ATTRACTING ADDITIONAL INFORMATION FOR ENHANCING THE UNCERTAINTY MODEL

Towards Improved Risk Assessments

V.G. KRYMSKY

Industrial Electronics Department, Ufa State Aviation Technical University
12 K. Marx Street
Ufa, Bashkortostan 450000 Russia
kvg@mail.rb.ru

Abstract: The paper outlines a reasonable modification of an approach developed in the framework of imprecise prevision theory and adapted to the available information about some features of probability density functions. This reduces the uncertainty associated with risk analysis operations and as a result leads to obtaining the close interval estimations of statistical characteristics necessary for decision support.

1. Introduction

Reasonable use of available information on factors and phenomena is indeed the root principle of obtaining adequate risk assessments for effective decision support. As risk is typically considered in the form of composed probabilities of events and their consequences, the statistical model of the situation (scenario) is of great importance for achieving correct analytical results. Everybody who deals with risk analysis confirms that the level of uncertainty can be very high (this is caused by the lack of initial statistical data; data collection is poor because the events are rare). The only option is to elicit subjective information from experts [1]. However, we would like to use the most reliable expert judgements to derive a model with acceptable accuracy. This means that suitable but inaccurate assumptions are not allowed.

If the uncertainty is so radical that nothing can be said even about the distribution families related to events or influencing factors, then we face a problem statement in which all the distributions are plausible. This problem statement falls in the scope of the imprecise prevision theory (IPT), established in fundamental publications by Walley and Kuznetsov [2, 3]. IPT is unique in searching for at least some conclusions about the performance

I. Linkov et al. (eds.), Real-Time and Deliberative Decision Making.

© Springer Science + Business Media B.V. 2008

of extremely uncertain characteristics. The main advantage of IPT is its capacity to combine both objective statistical and expert information to estimate the lower and upper bounds of probabilities and other relevant data. Such estimates can be obtained without any assumptions of a specific prior distribution law by solving linear programming problems.

As has been demonstrated [4], the impediment to previous IPT methodology is that optimal solutions are defined for a family of degenerated distributions (in other words, distributions composed of δ-functions). The existence of solutions for degenerated distributions often leads to high imprecision, negating the pragmatic value of the assessments of interest (especially for risk analysis applications).

The negative issues associated with attempts to quantify uncertainty via IPT algorithms can be reduced by incorporating some additional information on model features. This paper discusses a strategy of enhancing the estimation technique by means of 'economic' addition of available information, which allows computing more precise bounds of the intervals for the resulting assessments.

2. Imprecise Previsions: Traditional Problem Statement

Let $\rho(x)$ be unknown probability density function of a continuous random variable X distributed in the interval $[0, T]$. Traditional IPT problem formulation (one-dimensional case) [2–4] considers the following constraints:

$$\rho(x) \geq 0, \int_0^T \rho(x)dx = 1, \text{ and } \underline{a_i} \leq \int_0^T f_i(x)\rho(x)dx \leq \overline{a_i}, \quad i = 1, 2, ..., n. \quad (1)$$

Here $f_i(x)$ are the given real-valued positive functions ("gambles") and $\underline{a_i}$, $\bar{a}_i \in R_+$ are the given numbers.

Computing the coherent lower and upper previsions $\underline{M}(g)$ and $\overline{M}(g)$ for expectation $M(g)$ of any function $g(x)$, which is also a gamble, requires estimating

$$\inf_{\rho(x)} \int_0^T g(x)\rho(x)dx, \text{ as well as } \sup_{\rho(x)} \int_0^T g(x)\rho(x)dx, \quad (2)$$

subject to constraints (1).

As is known [2, 3], optimization problem (1), (2) is of the linear programming type. So the main approach to searching for a corresponding solution involves forming a dual of initial problem statement. In turn such a dual can be easily solved in many practical cases.

