MASARYK UNIVERSITY
Czech Republic

Jedovnice 2011

Computation of Information Value for Credit Scoring Models

Martin Řezáč, Jan Koláček

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

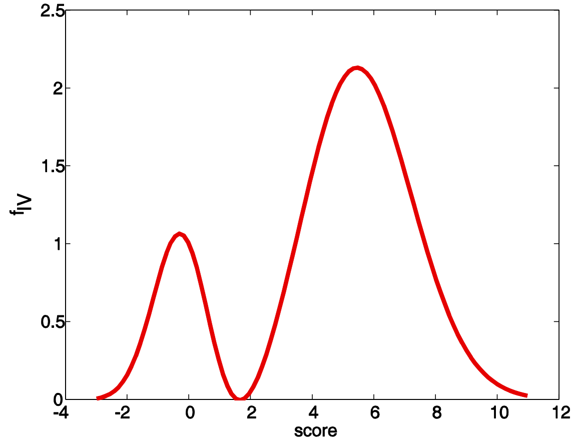
Information value

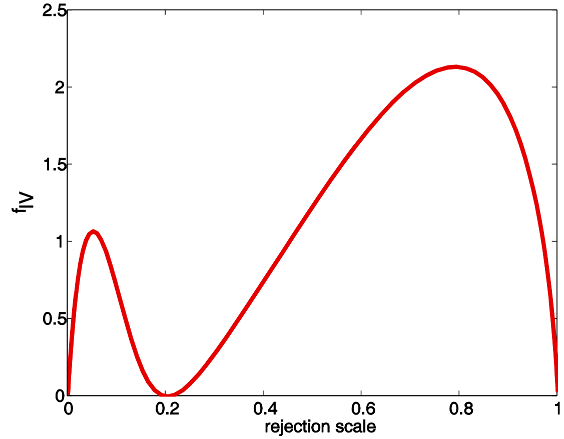
- The special case of Kullback-Leibler divergence given by:

$$I_{val} = \int_{-\infty}^{\infty} f_{IV}(x)dx \quad \text{where } f_{IV}(x) = (f_1(x) - f_0(x)) \ln\left(\frac{f_1(x)}{f_0(x)}\right)$$

  - f_0, f_1 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 $K_{h_i}(x) = \frac{1}{h_i} K\left(\frac{x}{h_i}\right), i = 0{,}1$ and K is the Epanechnikov kernel.

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

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

$$\Rightarrow \quad \widehat{I}_{val,KERN} = \frac{H - L}{2M}\left(\tilde{f}_{IV}(L) + 2\sum_{i=1}^{M-1} \tilde{f}_{IV}(x_i) + \tilde{f}_{IV}(H)\right).$$

Empirical estimate of I_{val}

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. Then the empirical Information value is calculated by

$$\widehat{I}_{val,DEC} = \sum_{j=1}^{r} \left(\frac{n_{1_j}}{n_1} - \frac{n_{0_j}}{n_0}\right) \ln\left(\frac{n_{1_j} n_0}{n_{0_j} n_1}\right).$$

Empirical estimate of I_{val}

- 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 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 client in each interval. This is the basic idea of all proposed algorithms.

Empirical estimate with supervised interval selection

- „the first” ESIS:

Set

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

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

Empirical estimate with supervised interval selection

- 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 Information value, 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 Information value is given by

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

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.

Simulation results

- Consider n clients, $100p_B$% of bad clients with $f_0 : N(\mu_0, \sigma_0)$ and $100(1-p_B)$% of good clients with $f_1 : N(\mu_1, \sigma_1)$.
- Because of normality we know $I_{val} = \left(\frac{\mu_1 - \mu_0}{\sigma} \right)^2$.
- Consider following values of parameters:
  - n = 100 000 , n = 1000
  - $\mu_0 = 0$
  - $\sigma_0 = \sigma_1 = 1$
  - $\mu_1 = 0.5, 1, 1.5$
  - $p_B = 0.02, 0.05, 0.1, 0.2$

Simulation results

1) Scores of bad and good clients were generated according to given parameters.

2) Estimates $\hat{I}_{val,DEC}$, $\hat{I}_{val,KERN}$, $\hat{I}_{val,ESIS}$ were computed.

3) Square errors were computed.

4) Steps 1)-3) were repeated one thousand times.

5) MSE was computed.

