Société de Calcul Mathématique SA
Tools for decision help
since 1995

Surveillance Points in High Dimensional Spaces

by Bernard Beauzamy

January 2016

Abstract

Let us consider any computer software, relying upon a large number of parameters, typically, 40 or more, used for any type of simulation. Such a tool will certainly be a very complicated software, therefore lengthy to execute, and the number of runs one may perform will be limited. One wants to perform these runs "at best", that is at specific places, called "Surveillance Points". These investigations will then be considered as representative of any situation.

These Surveillance Points may be interpreted as the center of balls, covering an hypercube which (after normalization) describes all the possible configurations of parameters. Our first result shows that these balls must be extremely large. In dimension 40, for 300 runs, their radius must be at least $2\sqrt{3}$, which means that the result obtained at a Surveillance Point should not be radically different from a result obtained at another point at this distance.

In other words, if one wants to monitor the results given by the computational code, using a small number of Surveillance Points, the computational code must be quite "stable" : its results should not be very different at points which are not too far. If, on the contrary, the computational code may take very different values under rather similar conditions (which is the case, for instance, if some kind of discontinuity occurs), then such a small number of Surveillance Points will not suffice.

In a second part, we give an explicit construction of the Surveillance Points. Basically, their number must be a power of 2, and the precision increases very slowly with the number of balls.

A comparison between the theoretical result and the practical construction shows that the orders of magnitude are correct and satisfactory. For instance, in dimension $K = 40$, if we use 256 balls in order to cover the hypercube, they must have a radius at least $2\sqrt{3} \approx 3.46$. On the other hand, we can construct explicitly 256 balls of radius $\sqrt{8.5} \approx 2.9$ which cover the hypercube.

The overall conclusion is that people who use computational codes in high dimensional spaces, that is depending on a large number of parameters, should be very cautious when they claim that a small number of runs suffices in order to evaluate the outputs of the code.

*****

Acknowledgements

The present paper comes from specific needs expressed originally by Framatome-ANP (2003), then by the Institut de Radioprotection et de Sûreté Nucléaire (since 2005) and more recently by EDF (2015).

Using a small number of runs or observations in order to reconstruct a global information was the topic of Olga Zeydina's thesis ; see the book "Probabilistic Information Transfer".

The present paper is a part of the general research topic "Malfunctions in Sensors Networks", developed jointly by IRSN and SCM.

*****

I. General presentation

Let, in the sequel, $y = C(x_1,...,x_K)$ be a simulation software, returning a real value y from K real valued parameters $x_1,....,x_K$. We may assume without loss of generality that each parameter takes its values between 0 and 1, replacing if necessary x by $\frac{x - x_{\min}}{x_{\max} - x_{\min}}$, where $x_{\min}$ and $x_{\max}$ are respectively the smallest and the largest value of the parameter.

Therefore, our function is defined on the $K-$ dimensional hypercube $H_K = [0,1]^K$.

A Surveillance Point A is a point in H_K at which we will try the code. Let $(a_1,...,a_K)$ be the coordinates of A. Let N be the number of runs we want to execute, and let $A_1,...,A_N$ be the corresponding Surveillance Points.

The points A_n are not distributed at random. On the contrary, we want them to be distributed as regularly as possible, so that any arbitrary point A in the hypercube should be at minimal distance from one of the A_n's. In mathematical terms, it means that the euclidean balls centered at the A_n's, with some radius r (same for all balls) should cover the hypercube H_K.

Here is a picture for $K = 2$; the hypercube is just a square, the balls are disks, and we cover the unit square with 9 balls, all having radius $\frac{\sqrt{2}}{6}$:

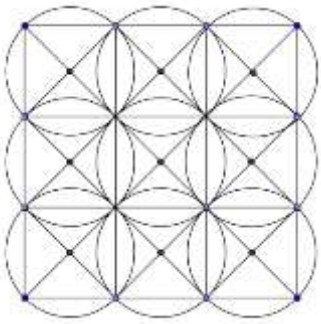

Figure 1 : covering the unit square with 9 equal disks

Here, our Surveillance Points will be the centers of the small circles. We note that there is always an overlap between the zones, which may be used in order to detect some malfunction or strange behavior of the sensors.

