MASARYK UNIVERSITY
Czech Republic

Bias and MSE of J-divergence Estimators for Scoring Models

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

evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
AMathNet
síť pro transfer znalostí v aplikované matematice

Introduction

- J-divergence 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 client's default.
- 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.
- To avoid these issues and to lower the bias, the empirical estimate with supervised interval selection (esis) can be used.
  - It is based on idea of constructing such intervals of scores which ensure to have sufficiently enough observations in each interval.
  - The quantile function F_0^{-1} is used for this purpose.
- The main aim of this paper is to give an alternative estimator of the J-divergence, which leads to lowered bias and MSE.

J-divergence

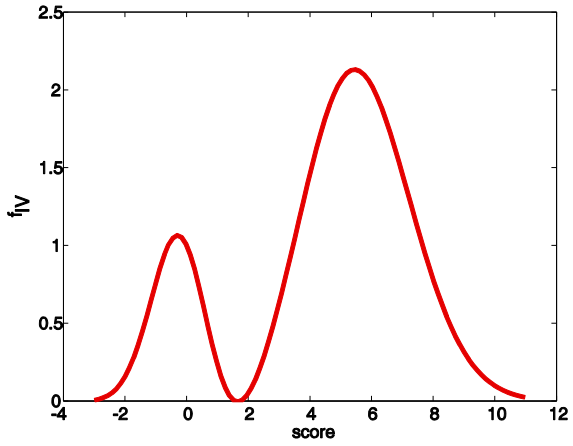
- Symetrized Kullback-Leibler divergence of two random variables X, Y given by:

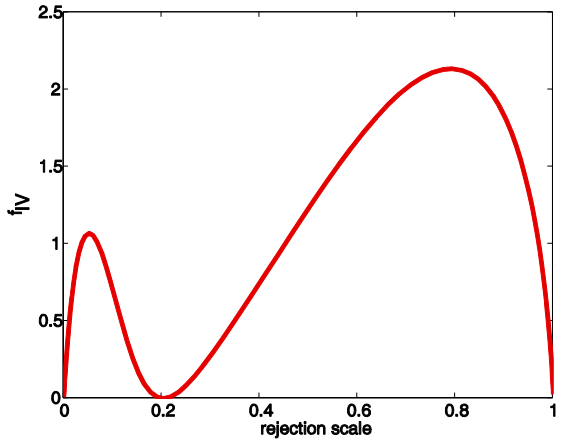
$$D_J(X,Y) = D_{KL}(X:Y) + D_{KL}(Y:X) = \int_{-\infty}^{\infty} f_{IV}(x)\,dx$$

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

- $f_1(x)$, $f_0(x)$ 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

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

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

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

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

- for given $M+1$ equidistant points $L=x_0, x_1, \ldots, x_M=H$

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

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}, $i = 1, \ldots, n_0$ for bad clients and n_1 score values s_{1_j}, $j = 1, \ldots, n_1$ 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 $[q_0, q_1], (q_1, q_2], \ldots, (q_{r-1}, q_r]$, where $q_0 = L, q_r = H$. Set

$$n_{0_j} = \sum_{i=1}^{n_0} I\left(s_{0_i} \in (q_{j-1}, q_j]\right)$$
$$n_{1_j} = \sum_{i=1}^{n_1} I\left(s_{1_i} \in (q_{j-1}, q_j]\right), \quad j = 1, \ldots, r$$

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 Information value index becomes infinite in cases when some of n_{0_j} or n_{1_j} 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_{0_j} or n_{1_j} .
- 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

$$q_0 = L - 1$$
$$q_i = \widehat{F_0^{-1}}\left(\frac{k \cdot i}{n_0}\right), i = 1, \ldots, \lfloor \frac{n_0}{k} \rfloor$$
$$q_{\lfloor \frac{n_0}{k} \rfloor + 1} = H,$$

where $\widehat{F_0^{-1}}(\cdot)$ is the empirical quantile function appropriate to the empirical cumulative distribution function of scores of bad clients.

Empirical estimate with supervised interval selection (ESIS)

- Usage of quantile function of scores of bad clients is motivated by the assumption, that number of bad clients is less than number of good clients.

