

Setting Cleanup Targets in a Probabilistic Assessment

W. Troy Tucker, David S. Myers, and Scott Ferson

Applied Biomathematics, 100 North Country Road, Setauket, NY 11733; (631) 751-4350; (631) 751-3435 (fax); troy@ramas.com

Abstract

Risks from environmental contaminants in water are estimated with a risk equation involving contaminant concentration and other factors. Because the intent of environmental remediation is to ensure that these risks are not intolerably large, some way is needed to backcalculate from tolerance limits on risk mandated by regulators to the allowable environmental concentration for the contaminant. It is now well known that the approach used in deterministic assessments of simply inverting the risk equation to compute the cleanup goal does not work in a probabilistic assessment. We introduce simple and efficient methods for computing cleanup goals satisfying multiple simultaneous criteria in the context of a probabilistic assessment. This approach uses probability bounds analysis to characterize the concentration distributions and can be used with multiple receptors with arbitrarily many constraints on percentiles of the target risk distribution. The calculations yield two kinds of bounds on concentration: a 'kernel' and 'shell'. If the concentration distribution is entirely inside the kernel, then the result surely obeys the prescribed constraints. If the concentration distribution is anywhere outside the shell, then the result certainly fails to comply with the prescribed constraints. If the concentration distribution is outside the kernel but inside the shell, then compliance must be determined by a forward calculation.

Introduction

Risk analysts estimate exposure to environmental contaminants or health risks from such exposures with equations that include contaminant concentrations and other input variables. Analysts charged with calculating cleanup targets need to compute contaminant concentrations that satisfy regulatory constraints while minimizing cleanup costs. In the past, point estimates were used as inputs to these risk assessments and safe environmental concentrations solved by simply inverting the dose or risk equation. In a probabilistic risk assessment, however, which replaces deterministic point estimated inputs with probability distributions, it is not possible to simply rearrange the equations to solve for these concentrations. Consider, for example, a highly simplified exposure model embodied in the equation

$$Dose = Concentration \times Intake.$$

This equation permits one to estimate the realized dose of a contaminant by multiplying the observed concentration of the contaminant in some environmental medium such as air or water by the intake rate of that medium such as by inhalation or imbibition. In a probabilistic risk assessment, the deterministic point estimates for concentration and intake are replaced by probability distributions to reflect the fact that concentration varies across space and through time, and that exposed people may have different patterns of breathing or drinking. Monte Carlo methods are often used in these assessments because they provide a convenient and flexible way

to estimate the distribution of dose that results from combining these distributions of concentration and intake.

But what about the backcalculation problem of determining the allowable environmental concentration that will ensure the received dose is not so large as to be hazardous? When the variables in the equation represent simple numbers, one can solve for the not-to-exceed concentration simply by dividing a not-to-exceed dose by the intake rate. However, when the variables are probability distributions, computing the concentration simply by dividing the dose distribution by the intake distribution yields a result that cannot be the answer (as is easy to show by substituting the result back into the original equation). In fact, without the essential pairwise information that ties together which dose goes with which intake, there is fundamentally no way to use straightforward Monte Carlo methods to estimate the desired distribution of concentrations. What is needed is a special backcalculation operation that solves the probabilistic equation.

The problem of performing backcalculation on variables with uncertainty can be seen as a specific case of mathematical deconvolution. Whereas the forward problem of adding or multiplying random variables can be considered an application of convolutions of distributions (Springer 1979; Morgan and Henrion 1990), the backward problem of deconvolution is not so simple. Although deconvolution methods to solve backcalculation problems are well known in many disciplines (e.g., Jansson 1984; Bernabini et al. 1987; Wunsch 1996), the difficulty inherent in untangling risk assessment equations involving probabilistic uncertainty was first articulated only recently (Ferson 1995; Ferson and Long 1995; Burmaster et al. 1995; Burmaster and Thompson 1995; Ferson 1996; Ferson and Long 1998). Recognition has now, however, become widespread within the risk analysis community that the defining equations cannot simply be rearranged to solve for tolerable concentrations of environmental contaminants.