Simulation results

n=100000, $\mu_1 - \mu_0$ = 0.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,000546	0,000310	0,000224	0,000168
IV_kern	0,000487	0,000232	0,000131	0,000076
IV_esis	0,000910	0,000384	0,000218	0,000127

n=1000, $\mu_1 - \mu_0$ = 0.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,025574	0,040061	0,026536	0,009074
IV_kern	0,038634	0,017547	0,009281	0,004737
IV_esis	0,038331	0,021980	0,016280	0,008028

- worst
- average
- best performance

n=100000, $\mu_1 - \mu_0$ = 1.0				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,006286	0,004909	0,004096	0,002832
IV_kern	0,003396	0,001697	0,001064	0,000646
IV_esis	0,002146	0,000973	0,000477	0,000568

n=1000, $\mu_1 - \mu_0$ = 1.0				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,186663	0,084572	0,043097	0,029788
IV_kern	0,117382	0,072381	0,045344	0,032131
IV_esis	0,150881	0,071088	0,036503	0,023609

n=100000, $\mu_1 - \mu_0$ = 1.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,056577	0,048415	0,034814	0,020166
IV_kern	0,019561	0,010789	0,006796	0,004862
IV_esis	0,013045	0,008134	0,007565	0,027943

n=1000, $\mu_1 - \mu_0$ = 1.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	1,663859	1,037778	0,535180	0,200792
IV_kern	0,529367	0,349783	0,266912	0,196856
IV_esis	0,609193	0,352151	0,172931	0,194676

Adjusted empirical estimate with supervised interval selection (AESIS)

- It is obvious that the choice of parameter *k* is crucial.
- So the question is:
  - Is the choice $k = \lceil \sqrt{n_0} \rceil$ optimal (according to MSE)?
  - What effect has n_0 on the optimal *k*?
  - And what effect, if any, has the difference of means $\mu_1 - \mu_0$?

Simulation results

- Consider 10000 clients, 100p_B% of bad clients with $f_0 : N(\mu_0, 1)$ and 100(1-p_B)% of good clients with $f_1 : N(\mu_1, 1)$. Set $\mu_0 = 0$ and consider $\mu_1 = 0.5, 1$ and 1.5, $p_B = 0.02, 0.05, 0.1$ and 0.2.

$$MSE = E((\hat{I}_{val} - I_{val})^2)$$ → $$k_{MSE} = \underset{k}{argmin}\, MSE.$$

k_{MSE}		p_B 0.02	0.05	0.1	0.2
$\|\mu_1 - \mu_0\|$	0.5	29	42	62	84
	1	12	18	23	32
	1.5	6	9	8	9
$k = \lceil \sqrt{n_0} \rceil$		15	23	32	45

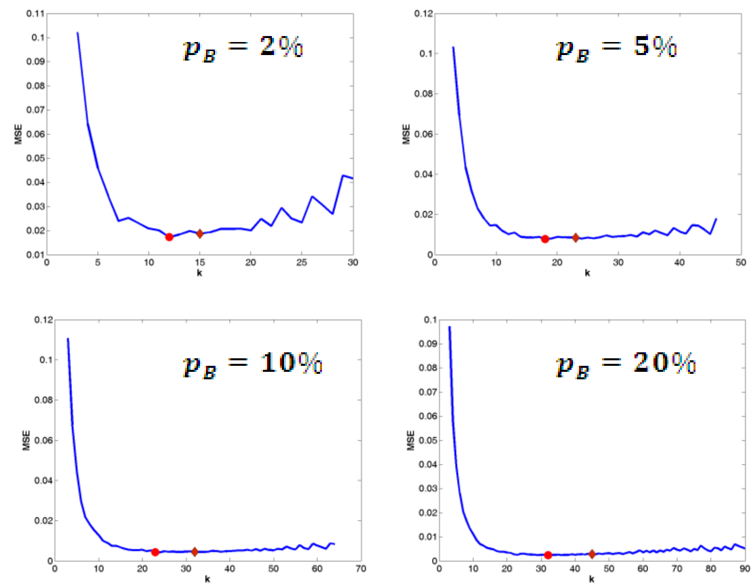
→

avg. # of bins		p_B 0.02	0.05	0.1	0.2
$\|\mu_1 - \mu_0\|$	0.5	8,00	13,00	18,00	24,90
	1	18,00	28,80	42,76	51,88
	1.5	33,62	50,20	95,96	127,67

Simulation results

- **Dependence of MSE on *k*,** $\mu_1 - \mu_0 = 1$.