The dual for optimization problem (1), (2) follows:

$$\underline{M}(g) = \sup_{c_0, c_i, d_i} \left(c_0 + \sum_{i=1}^{n} (c_i \underline{a}_i - d_i \bar{a}_i) \right) \tag{3}$$

subject to $c_0 \in \mathbf{R}$, c_i, $d_i \in \mathbf{R}_+$ and for any $x \geq 0$, $i = 1,2,\ldots,n$,

$$c_0 + \sum_{i=1}^{n} (c_i - d_i) f_i(x) \leq g(x). \tag{4}$$

and

$$\overline{M}(g) = \inf_{c_0, c_i, d_i} \left(c_0 + \sum_{i=1}^{n} (c_i \bar{a}_i - d_i \underline{a}_i) \right), \tag{5}$$

subject to $c_0 \in \mathbf{R}$, $c_i, d_i \in \mathbf{R}_+$ and for any $x \geq 0$, $i = 1,2,\ldots,n$,

$$c_0 + \sum_{i=1}^{n} (c_i - d_i) f_i(x) \geq g(x). \tag{6}$$

Investigation [4] shows that function $\rho(x)$ for which $M(g)$ attains the values of $\underline{M}(g)$ or $\bar{M}(g)$ belongs to a family of degenerated distributions (this density is composed of δ-functions). This undesirable fact is like a "payment" for reasoning under too high a level of uncertainty. Very often we may incorporate some limited additional information (typically elicited from experts), which has the capacity to provide more valuable analytical results. One possible method has been described previously [5].

3. The Case of Bounded Densities

The first portion of additional information which allows achieving improvement when solving the optimization problem (1), (2) is presented in the form of the bounded probability densities. To get these data we have to ask an expert questions like "What is the largest possible percentage of accidents per year/decade for a given plant with definite age?" The resulting judgement is reflected by inequality:

$$\rho(x) \leq K = const, \tag{7}$$

where K is a real positive number satisfying the condition $KT \geq 1$.

New problem formulation requires optimizing the objective function (2) subject to constraints (1), (7). This problem can be solved via the methods of the calculus of variations [5]. The resulting optimal density function

becomes a member of a family of step-functions equal to either zero or K (so degenerated solutions are eliminated). This leads to much more precise previsions (numerical examples confirm improvement of 50% in estimating the upper and the lower bounds of $M(g)$).

The knowledge of the solution type creates an opportunity for reducing the initial problem that belongs to scope of the calculus of variations to the easier-to-solve problem of optimizing a multivariable function subject to algebraic constraints.

Indeed, denote the intervals $[x_0,x_1)$, $[x_2,x_3)$, $[x_4,x_5)$,…, where $\rho(x) = K \neq 0$. Also denote

$$G(x_j, x_{j+1}) = \int_{x_j}^{x_{j+1}} g(x)dx; \tag{8}$$

$$\Phi_i(x_j, x_{j+1}) = \int_{x_j}^{x_{j+1}} f_i(x)dx, \quad i = 1,2,...,n. \tag{9}$$

Then we can reformulate our optimization problem:

$$\min_{x_0,x_1,\ldots}\left\{K \cdot \sum_{j=0}^{m} G(x_{2j}, x_{2j+1})\right\} \text{and} \max_{x_0,x_1,\ldots}\left\{K \cdot \sum_{j=0}^{m} G(x_{2j}, x_{2j+1})\right\} \tag{10}$$

$$K \cdot \sum_{j=0}^{m} (x_{2j+1} - x_{2j}) = 1; \tag{11}$$

$$\underline{a_i} \le K \cdot \sum_{j=0}^{m} \Phi_i(x_{2j}, x_{2j+1}) \le \overline{a_i}, \quad i = 1,2,...,n. \tag{12}$$

To solve such multivariable optimization problems in the general case, we can apply a lot of numerical methods like gradient algorithms, simplex-planning search, and genetic algorithms. In some simple situations, a solution can be reached in analytical form.

The remaining question is, how to choose the value of m? Very often we don't know this value *a priori*.

The recommendation for these situations is as follows: start from small values of m (e.g., set $m = 0$) to solve the optimization problem. The value of m can be increased ($m = 1$), continuing to solve the problem. The process can be stopped if the step-function for $\rho(x)$ begins retaining its form (this means that newly introduced intervals become the same as for the previous value of m). This finalizes the process of seeking the resulting assessment.

4. The Case of Bounded Modules of Density Derivatives

The next additional portion of information can be represented by constraints related to the maximum values of the density derivatives [6]. Sometimes it is realistic to elicit these data from experts by asking them a question like "What is the largest possible difference between the percentages of accidents computed for two neighboring years/decades for a given plant with definite age?"

