juliacon

ProbabilityBoundsAnalysis.jl: an arithmetic of probabilities and sets of probabilities

Ander Gray[1], Scott Ferson[1], and Edoardo Patelli[1, 2]

[1]Institute for Risk and Uncertainty, University of Liverpool
[2]Centre for Intelligent Infrastructure, University of Strathclyde

ABSTRACT

Probability Bounds Analysis combines interval arithmetic with probability theory, and provides a representation of sets of distributions in structures called probability boxes (p-boxes), which generalise both distribution functions and intervals. P-boxes generally return interval bounds on all probabilistic quantities, for example samples, cdfs, and probability measures are all intervals. This framework also allows for the comprehensive propagation of probabilities through calculations in a rigorous way, in a similar fashion that interval arithmetic does for sets. As such, ProbabilityBoundsAnalysis.jl gives a rigorous arithmetic of random variables, where both marginal (univariate) and dependency information can be known, partially known or missing completely.

Keywords

Julia, Probability, Uncertainty, Interval Arithmetic, Probabilistic Arithmetic

1. Introduction

An arithmetic of probability distributions has held a long interest among mathematicians and scientists:

> A question asked by Kolmogorov,
> answered for the sum by Marakov,
> partially answered by Sklar and Frank (for which the copula was invented),
> made algorithmically available by Williamson,
> and generalised by others.

Indeed it was Kolmogorov who originally asked what the result of a sum of two distributions without knowing their joint distribution. This was answered for the sum by Marakov [14], who showed the result was a set of distributions and was able to provide bounds on this function. Sklar, Schweizer and Frank generalised this result to other (positive) binary operations [9, 17]. In this pursuit they created copulas, a general way to encode probabilistic dependence independently from marginals, and a now essential object used in probabilistic modelling. In his seminal dissertation [21], Williamson described an algorithm for efficiently performing these arithmetic operations, which can give guaranteed bounds on probability distributions in terms of an upper and lower cdf. He called his method *Probabilistic Arithmetic* and his sets of distributions *Dependency Bounds*. Since then the method has been generalised [8, 6, 7] to most of the base binary and unary operations that would be present in a programming language. Probability Boxes (p-boxes) are the name now given to these structures, and Probability Bounds Analysis (PBA) the name of the method.

The goal of PBA can be stated as **to compute guaranteed bounds on functions of random variables given only partial knowledge of the input probability distributions and their dependencies**. That is to compute with partial knowledge about the input joint distribution. Ideally all of the available information about random variables should be used, but no more than what actually is available.

The idea of bounding probability has a very long tradition throughout the history of probability theory. George Boole [2, 11] used the notion of interval bounds on probability. Chebyshev [3] described bounds on a distribution when only the mean and variance of the variable are known, and Markov [15] found bounds on a positive variable when only the mean is known. Fréchet [10] discovered how to bound joint distributions solely from knowing the marginal distributions, without making independence assumptions. Bounding probabilities has continued to the present day, culminating into the modern theory of Imprecise Probabilities [20, 13, 19, 1].

Imprecise probabilities is effectively a generalisation of probability theory where uncertainty can be expressed about the probability measure. This is particularly relevant when information is scarce, unreliable, vague, conflicting or imprecise. In such cases defining a unique probability distribution is difficult. P-boxes are one of many ways to describe a set of distributions, others include: Dempster-Shafer structures [4, 18], random sets [16], possibility distributions [22, 5, 12] and lower previsions [19]. These structures were discovered independently, but are often synonymous and can be translated from one to another, with different degrees of generality. Imprecise probabilities links all these theories into one. For a comprehensive overview of the theory, and a for a formal description of uncertainty and information in terms of these structures, [13] is recommended. In that sense PBA is a part of imprecise probabilities but provides a framework for computing with p-boxes.

2. Probability Boxes

A probability box defines a set of distributions with the following three constrains: (1) interval bounds on the cumulative distribution function (cdf), (2) interval bound on the mean and variance, and (3) a collection of distribution families:

(1) $\underline{F}(x) \leq F(x) \leq \overline{F}(x)$

(2) $\mu \in [\underline{\mu}, \overline{\mu}]$

$\sigma^2 \in [\underline{\sigma}^2, \overline{\sigma}^2]$

(3) $F \in \mathbf{F}$

That is, a random variable is a member of a p-box if its cdf F falls within the cdf bounds of the p-box $F(x) \in [\underline{F}(x), \overline{F}(x)]$ for all x, its moments are inside the interval moments of the p-box, and it belongs to a family of distribution functions (e.g. normal, uniform) considered by the p-box. Some of the constraints may be missing. For example, if the distribution family is unknown then the set is defined solely from the cdf and moment bounds. Such p-boxes are sometimes called non-parametric, since its members do not have to belong to any particular class of distribution. Some constraints may also be inferred from others. For example, the interval moments may be bounded from the cdf bounds, and cdf may be bounded from moment information (explored further in section *2.2.2*).
All of a p-box's probabilistic quantities are intervals. The cdf of a p-box may be found by:

