Fitting a Normal Distribution to Interval and Fuzzy Data

Gang Xiang and Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
El Paso, Texas 79968, USA
gxiang@utep.edu, vladik@utep.edu

Scott Ferson
Applied Biomathematics
100 North Country Road
Setauket, NY 11733, USA
email scott@ramas.com

1. Formulation of the Problem

- *Known empirical fact:* in many real-life situations, the actual distribution is normal (Gaussian).
- *Known mathematical fact:* a normal distribution is uniquely determined by its mean a and its standard deviation σ: its cdf is $F(x) = \Phi\left(\dfrac{x-a}{\sigma}\right)$.
- *How to find a and σ:* main idea of method of moments is to equate moments of the distribution with sample moments.
- *Resulting algorithm:* compute

$$a = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i, \quad \sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - a)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - a^2,$$

and consider the normal distribution with these values a and σ as "fitted" to the data $x_1, \ldots, x_n$.

2. Case of Interval Uncertainty

- *Traditional statistical approach:* assumes that we have the exact values x_i forming the sample.
- *In practice:* we often only know the intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ of possible values of x_i.
- *Fact:* different values of $x_i \in \mathbf{x}_i$ lead, in general, to different values of a and σ.
- *Conclusion:* for every x, different values of $x_i \in \mathbf{x}_i$ lead, in general, to different values of the cdf $F(x) = \Phi\left(\frac{x-a}{\sigma}\right)$.
- *Objective:* it is therefore desirable to find the interval $[\underline{F}(x), \overline{F}(x)]$ of possible values of the cdf.
- *Objective reformulated:* find a *p-box* that contains all possible cumulative distribution functions.

3. Case of Fuzzy Uncertainty

- *Expert estimates:* often, knowledge comes in terms of uncertain expert estimates.
- *Fuzzy techniques:* to describe this uncertainty, for each value of estimation error Δx_i, we describe the degree $\mu_i(\Delta x_i)$ to which this value is possible.
- *α-cuts – general case:* for each degree of certainty α, we can determine the set of values of Δx_i that are possible with at least this degree of certainty – the *α-cut* $\{\Delta x_i \,|\, \mu_i(\Delta x_i) \geq \alpha\}$ of the original fuzzy set.
- *α-cuts – interval case:* in most cases, this α-cut is an interval – i.e., we have fuzzy numbers.
- *Conclusion:* fuzzy numbers $\mu_i(x)$ can be viewed as a family of nested intervals $\mathbf{x}_i(\alpha)$ – α-cuts of the given fuzzy sets.

4. From the Computational Viewpoint, Processing Fuzzy Uncertainty Reduces to Processing of Interval Uncertainty

- *Reminder:* a fuzzy number $\mu_i(x)$ can be viewed as a family of nested intervals $\mathbf{x}_i(\alpha)$ (its α-cuts).
- *Our objective:* to compute the fuzzy number corresponding to $y = f(x_1, \ldots, x_n)$.
- *Fact:* the α-cut of the desired fuzzy number is the range of f over the α-cuts $\mathbf{x}_i(\alpha)$ of the fuzzy sets $\mu_i(x)$.
- *Typical situation:* we want to describe 10 different levels of uncertainty.
- *Solution:* solve 10 interval computation problems.
- *Conclusion:* from the computational viewpoint, it is sufficient to produce an efficient algorithm for the interval case.

5. What Is Known

- *Reminder:* $F(x) = \Phi(r)$, where $r \stackrel{\text{def}}{=} \dfrac{x-a}{\sigma}$.
- *Reasonable idea:* compute $[\underline{a}, \overline{a}]$ and $[\underline{\sigma}, \overline{\sigma}]$.
- *Mean:* $\underline{a} = \dfrac{1}{n} \cdot \sum_{i=1}^{n} \underline{x}_i$, $\overline{a} = \dfrac{1}{n} \cdot \sum_{i=1}^{n} \overline{x}_i$.
- *Standard deviation:* in general, NP-hard.
- *Exponential time algorithm:* try all 2^n combinations of $x_i = \overline{x}_i$ and $x_i = \underline{x}_i$.
- *No-nesting case:* when $[\underline{x}_i, \overline{x}_i] \not\subseteq (\underline{x}_j, \overline{x}_j)$, there is an efficient algorithm.
- *Problem:* a and σ are dependent, so the resulting bound on $r = \dfrac{x-a}{\sigma}$ contains excess width.
- *Our objective:* find exact range for $F(x)$.

6. Algorithm for Computing the Exact Range

- Consider all 3^n partitions $\{1,\ldots,n\} = I^- \cup I^+ \cup I_0$.
- Set $x_i = \underline{x}_i$ for $i \in I^-$, $x_i = \overline{x}_i$ for $i \in I^+$, and $x_i = a_0$ for $i \in I_0$, where a_0 is computed as follows:
- $\widetilde{a} \stackrel{\text{def}}{=} \sum\limits_{i\in I^-} \underline{x}_i + \sum\limits_{j\in I^+} \overline{x}_j$, $\widetilde{m} \stackrel{\text{def}}{=} \sum\limits_{i\in I^-} (\underline{x}_i)^2 + \sum\limits_{j\in I^+} (\overline{x}_j)^2$.
- Find the value a from the quadratic equation

$$\widetilde{m}+\frac{1}{\#(I_0)}\cdot(n\cdot a-\widetilde{a})^2 = n\cdot\left(a-\frac{n\cdot a-\widetilde{a}}{\#(I_0)}\right)\cdot(x-a)+n\cdot a^2.$$

- Take $a_0 \stackrel{\text{def}}{=} \dfrac{n\cdot a-\widetilde{a}}{N_0}$.
- Compute $\sigma = \sqrt{(a-a_0)\cdot(x-a)}$ and $r = \dfrac{x-a}{\sigma}$.
- The smallest of these values r is returned as $\underline{r}$, and the largest is returned as $\overline{r}$; then, $[\underline{F}(x), \overline{F}(x) = [\Phi(\underline{r}), \Phi(\overline{r})]$.