   - The highlighted circles correspond to values of *k*, where minimal value of the *MSE* is obtained. The diamonds correspond to values of *k* given by $k = \lceil \sqrt{n_0} \rceil$.



$$k = \left\lceil \frac{\frac{2}{3}\sqrt{p_B \cdot n} + 2}{|\widehat{\mu_1} - \widehat{\mu_0}|^{1.4}} \right\rceil$$

$k = \left\lceil \frac{\frac{2}{3}\sqrt{p_B \cdot n}+2}{\lvert\widehat{\mu_1}-\widehat{\mu_0}\rvert^{1.4}} \right\rceil$		p_B			
		0.02	0.05	0.1	0.2
$\mu_1 - \mu_0$	0.5	31	45	61	84
	1	12	17	24	32
	1.5	7	10	14	19

k_{MSE}		p_B			
		0.02	0.05	0.1	0.2
$\mu_1 - \mu_0$	0.5	29	42	62	84
	1	12	18	23	32
	1.5	6	9	8	9

where $\widehat{\mu_1}$ and $\widehat{\mu_0}$ are suitable estimates of means of scores of good and bad clients, $p_B = \frac{n_0}{n}$ is the proportion of bad clients.

$$\hat{I}_{val,AESIS}$$

Simulation results

n=1000, $\mu_1 - \mu_0$ = 0.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,025574	0,040061	0,026536	0,009074
IV_kern	0,038634	0,017547	0,009281	0,004737
IV_esis	0,038331	0,021980	0,016280	0,008028
IV_aesis	0,042409	0,030808	0,015558	0,007223

n=1000, $\mu_1 - \mu_0$ = 1.0				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,186663	0,084572	0,043097	0,029788
IV_kern	0,117382	0,072381	0,045344	0,032131
IV_esis	0,150881	0,071088	0,036503	0,023609
IV_aesis	0,256181	0,093932	0,043860	0,027467

n=1000, $\mu_1 - \mu_0$ = 1.5				
MSE				
	0.02	0.05	0.1	0.2
IV_decil	1,663859	1,037778	0,535180	0,200792
IV_kern	0,529367	0,349783	0,266912	0,196856
IV_esis	0,609193	0,352151	0,172931	0,194676
IV_aesis	0,553650	0,205889	0,135187	0,089354

- The classical estimate $\widehat{I}_{val,DEC}$ had the lowest MSE in case of weak model and extremely low number (namely 20) of bad clients. If the number of bad client is slightly higher, $\widehat{I}_{val,KERN}$ had the best performance. Considering models with higher predictive power, $\widehat{I}_{val,ESIS}$ and $\widehat{I}_{val,AESIS}$ had the best performance.

Simulation results

n=100000, $\mu_1 - \mu_0$ = 0.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,000546	0,000310	0,000224	0,000168
IV_kern	0,000487	0,000232	0,000131	0,000076
IV_esis	0,000910	0,000384	0,000218	0,000127
IV_aesis	0,000603	0,000253	0,000135	0,000079

n=100000, $\mu_1 - \mu_0$ = 1.0				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,006286	0,004909	0,004096	0,002832
IV_kern	0,003396	0,001697	0,001064	0,000646
IV_esis	0,002146	0,000973	0,000477	0,000568
IV_aesis	0,002446	0,001157	0,000552	0,000311

n=100000, $\mu_1 - \mu_0$ = 1.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,056577	0,048415	0,034814	0,020166
IV_kern	0,019561	0,010789	0,006796	0,004862
IV_esis	0,013045	0,008134	0,007565	0,027943
IV_aesis	0,006140	0,002688	0,002502	0,011472

- We can see that the classical estimate $\widehat{I}_{val,DEC}$ was out-performed by all other estimates of the Information value in all considered cases when the number of clients was high (100000 in our case). The best performance was achieved by $\widehat{I}_{val,KERN}$ in case of weak scoring models, by $\widehat{I}_{val,ESIS}$ in case of models with high performance and by $\widehat{I}_{val,AESIS}$ for models with very high performance.

ESIS.1

➢ Algorithm for the modified ESIS:

$\hat{I}_{val,ESIS1}$

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

AESIS.1 – Simulation results

- Consider 1000 and 10000 clients, 100p$_B$% of bad clients with $f_0 : N(\mu_0, 1)$ and 100(1-p$_B$)% of good clients with $f_1 : N(\mu_1, 1)$. Set $\mu_0 = 0$ and consider $\mu_1 = 0.5, 1 \text{ and } 1.5$, $p_B = 0.02, 0.05, 0.1 \text{ and } 0.2$.

$$MSE = E((\hat{I}_{val} - I_{val})^2) \Rightarrow k_{MSE} = \underset{k}{argmin}\, MSE.$$

$n = 1000$

k_{MSE}		p_B			
		0.02	0.05	0.1	0.2
$\mu_1 - \mu_0$	0.5	5	9	10	22
	1	2	3	4	6
	1.5	1	2	3	2
$k = \lceil \sqrt{n_0} \rceil$		5	8	10	15

