



MASARYK UNIVERSITY
Czech Republic

BETA DISTRIBUTED CREDIT SCORE - ESTIMATION OF ITS J-DIVERGENCE

Martin Řezáč

Dept. of Mathematics and Statistics, Faculty of Science,
Masaryk University

International Days of Statistics and Economics
September 19–21, 2013

Introduction

- J-divergence (beside Hellinger and Jensen-Shannon) is widely used to describe the difference between two probability distributions F_0 and F_1 .
- It is also called the Information value for the purpose of scoring models.
 - I.e. models that try to predict a probability of an event, e.g. client's default.
 - In this case, the J-divergence is widely used (beside Gini index and K-S statistic) for assessment of the scoring models.
- Empirical estimate using deciles of scores is the common way how to compute it.
 - However, it may lead to strongly biased results.
 - Moreover, there are some computational issues to solve.

Introduction

- Two alternative methods to this approach can be used. First, it is the kernel smoothing theory, which allows estimating unknown densities and consequently, using some numerical method for integration, to estimate value of the J-divergence.

- ESIS (empirical estimate with supervised interval selection) estimator is the second alternative.
 - It is based on idea of constructing such intervals of scores which ensure to have sufficiently enough observations in each interval.
 - The quantile functions F_0^{-1} and F_1^{-1} are used for this purpose.

- The main objective of this paper is to describe the behaviour of the J-divergence estimates of credit scoring models with Beta distributed score.

J-divergence

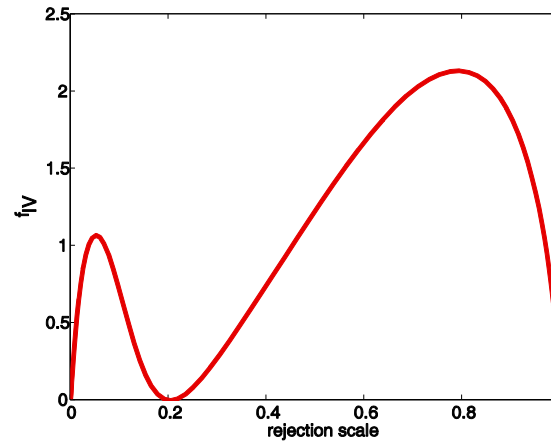
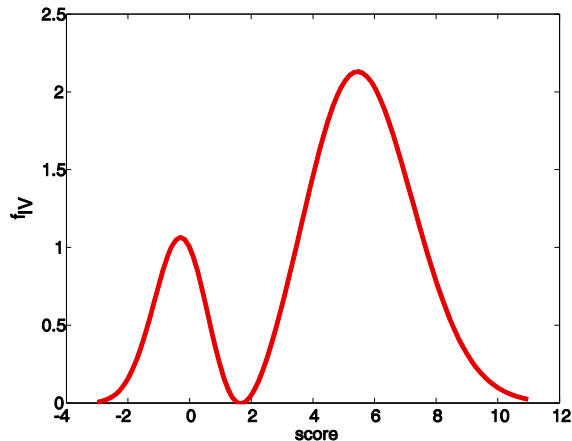
➤ Symetrized Kullback-Leibler divergence of two random variables X, Y is given by:

$$D_J(X, Y) = D_{KL}(X : Y) + D_{KL}(Y : X) = \int_{-\infty}^{\infty} f_{IV}(x) dx$$

where $f_{IV}(x) = (f_1(x) - f_0(x)) \cdot \ln\left(\frac{f_1(x)}{f_0(x)}\right)$ $D_{KL}(X : Y) = \int_{-\infty}^{\infty} f_1(x) \cdot \ln\left(\frac{f_1(x)}{f_0(x)}\right) dx$

- $f_1(x), f_0(x)$ are densities of scores of bad and good clients.