7. Quadratic-Time Algorithm

- *Applicable:* to find $\overline{r}$ for $x \ge \underline{a}$ (and $\underline{r}$ for $x \le \overline{a}$).
- First, we sort all $2n$ values $\underline{x}_i$ and $\overline{x}_i$ into a sequence

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(2n)}.$$

- Thus, we subdivide the real line into $2n+1$ zones $[x_{(0)}, x_{(1)}]$, $[x_{(1)}, x_{(2)}]$, $\ldots$, where $x_0 \stackrel{\text{def}}{=} -\infty$ and $x_{(2n+1)} \stackrel{\text{def}}{=} +\infty$.
- For each zone $[x_{(k)}, x_{(k+1)}]$, we partition indices i into

$$I^- = \{i : x_{(k+1)} \le \underline{x}_i\}, \quad I^+ = \{i \notin I^- : x_{(k)} \ge \overline{x}_i\},$$

$$I_0 = \{1, \ldots, n\} - I^- - I^+.$$

- Based on this partition, we compute $\widetilde{a}$, $\widetilde{m}$, a, a_0, and r as before.
- The largest of the resulting values r is the desired $\overline{r}$.

8. Cubic-Time Algorithm for the No-Nesting Case

- *No-nesting case (reminder):* $[\underline{x}_i, \overline{x}_i] \not\subseteq (\underline{x}_j, \overline{x}_j)$.
- *Result:* we can compute $\underline{r}$ for $x \ge \underline{a}$ (and $\overline{r}$ for $x \le \overline{a}$).
- *Algorithm:*
  - First, we sort all n intervals $[\underline{x}_i, \overline{x}_i]$ in the lexicographic order:

$$\underline{x}_1 \le \underline{x}_2 \le \ldots \le \underline{x}_n, \quad \overline{x}_1 \le \overline{x}_2 \le \ldots \le \overline{x}_n.$$

  - For each $n^- < n^+ \le n$, partition indices into $I^- = \{1, \ldots, n^-\}$ and $I^+ = \{n^+, \ldots, n\}$.
  - Based on this partition, we compute $\widetilde{a}$, $\widetilde{m}$, a, a_0, and r as before.
  - The smallest of the resulting values r is the desired $\underline{r}$.
- *Comment:* we have n^2 cases of $O(n)$ time algorithm.

9. Quadratic-Time Algorithm for the No-Nesting Case

- *Reminder:* we need to compute the values

$$\widetilde{a} = \sum_{i=1}^{n^-} \underline{x}_i + \sum_{j=n^+}^{n} \overline{x}_j, \quad \widetilde{m} = \sum_{i=1}^{n^-} (\underline{x}_i)^2 + \sum_{j=n^+}^{n} (\overline{x}_j)^2.$$

- *Main idea:* compute the values $\sum\limits_{i=1}^{n^-} \underline{x}_i$ and $\sum\limits_{i=1}^{n^-} (\underline{x}_i)^2$ consequently:
  - we start with 0 and
  - we consequently add, correspondingly, $\underline{x}_i$ or $(\underline{x}_i)^2$.
- *Number of steps:* constant for each new n^-.
- *Similarly:* to compute the n^+-sum, we need a constant number of step for each n^+.
- *Result:* $O(n^2)$ algorithm.

10. Main Idea of the Proofs

- *Known fact from calculus:* if a function $f(x)$ attains its maximum on $[\underline{x}, \overline{x}]$ at x_0, then:
  - if $x_0 \in (\underline{x}, \overline{x})$, then $\dfrac{\partial f}{\partial x_i} = 0$;
  - if $x_0 = \underline{x}$, then $\dfrac{\partial f}{\partial x_i} \le 0$;
  - if $x_0 = \overline{x}$, then $\dfrac{\partial f}{\partial x_i} \ge 0$.
- *Our case:* the sign of $\dfrac{\partial r}{\partial x_i}$ is the same as the sign of $p_i \stackrel{\text{def}}{=} -(x_i - a) \cdot (x - a) - \sigma^2$.
- *Conclusion:* for all i for which $p_i = 0$, we have the same x_i.
- *Related partition:* I^- is when $x_i = \underline{x}_i$, I^+ is when $x_i = \overline{x}_i$, all others have the same $x_i = a_0$.

Title Page
Page 11 of 12
Go Back
Full Screen
Close
Quit

11. Acknowledgments

This work was supported in part:

- by NSF grants EAR-0225670 and DMS-0532645 and
- by Texas Department of Transportation grant No. 0-5453.