In dimension 3, we cover the same way the unit cube by balls, and the description is the same in high dimensional spaces. However, in such situations, the results become highly counter-intuitive, as we will see. For a general introduction to the geometry of high dimensional spaces, see the book [BB1].

Assume that a set of balls $B_1,...,B_N$, with centers $A_1,...,A_N$ and with same radius r, covers the hypercube. Then these balls must certainly contain in particular the summits of the hypercube. A summit is a point whose coordinates are 0 or 1 : they are the extreme points of the hypercube. For the square in dimension 2, the summits are simply the four corners of the square. In dimension K, there are 2^K summits, since there are K coordinates and each coordinate may take the values 0 or 1.

II. Covering the Hypercube

A. A general result

We will prove :

Theorem 1. – *Let* K *be the dimension of the space and let* H_K *be the* $K-$ *dimensional hypercube* $[0,1]^K$ *. Let* N *be any number of balls of same radius, covering the hypercube. Let* j_0 *be the largest integer satisfying :*

$$1+\binom{K}{1}+\cdots+\binom{K}{j}<\frac{2^K}{N} \tag{1}$$

Then the radius of the balls satisfies :

$$r\geq\sqrt{j_0+1}$$

In the case $K=40$ and $N=300,$ we find $r\geq\sqrt{12}=2\sqrt{3}\approx 3.46.$

Proof of Theorem 1

Assume we have constructed balls which cover the hypercube. We have N balls and 2^K summits. Since the balls contain all the summits, one of the balls must contain a number of summits which is $\geq\frac{2^K}{N}$. Let us denote by B_0 this ball.

Let $S=(\varepsilon_1,...,\varepsilon_K)$ be any summit, with $\varepsilon_k=0$ or 1. We say that another summit $S'=(\varepsilon'_1,...,\varepsilon'_K)$ is a neighbour of S of order 1 if the coordinates of S and the coordinates of S' differ by one item only. For instance, if $S=(0,...,0)$, its neighbours of order 1 are the K points $(1,0,...,0)$, $(0,1,0,...,0)$, …, $(0,...,0,1)$. Each summit has K neighbours of order 1.

The same way, S' is a neighbour of S of order 2 if their coordinates differ by 2 items only. The neighbours of $(0,...,0)$ of order 2 are the points with 2 coordinates equal to 1, the others being 0. There are $\binom{K}{2}=\frac{K!}{(K-2)!2!}=\frac{K(K-1)}{2}$ neighbours of type 2.

Similarly again, S' is a neighbour of S of order j if their coordinates differ by j items only. The neighbours of $(0,...,0)$ of order j are the points with j coordinates equal to 1, the others being 0. There are $\binom{K}{j}$ neighbours of type j.

Take the ball B_0 defined above : the one with largest number of summits. Assume $1+\binom{K}{1}<\frac{2^K}{N}$. Certainly, this ball cannot be such that it contains simply a point and its neighbours of order 1, and no other summit, because in this case the number of points it contains is $1+\binom{K}{1}$, which is smaller than the number of points it should contain, namely $\frac{2^K}{N}$.

For instance, if $K=40,\quad N=300,\ 1+\binom{K}{1}=K+1=41$ and $\frac{2^K}{N}=\frac{2^{40}}{300}\approx 3\,665\,038\,759.$

The same way, if K and N are such that $1+\binom{K}{1}+\binom{K}{2}<\frac{2^K}{N}$, this ball cannot be such that it contains simply a point, its neighbours of order 1, its neighbours of order 2, and no other summit.