There are, of course, special cases for which solutions are known. For example, if the right-hand side of the equation is a product of independent lognormal distributions, the solution is also a lognormal distribution and it can be specified immediately. Likewise, if the right-hand side is a sum of normal distributions, the solution can simply be written down. For instance, suppose we're trying to solve for B in the equation $C = A + B$ and all the variables are normally distributed with A independent of B . Given that C has a larger variance than A , the answer has a mean of $E(C) - E(A)$ and variance of $\text{var}(C) - \text{var}(A)$ where E denotes expectation and var denotes variance. Since it's normal, these fully specify the distribution of B . An answer exists so long as the distribution of C is broader than that of A . More generally, an equation in any stable family of distributions can be solved quite easily. It is also possible to solve the equations if the dependence between any two of the three variables is linear and known. Unfortunately, these special cases do not generalize to nonlinear dependencies, or to arbitrary distribution families, or to combinations of distributions from different families. This therefore strongly limits their utility in real-world problems where complex distribution shapes and nonlinear dependencies are very common.

Probability Bounds Analysis

Extending the approach of Frank et al. (1987), Williamson and Downs (1990) developed a semi-analytical approach that computes rigorous bounds on the cumulative distribution functions of

convolutions without necessarily assuming independence between the operands. They demonstrated how the method can be used to estimate convolutions in the strict sense (addition) as well as other convolution-like operations (subtraction, multiplication and division). We implemented their approach in software (Ferson 2002) and extended it to transformations such as logarithms and square roots, and other convolutions such as minimum, maximum and powers. Because their approach uses bounds to represent discretization and dependency errors, it can also account for uncertainty about the shape of input distributions themselves. Because it does not require an analyst to assume independence when it is not warranted (Ferson and Long 1995) or to specify the precise shapes of input distributions when they are difficult to estimate, this approach is more appropriate for use in risk assessments than analogous Monte Carlo methods when empirical information is sparse.

We have developed efficient algorithms that can be used to backcalculate equations like $Dose = Concentration \times Intake$. The algorithms untangle the analytical convolution proposed by Williamson and Downs (1990), and are easily tested by applying the Williamson and Downs convolutions in the forward direction. Thus, we can check whether the computed bounds on the concentration distribution result in any probabilities of doses at any magnitude that exceed the planned constraints for the dose distribution. The algorithms are quite general in that they work for products, sums, differences or quotients and do not require assumptions about the dependency or correlations between any of the variables. Furthermore, and most importantly, they can handle arbitrarily many simultaneous constraints on the dose distribution. This means we can control all percentiles of the dose distribution to ensure a tolerable result.

Backcalculation with Probability Bounds

Distributions may be characterized in the context of a probabilistic assessment by any number of simultaneous constraints. For instance, suppose that public health officials believe that the median dose received by humans should be no larger than 10 units, that the 95th percentile of doses should be no larger than 30, and that no one in the population should receive a dose larger than 100. There is no further guidance available, except that smaller doses are better. Such constraints are typical of the kind of specifications made in human health risk assessments today. The region depicted in Figure 1 represents bounds on the distributions of dose that meet the

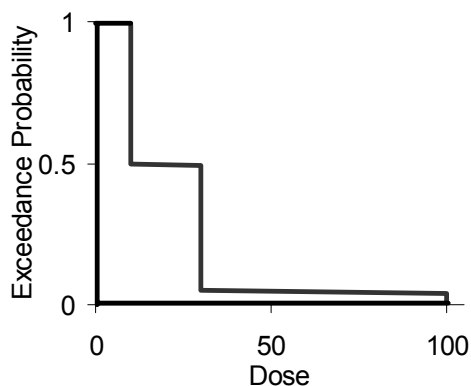


Figure 1. P-box around all allowable dose distributions.

specified constraints. The area enclosed by these bounds is called a *probability-box*, or *p-box* for short (Ferson 1997, 2002). The p-box in the figure is drawn in complementary cumulative distribution form. This means that the vertical axis gives one minus the cumulative probability; it is the chance the variate is larger than the corresponding value on the horizontal axis. Any actual dose distribution whose complementary cumulative distribution function resides entirely within the p-box would be acceptable according to the given dose constraints.