Let us denote $M \in \mathbf{R}_+$ an upper bound on the values of the probability density derivative module; i.e., for $\forall x$

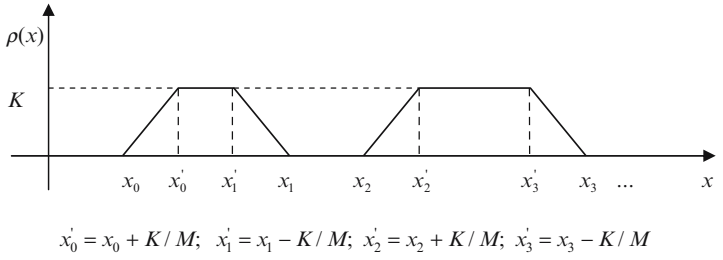
$$\left| d\rho(x)/dx \right| \le M = const. \tag{13}$$

Now we have to optimize the objective function (2) subject to constraints (1), (7) and (13). This is also a problem that can be solved via the methods of the calculus of variations (very similar to the approach described in [5]). This shows that optimal density functions belong to a family of trapezoid—or triangular—functions (Figure 1). Correspondingly the intervals for the final assessments are expected to be closer as the speed of density change is constrained. Another effect of recognizing the form of the optimal solution is the possibility of reducing the initial problem to an easier-to-solve optimization of a multivariable function subject to algebraic constraints (as was done above).

Indeed, let $[x_0,x_1),[x_2,x_3),[x_4,x_5),\ldots,[x_{2m},x_{2m+1})$ be the intervals that play the role of the trapezoid lower bases. It is easy to see that the trapezoid upper bases for which $\rho(x) = K$ are located within the intervals

$$[x_0 + K/M, x_1 - K/M),[x_2 + K/M, x_3 - K/M),\ldots,[x_{2m} + K/M, x_{2m+1} - K/M)$$

Let $[x_1, x_2)$, $[x_3, x_4)$, $[x_5, x_6),\ldots,[x_{2m+1},x_{2m+2})$ be the intervals on which $\rho(x) = 0$ (Figure 1).



$$x_0' = x_0 + K/M;\ \ x_1' = x_1 - K/M;\ \ x_2' = x_2 + K/M;\ \ x_3' = x_3 - K/M$$

Figure 1. The Plot of Optimal Bounded Density with Bounded Module of Derivative.

Then the problem statement can be easily reformulated in relation to optimizing the multivariate function, which depends on the variables $x_0, x_1, \ldots, x_{2m+1}$.

The choice of a value for m can be accomplished as was done above. Specifically, we can start from minimum values of m (e.g., $m = 0$) and then try to increase it with testing in parallel if change to m is followed by a change to optimal density.

The numerical examples confirm the improvement of up to 75% in resulting accuracy. However, the incentive to find a more promising model for uncertainty remains, as the bounded densities and bounded density derivatives lead to partially unrealistic solutions: the densities are equal to zero for some argument intervals, which means that the 'probabilistic mass' is concentrated in separate zones. This explains the desire to restrict ourselves by considering only the class of smooth differentiable density functions.

5. Application of Generalized Distribution Family

Let us introduce a family of distribution densities described by smooth differentiable functions as

$$\rho(x) = \left(\sum_{k=1}^{n} C_k \exp(-\alpha_k x) \right)^2 = \sum_{k=1}^{n} C_k^2 \exp(-2\alpha_k x) + 2 \sum_{\substack{l \neq r \\ l=1,r=1}}^{n} C_l C_r \left(\exp -(\alpha_l + \alpha_r) x \right), \quad (14)$$

in which C_k, $\alpha_k \geq 0$, $k = 1,2,\ldots,n$, are real numbers satisfying the condition

$$\sum_{k=1}^{n} C_k^2 / (2\alpha_k) + 2 \sum_{\substack{l \neq r \\ l=1,r=1}}^{n} C_l C_r / (\alpha_l + \alpha_r) = 1. \quad (15)$$

It is easy to verify that $\rho(x) \geq 0$ and $\int_0^\infty \rho(x) dx = 1$.

Now consider the previsions that may be given for $\int_0^\infty f_i(x) \rho(x) dx, i = 1,2,\ldots,m.$ In the case where $\rho(x)$ satisfies (14) we obtain

$$\int_0^\infty f_i(x) \rho(x) dx = \sum_{k=1}^{n} C_k^2 F_i(s)\Big|_{s=2\alpha_k} + 2 \sum_{\substack{l \neq r \\ l=1,r=1}}^{n} C_l C_r F_i(s)\Big|_{s=\alpha_l+\beta_r}, i = 1,2,\ldots,m, \quad (16)$$

in which $F_i(s)$, $i = 1,2,\ldots,m$, are Laplace transformed functions $f_i(x)$.