If $K=40,\ N=300,\quad 1+\binom{K}{1}+\binom{K}{2}=821.$

More generally, this ball cannot be such that it contains simply a point, its neighbours of order 1, …, its neighbours of order j, and no other summit, as long as :

$$1+\binom{K}{1}+\cdots+\binom{K}{j}<\frac{2^K}{N} \tag{1}$$

For $K=40,\ N=300,\ j=11,$ the left hand side of (1) takes the value $3\,533\,047\,572<\frac{2^K}{N}$ and for $j=12,$ the value $9\,119\,901\,052>\frac{2^K}{N}$. So the largest j for which (1) holds, denoted by j_0, has the value $j_0=11.$

From (1) follows that the ball B_0 must contain a point, which is a neighbour of order $\geq j_0+1,$ of its center. But in this case, their euclidean distance is at least $d=\sqrt{j_0+1}$ and the radius of the ball is at least $r=\sqrt{j_0+1}.$

For $K=40,\ N=300,$ they have at least 12 coordinates which differ.
Our Theorem is proved.

In practice, evaluations of sums of binomial coefficients are needed:

Corollary 2. – *The theoretical radius* r_{th} *can be found from the formula:*

$$r_{th}^2 \approx \frac{K}{2} + \varphi\left(\frac{1}{N}\right)\frac{\sqrt{K}}{2}$$

where φ *is the inverse function of the Gaussian repartition function, that is* $\varphi(u) = v$ *if* $H(v) = u$*, with:*

$$H(x) = \int_{-\infty}^{x} e^{-t^2/2}\frac{dt}{\sqrt{2\pi}}$$

Proof of Corollary 2

The condition:

$$1 + \binom{K}{1} + \cdots + \binom{K}{j} < \frac{2^K}{N}$$

can be rewritten:

$$\frac{1}{2^K}\sum_{i=0}^{j}\binom{K}{i} \le \varepsilon \tag{1}$$

with $\varepsilon = \frac{1}{N}$.

Let X be a random variable with binomial law, of parameters $\left(K, \frac{1}{2}\right)$. Condition (1) may be rewritten as:

$$P(X \le j) \le \varepsilon \tag{2}$$

But such a random variable has expectation equal to $K/2$ and variance equal to $K/4$. Condition (2) may be rewritten:

$$P\left(\frac{X - K/2}{\sqrt{K/4}} \le \frac{j - K/2}{\sqrt{K/4}}\right) \le \varepsilon \tag{3}$$

Using the approximation of the binomial law by a normal law (which is legitimate here), we may write (3) under the form :

$$P\left(Z \le \frac{j - K/2}{\sqrt{K/4}}\right) \le \varepsilon \tag{4}$$

where Z is a normalized Gaussian random variable, that is $E(Z)=0,\ \mathrm{var}(Z)=1$.

Using the repartition function of the normal law, (4) is equivalent to :

$$H\left(\frac{j-K/2}{\sqrt{K/4}}\right)\le\varepsilon \tag{5}$$

$$\frac{j-K/2}{\sqrt{K/4}}\le\varphi(\varepsilon) \tag{6}$$

That is:

$$j\le\frac{K}{2}+\varphi(\varepsilon)\frac{\sqrt{K}}{2} \tag{7}$$

The value of φ is given by tables of the normal law, and we take:

$$j_0=\mathrm{int}\left(\frac{K}{2}+\varphi(\varepsilon)\frac{\sqrt{K}}{2}\right) \tag{8}$$

which gives a theoretical radius with:

$$r_{th}^2\approx\frac{K}{2}+\varphi(\varepsilon)\frac{\sqrt{K}}{2} \tag{9}$$

which proves Corollary 2.

