On the problem of variability of interval data

Václav Finěk[♠], Ctirad Matonoha[♠♡]

[♠] Technical University of Liberec, Faculty of Science, Humanities and Education,
Studentská 2, 461 17 Liberec, Czech Republic

[♡] Institute of Computer Science, Academy of Sciences of the Czech Republic,
Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic

ICPM'12, June 21 - 22, 2012, Liberec, Czech Republic

Introduction

In some cases, we have only intervals $[a_i, b_i]$ of possible values of x_i instead of the actual value x_i. Example:

- Measured values usually include some measurement error with known upper bounds.

Then, the actual value of x_i is unknown and we only know that its value is located within the interval determined by the upper bound of measurement error.

Therefore, we should work rather with these intervals than with these single values. Consequently, possible values of their average and their variance are also intervals. While the computation of lower and upper bounds for average of interval data is straightforward, the computation of lower and upper bounds for their variance is significantly complicated.

We present some theoretical results concerning the solution of maximization problem and introduce preliminary algorithms for solving both problems which use their special structure.

Outline

Outline

The problem

We consider n intervals $I_i = [a_i, b_i]$ and define

$$K = I_1 \otimes I_2 \otimes I_2 \otimes \cdots \otimes I_n. \tag{1}$$

We would like to find:

$$\mathbf{x}^{min} = \arg\min_{x \in K} \frac{1}{n} F(x), \tag{2}$$

$$\mathbf{x}^{max} = \arg\max_{x \in K} \frac{1}{n} F(x), \tag{3}$$

where $F(x) = \sum_{i=1}^{n} (x_i - \overline{\mathbf{x}})^2$ with $\overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Function F

Function F can be written in the following forms:

$$\begin{aligned} F(x) &= \sum_{i=1}^{n}(x_i - \overline{\mathbf{x}})^2 = \sum_{i=1}^{n} x_i^2 - \frac{(x_1 + \cdots + x_n)^2}{n} = \sum_{i=1}^{n} x_i^2 - n\overline{\mathbf{x}}^2 \\ &= \frac{1}{2n}\left[\sum_{i,j=1}^{n}(x_i - x_j)^2\right] = \frac{1}{n}\left[\sum_{j<i}(x_i - x_j)^2\right] \end{aligned}$$

Matrix representation

Define

$$\mathbf{Y} = \begin{pmatrix} x_1 - x_1 & x_1 - x_2 & \dots & x_1 - x_n \\ x_1 - x_2 & x_2 - x_2 & \dots & x_2 - x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_1 - x_n & x_2 - x_n & \dots & x_n - x_n \end{pmatrix}$$

Matrix $\mathbf{Y}$ is symmetric, indefinite, zero diagonal, $\sum_{i=1}^{n} \lambda_i = 0$. Now

$$\begin{aligned} F(x) &= \frac{1}{2n} \sum_{i,j=1}^{n} (x_i - x_j)^2 = \frac{1}{2n} \|\mathbf{Y}\|_F^2 = \frac{1}{2n} \mathrm{tr}(\mathbf{Y}^2) \\ &= \frac{1}{2n} \sum_{i=1}^{n} \mu_i = \frac{1}{2n} \sum_{i=1}^{n} \lambda_i^2 = \frac{1}{2n} \sum_{i=1}^{n} \sigma_i^2 \end{aligned}$$

where

- μ_i are eigenvalues of $\mathbf{Y}^2$
- λ_i are eigenvalues of $\mathbf{Y}$
- σ_i are singular values of $\mathbf{Y}$

The problem

In this contribution we give some theoretical results of problems (2)-(3) and outline the algorithms for seeking the solutions. We will consider the following structure of problems (2)-(3):

$$\mathbf{x}^{min} = \arg\min_{x \in \mathcal{R}^n} F(x), \tag{4}$$
$$\text{subject to} \quad x_i \in [a_i, b_i], \quad i = 1 \ldots n$$

$$\mathbf{x}^{max} = \arg\max_{x \in \mathcal{R}^n} F(x), \tag{5}$$
$$\text{subject to} \quad x_i \in [a_i, b_i], \quad i = 1 \ldots n$$

The following quantities are used throughout the paper:

$$c_i = \frac{a_i + b_i}{2}, \quad d_i = b_i - a_i > 0, \quad i = 1 \ldots n,$$

$$p_a = \frac{a_1 + \cdots + a_n}{n}, \quad p_b = \frac{b_1 + \cdots + b_n}{n}.$$

Properties

There are several apparent properties:

- If $\cap_i I_i \neq \emptyset$ then there exist either one or infinitely many solutions to (4). All elements of the solution are the same and belong to $\cap_i I_i$. Consequently, $F = 0$.
- The solution to (5) lies on the vertex of K.
- It holds $\sum_{i=1}^{n}(x_i - \overline{\mathbf{x}})^2 < \sum_{i=1}^{n}(x_i - d)^2$ for any $d \neq \overline{\mathbf{x}}$.