- If n_0 is not divisible by k, it is necessary to adjust our intervals, because we obtain number of scores of bad clients in the last interval, which is less than k. In this case, we have to merge the last two intervals.

- Furthermore we need to ensure, that the number of scores of good clients is as required in each interval. To do so, we compute n_{1_j} for all actual intervals. If we obtain $n_{1_j} < k$ for j^{th} interval, we merge this interval with its neighbor on the right side.

- This can be done for all intervals except the last one. If we have $n_{1_j} < k$ for the last interval, than we have to merge it with its neighbor on the left side, i.e. we merge the last two intervals.

Empirical estimate with supervised interval selection

- 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-divergence is given by

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

where n_{0_j} and n_{1_j} correspond to observed counts of good and bad clients in intervals created according to the described procedure.

ESIS.1

- Algorithm for the modified ESIS:

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

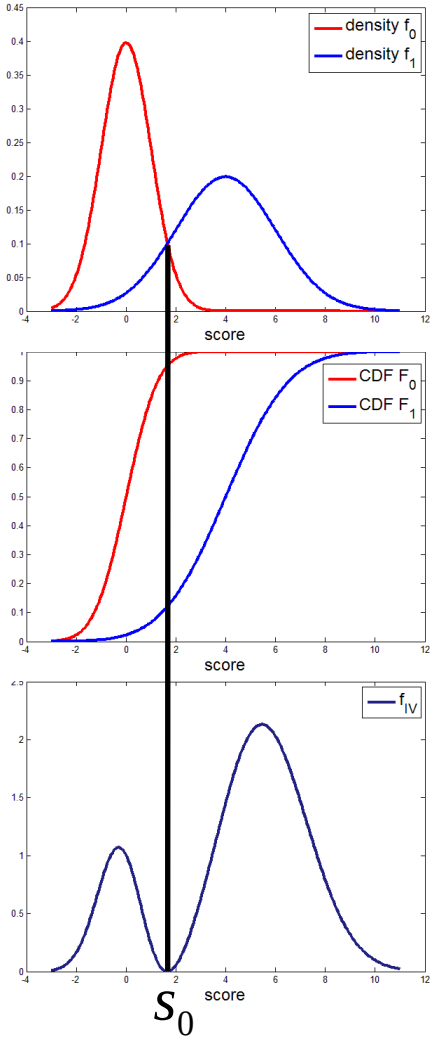
where $k = \lceil \sqrt{n_0} \rceil$

1) $\mathbf{q} = []$
2) $q_{j1} = F_1^{-1}\left(\frac{k}{n_1}\right) \qquad q_{j0} = F_0^{-1}\left(\frac{k}{n_0}\right)$
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) $\mathbf{q} = [\min(score) - 1, \mathbf{q}]$

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

ESIS.2

- The original ESIS often merge calculated intervals in the second phase of the algorithm.
- Only $F_0^{-1}(.)$ is used for computation.
- But it is clear that in order to meet the condition $n_{1_1} > k$, the border of the first interval has to be greater or equal to $F_1^{-1}\left(\frac{k}{n_1}\right)$.
- This directly leads to idea to use F_1 firstly, and then, from some value of the score, to use F_0.
- A suitable value of the score for this purpose would be the value of s_0, in which intersect the density functions of the scores, difference of distribution functions of the scores takes its maximum value and also the function F_{IV} becomes zero.



Point of intersection of densities

=

Point of maximal difference of CDFs

=

Point of zero value of f_{IV}

ESIS.2

- Algorithm for the modified ESIS:

1) $s_0 = mean\left(\arg\max_s \left|\hat{F}_1(s) - \hat{F}_0(s)\right|\right)$

it is necessary to solve the case when $\left|\hat{F}_1(s) - \hat{F}_0(s)\right|$ takes its maximal value for more than one value s.

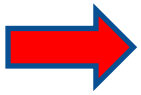
2) $q_{1_j} = F_1^{-1}\left(\frac{j \cdot k}{n_1}\right),\ j = 1, \ldots, \left\lfloor \frac{n_1}{k} \cdot F_1(s_0) \right\rfloor$

3) $q_{0_j} = F_0^{-1}\left(\frac{j \cdot k}{n_0}\right),\ j = \left\lceil \frac{n_0}{k} \cdot F_0(s_0) \right\rceil, \ldots, \left\lfloor \frac{n_0}{k} \right\rfloor - 1$

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.