We deduce from Chernoff's bound (see [Chernoff]) an explicit estimate:

Corollary 3. – *The theoretical radius satisfies:*

$$r_{th}^2\approx\frac{1}{2}\left(K-\sqrt{2K\ Log(N)}\right)+1$$

Proof of Corollary 3

We have, from Chernoff's inequality:

$$\sum_{i=0}^{j}\binom{K}{j}\le 2^K\exp\left(\frac{-(K-2j)^2}{2K}\right)$$

So the condition:

$$1+\binom{K}{1}+\cdots+\binom{K}{j}<\frac{2^K}{N}$$

is satisfied if :

$$\exp\left(\frac{-(K-2j)^2}{2K}\right)\le\frac{1}{N}$$

which is equivalent to:

$$j\le\frac{1}{2}\left(K-\sqrt{2KLog(N)}\right)$$

This gives a theoretical radius with:

$$r_{th}^2\approx\frac{1}{2}\left(K-\sqrt{2K\ Log(N)}\right)+1$$

which proves Corollary 3.

B. Simple cases

1. Covering the hypercube with a single ball

The Theorem asks for the largest j such that $1+\binom{K}{1}+\cdots+\binom{K}{j}<2^K$, which is obviously $j_0=K-1.$ Then the radius given by the theorem is $r_{th}^2=j_0+1=K.$

In practice, we may cover the hypercube with a single ball, centered at $C=\left(\frac{1}{2},...,\frac{1}{2}\right)$. Then, we have :

$$r_{obs}^2=\frac{K}{4}$$

(where the subscript "obs" stands for "observed"). So the prevision of the Theorem is pessimistic, but the order of magnitude is correct.

2. Case of two balls

Using the Theorem, we need to find the largest j such that :

$$1+\binom{K}{1}+\cdots+\binom{K}{j}<2^{K-1}$$

which gives $j_0 = \frac{K}{2} - 1$ if K is even, and $j_0 = \left[\frac{K}{2}\right]$ if K is odd. So, in both cases, we obtain approximately :

$$r_{th}^2 \approx \frac{K}{2}$$

Now, in practice, we can cover the hypercube with the two balls of centers:

$$\left(\frac{1}{4}, \frac{1}{2}, ..., \frac{1}{2}\right), \left(\frac{3}{4}, \frac{1}{2}, ..., \frac{1}{2}\right)$$

with:

$$r_{obs}^2 = \frac{1}{16} + \frac{K-1}{4}$$

so again the order of magnitude given by the Theorem is correct.

C. Volume considerations

The volume of the hypercube is obviously 1, no matter what the dimension is. The volume of a ball of radius r in dimension K (K even) is given by the formula :

$$V_K(r) = \frac{\pi^{K/2} r^K}{(K/2)!}$$

If $K = 40$, $r = 2\sqrt{3}$, we get :

$$V_K(r) \approx 10^{13}$$

This means that there is a considerable loss in volume : we need $N = 300$ balls of radius 10^{13} in order to cover something of volume 1.

We may wonder about the volume of the ball, centered at the middle point of the hypercube (that is $\left(\frac{1}{2}, ..., \frac{1}{2}\right)$), containing all summits. This ball has radius $r = \frac{\sqrt{K}}{2}$ and, by the formula above, volume 0.36×10^{12}. So it is extremely large. Recall that, in dimension K, the length of the diagonal of the hypercube is $\sqrt{K}$. The volume of the hypercube is small (equal to 1), but the diagonal is large, which is very counter-intuitive. In fact, the hypercube extends in many directions (or, more exactly, in many dimensions).

D. Chosing the number of balls

Assume now that we fix the radius of all balls ; how many balls do we need ? This is clear from formula (1). If r is fixed, we define j_0 by :

$$j_0 = r^2 - 1$$

and the number N is given by :

$$N = 1 + \mathrm{int}\left(\frac{2^K}{1 + \binom{K}{1} + \cdots + \binom{K}{j_0}} \right) ,$$

where $\mathrm{int}(x)$ denotes the integral part of x. This means that :

$$\frac{2^K}{\sigma_{j_0+1}} \le N < \frac{2^K}{\sigma_{j_0}} \tag{2}$$

where we write :

$$\sigma_j = 1 + \binom{K}{1} + \cdots + \binom{K}{j} \tag{3}$$

Any change of N in the interval (2) is useless, so one should take the N which is as small as possible, that is the left bound of the interval. For $K = 40$ and $j_0 = 11$, we find the interval :

$$120 \le N < 311$$

which means that we can achieve the same result with 120 balls as with 300 balls : this is important in practice.