Suppose that the distribution of intake rates looks like that shown in Figure 2 (also drawn in complementary cumulative form). The information in the two graphs

can be combined to draw conclusions about what the concentration distribution must be like. A distribution of concentrations is a ‘solution’ if it results in a distribution of doses that satisfy the dose constraints. Obviously, there can be more than one solution to this problem. Instead of picking one solution as representative of all solutions or some other ad hoc method, a p-box can be formed which envelopes all the possible (complementary) cumulative distribution functions.

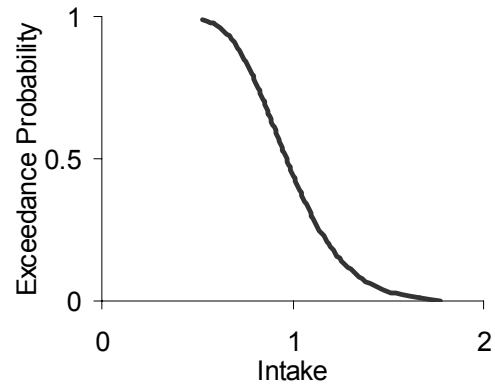


Figure 2. Intake distribution.

Two kinds of p-boxes can bound the concentration distribution: a ‘kernel’ and a ‘shell’. The shell is calculated numerically by a forward application of probability bounds analysis (Williamson and Downs 1990; Ferson 2002). This region is unique and encompasses the entire space of all solutions to the problem. The shell associated with the dose constraints specified in Figure 1 and the intake distribution shown in Figure 2 is depicted as the thin curve in Figure 3. All solutions will have distribution functions that lie within the shell, that is, below or to the left of the thin line in Figure 3. It is not the case, however, that every distribution inside the shell will be a solution. In general there could be many concentration distributions within the shell that fail to satisfy the dose constraints.

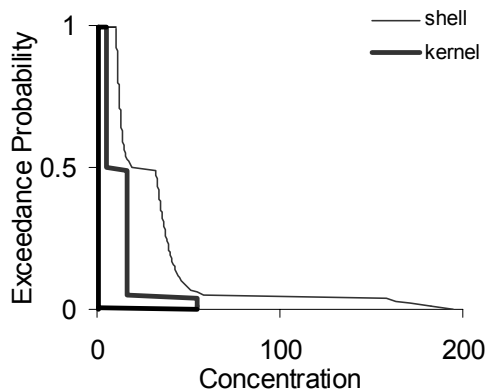


Figure 3. The shell and the kernel backcalculations of Concentration from Dose and Intake.

The kernel is a different kind of answer to the backcalculation problem. By definition, any distribution drawn from within the kernel is sure to satisfy the prescribed constraints on dose and is therefore certain to be a solution to the problem. The kernel is of most interest to those regulating environmental contaminants because if the actual concentration distribution can be shown to be within the kernel, there is a guarantee of compliance with the dose constraints. Figure 3 also depicts the heavy-tailed kernel (bold line). Although any distribution entirely within the kernel is sure to be a solution, not all solutions to the problem are within the kernel. Some

concentration distributions lie at least partially outside the kernel and yet still yield distributions of dose that do satisfy the prescribed constraints.

Figure 4 depicts a check on the kernel and shell concentration p-boxes obtained by calculating the p-boxes enclosing the dose distribution in the forward direction via the equation for dose. The forward calculation using any concentration drawn within the shell produces the thin line in Figure 4. While the shell p-box clearly contains all concentration distributions that satisfy the dose constraint, it also contains concentration distributions leading to doses that are well over the prescribed constraints on dosage.