The example of $f_{IV}(x)$ for 10% of bad clients with $f_0 \sim N(0, 1)$ and 90% of good clients with $f_1 \sim N(4, 2)$



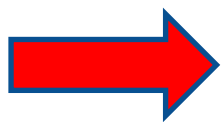
Kernel estimate

- The kernel density estimates are defined by

$$\tilde{f}_0(x, h_0) = \frac{1}{n_0} \sum_{i=1}^{n_0} K_{h_0}(x - s_{0i})$$

$$\tilde{f}_1(x, h_1) = \frac{1}{n_1} \sum_{i=1}^{n_1} K_{h_1}(x - s_{1i})$$

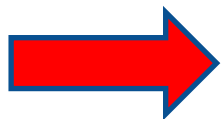
where $K_{h_i}(x) = \frac{1}{h_i} K\left(\frac{x}{h_i}\right)$, $i = 0, 1$, K is a kernel function, for instance the Epanechnikov kernel and h_i is a smoothing parameter.



$$\tilde{f}_{IV}(x) = (\tilde{f}_1(x, h_1) - \tilde{f}_0(x, h_0)) \ln \left(\frac{\tilde{f}_1(x, h_1)}{\tilde{f}_0(x, h_0)} \right).$$

- Using the composite trapezoidal rule with given $M+1$ equidistant points

$L = x_0, x_1, \dots, x_M = H$ we have:



$$\hat{D}_{J, KERN} = \frac{H - L}{2M} \left(\tilde{f}_{IV}(L) + 2 \sum_{i=1}^{M-1} \tilde{f}_{IV}(x_i) + \tilde{f}_{IV}(H) \right).$$

Empirical estimate of D_j

The main idea of this approach is to replace unknown densities by their empirical estimates. Let's have n score values, of which n_0 score values s_{0_i} , $i = 1, \dots, n_0$ for bad clients and n_1 score values s_{1_j} , $j = 1, \dots, n_1$ for good clients and denote L (resp. H) as the minimum (resp. maximum) of all values. Let's divide the interval $[L, H]$ up to r equal subintervals $[q_0, q_1], (q_1, q_2], \dots, (q_{r-1}, q_r]$, where $q_0 = L, q_r = H$. Set

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

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

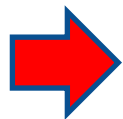
observed counts of bad or good clients in each interval.



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

Empirical estimate of D_j

- However in practice, there could occur computational problems. The J-divergence becomes infinite in cases when some of n_{0j} or n_{1j} are equal to 0.
- 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.



Empirical estimate using deciles of scores

Empirical estimate with supervised interval selection (ESIS)

- We want to avoid zero values of n_{0j} or n_{1j} .
- We propose to require to have at least k , where k is a positive integer, observations of scores of both good and bad clients in each interval.
- Intervals of score are given by

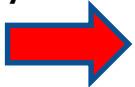
$$\begin{aligned} q_0 &= L - 1 \\ q_i &= \widehat{F}_0^{-1} \left(\frac{k \cdot i}{n_0} \right), i = 1, \dots, \lfloor \frac{n_0}{k} \rfloor \\ q_{\lfloor \frac{n_0}{k} \rfloor + 1} &= H, \end{aligned}$$

where $\widehat{F}_0^{-1}(\cdot)$ is the empirical quantile function appropriate to the empirical cumulative distribution function of scores of bad clients.

- Very important is the choice of k . If we choose too small value, we get overestimated value of the J-divergence, and vice versa. As a reasonable compromise seems to be adjusted square root of number of bad clients given by

$$k = \lceil \sqrt{n_0} \rceil.$$

- The estimate of the J-div. is given by



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

ESIS1

➤ Algorithm for the modified ESIS:

1) $\mathbf{q} = []$

2) $q_{j1} = F_1^{-1}\left(\frac{k}{n_1}\right) \quad q_{j0} = F_0^{-1}\left(\frac{k}{n_0}\right)$

3) $s_{\max} = \max(q_{j1}, q_{j0})$

4) Add s_{\max} to the sequence, i.e. $\mathbf{q} = [\mathbf{q}, s_{\max}]$

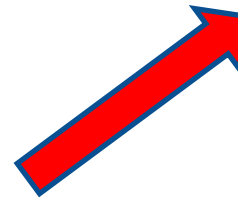
5) Erase all scores $\leq s_{\max}$

6) While n_0 and n_1 are greater than $2*k$, repeat step 2) – 5)

7) $\mathbf{q} = [\min(score) - 1, \mathbf{q}]$

$$\hat{D}_{J,ESIS1}$$

where $k = \lceil \sqrt{n_0} \rceil$

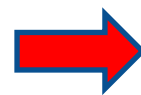


passing data just once with no need to merge the constructed intervals.