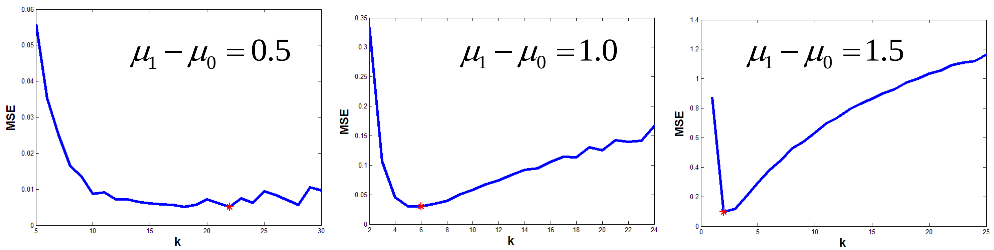
$n = 10000$

k_{MSE}		p_B			
		0.02	0.05	0.1	0.2
$\mu_1 - \mu_0$	0.5	12	18	32	51
	1	4	7	10	13
	1.5	2	3	4	5
$k = \lceil \sqrt{n_0} \rceil$		15	23	32	45

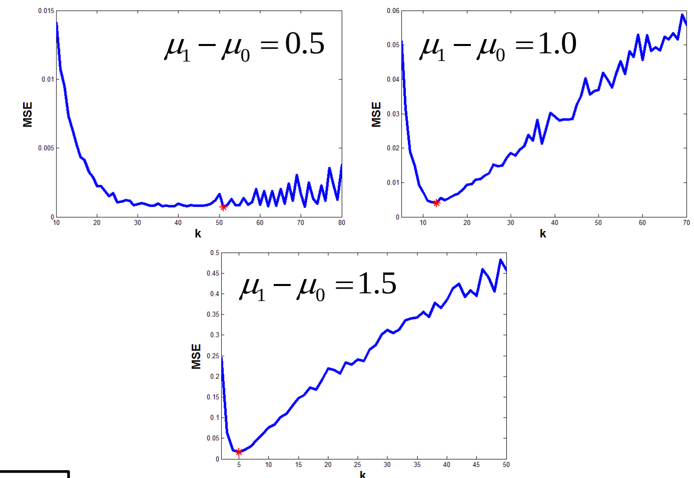
Simulation results

Dependence of MSE on *k*.

$n = 1000,\ p_B = 0.2$



$n = 10000,\ p_B = 0.2$



$$k = \left\lceil \frac{\frac{1}{3}\sqrt{p_B \cdot n}}{|\hat{\mu}_1 - \hat{\mu}_0|^{1.4}} \right\rceil$$

$n = 1000$

$k = \left\lceil \frac{\frac{1}{3}\sqrt{p_B \cdot n}}{\lvert\hat{\mu}_1 - \hat{\mu}_0\rvert^{1.4}} \right\rceil$		p_B			
		0.02	0.05	0.1	0.2
$\mu_1 - \mu_0$	0.5	4	7	9	13
	1	2	3	4	5
	1.5	1	2	2	3

k_{MSE}		p_B			
		0.02	0.05	0.1	0.2
$\mu_1 - \mu_0$	0.5	5	9	10	22
	1	2	3	4	6
	1.5	1	2	2	2

$n = 10000$

$k = \left\lceil \frac{\frac{1}{3}\sqrt{p_B \cdot n}}{\lvert\hat{\mu}_1 - \hat{\mu}_0\rvert^{1.4}} \right\rceil$		p_B			
		0.02	0.05	0.1	0.2
$\mu_1 - \mu_0$	0.5	13	20	28	39
	1	5	8	11	15
	1.5	3	5	6	9

k_{MSE}		p_B			
		0.02	0.05	0.1	0.2
$\mu_1 - \mu_0$	0.5	12	18	32	51
	1	4	7	10	13
	1.5	2	3	4	5

$$\hat{I}_{val,AESIS1}$$

Simulation results

n=100000, $\mu_1 - \mu_0$ = 0.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,000546	0,000310	0,000224	0,000168
IV_kern	0,000487	0,000232	0,000131	0,000076
IV_esis	0,000910	0,000384	0,000218	0,000127
IV_aesis	0,000603	0,000253	0,000135	0,000079
IV_esis1	0,000550	0,000273	0,000480	0,000179
IV_aesis1	0,000938	0,000381	0,000203	0,000123

n=100000, $\mu_1 - \mu_0$ = 1.0				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,006286	0,004909	0,004096	0,002832
IV_kern	0,003396	0,001697	0,001064	0,000646
IV_esis	0,002146	0,000973	0,000477	0,000568
IV_aesis	0,002446	0,001157	0,000552	0,000311
IV_esis1	0,031965	0,014877	0,017427	0,011134
IV_aesis1	0,009763	0,004326	0,002494	0,001534