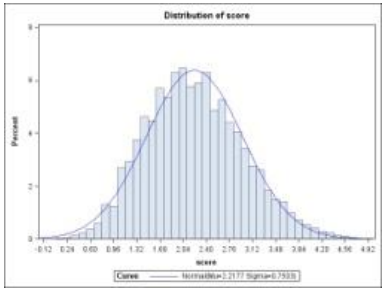
$\hat{D}_{J,ESIS2}$

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

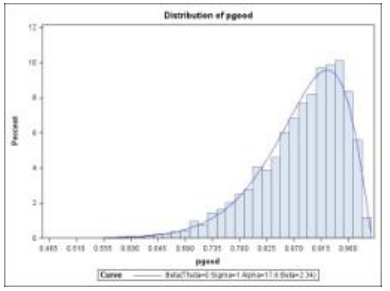
Simulation settings

- Consider n clients, $100p_B$% of bad and $100(1-p_B)$% of good clients with 1) normally, 2) beta and 3) gamma distributed scores:

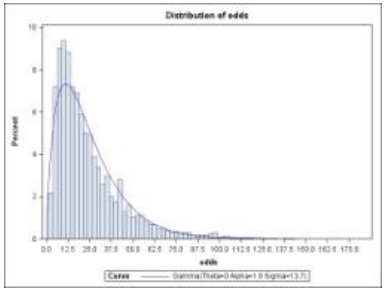
1) $$f_0(x) = \frac{1}{\sqrt{2\pi}\sigma_0}\exp\left(\frac{-(x-\mu_0)^2}{2\sigma_0^2}\right) \qquad f_1(x) = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right)$$



2) $$f_0(x) = \frac{1}{B(\alpha_0,\beta_0)}x^{\alpha_0-1}(1-x)^{\beta_0-1} \qquad f_1(x) = \frac{1}{B(\alpha_1,\beta_1)}x^{\alpha_1-1}(1-x)^{\beta_1-1}$$



3) $$f_0(x) = \frac{\lambda_0^{\alpha_0}}{\Gamma(\alpha_0)}x^{\alpha_0-1}\exp(-\lambda_0 x) \qquad f_1(x) = \frac{\lambda_1^{\alpha_1}}{\Gamma(\alpha_1)}x^{\alpha_1-1}\exp(-\lambda_1 x)$$



Gamma function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ Beta function $B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$

Simulation settings

- The exact value is then

1) $$D_J = (A+1)D^2 + A - 1, \text{ where } A = \frac{1}{2}\left(\frac{\sigma_0^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_0^2}\right), D = \frac{\mu_1 - \mu_0}{\sqrt{\sigma_0^2 + \sigma_1^2}}$$

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

3) $$D_J = (\alpha_1 - \alpha_0)\left(\psi(\alpha_1) - \psi(\alpha_0) + \ln\left(\frac{\lambda_0}{\lambda_1}\right)\right) + \alpha_0\left(\frac{\lambda_1}{\lambda_0} - 1\right) + \alpha_1\left(\frac{\lambda_0}{\lambda_1} - 1\right)$$