The dark bold line in Figure 4 depicts upper bounds on the complementary cumulative distribution function of dose for any concentration drawn entirely within the kernel. Notice that it lies entirely below the regulatory constraints (gray line), as required. Because the upper bound is known to be rigorous (Williamson and Downs 1990; Frank et al. 1987), any distribution within the kernel will satisfy the constraints on dose. Note, however, that the bounds defining a kernel are rarely best possible. A backcalculation problem generally does not have a unique kernel. Indeed, any single solution itself constitutes a kernel since any distribution taken from it will be a solution. A unique selection from among the possible kernels that is of special interest in risk analysis is the kernel that has greatest weight in the tails (i.e., at each percentile from 100 down to zero, selecting only those kernels that have largest value at that point). We have derived a new backcalculation algorithm for finding the kernel with the heaviest tails when it exists.

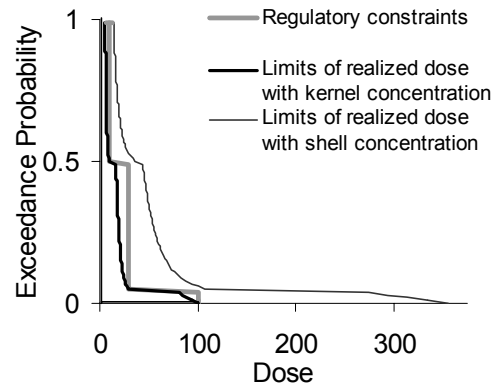


Figure 4. Checks on the backcalculated shell and kernel around Concentration obtained by forward calculation of Dose = Concentration × Intake.

The Existence of a Kernel

The shell always exists. The kernel may or may not exist. For the interval equation $C = A + B$, where A and C are distributions constrained only in terms of their supports, a kernel solution exists whenever C is equal to or wider than A . For instance, if A equals the interval $[1,2]$ and $C = [2,6]$, the shell solution is $B = C - A = [0,5]$. The kernel solution is $B = \text{decon}(A,C) = [1,4]$, where “decon()” represents the additive backcalculation found by applying the backcalculation algorithm described in the next section. The solution space for C is graphed as a function of A and B in Figure 5 on the left. The dark gray parallelogram encloses all combinations of A and B satisfying the constraint, C . The shell for B , $[0,5]$, includes all values of B for which any solution at all exists. The light gray area inside it encloses all solutions within the kernel for B , $[1,4]$. Note that all values of B within the kernel are solutions meeting the constraint for *all* values of A , whereas some values of B in the shell are not solutions when paired with some values of A . The middle graph in Figure 5 shows the analogous case for the multiplicative constraint $C = A \times B$, where the shell is $B = [1,6]$ and the kernel is $B = [2,3]$.

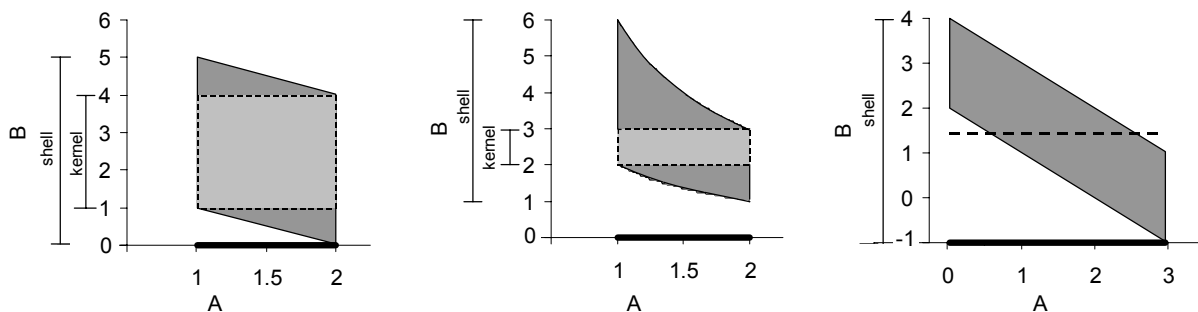


Figure 5. Shells and kernels. Left graph shows the solution spaces for $C = A + B$. The middle graph shows the solution spaces for $C = A \times B$. The right graph shows a case where the kernel does not exist.