n=100000, $\mu_1 - \mu_0$ = 1.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,056577	0,048415	0,034814	0,020166
IV_kern	0,019561	0,010789	0,006796	0,004862
IV_esis	0,013045	0,008134	0,007565	0,027943
IV_aesis	0,006140	0,002688	0,002502	0,011472
IV_esis1	0,435297	0,296417	0,219711	0,169501
IV_aesis1	0,069952	0,043534	0,032264	0,023526

n=1000, $\mu_1 - \mu_0$ = 0.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,025574	0,040061	0,026536	0,009074
IV_kern	0,038634	0,017547	0,009281	0,004737
IV_esis	0,038331	0,021980	0,016280	0,008028
IV_aesis	0,042409	0,030808	0,015558	0,007223
IV_esis1	0,021719	0,015727	0,008051	0,006886
IV_aesis1	0,060838	0,027738	0,018239	0,010903

n=1000, $\mu_1 - \mu_0$ = 1.0				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,186663	0,084572	0,043097	0,029788
IV_kern	0,117382	0,072381	0,045344	0,032131
IV_esis	0,150881	0,071088	0,036503	0,023609
IV_aesis	0,256181	0,093932	0,043860	0,027467
IV_esis1	0,289062	0,144170	0,159419	0,098609
IV_aesis1	0,260596	0,129346	0,071732	0,036574

n=1000, $\mu_1 - \mu_0$ = 1.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	1,663859	1,037778	0,535180	0,200792
IV_kern	0,529367	0,349783	0,266912	0,196856
IV_esis	0,609193	0,352151	0,172931	0,194676
IV_aesis	0,553650	0,205889	0,135187	0,089354
IV_esis1	1,510276	1,058317	0,922429	0,860921
IV_aesis1	0,613465	0,293109	0,211057	0,137516

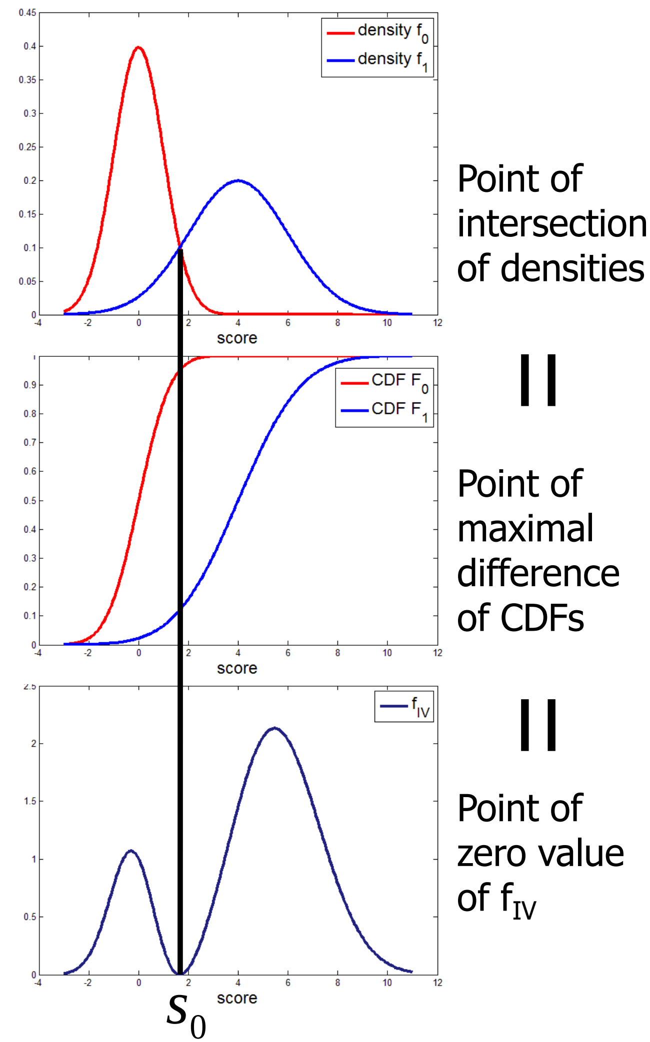
Simulation results

- The new algorithm (ESIS.1) ended disastrously in most cases. The only exception was a situation where n=1000 a μ_1-μ_0=0.5. This corresponds to a scoring model with very poor discriminatory power and a very small number of observed bad clients (20-200).

- AESIS.1 ranked among the average.