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

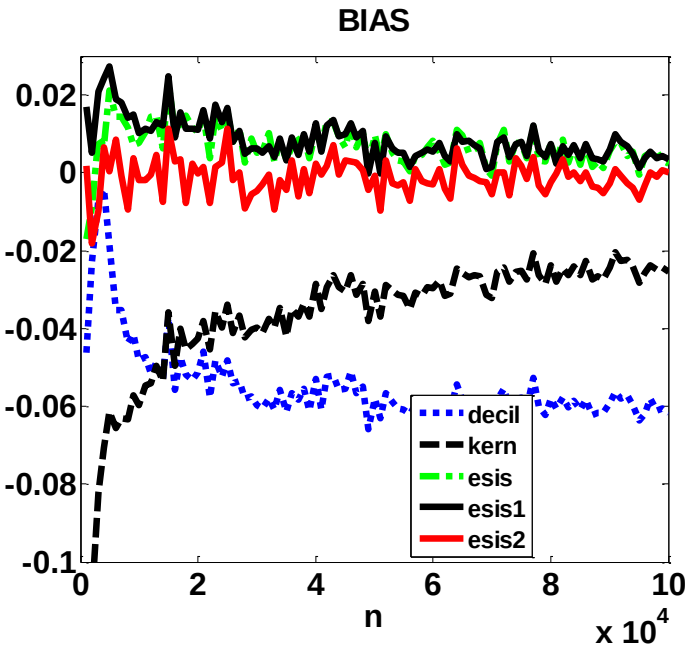
Simulation settings

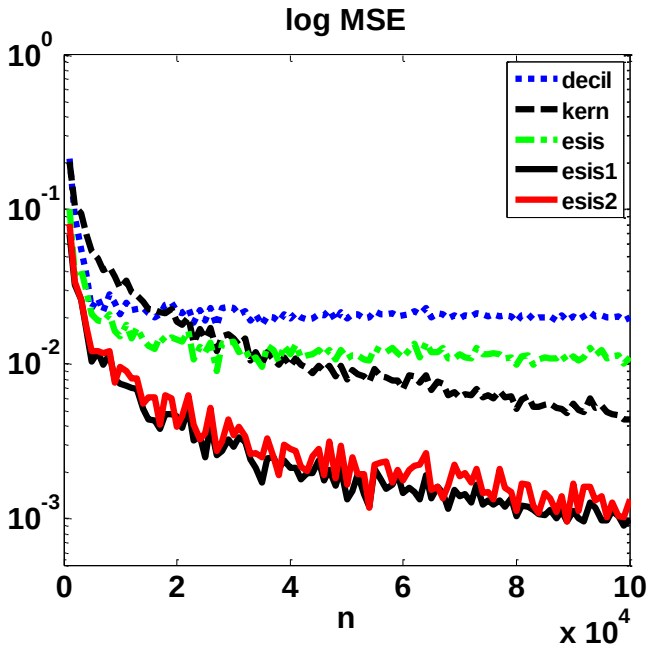
- Consider following values of parameters:
  - n = 1000 to 100 000
    - **Very small to large data set.**
  - $(\mu_0, \mu_1, \sigma_0, \sigma_1)$, $(\alpha_0, \alpha_1, \beta_0, \beta_1)$, $(\alpha_0, \alpha_1, \lambda_0, \lambda_1)$ that lead to $D_J = 0.25, 1, 2.25$
    - **Weak to very high performance 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.
  - $$Bias = E\left(\hat{I}_{val} - I_{val}\right)$$
  - $$\log MSE = \log\left(E(\hat{I}_{val} - I_{val})^2\right)$$

Simulation results

- Bias and MSE for normally distributed scores:
  - p_B=0.1, D_J=1 for bias, p_B=0.2, D_J=2.25 for MSE

