$, $\alpha_1 > 0$, $\lambda_0 > 0$ and $\lambda_1 > 0$ are real constants.

Lemma 4.1. *For real constants* $\alpha_0 > 0$, *and* $\lambda_0 > 0$ *holds*

$$\int_0^\infty \frac{\lambda_0^{\alpha_0}}{\Gamma(\alpha_0)} x^{\alpha_0 - 1} \cdot e^{-\lambda_0 x} \cdot \ln(x) dx = \psi(\alpha_0) - \ln(\lambda_0). \tag{4.3}$$

Proof. Follows from formula

$$\int_0^\infty x^{\nu - 1} \cdot e^{\mu x} \cdot \ln(x) dx = \frac{1}{\mu^\nu} \Gamma(\nu)[\psi(\nu) - \ln(\mu)],$$

which holds for all real constants $\mu > 0$ and $\nu > 0$ (can be found in [3, p. 604, eq. 4.352.1]), if we choose $\mu = \lambda_0$ and $\nu = \alpha_0$. □

Provided the previous lemma, we can get a formula for D_{KL} for the case of gamma distributed random variables, such as *odds*.

Lemma 4.2. *Consider X, Y to be random variables. Let* $f_0(x)$ *and* $f_1(x)$ *be density functions given by* (4.1) *and* (4.2), *respectively. Then it holds*

$$D_{KL}(X : Y) = (\alpha_0 - \alpha_1) \cdot \psi(\alpha_0) + \ln\left(\frac{\Gamma(\alpha_1)}{\Gamma(\alpha_0)}\right)$$
$$+ \alpha_1 \ln\left(\frac{\lambda_0}{\lambda_1}\right) + \alpha_0 \cdot \left(\frac{\lambda_1}{\lambda_0} - 1\right). \tag{4.4}$$

Proof. One can find this result in Gupta and Kundu [4]. However, it is easy to show that using the relation $\Gamma(x + 1) = x\Gamma(x)$ and Lemma 4.1, we obtain (4.4). □

Now, when we know the formula for D_{KL}, it is easy to get formula for D_J.

Theorem 4.1. *Consider X, Y to be random variables. Let* $f_0(x)$ *and* $f_1(x)$ *be density functions given by* (4.1) *and* (4.2), *respectively. Then it holds*

$$D_J(X, Y) = (\alpha_1 - \alpha_0) \cdot \left(\psi(\alpha_1) - \psi(\alpha_0) + \ln\left(\frac{\lambda_0}{\lambda_1}\right)\right) + \alpha_0 \cdot \left(\frac{\lambda_1}{\lambda_0} - 1\right) + \alpha_1 \cdot \left(\frac{\lambda_0}{\lambda_1} - 1\right). \quad (4.5)$$

Proof. Following Definition 1.1 and the previous Lemma 4.2, we have

$$\begin{aligned} D_J(X, Y) &= D_{KL}(X : Y) + D_{KL}(Y : X) \\ &= \alpha_0 \ln(\lambda_0) - \alpha_1 \ln(\lambda_1) + \ln\left(\frac{\Gamma(\alpha_1)}{\Gamma(\alpha_0)}\right) \\ &\quad + (\alpha_0 - \alpha_1) \cdot (\psi(\alpha_0) - \ln(\lambda_0)) + \alpha_0 \cdot \left(\frac{\lambda_1}{\lambda_0} - 1\right) \\ &\quad + \alpha_1 \ln(\lambda_1) - \alpha_0 \ln(\lambda_0) + \ln\left(\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)}\right) \\ &\quad + (\alpha_1 - \alpha_0) \cdot (\psi(\alpha_1) - \ln(\lambda_1)) + \alpha_1 \cdot \left(\frac{\lambda_0}{\lambda_1} - 1\right) \\ &= (\alpha_1 - \alpha_0) \cdot \left(\psi(\alpha_1) - \psi(\alpha_0) + \ln\left(\frac{\lambda_0}{\lambda_1}\right)\right) \\ &\quad + \alpha_0 \cdot \left(\frac{\lambda_1}{\lambda_0} - 1\right) + \alpha_1 \cdot \left(\frac{\lambda_0}{\lambda_1} - 1\right). \end{aligned}$$