ESIS.2

- U původního ESIS často dochází ke slučování vypočtených intervalů ve druhé fázi algoritmu.
- Pro výpočet se používá jen $F_0^{-1}(.)$.
- Aby byla splněna podmínka $n_{11}>k$, je zřejmě nutné, aby hranice prvního intervalu byla větší než $F_1^{-1}\left(\frac{k}{n_1}\right)$.
- To vede k myšlence použít ke konstrukci intervalů nejprve $F_1^{-1}(.)$ a následně, od nějaké hodnoty skóre $F_0^{-1}(.)$.
- Jako vhodná hodnota skóre pro tento účel se jeví hodnota s_0, ve které se protínají hustoty skóre, rozdíl distribučních funkcí skóre nabývá své maximální hodnoty a také platí, že funkce f_{IV} nabývá nulové hodnoty.



ESIS.2

- Algorithm for the modified ESIS:

1) $s_0 = \arg\max_s |F_1 - F_0|$

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) $\mathbf{q} = [\min(\,score\,) - 1, \mathbf{q}_1, \mathbf{q}_0, \max(\,score\,) + 1]$

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*.

$\Rightarrow$ $\hat{I}_{val,ESIS2}$

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

AESIS.2 – Simulation results

- Consider 1000, 10000 and 100000 clients, $100p_B$% of bad clients with $f_0 : N(\mu_0, 1)$ and $100(1-p_B)$% of good clients with $f_1 : N(\mu_1, 1)$. Set $\mu_0 = 0$, $\mu_1 = 0.5, 1$ and 1.5 and consider $p_B = 0.02, 0.05, 0.1$ and 0.2.

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

→

$$k_{MSE} = \underset{k}{argmin}\, MSE.$$

$n = 1000$

k_{MSE}		p_B 0.02	0.05	0.1	0.2
$\mu_1 - \mu_0$	0.5	15	19	22	45
	1	3	8	11	16
	1.5	2	3	6	7
$k = \lceil \sqrt{n_0} \rceil$		5	8	10	15

$n = 10000$

k_{MSE}		p_B 0.02	0.05	0.1	0.2
$\mu_1 - \mu_0$	0.5	29	51	77	112
	1	15	24	28	45
	1.5	6	11	11	14
$k = \lceil \sqrt{n_0} \rceil$		15	23	32	45

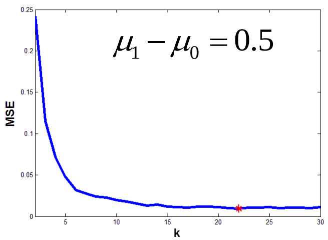
$n = 100000$

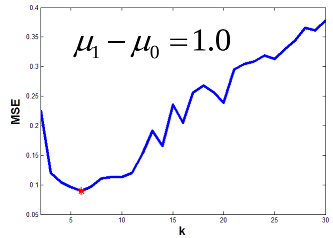
k_{MSE}		p_B 0.02	0.05	0.1	0.2
$\mu_1 - \mu_0$	0.5	118	198	298	371
	1	50	61	106	141
	1.5	17	28	32	48
$k = \lceil \sqrt{n_0} \rceil$		5	8	10	15