The shell always exists for any constraint where $C = f(A,B)$, and $f(A,B)$ is solvable for B . In the additive case when a shell exists, the kernel only exists if C is wider than A and B . The right graph in Figure 5 shows a case where the shell exists but the kernel does not exist. In that graph, $A = [0,3]$, $C = [2,4]$, and the constraint is $C = A + B$. Note that the width of A is 3 while the width of C is only 2. The shell solution for B is $B = [-1,4]$, however, as the dashed line indicates, there is no value of B , not even a point value, for which all values of A are solutions that satisfy the constraint. In all cases under addition and subtraction operations, if the width of C is less than the width of A no kernel solution exists. The existence condition for multiplication and division is that the width of $\log(C)$ must be greater than the width of $\log(A)$ for a kernel solution to exist. Furthermore, under multiplication and division, the ranges of C and A must not span zero. This is not generally a problem in real-world risk analysis applications, however, as the variables are almost universally all non-negative.

Backcalculation Algorithms

The computation of the shell follows from a simple (forward) application of probability bounds analysis (Williamson and Downs 1990). The computation of the kernel is in general a more complex affair. The method of Williamson and Downs (1990) does not produce a kernel, but their algorithm can be reformulated to obtain a method that does. We developed a small set of simple and efficient algorithms that can be used to obtain backcalculations in a general way for simple equations such as $Dose = Concentration \times Intake$. Algorithms were developed for the cases of backcalculating multiplication, division, addition, and subtraction without any assumption about the dependence among the variables, by untangling the algorithms used in the forward calculations and uniquely selecting the kernel with heaviest tail. The Appendix contains pseudocode for the addition algorithm.

In the case of bounds on probability distributions, there are usually multiple kernels. The algorithms we have developed select bounds around the region with the thickest tails. Thus the upper bound on allowable concentrations is only as strict as is absolutely necessary to guarantee that no one receives a dose higher than is allowed.

Discussion

As a part of a widespread trend toward realism, the deterministic expressions traditionally used in risk assessments are being replaced with probabilistic expressions that represent natural heterogeneity in human populations and variability among possible exposure scenarios (e.g. Karstadt 1988; McKone and Ryan 1989; Roberts 1990; Burmaster and Harris 1993; von Stackelberg 2002). While this is a laudable and important advance, it has led to certain complications. So long as only point estimates were used in these quantitative risk assessments, it was straightforward to solve for the environmental concentration that, if not exceeded, would ensure that resulting doses received by humans were below tolerable limits. Now that the mathematical equations involve probability distributions, however, it is not possible to simply rearrange the equations to solve for these concentrations. Recognition has now become widespread within the risk analysis community that the defining equations cannot simply be inverted to solve for tolerable concentrations of environmental contaminants. Unfortunately, many risk analysts take this news as justification for returning to the old and discredited deterministic approaches of the past. To sustain the methodological advances achieved by the proponents of probabilistic methods, it is critical that flexible approaches are developed to solve

backcalculation problems in the probabilistic setting. We present one such approach in this paper.