□

When the gamma distribution is considered, one can use (4.5) for parametric estimation of D_J with moment or MLE estimates of the parameters. Consider $X_1^j, ..., X_{n_j}^j \sim Gamma(\alpha_j, \lambda_j)$, $j = 0, 1$ is a random sample of scores. MLE estimates of α_j and λ_j (see [6, pp. 360-367]) can be computed by

$$n_j^{-1}\sum_{i=1}^{n_j}\log x_i^j = \psi(\hat{\alpha}_j) - \log\hat{\lambda}_j, \qquad n_j^{-1}\sum_{i=1}^{n_j} x_i^j = \frac{\hat{\alpha}_j}{\hat{\lambda}_j}, \tag{4.6}$$

which leads to solving

$$R_{n,j} = \ln\left(\frac{\bar{x}_j}{\tilde{x}_j}\right) = \ln\hat{\alpha}_j - \psi(\hat{\alpha}_j), \tag{4.7}$$

where $\bar{x}_j$ is the arithmetic mean and $\tilde{x}_j$ is the geometric mean of appropriate realizations of $X_1^j, ..., X_{n_j}^j$, $j = 0, 1$. If we use the approximation $\psi(x) \approx \ln(x - 0.5)$, then we can get

$$\hat{\alpha}_j \approx \frac{\bar{x}_j}{2(\bar{x}_j - \tilde{x}_j)}. \tag{4.8}$$

Further approximation can be found in [6]:

$$\hat{\alpha}_j \approx \frac{1 + \sqrt{1 + \frac{4}{3}R_{n,j}}}{4R_{n,j}}. \tag{4.9}$$

5. Empirical Estimates of J-divergence

It is natural to ask what is the performance of parametric estimates of J-divergence compared to some standard estimate from practice. Empirical estimate using deciles of scores, i.e., discretization of data into bins using deciles with requirement on the nonzero number of cases for all bins, is the classical way how to compute the J-divergence (Information value) for credit scoring models.

The main idea of the empirical estimate approach is to replace unknown densities by their empirical estimates. Let us have n score values, of which n_0 score values s_{0_i}, $i = 1, ..., n_0$ for bad clients and n_1 score values s_{1_j}, $j = 1, ..., n_1$ for good clients and denote L (resp. H) as the minimum (resp.

maximum) of all values. Let us divide the interval $[L, H]$ up to r equal subintervals $[q_0, q_1]$, $(q_1, q_2]$, ..., $(q_{r-1}, q_r]$, where $q_0 = L$, $q_r = H$. Set

$$n_{0_j} = \sum_{i=1}^{n_0} I(s_{0_i} \in (q_{j-1}, q_j]),$$

$$n_{1_j} = \sum_{i=1}^{n_1} I(s_{1_i} \in (q_{j-1}, q_j]), \quad j = 1, ..., r \tag{5.1}$$

observed counts of bad or good clients in each interval. Then the empirical estimate of the J-divergence is calculated by

$$\hat{D}_{J,EMP} = \sum_{j=1}^{r} \left(\frac{n_{1_j}}{n_1} - \frac{n_{0_j}}{n_0} \right) \ln \left(\frac{n_{1_j} n_0}{n_{0_j} n_1} \right). \tag{5.2}$$