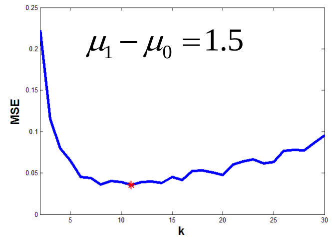
Simulation results

- **Dependence of MSE on *k*.**

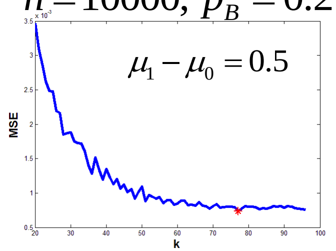
$n=1000,\ p_B=0.2$

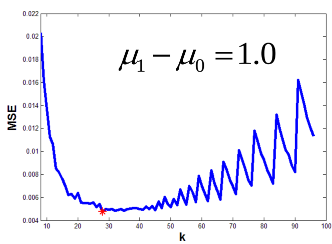


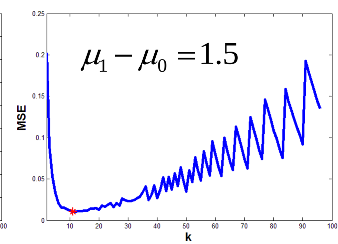




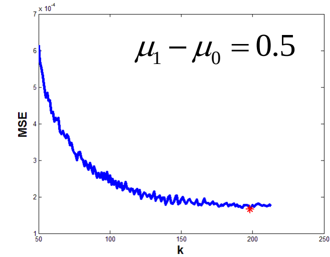
$n=10000,\ p_B=0.2$

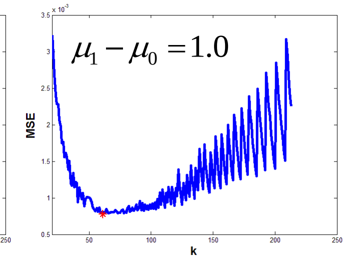


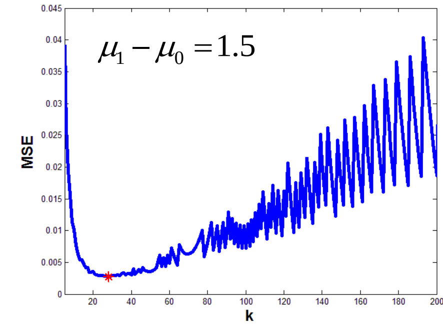




$n=100000,\ p_B=0.05$







$$k=\left\lceil \frac{\sqrt{p_B\cdot n}}{|\hat{\mu}_1-\hat{\mu}_0|^{\sqrt{2}}} \right\rceil$$

$$\hat{I}_{val,AESIS2}$$

$n=10000$

$k=\left\lceil \frac{\sqrt{p_B\cdot n}}{\lvert\hat{\mu}_1-\hat{\mu}_0\rvert^{\sqrt{2}}} \right\rceil$		p_B			
		0.02	0.05	0.1	0.2
$\mu_1-\mu_0$	0.5	38	60	85	120
	1	15	23	32	45
	1.5	8	13	18	26

k_{MSE}		p_B			
		0.02	0.05	0.1	0.2
$\mu_1-\mu_0$	0.5	29	51	77	112
	1	15	24	28	45
	1.5	6	11	11	14

Simulation results

n=1000, $\mu_1 - \mu_0$ = 0.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,025574	0,040061	0,026536	0,009074
IV_kern	0,038634	0,017547	0,009281	0,004737
IV_esis	0,038331	0,021980	0,016280	0,008028
IV_aesis	0,042409	0,030808	0,015558	0,007223
IV_esis1	0,021719	0,015727	0,008051	0,006886
IV_aesis1	0,060838	0,027738	0,018239	0,010903
IV_esis2	0,038112	0,025568	0,019098	0,009540
IV_esis2a	0,048697	0,027729	0,014114	0,007988
IV_esis2b	0,091599	0,043529	0,026044	0,014985
IV_aesis2	0,051170	0,026518	0,014131	0,007838

n=1000, $\mu_1 - \mu_0$ = 1.0				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,186663	0,084572	0,043097	0,029788
IV_kern	0,117382	0,072381	0,045344	0,032131
IV_esis	0,150881	0,071088	0,036503	0,023609
IV_aesis	0,256181	0,093932	0,043860	0,027467
IV_esis1	0,289062	0,144170	0,159419	0,098609
IV_aesis1	0,260596	0,129346	0,071732	0,036574
IV_esis2	0,171946	0,074200	0,041890	0,022861
IV_esis2a	0,213419	0,089678	0,048128	0,031476
IV_esis2b	0,357268	0,168690	0,097523	0,064290
IV_aesis2	0,209890	0,091614	0,050609	0,028864

- worst
- average
- best performance

	algoritmus	k
esis2	ESIS.2	$k = \lceil \sqrt{n_0} \rceil$
esis2a	ESIS.2	$k = \left\lceil \frac{\frac{2}{3}\sqrt{p_B \cdot n} + 2}{\lvert \widehat{\mu_1} - \widehat{\mu_0} \rvert^{1.4}} \right\rceil$
eis2b	ESIS.2	$k = \left\lceil \frac{\frac{1}{3}\sqrt{p_B \cdot n}}{\lvert \hat{\mu}_1 - \hat{\mu}_0 \rvert^{1.4}} \right\rceil$
aesis2	ESIS.2	$k = \left\lceil \frac{\sqrt{p_B \cdot n}}{\lvert \hat{\mu}_1 - \hat{\mu}_0 \rvert^{\sqrt{2}}} \right\rceil$

n=1000, $\mu_1 - \mu_0$ = 1.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	1,663859	1,037778	0,535180	0,200792
IV_kern	0,529367	0,349783	0,266912	0,196856
IV_esis	0,609193	0,352151	0,172931	0,194676
IV_aesis	0,553650	0,205889	0,135187	0,089354
IV_esis1	1,510276	1,058317	0,922429	0,860921
IV_aesis1	0,613465	0,293109	0,211057	0,137516
IV_esis2	0,838666	0,244379	0,133832	0,116534
IV_esis2a	0,577075	0,183188	0,128374	0,067026
IV_esis2b	0,737516	0,261110	0,202780	0,092417
IV_aesis2	0,575143	0,184840	0,125203	0,073825