$$[\underline{F}(x), \overline{F}(x)],$$

a sample of the p-box may be drawn using the inverse cdfs:

$$[\underline{F}^{-1}(\alpha), \overline{F}^{-1}(\alpha)]$$

where $\alpha \sim U(0,1)$ is a sample from a uniform distribution, and the probability measure[1] on some interval $U = [a, b]$ is bounded as follows:

$$\underline{\mathbb{P}}(U) = \max(0, \underline{F}(b) - \overline{F}(a))$$
$$\overline{\mathbb{P}}(U) = \overline{F}(b) - \underline{F}(a),$$

where the max operator is required when $\underline{F}(b) < \overline{F}(a)$. Note the same can be achieved by using the standard formula for finding the measure from the cdf, $F(b) - F(a)$, and using interval arithmetic. P-boxes generalise precise distributions and intervals in the following way. A distribution is a p-box with a precise cdf and moments, i.e. when:

$$\underline{F}(x) = \overline{F}(x),$$
$$\underline{\mu} = \overline{\mu},$$
$$\underline{\sigma}^2 = \overline{\sigma}^2$$

and an interval $[a, b]$ is a p-box whose bounds are step functions:

$$\underline{F}(x) = \epsilon_b(x)$$
$$\overline{F}(x) = \epsilon_a(x),$$

where ϵ_k is:

$$\epsilon_k(x) = \begin{cases} 0 & \text{when } x < k \\ 1 & \text{when } x \geq k \end{cases}$$

Moreover, theoretical bounds on the mean and variance of an interval can be found [CITE SCOTT]:

$$\mu \in [a, b]$$
$$\sigma^2 \in [0, (b-a)^{2/4}]$$

[1]The probability measure is a function which returns the probability that the random variable is in some set

That is, it is not possible to find a distribution whose range is in $[a, b]$ and whose variance is greater than $(b-a)^{2/4}$. The lower bound on the variance is zero, since scalars are also included in the interval. Under this definition of an interval, a random sample will always be the interval $[a, b]$, the cdf returns 0 when $x < a$, the interval $[0, 1]$ when $a <= x < b$, and 1 when $b <= x$. Further, the probability measure of the interval $X = [a, b]$ will be:

$$\mathbb{P}_X(U) = \begin{cases} 0 & \text{when } U \cap [a,b] = \emptyset \\ 1 & \text{when } [a,b] \subseteq U \\ [0,1] & \text{otherwise} \end{cases}$$

That is, if the set U does not intersect the interval X, $\mathbb{P}_X = 0$, i.e. U certainly does not contain any of the random variables. If the X is fully contained in U, $\mathbb{P}_X = 1$, i.e. U certainly contains all of the random variables. And finally if U intersects (but does not contain) X, the probability measure is the vacuous probability interval $\mathbb{P}_X = [0, 1]$, i.e. we are completely uncertain about the containment. Note that the $P_X = \{0, 1, [0, 1]\}$ is due to the interval bounds being step function. Generally p-boxes may yield any probability interval.

2.1 Outer representations of p-boxes

An important feature of probability bounds analysis is how distribution functions and p-boxes are represented. Generally, analytical solutions for cdfs are not readily available, for example the normal distribution's cdf can only be found by integrating the density, usually done with Gaussian quadrature. However, even if the functions were available analytically, when the variables are used in an arithmetic operation the output distribution will not necessarily belong to the same family. Therefore we require a representation of these continuous functions which is not dependent on the distribution family (for example Polynomial chaos expansion), and ideally is robust.

2.2 Where do p-boxes come from?

In this section we discuss some situations where p-boxes arise naturally.