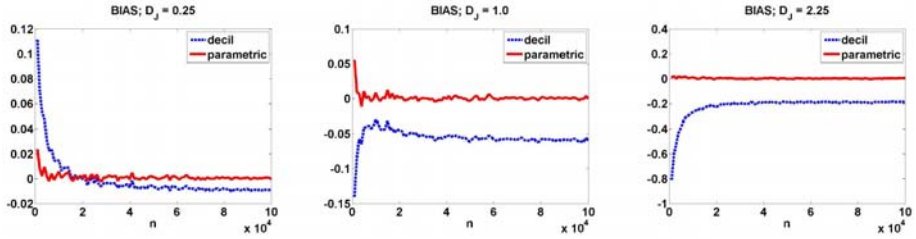
However, in practice, there could occur computational problems. The J-divergence becomes infinite in cases when some of n_{0_i} or n_{1_i} are equal to 0. When this arises there are numerous practical procedures for preserving finite results. For example we replace the zero entry of numbers of goods or bads by a minimum constant of say 0.0001. Choosing of the number of bins is also very important. In the literature and also in many applications in credit scoring, the value $r = 10$ is preferred.

Considering $r = 10$, i.e., the J-divergence (Information value) is computed using deciles of the score, we denote the estimate by $\hat{D}_{J,DEC}$.

6. Simulation Study

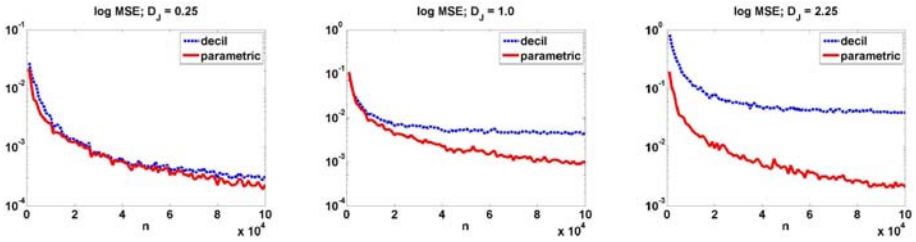
To compare all listed estimates of the J-divergence, a simulation study was done. Consider *n* clients, np_B of bad clients $n(1 - p_B)$ of good clients, $p_B = 0.02, 0.05, 0.1$ and 0.2. Further consider scores that are normally, beta and gamma distributed with parameters that lead to $D_J = 0.25, 1$ and 2.25. These choices represent models that have weak to very high performance and

portfolios with very low risk (2% bad rate $(p_B = 0.02)$, e.g., mortgages before current crisis) to very high risk, e.g., subprime cash loans. The number of clients was set from $n = 1000$ to $n = 100000$, which represents a very small to a large data set. Given the choice of distributions of the scores and the choice of parameters we know the exact value of D_J, thus we can assess quality of the J-divergence estimates. Very common way how to do it is to compute bias and mean square error (MSE). For this purpose, the simulation was done in the following way. First, scores of bad and good clients were generated according to given distribution and its parameters. Second, all mentioned estimates were computed. Third, square error for all estimates was computed. These three steps were repeated one thousand times. Finally, MSE was computed as average of appropriate square errors and the bias was computed as average of appropriate estimate minus the theoretical value of the J-divergence. The results are presented in the following Figures 4 to 9.



(a) (b) (c)

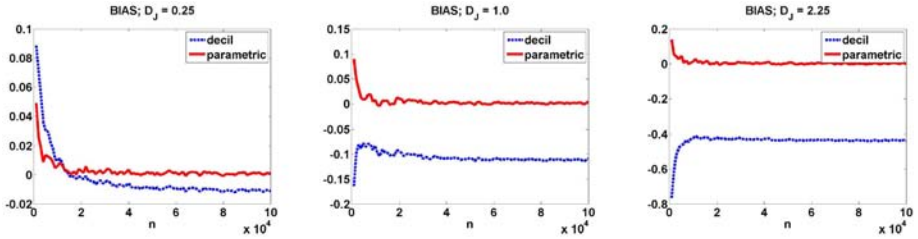
Figure 4. Bias for parametric and decile estimate for normally distributed scores: (a) $D_J = 0.25$, (b) $D_J = 1.0$, (c) $D_J = 2.25$.