MASARYK UNIVERSITY
Czech Republic

Simulation results

n=100000, $\mu_1 - \mu_0$ = 0.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,000546	0,000310	0,000224	0,000168
IV_kern	0,000487	0,000232	0,000131	0,000076
IV_esis	0,000910	0,000384	0,000218	0,000127
IV_aesis	0,000603	0,000253	0,000135	0,000079
IV_esis1	0,000550	0,000273	0,000480	0,000179
IV_aesis1	0,000938	0,000381	0,000203	0,000123
IV_esis2	0,000905	0,000353	0,000222	0,000113
IV_esis2a	0,000610	0,000231	0,000135	0,000072
IV_esis2b	0,001152	0,000419	0,000252	0,000139
IV_aesis2	0,000570	0,000211	0,000119	0,000064

n=100000, $\mu_1 - \mu_0$ = 1.0				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,006286	0,004909	0,004096	0,002832
IV_kern	0,003396	0,001697	0,001064	0,000646
IV_esis	0,002146	0,000973	0,000477	0,000568
IV_aesis	0,002446	0,001157	0,000552	0,000311
IV_esis1	0,031965	0,014877	0,017427	0,011134
IV_aesis1	0,009763	0,004326	0,002494	0,001534
IV_esis2	0,002158	0,000905	0,000484	0,000285
IV_esis2a	0,002578	0,001114	0,000582	0,000315
IV_esis2b	0,005844	0,002378	0,001366	0,000725
IV_aesis2	0,002317	0,000945	0,000497	0,000294

- worst
- average
- best performance

n=100000, $\mu_1 - \mu_0$ = 1.5				
MSE	p_B			
	0.02	0.05	0.1	0.2
IV_decil	0,056577	0,048415	0,034814	0,020166
IV_kern	0,019561	0,010789	0,006796	0,004862
IV_esis	0,013045	0,008134	0,007565	0,027943
IV_aesis	0,006140	0,002688	0,002502	0,011472
IV_esis1	0,435297	0,296417	0,219711	0,169501
IV_aesis1	0,069952	0,043534	0,032264	0,023526
IV_esis2	0,012158	0,006407	0,002796	0,003459
IV_esis2a	0,006033	0,002356	0,001235	0,000727
IV_esis2b	0,011927	0,004857	0,002937	0,001373
IV_aesis2	0,008045	0,003735	0,001760	0,001244

Simulation results

n=1000, $\mu_1 - \mu_0$ = 0.5				
rMSE	p_B			
	0.02	0.05	0.1	0.2
IV_kern	0,154537	0,070189	0,037124	0,018946
IV_esis1	0,086877	0,062907	0,032206	0,027545
IV_aesis2	0,204681	0,106072	0,056524	0,031350

n=100000, $\mu_1 - \mu_0$ = 0.5				
rMSE	p_B			
	0.02	0.05	0.1	0.2
IV_kern	0,001947	0,000928	0,000524	0,000306
IV_esis1	0,002199	0,001091	0,001919	0,000715
IV_aesis2	0,002280	0,000844	0,000475	0,000255

n=1000, $\mu_1 - \mu_0$ = 1.0				
rMSE	p_B			
	0.02	0.05	0.1	0.2
IV_kern	0,117382	0,072381	0,045344	0,032131
IV_esis	0,150881	0,071088	0,036503	0,023609
IV_esis2	0,171946	0,074200	0,041890	0,022861

n=100000 $\mu_1 - \mu_0$ = 1.0				
rMSE	p_B			
	0.02	0.05	0.1	0.2
IV_kern	0,003396	0,001697	0,001064	0,000646
IV_esis	0,002146	0,000973	0,000477	0,000568
IV_esis2	0,002158	0,000905	0,000484	0,000285

n=1000, $\mu_1 - \mu_0$ = 1.5				
rMSE	p_B			
	0.02	0.05	0.1	0.2
IV_kern	0,235274	0,155459	0,118628	0,087492
IV_esis2a	0,256478	0,081417	0,057055	0,029789
IV_aesis2	0,255619	0,082151	0,055646	0,032811

n=100000 $\mu_1 - \mu_0$ = 1.5				
rMSE	p_B			
	0.02	0.05	0.1	0.2
IV_kern	0,008694	0,004795	0,003020	0,002161
IV_esis2a	0,002682	0,001047	0,000549	0,000323
IV_aesis2	0,003576	0,001660	0,000782	0,000553

- Relativní MSE (rMSE)…jde o MSE z předchozích tabulek vydělené příslušnou teoretickou hodnotou IV.
- Umožní lepší porovnání – eliminuje vliv absolutní výše IV.

Conclusions

- The classical way of computation of the information value, i.e. empirical estimate using deciles of scores, may lead to strongly biased results.

- We conclude that kernel estimates and empirical estimates with supervised interval selection (ESIS, AESIS, ESIS.1, ESIS.2 and AESIS.2) are much more appropriate to use.

Děkuji za pozornost.