ESIS2

➤ Algorithm for the modified ESIS:

- 1) $s_0 = \text{mean} \left(\arg \max_s \left| \hat{F}_1(s) - \hat{F}_0(s) \right| \right)$ it is necessary to solve the case when $|\hat{F}_1(s) - \hat{F}_0(s)|$ takes its maximal value for more than one value s .
- 2) $q_{1_j} = F_1^{-1} \left(\frac{j \cdot k}{n_1} \right), j = 1, \dots, \left\lfloor \frac{n_1}{k} \cdot F_1(s_0) \right\rfloor$
- 3) $q_{0_j} = F_0^{-1} \left(\frac{j \cdot k}{n_0} \right), j = \left\lceil \frac{n_0}{k} \cdot F_0(s_0) \right\rceil, \dots, \left\lfloor \frac{n_0}{k} \right\rfloor - 1$
- 4) $\mathbf{q} = [\min(s) - 1, \mathbf{q}_1, \mathbf{q}_0, \max(s) + 1]$
- 5) Merge intervals given by \mathbf{q}_1 where number of bads is less than k .
- 6) Merge intervals given by \mathbf{q}_0 where number of goods is less than k .



$$\hat{D}_{J,ESIS2}$$

where $k = \lceil \sqrt{n_0} \rceil$

Beta distributed scores

➤ Densities of Beta distributed scores:



$$f_0(x) = \begin{cases} \frac{1}{B(\alpha_0, \beta_0) \cdot \sigma_0^{\alpha_0 + \beta_0 - 1}} (x - \mathcal{G}_0)^{\alpha_0 - 1} \cdot (\sigma_0 + \mathcal{G}_0 - x)^{\beta_0 - 1} & \text{for } \mathcal{G}_0 < x < \mathcal{G}_0 + \sigma_0 \\ 0 & \text{for } x \leq \mathcal{G}_0 \text{ or } x \geq \mathcal{G}_0 + \sigma_0 \end{cases}$$

$$f_1(x) = \begin{cases} \frac{1}{B(\alpha_1, \beta_1) \cdot \sigma_1^{\alpha_1 + \beta_1 - 1}} (x - \mathcal{G}_1)^{\alpha_1 - 1} \cdot (\sigma_1 + \mathcal{G}_1 - x)^{\beta_1 - 1} & \text{for } \mathcal{G}_1 < x < \mathcal{G}_1 + \sigma_1 \\ 0 & \text{for } x \leq \mathcal{G}_1 \text{ or } x \geq \mathcal{G}_1 + \sigma_1. \end{cases}$$

• transformation $y = \frac{x - \mathcal{G}_i}{\sigma_i}$ leads to densities:



$$g_0(y) = \begin{cases} \frac{1}{B(\alpha_0, \beta_0)} y^{\alpha_0 - 1} \cdot (1 - y)^{\beta_0 - 1} & \text{for } 0 < y < 1 \\ 0 & \text{otherwise,} \end{cases}$$

$$g_1(y) = \begin{cases} \frac{1}{B(\alpha_1, \beta_1)} y^{\alpha_1 - 1} \cdot (1 - y)^{\beta_1 - 1} & \text{for } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Gamma function $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$

Beta function $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$

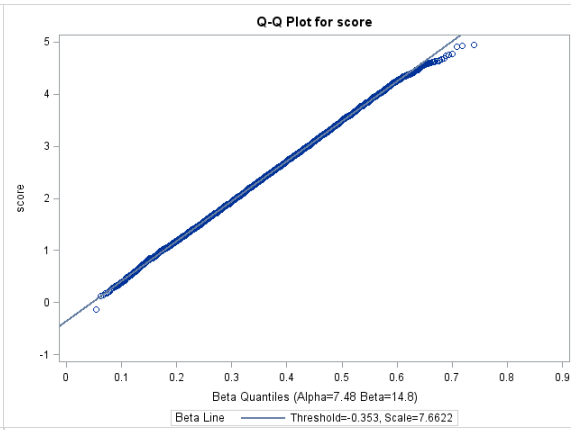
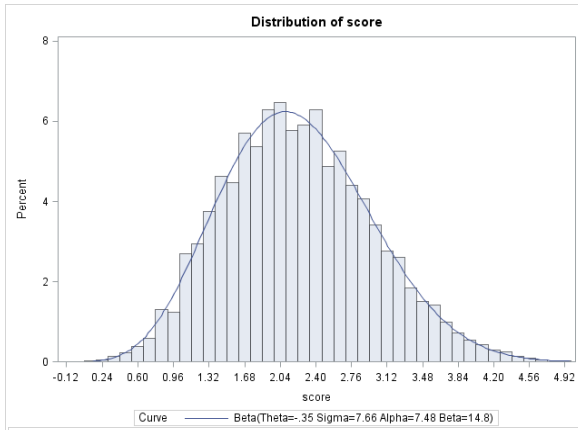
Beta distributed scores