Outline

Basic optimization method

Both problems (4)-(5) are classical optimization problems with simple bounds that can be solved by a suitable optimization method.
Basic optimization method is an iteration process starting from an initial point $x^{(0)}$ and generating a sequence of points $x^{(1)}, x^{(2)}, \ldots$ leading to a solution x^* such that

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)},$$

where

- $d^{(k)}$ is a direction vector – is determined on the basis of values

$$x^{(j)},\ F(x^{(j)}),\ F'(x^{(j)}),\ F''(x^{(j)}),\ 0 \leq j \leq k,$$

- $\alpha^{(k)} > 0$ is a step-length – is determined on the basis of behavior of function F in the neighborhood of $x^{(k)}$.

We use a suitable optimization method (line-search or trust-region method) from the so-called UFO system

http://www.cs.cas.cz/luksan/ufo.html

Making use of a special structure

Unfortunately, our problem is dense (the Hessian matrix of F is dense), so it is hard to solve problems (4)-(5) for large n by a general optimization method. We will make use of a special structure of the problem and introduce algorithms that take into consideration this structure (the solution satisfies certain properties).

The problem of minimization (4) is simple and no significant difficulties arise. We do not need know anything special about the solution and the following algorithm works well on the test problems.

Algorithm 1 *Solving problem (4).*

1. Set

$$p_a = \emptyset a_i,\ p_b = \emptyset b_i,\ p = \frac{p_a + p_b}{2},\ A^{\max} = \max\{a_i\},\ B_{\min} = \min\{b_i\}$$

and $s_i = 0\ \forall i$ (indicator of the solution x_i^*: 0 - not found, 1 - found).

2. If $A^{\max} \le B_{\min}$, then the solution $x_i^*\ \forall i$ is an arbitrary number in $[A^{\max}, B_{\min}]$ and $F = 0$.

3. For $i = 1 \dots n$ such that $s_i = 0$ perform:
   1. If $p \in [a_i, b_i]$, then $x_i = p$, otherwise $x_i = a_i$ or b_i according to if p is closer to p_a or p_b.
   2. We reduce the intervals $[a_i, b_i]$ so that $[a_i, b_i] \subseteq [p_a, p_b]$.
   3. If $[a_i, b_i] \cap [p_a, p_b] = \emptyset$, then $x_i^* = x_i$ and $s_i = 1$.

4. Set

$$p_a^{old} = p_a,\ p_b^{old} = p_b,\ p_a = \emptyset a_i,\ p_b = \emptyset b_i,\ p = \frac{p_a + p_b}{2}$$

5. If $\max\{|p_a - p_a^{old}|, |p_b - p_b^{old}|\} > \varepsilon$, go to Step 3.

Outline

It is harder

The problem of maximization (5) is much harder than the problem of minimization (4), so we will focus on the theoretical analysis and find useful information concerning the solution.

The solution x^* to (5) lies on the vertex of K, see the second apparent property above or Lemma 1 below. Thus in the subsequent analysis we assume that x^* has components x_i^* equal to either a_i or b_i. First, we will study what the solution must satisfy.

About the components x_i^*

Lemma 1

The solution x^* to (5) lies on the vertex of K, i.e. $x^* = (x_1^* \dots x_n^*)$, where $x_i^* = a_i$ or $x_i^* = b_i$.

Lemma 2

The solution x^* to (5) has the property that there exists at least one i and at least one j, $i \neq j$ such that $x_i^* = a_i$ and $x_j^* = b_j$.

Lemma 3

Let $x \in \mathcal{R}^n$ and $\overline{\mathbf{x}}$ be the average value of points $x_1 \dots x_n$. Suppose that there exists j such that $x_j = \overline{\mathbf{x}}$. Then $F(x) < F(x_\tau)$, where

$$x_\tau = (x_1 \dots x_{j-1},\ x_j + \tau,\ x_{j+1} \dots x_n)$$

for $\tau \neq 0$. In other words, no component x_i^* of the solution x^* is equal to the average value $\overline{\mathbf{x}^*}$.

About the difference $F(x) - F(y)$

The following analysis concerns the differences between function values. The first lemma stands for the most general case.

