SAMPLE VARIANCE, INTERVAL DATA AND GENETIC ALGORITHMS

Jaromír Antoch

ROBUST'14
Jetřichovice, January 20, 2014

GOALS

- When we have interval values for the observations instead of exact sample values, what is the interval of possible values for the variance of these interval observations?
- What is the situation when calcultaion, based on the same type of the data, another statistical characteristics as, e.g., covariances, information, etc.
- What is the inpact on regression etc.

How tall is the tree?

Main task

Let us have n intervals $I_i = [a_i, b_i],\ a_i \leq b_i,\ i = 1, \dots, n$
let $\mathcal{K} = I_1 \otimes \dots \otimes I_n = [a_1, b_1] \otimes \dots \otimes [a_n, b_n] \subset \mathbb{R}^n$
Main tasks are:

P1 To find among all the vectors falling into $\mathcal{K}$ that one which has maximal variance, i.e. to find

$$x^{max} = \underset{x \in \mathcal{K}}{\arg\max}\ \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x}_n\right)^2,$$

P2 To find among all the vectors falling into $\mathcal{K}$ that one which has minimal variance, i.e. to find

$$x^{min} = \underset{x \in \mathcal{K}}{\arg\min}\ \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x}_n\right)^2.$$

Main task (cont.)

P3 To find among all the vectors falling into $\mathcal{K}$

$$\widetilde{x}^{max} = \arg\max_{x \in \mathcal{K}} \frac{1}{n} \sum_{i=1}^{n} \left| x_i - \operatorname*{med}_{1 \leq j \leq n} x_j \right|,$$

P4 To find among all the vectors falling into $\mathcal{K}$

$$\widetilde{x}^{min} = \arg\min_{x \in \mathcal{K}} \frac{1}{n} \sum_{i=1}^{n} \left| x_i - \operatorname*{med}_{1 \leq j \leq n} x_j \right|$$

P5 To find among all the vectors falling into $\mathcal{K}$

$$\overset{\approx}{x}^{max} = \arg\max_{x \in \mathcal{K}} \frac{1}{n} \sum_{i=1}^{n} \left| x_i - \overline{x}_n \right|$$

P6 To find among all the vectors falling into $\mathcal{K}$

$$\overset{\approx}{x}^{min} = \arg\min_{x \in \mathcal{K}} \frac{1}{n} \sum_{i=1}^{n} \left| x_i - \overline{x}_n \right|$$

Solution to P2, P4 and P6

Assertion If the intersection of all intervals is not empty ($\mathcal{I} = \cap_i I_i \neq \emptyset$) then there exist either one or infinitely many solutions of P2, P4 and P6 consisting of all $x \in \mathcal{K}$ for which $x_1 \in \mathcal{I}$ and $x_1 = \ldots = x_n$. One solution exists if $\mathcal{I}$ contains only one point while infinitely many solutions exist if $\mathcal{I}$ is an interval.

$$x = \left(x_1, \ldots, x_n\right)' \in \mathcal{K}, \quad a_{max} \leq x_1 \leq b_{min} \quad \text{and} \quad x_i = x_1 \quad \imath = 2, \ldots, n,$$

where

$$a_{max} = \max_{1 \leq i \leq n} \ a_i \quad \& \quad b_{min} = \min_{1 \leq i \leq n} \ b_i$$

Solution to P2, P4 and P6

Assertion If the intersection of all intervals is not empty ($\mathcal{I} = \cap_i I_i \neq \emptyset$) then there exist either one or infinitely many solutions of P2, P4 and P6 consisting of all $x \in \mathcal{K}$ for which $x_1 \in \mathcal{I}$ and $x_1 = \ldots = x_n$. One solution exists if $\mathcal{I}$ contains only one point while infinitely many solutions exist if $\mathcal{I}$ is an interval.

$$x = \left(x_1, \ldots, x_n\right)' \in \mathcal{K}, \quad a_{max} \leq x_1 \leq b_{min} \quad \text{and} \quad x_i = x_1 \quad \imath = 2, \ldots, n,$$

where

$$a_{max} = \max_{1 \leq i \leq n} \; a_i \quad \& \quad b_{min} = \min_{1 \leq i \leq n} \; b_i$$