➤ The J-divergence is not generally invariant with respect to transformations. Indeed, this holds for transformations used for converting four-parameters Beta distributed variables given by ♥ to two-parameters Beta distributed variables given by ✨. Nevertheless, when comparing the discriminative power of several credit scoring models on the same data, then this property (disadvantage) does not matter. And what is quite important, estimation of parameters in ♥ and consequent computation of the J-divergence is quite complicated. From this perspective, it seems to be appropriate to use parametric estimate given by

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

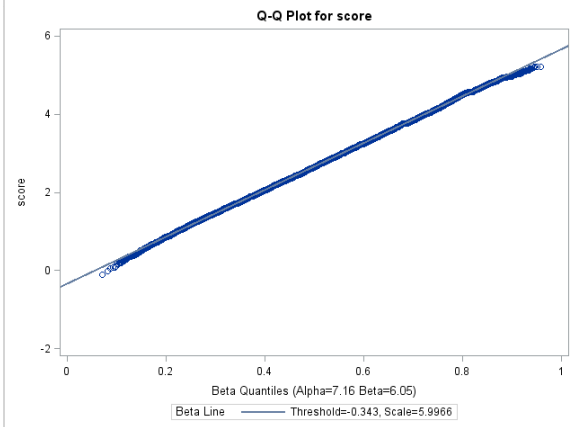
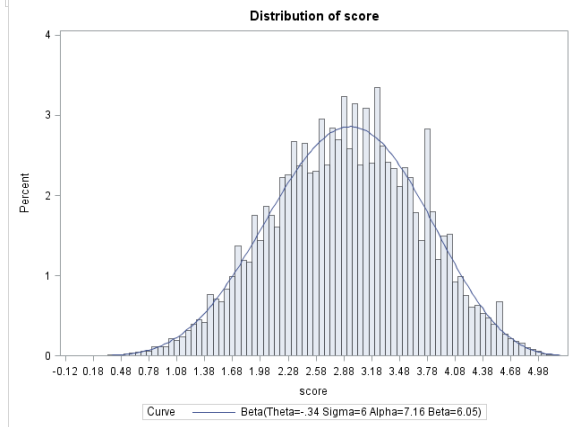
Digamma function $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$

Real data – fitted Beta distributions of scores



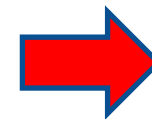
Parameters for Beta Distribution

Parameter	Symbol	Estimate
Threshold	Theta	-0.35327
Scale	Sigma	7.662249
Shape	Alpha	7.479128
Shape	Beta	14.81027
Mean		2.217774
Std Dev		0.749697



Parameters for Beta Distribution

Parameter	Symbol	Estimate
Threshold	Theta	-0.34301
Scale	Sigma	5.996635
Shape	Alpha	7.15525
Shape	Beta	6.05231
Mean		2.905692
Std Dev		0.792681



	D_j
decil	2.508551
kern	2.797372
esis	2.945658
esis1	3.117013
esis2	2.967163
param	3.403594

Simulation settings

- Consider following values of parameters:
 - $n = 1000$ to $100\ 000$
 - **Very small to large data set.**
 - $(\alpha_0, \alpha_1, \beta_0, \beta_1)$ that lead to $D_j = 0.25, 1, 2.25$
 - **Weak to very high performance of a model.**
 - $p_B = 0.02, 0.05, 0.1, 0.2$
 - **Portfolios with very low risk (mortgages) to very high risk (subprime cash loans).**