Lemma 4

Let $x, y \in \mathcal{R}^n$ be arbitrary. Denote p_{xy} the average value of all points $x_i, y_i,\ i = 1 \ldots n,$ and define the following sets:

- $N_{aa} = \{i :\ x_i = a_i,\ y_i = a_i\},$
- $N_{ab} = \{i :\ x_i = a_i,\ y_i = b_i\},$
- $N_{ba} = \{i :\ x_i = b_i,\ y_i = a_i\},$
- $N_{bb} = \{i :\ x_i = b_i,\ y_i = b_i\}.$

It is evident that $N_{aa} \cup N_{ab} \cup N_{ba} \cup N_{bb} = \{1 \ldots n\}.$ Then it holds

$$F(x) - F(y) = 2 \left[\sum_{i \in N_{ba}} d_i (c_i - p_{xy}) - \sum_{i \in N_{ab}} d_i (c_i - p_{xy}) \right]$$

About the difference $F(\{x_i\}_{i\neq j}, b_j) - F(\{x_i\}_{i\neq j}, a_j)$

Suppose that we have some combination of points $x_1 \dots x_n$. Now we fix some j and ask if the function value is larger for $x_j = a_j$ or $x_j = b_j$.

Lemma 5

Let $x \in \mathcal{R}^n$ and take an arbitrary index j. Denote p_{x_i} the average value of points $\{x_i\}_{i\neq j}$. Then

$$F(\{x_i\}_{i\neq j}, b_j) - F(\{x_i\}_{i\neq j}, a_j) = 2\,\frac{n-1}{n}\,d_j(c_j - p_{x_i})$$

The consequence is that:

- $F(\{x_i\}_{i\neq j}, b_j) > F(\{x_i\}_{i\neq j}, a_j) \quad \Leftrightarrow \quad c_j > p_{x_i}$,
- $F(\{x_i\}_{i\neq j}, b_j) < F(\{x_i\}_{i\neq j}, a_j) \quad \Leftrightarrow \quad c_j < p_{x_i}$,
- $F(\{x_i\}_{i\neq j}, b_j) = F(\{x_i\}_{i\neq j}, a_j) \quad \Leftrightarrow \quad c_j = p_{x_i}$,

Special case of previous lemma

A special case of the previous lemma is the following result.

Lemma 6

Consider sets $\{a_i\}, \{b_i\}$ and let j be an arbitrary index. Then

$$F(\{a_i\}_{i\neq j}, b_j) - F(\{a_i\}) = 2d_j(c_j - p_a - \frac{1}{2n}d_j),$$

$$F(\{b_i\}_{i\neq j}, a_j) - F(\{b_i\}) = 2d_j(p_b - c_j - \frac{1}{2n}d_j).$$

Both numbers on the right-hand side are equal if and only if $c_j = p := \frac{p_a+p_b}{2}$.
In general, if we compare both numbers on the right-hand side, we will derive relations

- $F(\{a_i\}_{i\neq j}, b_j) - F(\{a_i\}) > F(\{b_i\}_{i\neq j}, a_j) - F(\{b_i\}) \quad \Leftrightarrow \quad c_j > p$
- $F(\{a_i\}_{i\neq j}, b_j) - F(\{a_i\}) < F(\{b_i\}_{i\neq j}, a_j) - F(\{b_i\}) \quad \Leftrightarrow \quad c_j < p$
- $F(\{a_i\}_{i\neq j}, b_j) - F(\{a_i\}) = F(\{b_i\}_{i\neq j}, a_j) - F(\{b_i\}) \quad \Leftrightarrow \quad c_j = p$

Consequence of previous result

Lemma 7

As the solution x^* must contain at least one a_i and at least one b_j, the consequence is that if we consider the initial iteration $x^{(0)}$, then

- a_j can be replaced with b_j if and only if $p_a < c_j - \frac{1}{2n} d_j$;
- b_j can be replaced with a_j if and only if $p_b > c_j + \frac{1}{2n} d_j$.

Comparison of $F(a)$ and $F(b)$

The following lemma gives the comparison of function values on each side of the set K.