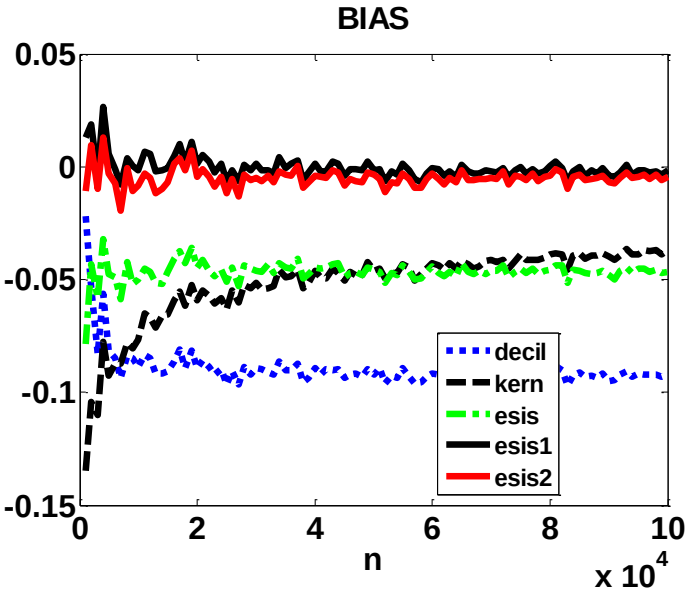


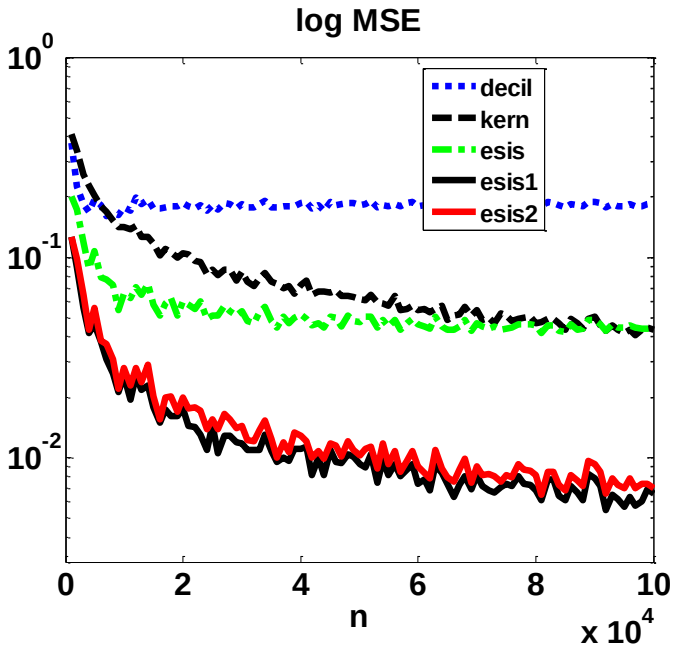


- Estimator using deciles is strongly biased – D_J was underestimated – for majority of considered values of parameters.
- Overall, the proposed algorithm ESIS2 had the best performance, followed by ESIS1 and ESIS estimator.

Simulation results

- Bias and MSE for beta distributed scores:
  - p_B=0.2, D_J=1 for bias, p_B=0.1, D_J=2.25 for MSE