2.2.1 Partial distributional information.

2.2.2 Partial moment information.

2.2.3 Operations involving intervals and distributions.

2.2.4 When dependencies between distributions are uknown.

2.2.5 Outer approximations of precise distributions.

2.2.6 Inferential methods where data is limited or bad.

2.3 Relationship to other ideas

2.3.1 Random set theory.

2.3.2 Possibility theory.

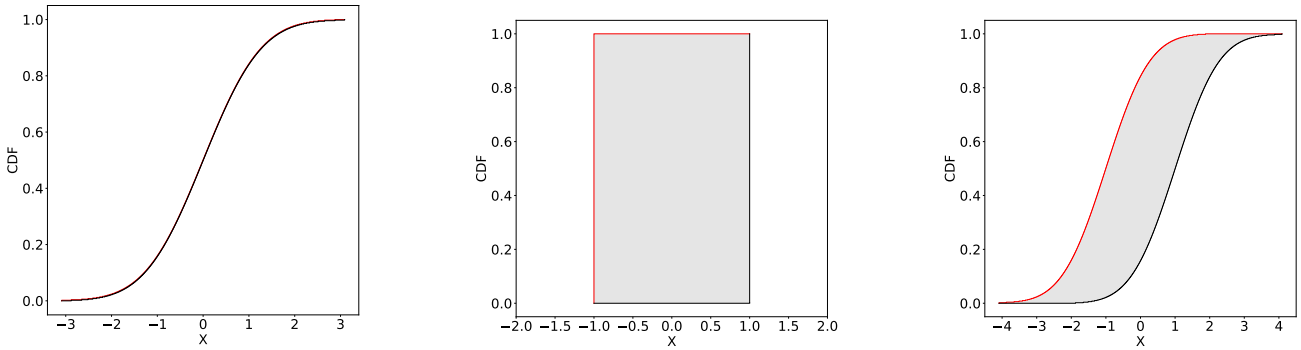

Fig. 1: A precise distribution, an interval and a probability box

2.4 Bivariate p-boxes

3. P-box arithmetic

3.1 Uniary operations

4. An uncertain programming language

The long term goal of such a framework is to create a fully uncertain programming language, where any computer variable may be represented as an interval, distribution, p-box or other uncertain quantity. Such a framework would allow for uncertain extensions of deterministic functions to be computed in an automatic, rigorous and tight fashion. In this section we argue why Julia is an ideal target language for such a framework, and discuss some of the remaining theoretical tasks required to make such a goal a reality.

Acknowledgements

The authors would like to thank the support of this work from the Engineering Physical Sciences Research Council (EPSRC) iCase studentship award. This research is funded by the EPSRC with grant number EP/R006768/1, which is acknowledged for its funding and support. This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014–2018 and 2019–2020 under grant agreement number 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission

5. References

[1] Thomas Augustin, Frank PA Coolen, Gert De Cooman, and Matthias CM Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.

[2] George Boole. *An investigation of the laws of thought: on which are founded the mathematical theories of logic and probabilities*, volume 2. Walton and Maberly, 1854.

[3] Pafnutii L. Chebyshev [Tchebichef]. *Sur les valeurs limites des intégrales*. Imprimerie de Gauthier-Villars, 1874.

[4] Arthur P Dempster. Upper and lower probabilities induced by a multivalued mapping. In *Classic works of the Dempster-Shafer theory of belief functions*, pages 57–72. Springer, 2008.

[5] D Dubois and H Prade. possibility theory: Approach to computerized processing of uncertainty, plennm n. 4. 1988.

[6] Scott Ferson, Lev Ginzburg, and Resit Akçakaya. Whereof one cannot speak: when input distributions are unknown. *Risk Analysis*, 1996.

[7] Scott Ferson and Janos G Hajagos. Arithmetic with uncertain numbers: rigorous and (often) best possible answers. *Reliability Engineering & System Safety*, 85(1-3):135–152, 2004.

[8] Scott Ferson, Vladik Kreinovich, Lev Grinzburg, Davis Myers, and Kari Sentz. Constructing probability boxes and dempster-shafer structures. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2015.

[9] Maurice J Frank, Roger B Nelsen, and Berthold Schweizer. Best-possible bounds for the distribution of a sum—a problem of kolmogorov. *Probability theory and related fields*, 74(2):199–211, 1987.

[10] Maurice Fréchet. Généralisation du théoreme des probabilités totales. *Fundamenta mathematicae*, 1(25):379–387, 1935.

[11] Theodore Hailperin. *Boole's logic and probability: a critical exposition from the standpoint of contemporary algebra, logic and probability theory*. Elsevier, 1986.

[12] Dominik Hose and Michael Hanss. Possibilistic calculus as a conservative counterpart to probabilistic calculus. *Mechanical Systems and Signal Processing*, 133:106290, 2019.

[13] George J Klir and Mark J Wierman. *Uncertainty-based information: elements of generalized information theory*, volume 15. Physica, 2013.

[14] GD Makarov. Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory of Probability & its Applications*, 26(4):803–806, 1982.

[15] A Markov [Markoff]. Sur une question de maximum et de minimum: Proposée par m. tchebycheff. *Acta Mathematica*, 9:57–70, 1900.

[16] Ilya Molchanov and Ilya S Molchanov. *Theory of random sets*, volume 19. Springer, 2005.

[17] Berthold Schweizer and Abe Sklar. *Probabilistic metric spaces*. Courier Corporation, 2011.

[18] Glenn Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

[19] Matthias CM Troffaes and Gert De Cooman. *Lower previsions*. John Wiley & Sons, 2014.

[20] Peter Walley. *Statistical reasoning with imprecise probabilities*. Chapman and Hall, London, 1991.

[21] Robert Charles Williamson et al. *Probabilistic arithmetic*. PhD thesis, University of Queensland Australia, 1989.

[22] Lotfi Asker Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems*, 1(1):3–28, 1978.