The use of these algorithms for backcalculation in practical determinations in risk assessments is straightforward. They can be used with multiple receptors (e.g., different susceptible subpopulations) and with arbitrarily many constraints on percentiles or moments of the target risk. If an observed concentration distribution is entirely inside the kernel, then the result surely obeys the prescribed constraints. If the concentration distribution is anywhere outside the shell, then the result certainly fails to comply with the prescribed constraints. If the concentration distribution is outside the kernel but inside the shell, then compliance must be determined by a forward calculation. Note that although the kernel is essentially comparable to the screening level familiar from deterministic assessments, the shell cannot be similarly analogized with an action level. A similar duality arises in interval analysis (Moore 1966) between the general or naïve solution (shell) and the tolerance solution (kernel). In interval analysis, however, the kernel is always unique if it exists.

The shell is guaranteed to enclose all solutions, and any distribution inside the kernel is guaranteed to be a solution, no matter what stochastic dependence there may be between the variables. In other situations, an analyst may wish to assume that concentration and intake are stochastically independent. Knowledge of complete independence between variables is a strong claim, and has important consequences for the analysis. When warranted, an independence assumption can substantially increase the size of the kernel solution. This translates to higher allowable concentrations (and potentially lower cleanup costs) while still guaranteeing that the dose constraint is met. This is contingent, of course, on the existence of independence between concentration and intake, which may or may not be a reasonable assumption depending on the particulars of any given case. To compute the kernel when variables A and B are independent, we have explored a “region growing” algorithm. This algorithm starts from a kernel derived without independence assumptions and recursively enlarges the region and tests whether the increment is small enough so that all doses are necessarily within the prescribed constraints. This approach provides a solution, but is computationally expensive. Further research is necessary to investigate more efficient solutions by exploring the problem analytically. Solutions to backcalculations on variables that exhibit more complex dependency structures also require development.

Acknowledgements

We thank Lev Ginzburg and two anonymous reviewers for valuable comments. This work was supported in part by a National Institute of Environmental Health Sciences (NIEHS) grant to Applied Biomathematics (1R43ES10511-01). Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of NIEHS.

References

Bernabini, M., Carrion, P., Jacovitti, G., Rocca, F., and Treitel, S. (1987). *Deconvolution and Inversion*, Blackwell Scientific Publications, Oxford.

- Burmester, D.E., Lloyd, K.J., and Thompson, K.M. (1995). "The need for new methods to backcalculate soil cleanup targets in interval and probabilistic cancer risk assessments." *Human and Ecological Risk Assessment* 1:89–100.
- Burmester, D.E. and Thompson, K.M. (1995). "Backcalculating cleanup targets in probabilistic risk assessments when the acceptability of cancer risk is defined under different risk management policies." *Human and Ecological Risk Assessment* 1:101-120.
- Burmester, D.E. and Harris, R.H. (1993). "The magnitude of compounding conservatisms in Superfund risk assessments." *Risk Analysis* 13: 131.
- Ferson, S. (1995). "Using approximate deconvolution to estimate cleanup targets in probabilistic risk analyses." *Hydrocarbon Contaminated Soils*, P. Kostecki (ed). Amherst Scientific Press, Amherst, Massachusetts, pp. 239–248.
- Ferson, S. (1996). "What Monte Carlo methods cannot do." *Human and Ecological Risk Assessment* 2: 990–1007.
- Ferson, S. (1997). "Probability bounds analysis." *Computing in Environmental Resource Management. Proceedings of the Conference*, A. Gertler (ed.), Air and Waste Management Association and the U.S. Environmental Protection Agency, Pittsburgh, Pennsylvania, pp. 669–678.
- Ferson, S. (2002) *RAMAS Risk Calc 4.0 Software: Risk Assessment with Uncertain Numbers*. Lewis Publishers, Boca Raton, Florida.
- Ferson, S. and Long, T.F. (1995). "Conservative uncertainty propagation in environmental risk assessments." *Environmental Toxicology and Risk Assessment*, Third Volume, ASTM STP 1218, J.S. Hughes, G.R. Biddinger and E. Mones (eds.), ASTM, Philadelphia, pp. 97–110.
- Ferson, S. and Long, T.F. (1998). "Deconvolution can reduce uncertainty in risk analyses." *Risk Assessment: Measurement and Logic*, M. Newman and C. Strojan (eds.), Ann Arbor Press.
- Frank, M.J., Nelsen, R.B. and Schweizer, B. (1987). "Best-possible bounds for the distribution of Jansson, P.A. (ed.) (1984). *Deconvolution with Applications in Spectroscopy*, Academic Press, Orlando, Florida.
- Karstadt, M. (1988). "Quantitative risk assessment: qualms and questions." *Teratogenesis, Carcinogenesis, and Mutagenesis* 8:137.
- McKone, T.E. and Ryan, P.B. (1989). "Human exposures to chemicals through food chains: an uncertainty analysis." *Environmental Science and Technology* 23:1154–1163.
- Moore, R.E. (1966). *Interval Analysis*. Prentice-Hall, Englewood Cliffs, New Jersey.
- Morgan, M.G. and Henrion, M. (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press, Cambridge, England.
- Roberts, L. (1990). "Risk assessors taken to task." *Science* 248:1173.
- Springer, M.D. (1979). *The Algebra of Random Variables*, Wiley, New York.
- von Stackelberg, K.E., Burmistrov, D. Vorhees, D.J. and Bridges, T.S. (2002). "Importance of uncertainty and variability to predicted risks from trophic transfer of PCBs in dredged sediments." *Risk Analysis* 22:499–512.
- Williamson, R.C. and Downs, T. (1990). "Probabilistic arithmetic I: Numerical methods for calculating convolutions and dependency bounds." *International Journal of Approximate Reasoning* 4: 89-158.
- Wunsch, C. (1996). *The Ocean Circulation Inverse Problem*, Cambridge University Press, New York.