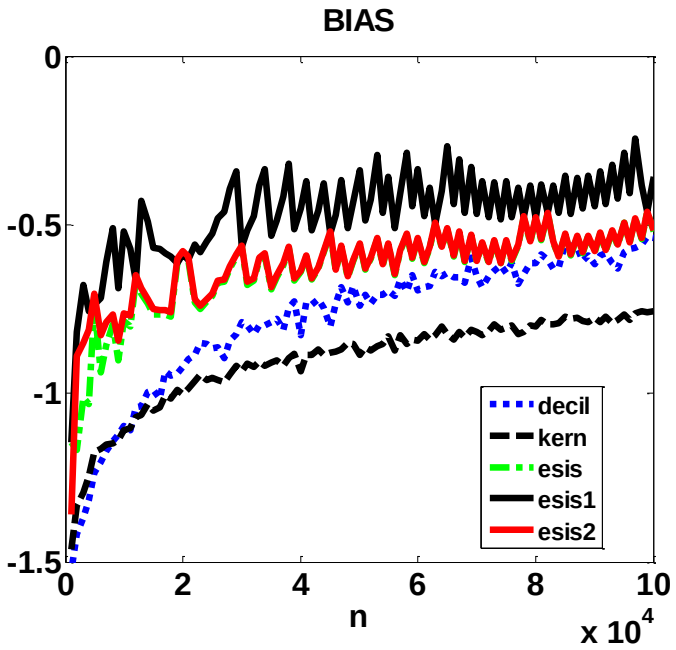


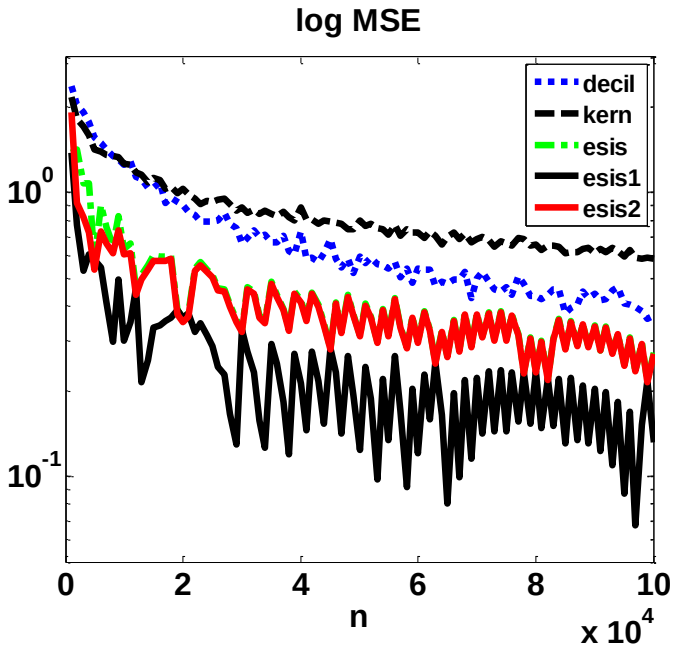


- Estimator using deciles is even worse than for normally distributed scores.
- The proposed algorithms ESIS1 and ESIS2 had the best performance again.

Simulation results

- Bias and MSE for gamma distributed scores:
  - p_B=0.02, D_J=2.25 for bias, p_B=0.02, D_J=2.25 for MSE

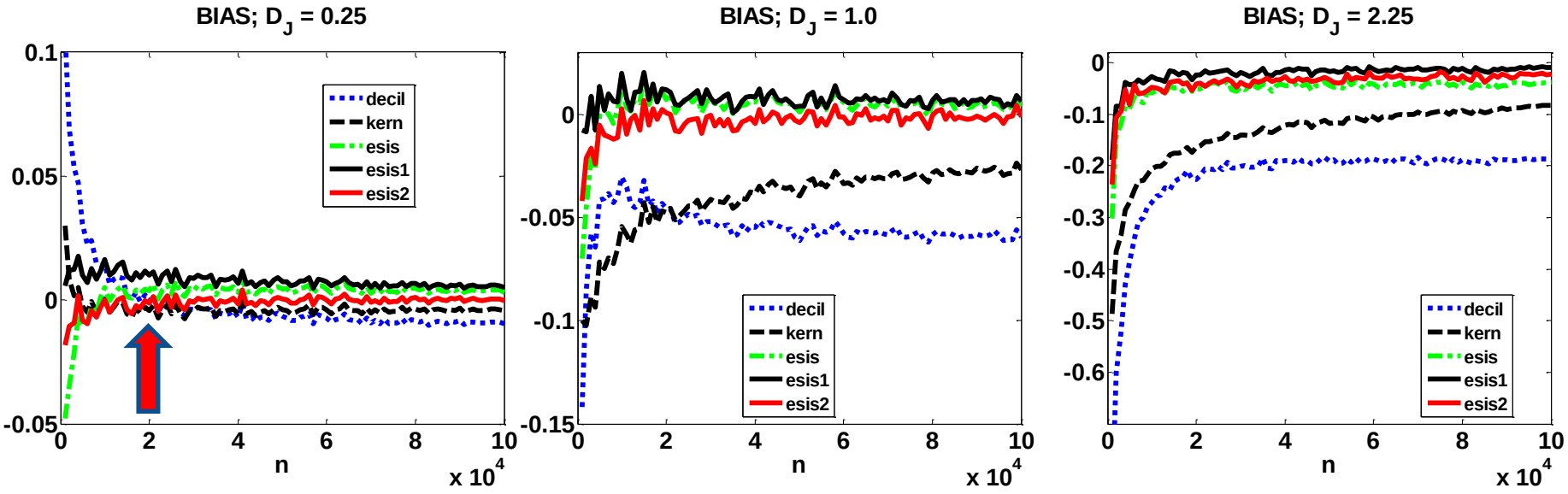




- ESIS and ESIS2 had very similar performance considering both bias and MSE. However, all estimators were biased in case of gamma distributed scores – D_J was underestimated for majority of considered values of parameters.
- Overall, considering bias and MSE, ESIS1 had the best performance.