(a) (b) (c)

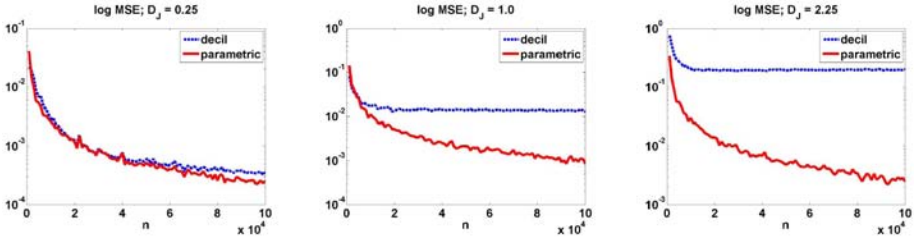
Figure 5. MSE for parametric and decile estimate for normally distributed scores: (a) $D_J = 0.25$, (b) $D_J = 1.0$, (c) $D_J = 2.25$.

It can be seen from Figures 4 and 5 that the parametric estimate was significantly better than $\hat{D}_{J,DEC}$ in case of strong models (D_J greater or equal to 1). The parametric estimate and $\hat{D}_{J,DEC}$ were comparable according to MSE, but the parametric estimate was much more stable according to bias.



(a) (b) (c)

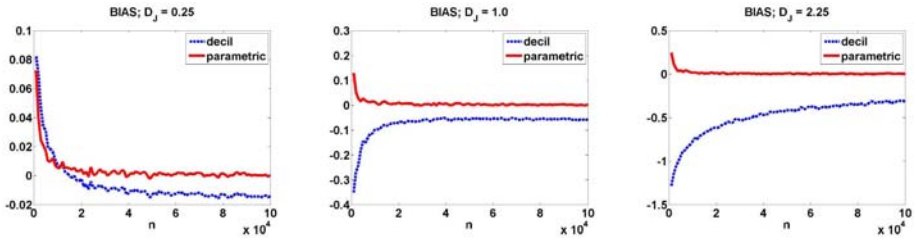
Figure 6. Bias for parametric and decile estimate for beta distributed scores: (a) $D_J = 0.25$, (b) $D_J = 1.0$, (c) $D_J = 2.25$.



(a) (b) (c)

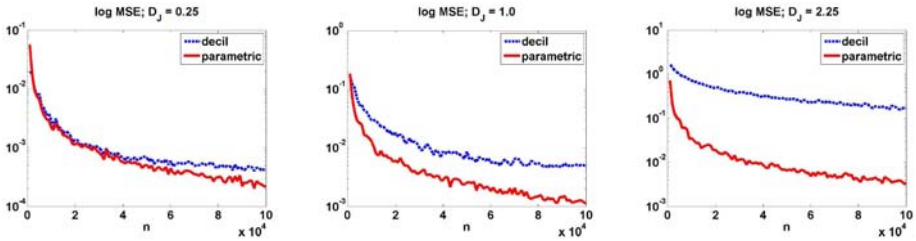
Figure 7. MSE for parametric and decile estimate for beta distributed scores: (a) $D_J = 0.25$, (b) $D_J = 1.0$, (c) $D_J = 2.25$.

We can see in Figures 6 and 7 that it holds the same statement as for normally distributed data. The only difference is in the fact that the parametric estimate is much more better than $\hat{D}_{J,DEC}$ for stronger models ($D_J \geq 1$).



(a) (b) (c)

Figure 8. Bias for parametric and decile estimate for gamma distributed scores: (a) $D_J = 0.25$, (b) $D_J = 1.0$, (c) $D_J = 2.25$.



(a) (b) (c)

Figure 9. MSE for parametric and decile estimate for gamma distributed scores: (a) $D_J = 0.25$, (b) $D_J = 1.0$, (c) $D_J = 2.25$.

Again, we can see, now in Figures 8 and 9 that $\hat{D}_{J,DEC}$ had comparable performance for weak models and low sample size. But once it was the case that a model was stronger or the sample size was bigger, the parametric estimate significantly outperformed $\hat{D}_{J,DEC}$.