Lemma 8

It holds

- $F(a_1, \ldots, a_n) < F(b_1, \ldots, b_n) \quad \Leftrightarrow \quad p < \frac{\sum(c_i d_i)}{\sum d_i}$,
- $F(a_1, \ldots, a_n) > F(b_1, \ldots, b_n) \quad \Leftrightarrow \quad p > \frac{\sum(c_i d_i)}{\sum d_i}$,
- $F(a_1, \ldots, a_n) = F(b_1, \ldots, b_n) \quad \Leftrightarrow \quad p = \frac{\sum(c_i d_i)}{\sum d_i}$,

Lemma 9

Let j be an arbitrary index. Then

- $F(\{a_i\}_{i \neq j}, b_j) < F(\{b_i\}_{i \neq j}, a_j) \quad \Leftrightarrow \quad (c_j - p)d_j < \sum_{i \neq j}[(c_i - p)d_i]$
- $F(\{a_i\}_{i \neq j}, b_j) > F(\{b_i\}_{i \neq j}, a_j) \quad \Leftrightarrow \quad (c_j - p)d_j > \sum_{i \neq j}[(c_i - p)d_i]$
- $F(\{a_i\}_{i \neq j}, b_j) = F(\{b_i\}_{i \neq j}, a_j) \quad \Leftrightarrow \quad (c_j - p)d_j = \sum_{i \neq j}[(c_i - p)d_i]$

Quadratic function approach

Another approach to determine previous results consists in that we fix a set of points $\{x_i\}_{i\neq j}$ and study the function values on $[a_j, b_j]$. Denote

$$x_j = a_j + \tau_j(b_j - a_j) = a_j + \tau_j d_j, \quad \tau_j \in [0,1].$$

Then the quadratic function $f(\tau_j)$ satisfies

$$f(\tau_j) \equiv F(\{x_i\}_{i\neq j}, x_j) = \sum_{i\neq j} x_i^2 + (a_j + \tau_j d_j)^2 - \frac{1}{n}\left(\sum_{i\neq j} x_i + a_j + \tau_j d_j\right)^2$$

$$f'(\tau_j) = 2d_j(a_j + \tau_j d_j) - \frac{2}{n} d_j \left(\sum_{i\neq j} x_i + a_j + \tau_j d_j\right)$$

$$f'(\tau_j) = 0 \quad \Leftrightarrow \quad \tau_j^* = \frac{p_{x_i} - a_j}{b_j - a_j}$$

$$f''(\tau_j) = 2\,\frac{n-1}{n}\, d_j^2 \quad \Rightarrow \quad f(\tau_j^*) = \min f(\tau_j)$$

General difference $F(\{x_i\}_{i\neq j}, b_j) - F(\{x_i\}_{i\neq j}, a_j)$

From here we obtain properties mentioned above, that is

$$\begin{aligned}
\tau_j^* < 0.5 \quad &\Leftrightarrow \quad c_j > p_{x_i} \quad \Leftrightarrow \quad F(\{x_i\}_{i\neq j}, b_j) > F(\{x_i\}_{i\neq j}, a_j) \\
\tau_j^* > 0.5 \quad &\Leftrightarrow \quad c_j < p_{x_i} \quad \Leftrightarrow \quad F(\{x_i\}_{i\neq j}, b_j) < F(\{x_i\}_{i\neq j}, a_j) \\
\tau_j^* = 0.5 \quad &\Leftrightarrow \quad c_j = p_{x_i} \quad \Leftrightarrow \quad F(\{x_i\}_{i\neq j}, b_j) = F(\{x_i\}_{i\neq j}, a_j)
\end{aligned}$$

Remark

Function $f(\tau_j)$ satisfies

$$\begin{aligned}
f(\tau_j) \quad = \quad & \frac{n-1}{n} d_j^2 \tau_j^2 \\
& - 2\,\frac{n-1}{n} d_j (p_{x_i} - a_j)\tau_j \\
& + \sum_{i\neq j} x_i^2 + a_j^2 - \frac{1}{n}\left(\sum_{i\neq j} x_i + a_j\right)^2
\end{aligned}$$

Determination of some components of the solution

We can immediately determine some components of the solution. Denote

- p_j^a = the average value of $\{a_i\}_{i \neq j},\ j = 1 \ldots n$
- p_j^b = the average value of $\{b_i\}_{i \neq j},\ j = 1 \ldots n$
- p_α = the average value of $\left\{\{a_i\}_{i \neq k}, b_k\right\}$, where k is such that $d_k = \min_i\{d_i\}$.
- p_β = the average value of $\left\{\{b_i\}_{i \neq k}, a_k\right\}$, where k is such that $d_k = \min_i\{d_i\}$.

Then it holds for $j = 1 \ldots n$

- If $c_j < \max\{p_j^a, p_\alpha\}$, then $x_j^* = a_j$.
- If $c_j > \min\{p_j^b, p_\beta\}$, then $x_j^* = b_j$.
- If $c_j \in P_j = [\max\{p_j^a, p_\alpha\}, \min\{p_j^b, p_\beta\}]$, an iteration process must be performed.