Simulation results

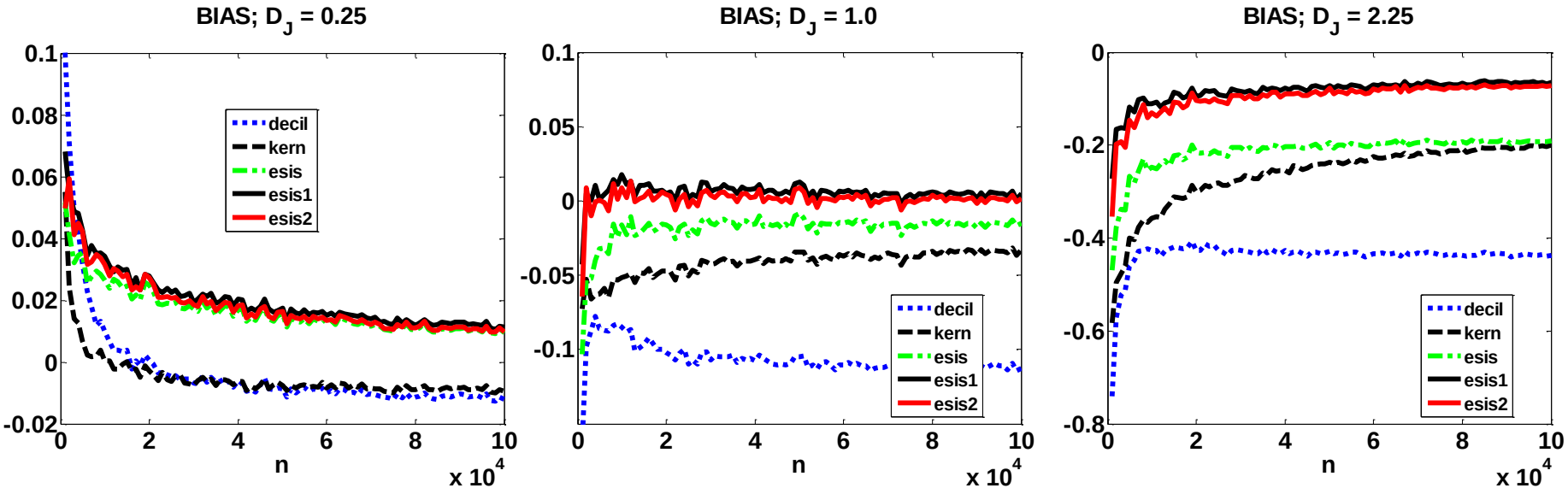
- Bias for normally distributed scores (according to D_J):



- Jde o zprůměrované hodnoty přes uvažované hodnoty parametru p_B.
- Nejstabilnější (z pohledu vychýlení) je ESIS2.
- Zajímavé je chování estimátoru „DECIL" pro slabé modely/proměnné (D_J=0.25) – pro malé počty pozorování je odhad nadhodnocený, pro cca 20 000 pozorování je odhad nevychýlený, pro vyšší počty pozorování je odhad podhodnocený. Značně podhodnocený je pro silnější modely/proměnné, a to bez ohledu na počet pozorování.

Simulation results

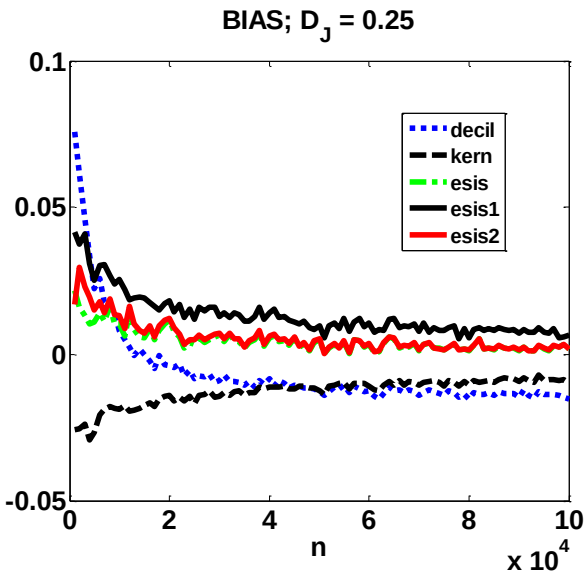
- Bias for beta distributed scores (according to D_J):