Idea

$$\text{var } x = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x}_n\right)^2 = \frac{1}{n(n-1)} \sum_{1 \leq i < j \leq n} \left(x_i - x_j\right)^2$$

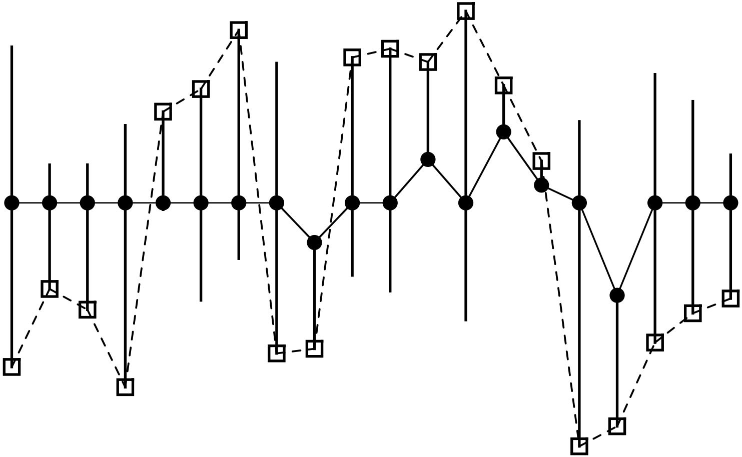
Solution to P1 and P2

Assertion Assume the above mentioned setup. Then solutions of both task P1 ("max") and task P2 ("min") are on the boundary of $\mathcal{K}$. Moreover, solution(s) of task P1 coincide with one (or more) vertex(es) of $\mathcal{K}$.

Remarks:

- Proof of assertion shows that in the case of task P1 ($\arg\max_{x\in\mathcal{K}}\ \operatorname{var} x$) we are looking for that corner of $\mathcal{K}$ which has the largest distance from the straight line passing through the origin and the point $(1,\ldots,1)'$.
- In the case of task P2 ($\arg\min_{x\in\mathcal{K}}\ \operatorname{var} x$) we are, analogously, looking for that point(s) from $\mathcal{K}$ which have the smallest distance.
- Due to the fact that the straight line passing through the origin and the point $(1,\ldots,1)'$ can have a nonempty intersection with $\mathcal{K}$, number of solutions of P2 can be infinite.

Typical example

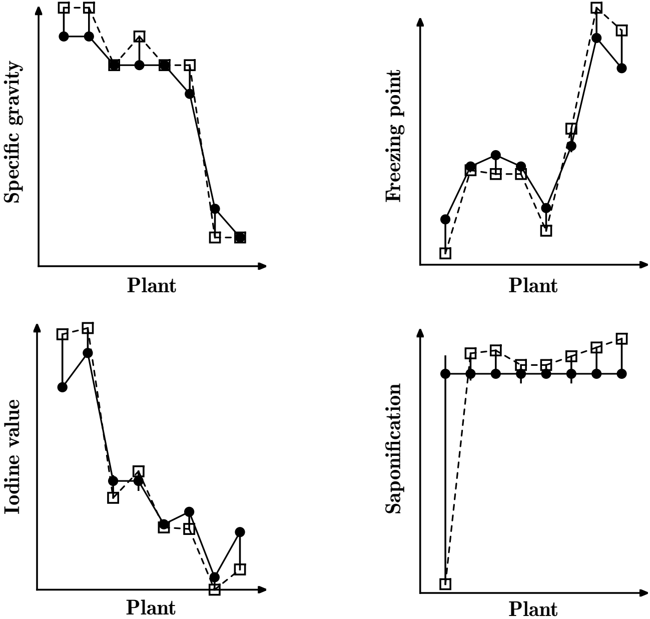


Vertical lines denote individual intervals, squares solution of task P1 ("max"), dots solution of task P2 ("min"). Solution of P1 ("max") is connected by a dashed line and solution of P2 ("min") by a solid line.