Note that Laplace transform $F_i(s)$ for any continuous function $f_i(x)$ is introduced by the expression

$$F_i(s) = \int_0^\infty f_i(x)\exp(-sx)dx,$$

in which s is the Laplace variable (which may take complex values in general case: $s = \mathrm{Re}(s) + j \cdot \mathrm{Im}(s)$; here $\mathrm{Re}(s)$, $\mathrm{Im}(s)$ denote real and imaginary parts of s respectively).

Meanwhile it is proven by D.V. Widder [7] that if we know the performances of $F_i(s)$ for real positive values of s then we have a unique expansion of its behavior to the whole complex plane of s values.

The tables containing results of the Laplace transformation for different functions are widely presented in relevant literature.

For instance, if $f_i(x) = x$ then $F_i(s) = 1 / s^2$; if $f_i(x) = x^2$ then $F_i(s) = 2 / s^3$, etc.

Hence, we can write the following general formulae for expectation and variation:

$$E[X] = \int_0^\infty x\rho(x)dx = \sum_{k=1}^n C_k^2/(4\alpha_k^2) + 2\sum_{\substack{l\neq r\\ l=1,r=1}}^n C_l C_r/(\alpha_l+\alpha_r)^2, \tag{17}$$

$$Var[X] = \int_0^\infty x^2\rho(x)dx = \sum_{k=1}^n C_k^2/(8\alpha_k^3) + 2\sum_{\substack{l\neq r\\ l=1,r=1}}^n C_l C_r/(\alpha_l+\alpha_r)^3. \tag{18}$$

In turn probability of the event $X \le x$ can be found as

$$P(X \le x) = \int_0^x \rho(x)dx = \sum_{k=1}^n \frac{C_k^2}{2\alpha_k}\left(1-\exp(-2\alpha_k x)\right) + 2\sum_{\substack{l\neq r\\ l=1,r=1}}^n \frac{C_l C_r}{\alpha_l+\alpha_r}\left(1-\exp\left(-(\alpha_l+\alpha_r)x\right)\right). \tag{19}$$

Formulae (17)–(19) allow the reduction of typical IPT problems to easier-to-solve standard problems that belong to the scope of optimizing nonlinear multivariable functions depending on the values of C_k, α_k, $k = 1,2,\ldots,n$.

For instance, if we would like to estimate $P(X \le x_0)$ for given x_0 on the basis of interval previsions for moments (17), (18), then we have to substitute x_0 instead of x into objective function (6) and search for $\max\limits_{C_k,\alpha_k} P(X \le x_0)$ and $\min\limits_{C_k,\alpha_k} P(X \le x_0)$ subject to constraints (1) as well as

$$\underline{E}[X] \le E[X] \le \overline{E}[X],\ \underline{Var}[X] \le Var[X] \le \overline{Var}[X]. \tag{20}$$

Here $\underline{E}[X], \underline{Var}[X], \bar{E}[X], \overline{Var}[X]$ are the lower and the upper bounds of the intervals for the values of the moments respectively.

An important particular case of the introduced distributions which can be obtained if we consider only two exponential terms in the sum for $\sqrt{\rho(x)}$ in equality (16) is analyzed below.

Consider the case in which

$$\rho(x) = \left(C_1 \exp(-\alpha x) + C_2 \exp(-\beta x)\right)^2 = C_1^2 \exp(-2\alpha x) + C_2^2 \exp(-2\beta x) + 2C_1C_2 \exp\left(-(\alpha+\beta)x\right). \quad (21)$$

Here $C_1, C_2, \alpha \geq 0$, $\beta \geq 0$ are the distribution parameters.

First, analyze which type of statistical characteristic behavior can be presented by Expression (21).

If C_1 and C_2 are of the same sign (i.e., $C_1C_2 \geq 0$), then Expression (21) corresponds to monotonic density functions (Figure 2).

Note that function behavior like of $\rho^{\bullet}(x)$ is more typical for different nonzero values of C_1, C_2 and $C_1C_2 \geq 0$; the behavior reflected by $\rho^{*\bullet}(x)$ ('pure' exponential type) takes place if $C_1 = 0$ or $C_2 = 0$. The last situation appears also if $\alpha = \beta$.

If C_1 and C_2 have different signs (i.e. $C_1C_2 \leq 0$) then Expression (21) may correspond to nonmonotonic density functions (Figure 3).