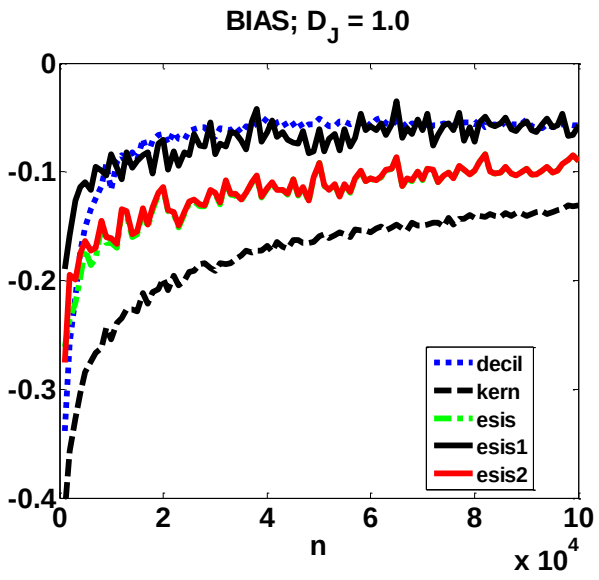


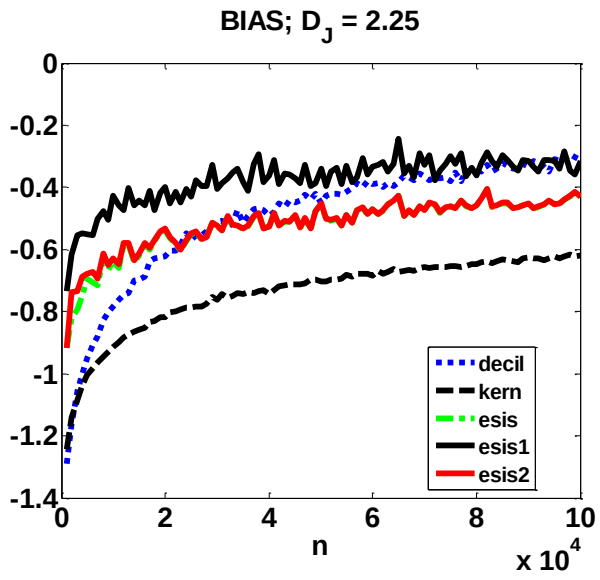
- Vlastnosti odhadů jsou velmi podobné jako předchozím případě. Pro D_J=1 a D_J=2.25 jsou nejlepší ESIS1 a ESIS2.
- Největší rozdíl je u D_J=0.25. Všechny tři algoritmy ESIS, ESIS1 i ESIS2 dávají nadhodnocený odhad D_J. Pro jádrový odhad a odhad decilový platí to, co pro decilový odhad v případě normálních dat.

Simulation results

- Bias for gamma distributed scores (according to D_J):







- ESIS a ESIS2 se chovají téměř identicky – pro D_J=0.25 velmi dobře, pro vyšší hodnoty D_J nic moc. Zde se jako nejlepší jeví ESIS1, popř. decilový odhad.
- Pro D_J=0.25 a decilový odhad platí opět totéž jako v předchozích případech.
- Celkově je zřejmé, že algoritmy ESIS se na datech s gamma rozložením chovají hůř než v případě normálních nebo beta rozložených dat.

Conclusions

- The classical way of computation of the J-divergence (Information value), i.e. empirical estimator using deciles of scores, is easy to implement, but may lead to strongly biased results. We conclude that kernel estimator and empirical estimators with supervised interval selection (ESIS) are much more appropriate to use. In total, the new algorithms ESIS1 and ESIS2 outperformed all other considered estimators.

- Consequently, ESIS1 and ESIS2 seem to be the right choice to estimate the J-divergence (Information value) when assessing discriminatory power of scoring models. Moreover the Information value is very often used to assess the discriminatory power of variables that enter into these models. This means that ESIS1/ESIS2 may lead to more appropriate, compared to the empirical estimator using deciles, filter for variable selection.

Thank you for your attention