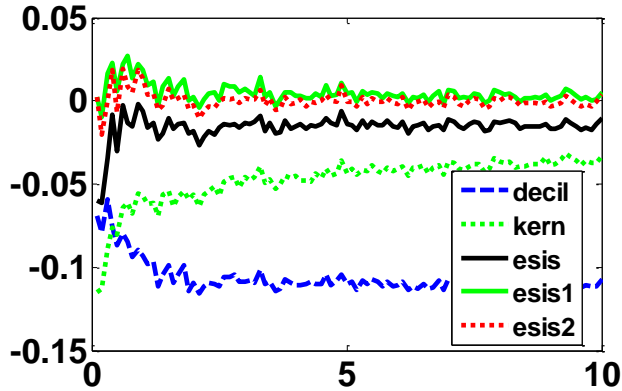
- Very common way how to assess quality of the J-divergence estimators is to compute bias and mean square error (MSE), or its logarithm.

- $$\text{Bias} = E(\hat{I}_{val} - I_{val})$$

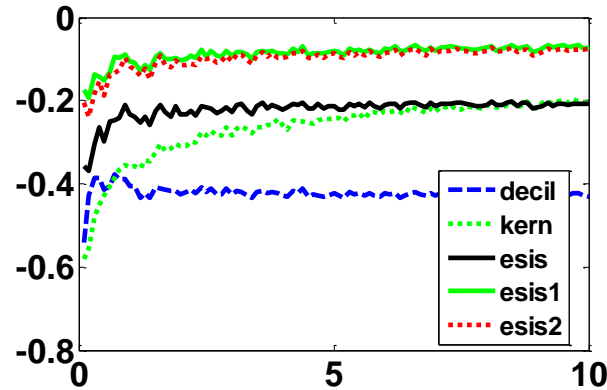
- $$\log MSE = \log(E(\hat{I}_{val} - I_{val})^2)$$

Simulation results

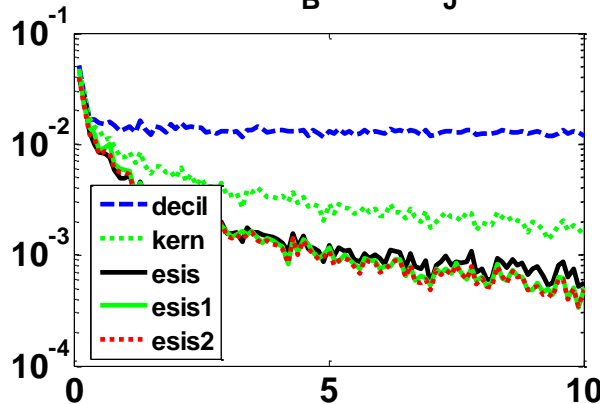
BIAS; $p_B=0,1$; $D_J = 1.0$



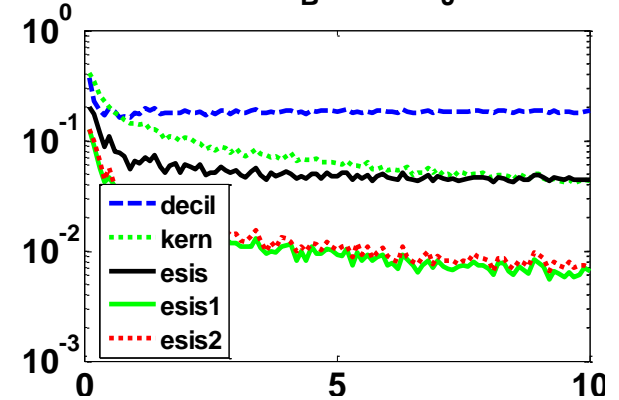
BIAS; $p_B=0,1$; $D_J = 2.25$



Log MSE; $p_B=0,1$; $D_J = 1.0$



Log MSE; $p_B=0,1$; $D_J = 2.25$



➤ From these figures it is apparent that the decile estimate is significantly biased, specifically undervalued. The value of log MSE became quite quickly stabilized and with increasing number of observations did not fall. Overall, this estimate is thus not very suitable. In contrast, algorithms ESIS1 and ESIS2 led in the case of a weaker model ($D_J = 1.00$) to almost unbiased estimate. For a stronger model ($D_J = 2.25$) are their properties worse. However, they were the best of all considered methods of estimating D_J .



Thank you for
your attention