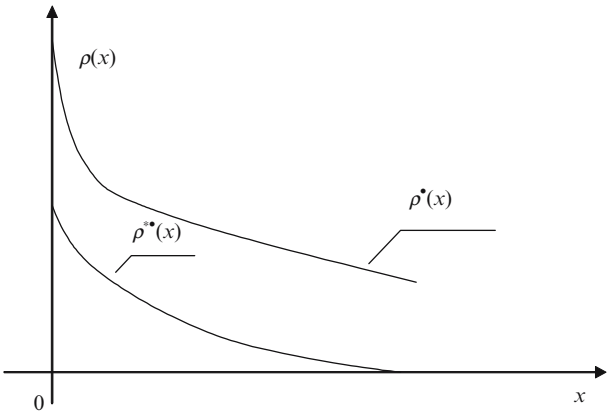


Figure 2. Types of Density Functions Presented by Expression (21) if $C_1C_2 \geq 0$.

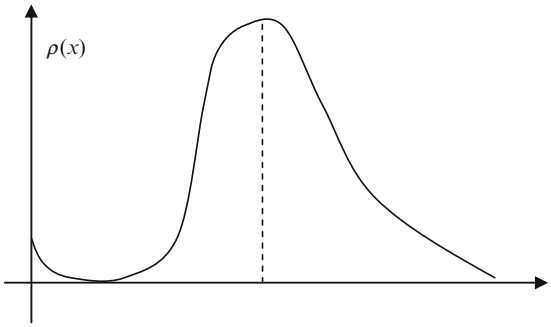


Figure 3. General form of Density Functions Presented by Expression (21) if $C_1C_2 < 0$.

It becomes clear that the introduced family of density functions covers a wide class of practically important distribution types (unimodal and even antimodal).

Using this generalized distribution family allows reduction of the problem to optimization of the objective function (2) as a multivariable function depending on the initially unknown parameters C_k, $\alpha_k \geq 0$, $k = 1,2,\ldots,n$.

The improvement achieved for numerical assessments becomes somewhat higher than for cases of nonsmooth density functions.

6. Concluding Remarks

Some reasonable data elicited from experts and accessible for verification can significantly improve the decisions made under the conditions of uncertainty. Adding information on density bounds, density derivative bounds, or any generalized form of distribution function with unknown parameters is, on the one hand, a kind of reasonable enhancement and, on other hand, does not actually restrict us in taking into account possible (probable) scenarios. Meanwhile, this technique provides promising modification of traditional approaches based on imprecise previsions and creates a bridge between the strict concepts of the corresponding theory and practical needs for assessment accuracy.

The proposed methodology opens the door for next steps associated with incorporating additional information elicited from experts. Thus it makes sense sometimes to ask an expert if s/he is ready to give preferences about some kinds of data. By presenting these preferences in the form of subjective probabilities, we have an opportunity to compute the expectations of the upper and the lower bounds derived for previsions. In turn this strengthens support for responsible decisions in the framework of risk analysis.

Acknowledgements

Dr. Igor Kozine of Risø National Laboratory, Denmark, introduced me to the world of interval-valued probabilities and helped me with his valuable comments. Dr. Igor Linkov of Intertox Inc., USA, was among the key organizers of NATO Advanced Research Workshop at Lisbon, so his kind invitation to participate was the necessary condition which allowed me to join the team of this book's authors. The contribution of both Igors to supporting my research activity is gratefully acknowledged.

References

1. Cooke R.M., *Experts in Uncertainty*, Oxford University Press, Oxford, 1991.
2. Walley P., *Statistical Reasoning with Imprecise Probabilities*, Chapman & Hall, New York, 1991.
3. Kuznetsov V., *Interval Statistical Models*, Radio & Sviaz, Moscow, 1991 (in Russian).
4. Utkin L.V., Kozine I. O., Different faces of the natural extension, *Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, ISIPTA '01, 2001, pp. 316–323.
5. Krymsky V.G., Computing Interval Bounds for Statistical Characteristics Under Expert-Provided Bounds on Probability Density Functions, *Applied Parallel Computing, State of the Art in Scientific Computing, Revised Selected Papers of 7th International Workshop PARA'04,* Lecture Notes in Computer Science, Springer, 2006, pp.151–160.
6. Kozine I.O., Krymsky V.G., Enhancement of Natural Extension, *Proceedings of the Fifth International Symposium on Imprecise Probabilities: Theories and Applications (ISIPTA'2007)*, Action M Agency, Prague, 2007, pp.253–262.
7. Widder D.V., *The Laplace Transform*, Princeton University Press, Princeton, NJ, 1972.