Algorithm 2 *Solving problem (5) – preliminary algorithm.*

1. Set

$$p_a = \emptyset a_i, \;\; p_b = \emptyset b_i, \;\; p = \frac{p_a + p_b}{2}, \;\; c_i = \frac{a_i + b_i}{2}, \;\; d_i = b_i - a_i,$$

$$p_j^a = \emptyset\{a_i\}_{i \neq j}, \quad p_j^b = \emptyset\{b_i\}_{i \neq j}, \quad j = 1 \ldots n,$$

$$p_\alpha = \emptyset\Big\{\{a_i\}_{i \neq k}, b_k\Big\}, \text{ where } k = \arg\min_i\{d_i\},$$

$$p_\beta = \emptyset\Big\{\{b_i\}_{i \neq k}, a_k\Big\}, \text{ where } k = \arg\min_i\{d_i\},$$

   and $s_i = 0 \;\forall i$ (indicator of the solution x_i^*: 0 - not found, 1 - found).

2. For $i = 1 \ldots n$ such that $s_i = 0$ perform:
   1. If $c_i < \max\{p_j^a, p_\alpha\}$, then $x_i^* = a_i$ and $s_i = 1$.
      If $c_i > \min\{p_j^b, p_\beta\}$, then $x_i^* = b_i$ and $s_i = 1$.
   2. If $c_i < p$, then $x_i = a_i$ and $b_i = \min\{b_i, p_a\}$.
      If $c_i > p$, then $x_i = b_i$ and $a_i = \max\{a_i, p_b\}$.
      If $p = c_i$ and $[p_a, p_b] \subseteq [a_i, b_i]$, then $b_i = p_a$ or $a_i = p_b$.
3. Set $p^{old} = p$ and compute new p_a, p_b, p.
4. If $|p - p^{old}| > \varepsilon$, go to Step 2.

Outline

Example 1 Let $n = 3$ and

$$[a_i,\ b_i] = [-3,\ 1],\ [-9/2,\ 3],\ [-1,\ 1/2]$$

It holds that

$$c_1 = -1, \quad d_1 = 4, \quad p_1^a = -11/4, \quad p_1^b = 7/4$$

$$c_2 = -3/4, \quad d_2 = 15/2, \quad p_2^a = -2, \quad p_2^b = 3/4$$

$$c_3 = -1/4, \quad d_3 = 3/2, \quad p_3^a = -15/4, \quad p_3^b = 2$$

$$p_\alpha = -7/3, \quad p_\beta = 1, \quad p = -2/3$$

We do not have immediately any component of the solution because $c_j \in P_j\ \forall j$. The solution satisfies

$$x^* = (-3,\ 3,\ -1), \quad \text{ø} x_i^* = -1/3, \quad F(x^*) = 6.\overline{2}$$

and the point just on the other sides of all interval satisfies

$$\tilde{x} = (1,\ -9/2,\ 1/2), \quad \text{ø}\tilde{x}_i = -1, \quad F(\tilde{x}) = 6.1\overline{6}$$

The function values are located close together which makes hard for the algorithm to identify the right solution.

Example 2

Example 2 Demonstration of making use of a special structure of our problem (the components of the solution lies on the edges of intervals).

We have the set of test intervals $[a_i, b_i]$ with $n = 40$.

- If we use a general optimization method, we obtain the results

$$\text{NIT} = 41 \quad \text{and} \quad F = 4709.79625$$

- If we use Algorithm 2, we obtain

$$\text{NIT} = 2 \quad \text{and} \quad F = 4911.99902$$

Conclusion

- We have presented some theoretical results and preliminary algorithms for computing variance of interval data described in the introduction.
- Although the problem can be solved using various optimization methods combining direction vectors and the step-length, developing a special algorithm taking into consideration a special structure of the problem is more advantageous.
- Concerning the maximization problem, where finding a solution is much harder than in case of the minimization problem, the solution satisfies useful properties which allow us to identify directly some components of the solution.

Main tasks

The main task is to develop more robust algorithm to identify the right solution and to deal with the unpleasant facts:

- what to do when $c_j \in P_j$?
- the component x_i^* does not depend on the relation between c_i and p (it can hold both $c_i < p$ and $c_i > p$)
- both $\min\{a_i\}$ and $\max\{b_i\}$ need not be components of the solution (but at least one of them must be - which one?)
- using the same sequence of several different intervals does not generate the same sequence of the solution (is there any regularity for the solution?)