Practical example

Data present eight different classes of oils described by four quantitative interval valued variables, i.e. *Specific gravity, Freezing point, Iodine value* and *Saponification*.

Plant number	Plant	Specific gravity		Freezing point		Iodine value		Saponification	
1	Linseed	0.93	0.94	-27	-18	170	204	118	196
2	Perilla	0.93	0.94	-5	-4	192	208	188	197
3	Cotton	0.92	0.92	-6	-1	99	113	189	198
4	Sesame	0.92	0.93	-6	-4	104	116	187	193
5	Camellia	0.92	0.92	-21	-15	80	82	189	193
6	Olive	0.91	0.92	0	6	79	90	187	196
7	Beef	0.86	0.87	30	38	40	48	190	199
8	Hog	0.86	0.86	22	32	53	77	190	202



Vertical lines denote individual intervals, squares solution of **P1**, dots solution of **P2**.

Numerical results

Point with minimal variance				Point with maximal variance			
Specific gravity	Freez. point	Iodine value	Saponification	Specific gravity	Freez. point	Iodine value	Saponification
0.93	-18.00	170.0	190-3	0.94	-27	204	118
0.93	-4.00	192.0	190-3	0.94	-5	208	197
0.92	-1.00	109.8	190-3	0.92	-6	99	198
0.92	-4.00	109.8	190-3	0.93	-6	116	193
0.92	-15.00	82.0	190-3	0.92	-21	80	193
0.91	1.42	90.0	190-3	0.92	6	79	196
0.87	30.00	48.0	190-3	0.86	38	40	199
0.86	22.00	77.0	190-3	0.86	32	53	202
0.000 735	278.82	2348.7	0	0.001 070	536.55	4086.7	786.29

Extreme points x^{min} and x^{max} solving **P1** and **P2**. Last line in the table corresponds to the variance of the corresponding column of data.

Numerical results

Point with minimal variance				Point with maximal variance			
Specific gravity	Freez. point	Iodine value	Saponification	Specific gravity	Freez. point	Iodine value	Saponification
0.93	-18.00	170.0	190-3	0.94	-27	204	118
0.93	-4.00	192.0	190-3	0.94	-5	208	197
0.92	-1.00	109.8	190-3	0.92	-6	99	198
0.92	-4.00	109.8	190-3	0.93	-6	116	193
0.92	-15.00	82.0	190-3	0.92	-21	80	193
0.91	1.42	90.0	190-3	0.92	6	79	196
0.87	30.00	48.0	190-3	0.86	38	40	199
0.86	22.00	77.0	190-3	0.86	32	53	202
0.000 735	278.82	2348.7	0	0.001 070	536.55	4086.7	786.29

Extreme points x^{min} and x^{max} solving **P1** and **P2**. Last line in the table corresponds to the variance of the corresponding column of data.

Danger for many statistics, including sample covariances or correlations is more than evident.

Genetical algorithms

Genetic	Optimization Problem
Individual	Candidate solution
Fitness of an individual	Objective function calculated on solution
Chromosome of an individual	Coding of a candidate solution
Gene (digit of a chromosome)	Piece of a candidate solution

For a genetic algorithm solving (maybe) the problem, user must set up:

1. Representation scheme (encoding of objects)
   usually most critical point
2. Measure of fitness
3. Parameters and variables controlling algorithm
4. Terminating criterion

Genetic algorithm – set up

Representation of variance (6) allows to concentrate only on the vertexes of the cube $\mathcal{K}$!

Idea: It is evident that there exist a 1-1 mapping between set of all vertexes of $\mathcal{K}$ and a set of vectors $\boldsymbol{\alpha} = \big(\alpha_1, \ldots, \alpha_n\big)' \in \{0, 1\}^n$, where: ‘

- $\alpha_i = 0$ corresponds to the choice $x_i = a_i$
- $\alpha_i = 1$ corresponds to the choice $x_i = b_i$. ‘

Fitness function is given by the variance calculated for the vertex corresponding to current $\boldsymbol{\alpha}$.