Appendix

Pseudocode for the addition algorithm is shown in the box below. In the algorithm, p-boxes are represented by N discrete intervals that correspond to probability levels (p-values) 0 through N-1.

Additive deconvolution of the kernel for B from the forward equation $C = A + B$, where the p-boxes for A and C are known

```
N = number of probability levels used in the representation; {e.g. 100}
{algorithm below finds the left side of the p-box around the distribution of B}
left limit on B at lowest p-value = left limit on C at lowest p-value – left limit on A at lowest p-value;
for i = 1 to N do begin
    set flag to "not done";
    for j = 0 to i-1 do begin
        if (left limit on C at  $i^{\text{th}}$  p-value)  $\leq$  (left limit on A at  $[i-j]^{\text{th}}$  p-value) + (left limit on B at the  $j^{\text{th}}$  p-value)
            then set flag to "done";
        end; {if}
        if flag is "done" then
            left limit on B at  $i^{\text{th}}$  p-value = left limit on B at  $[i-1]^{\text{th}}$  p-value {the one right below it}
        else left limit on B at  $i^{\text{th}}$  p-value = left limit on C at  $i^{\text{th}}$  p-value – left limit on A at  $0^{\text{th}}$  p-value;
        end; {if}
    end; {for j}
end; {for i, left bound}
{algorithm below finds the right side of the p-box around the distribution of B}
right limit on B at highest p-value = right limit on C at highest p-value – right limit on A at highest p-value;
for i = N-1 down to 0 do begin
    set flag to "not done";
    for j = N down to i+1 do begin
        if (right limit on C at  $i^{\text{th}}$  p-value)  $\geq$  (right limit on A at  $[i-j+N]^{\text{th}}$  p-value) + (right limit on B at the  $j^{\text{th}}$  p-value)
            then set flag to "done";
        end; {if}
        if flag is "done" then
            right limit on B at  $i^{\text{th}}$  p-value = right limit on B at  $[i+1]^{\text{th}}$  p-value {the one right above it}
        else right limit on B at  $i^{\text{th}}$  p-value = right limit on C at  $i^{\text{th}}$  p-value – right limit on A at  $N^{\text{th}}$  p-value;
        end; {if}
    end; {for j}
end; {for i, right bound}
end. {algorithm}
```