7. Conclusions

We considered three classes of distributions which typically occur in credit scoring. Namely, it was normal, beta and gamma distribution. We showed on the real data how these distributions can be achieved. Then the J-divergence, which is one of the widely used indexes for assessing credit scoring models, was introduced. Methods of estimation of the J-divergence were discussed.

The main aim of the paper was to provide formulas for normally, beta

and gamma distributed scores. This allows the parametric estimating of the J-divergence. Furthermore they allow to asses the quality of nonparametric estimators of the J-divergence via MSE, bias and so forth.

We made a simulation study to compare parametric estimates with the empirical estimator using deciles, which is a nonparametric estimator widely used in practice. As was expected, parametric estimates had much more better performance than the nonparametric one for strong credit scoring models or for large data sizes. However, the performance depended on data size in case of weak models - in some small range of data size had parametric and nonparametric estimates comparable performance. All these statements were valid for all three considered distribution classes. Overall, we can state that the parametric estimates are, according to MSE and bias, better to use for all sizes of data and all considered distribution classes.

References

[1] R. Anderson, The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation, Oxford University Press, Oxford, 2007.

[2] A. Cichocki and S. Amari, Families of alpha-beta-and gamma-divergences: flexible and robust measures of similarities, Entropy 12 (2010), 1532-1568.

[3] I. S. Gradshtein and I. M. Ryzhik, Tables of Integrals, Sums, Series and Products, Academic Press, New York, London, 1965.

[4] R. D. Gupta and D. Kundu, Closeness of gamma and generalized exponential distributions, Comm. Statist. Theory Methods 32(4) (2003), 705-721.

[5] D. J. Hand and W. E. Henley, Statistical classification methods in consumer credit scoring: a review, J. Royal Statistical Society, Series A 160(3) (1997), 523-541.

[6] N. L. Johnson, S. Kotz and N. Balakrishnan, Continuous Univariate Distributions, Vol. 1, 2nd ed., Wiley, New York, 1994.

[7] N. L. Johnson, S. Kotz and N. Balakrishnan, Continuous Univariate Distributions Vol. 2, 2nd ed., Wiley, New York, 1995.

[8] J. Koláček and M. Řezáč, Assessment of scoring models using information value, 19th International Conference on Computational Statistics, Paris, France, August 22-27, 2010 Keynote, Invited and Contributed Papers 1, 2010, pp. 1191-1198.

[9] D. Kundu and A. Manglick, Discriminating between the log-normal and gamma distributions, J. Applied Statistical Sciences 14 (2005), 175-187.

[10] L. A. Medina and V. H. Moll, The integrals in Gradshteyn and Ryzhik. Part 10: the digamma function, Scientia, Series A: Mathemaical Sciences 17 (2009), 45-66.

[11] M. Řezáč, Advanced empirical estimate of information value for credit scoring models, Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis LIX (2) (2011), 267-273.

[12] M. Řezáč and F. Řezáč, How to measure the quality of credit scoring models finance a úvěr, Czech J. Economics and Finance 61(5) (2011), 486-507.

[13] N. Siddiqi, Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring, Wiley, New Jersey, 2006.

[14] L. C. Thomas, A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers, Inter. J. Forecasting 16(2) (2000), 149-172.

[15] L. C. Thomas, Consumer Credit Models: Pricing, Profit, and Portfolio, Oxford University Press, Oxford, 2009.

[16] L. C. Thomas, D. B. Edelman and J. N. Crook, Credit Scoring and its Applications, SIAM Monographs on Mathematical Modeling and Computation, Philadelphia, 2002.

[17] B. L. Wiens, When log-normal and gamma models give different results: a case study, The American Statistician 53(2) (1999), 89-93.

[18] A. D. Wilkie, Measures for comparing scoring systems, Readings in Credit Scoring, L. C. Thomas, D. B. Edelman and J. N. Crook, eds., Oxford University Press, Oxford, 2004, pp. 51-62.