It is necessary to set

- population size S
- crossover k (alternatively $k_1,\ k_2$ etc.)
- mutation probability p_M
- stopping rule.

Choice of parameters

Other parameters of genetic algorithm used were chosen (empirically) as follows:

- Initial generation has been chosen randomly, i.e., the genes were simulated from the alternative distribution $Alt(1/2)$.
- Crossover scheme : single point crossover with $k \approx 0.6n$.
- Mutation probability $p_M \approx 0.01$.
- Fitness $f(\boldsymbol{\alpha}) = var\, x$, where x is that vertex of $\mathcal{K}$ that corresponds to the chromosome $\boldsymbol{\alpha}$.
- Population size $card(S) = 100$.
- Number of generations 300.
- Elitism was used, i.e., the best individual of a generation is cloned with the new one.

Sensitivity of proposed procedure

Basic conclusion is that the mutation probability considerably influences both the population size $card(S)$ and number of generations.

Other parameters do not play so important role.

More specifically:

- If we increase mutation probability, we must either considerably increase the number of generations or population size. For example, the choice $p_M = 0.025$ recommended by the literature required either to double the population size or to triple the number of generations.
- Choice of the initial generation does not have substantial impact on the speed to arrive to the optimal solution.
- Crossover scheme does not have an impact on the speed to arrive to the optimal solution. Single point crossover gave us practically the same results as two point or random crossover.

Moderately difficult example

Genetic algorithm for **P1** ("max") has been tested on many real and simulated datasets and converged very rapidly.

-47.50, 28.75	47.25, 91.75	38.50, 81.50	-53.50, 45.00	-11.00, -05.00
-46.50, 93.50	-98.25, -40.75	-34.50, -28.00	51.00, 64.00	-32.50, -13.50
40.75, 95.00	-47.00, 30.50	-95.75, -25.25	-34.00, -28.25	-88.50, -71.50
-01.25, 34.50	16.50, 81.25	-21.75, 39.75	30.25, 80.25	-14.50, -06.00
-10.00, 07.75	-81.50, -72.25	-94.00, -18.75	-65.50, 22.00	-93.00, -83.00
-03.25, 83.25	-90.25, -77.00	-24.75, -01.25	42.50, 79.75	-66.00, -61.25
27.75, 57.00	49.75, 85.50	12.75, 90.75	18.50, 85.50	-19.50, 06.75
-42.25, 20.00	-93.75, -33.00	-22.50, 01.25	-95.75, -52.00	-77.50, -42.75

An exhaustive search (that on a cluster of 16 processors took about 4 hours) revealed that the maximum for (4) was 4911.9990, which is exactly the same value (and corresponds to the same endpoints configuration) that the genetic algorithm found in less than 100 iterations (that took less than a second).

Solution

According to Ferson there does not exists for this type of the data other algorithms enabling to find x^{max} than exhaustive search. It took us about 4 hours on a cluster with 16 multi-kernel processors to reveal that x^{max} correspond to the point given in Table 2. Notice that we have found the same point using our genetic algorithm in less than several hundred iterations, taking less than a second of CPU on one of the processors.

-47.50	91.75	81.50	45.00	93.50	-98.25	-34.50	64.00
95.00	-47.00	-95.75	-34.00	34.50	81.25	39.75	80.25
7.75	-81.50	-94.00	-65.50	83.25	-90.25	-24.75	79.75
57.00	85.50	90.75	85.50	-93.00	-95.75	-66.00	-77.50
-88.50	-93.75	-14.50	-22.50	-11.00	-19.50	-32.50	-42.25

Table 2. Solution.

Some other approaches

1. Random sampling
2. Random walk on vertexes of hypercube

1. Random sampling – **does not work**
2. Random walk on vertexes of hypercube – **works better, but much more slowly**

GENERAL PROBLEM : Unlike systematic search, none of mentioned methods ensures finding global minima.