We have obtained:

Theorem 4.- *For a given value of r, and $j_0 = r^2 - 1$, all values of N in the interval $\frac{2^K}{\sigma_{j_0+1}} \le N < \frac{2^K}{\sigma_{j_0}}$ (where σ_j is defined by (3)) provide the same covering. Therefore, one should use the value $N = 1 + \mathrm{int}\left(\frac{2^K}{\sigma_{j_0+1}} \right)$ which represents the smallest number of balls which are required in order to obtain this covering. Putting more balls is a waste of time.*

III. Choosing the centers of the balls

We now indicate how to chose the balls. In this section, we give a constructive approach which, theoretically speaking, might not be best possible. However, the order of magnitude is correct, as Theorem 1 shows.

In what follows, the parameters are treated one by one. If we know nothing about their respective importance, the order is arbitrary. But if, for some practical reasons, we have a ranking upon the parameters, we should of course start with the most important one. This is often the case in practice.

Covering with 1 and 2 balls has been described above.

A. Three balls

A covering by 3 balls is obtained by dividing the interval for the first parameter into 3 subintervals, which gives the centers:

$$\left(\frac{1}{6},\frac{1}{2},...,\frac{1}{2}\right),\left(\frac{3}{6},\frac{1}{2},...,\frac{1}{2}\right),\left(\frac{5}{6},\frac{1}{2},...,\frac{1}{2}\right)$$

The radius is $r^2=\frac{1}{6^2}+\frac{K-1}{4}$

B. Four balls

A covering by 4 balls is obtained by dividing the intervals for the first and second parameters into 2, which gives the centers:

$$\left(\frac{1}{4},\frac{1}{4},\frac{1}{2},...,\frac{1}{2}\right),\left(\frac{3}{4},\frac{1}{4},\frac{1}{2},...,\frac{1}{2}\right),\left(\frac{1}{4},\frac{3}{4},\frac{1}{2},...,\frac{1}{2}\right),\left(\frac{3}{4},\frac{3}{4},\frac{1}{2},...,\frac{1}{2}\right)$$

The radius is $r^2=\frac{2}{4^2}+\frac{K-2}{4}$

C. General pattern

1. Description

Assume that the interval for the first parameter is divided into n_1 sub-intervals,..., the interval for the $K-$ th parameter is divided into n_K sub-intervals, then the centers are all the points of the form :

$$\left(\frac{2j_1+1}{2n_1},...,\frac{2j_K+1}{2n_K}\right), \text{ with } j_1=0,...,n_1-1,\ \ldots,\ j_K=0,...,n_K-1. \qquad (1)$$

and the radius satisfies:

$$r^2 = \sum_{k=1}^{K} \frac{1}{(2n_k)^2} = \frac{1}{4}\sum_{k=1}^{K} \frac{1}{n_k^2} \tag{2}$$

The number of balls in this case is $N = n_1 \times \cdots \times n_K$.

2. Obtaining the smallest radius

Given a number N of balls, we want the radius to be as small as possible. From formula (1) follows that we should divide as many parameters as possible. For instance, if $N = 4$, dividing the first interval into 4 gives $r^2 = \frac{1}{4}\left(\frac{1}{16} + K - 1\right)$ and dividing the first two intervals into 2 gives $r^2 = \frac{1}{4}\left(\frac{2}{4} + K - 2\right)$, which is obviously smaller.

3. Consequences

An obvious consequence is that a larger N does not necessarily provide a better solution. For instance, the value $N = 17$ allows the division of the first interval into 17 sub-intervals, giving:

$$r^2 = \frac{1}{4}\left(\frac{1}{17^2} + K - 1\right),$$

whereas the value $N = 16$ allows the division of 4 intervals into 2 pieces, giving:

$$r^2 = \frac{1}{4}\left(\frac{4}{2^2} + K - 4\right),$$

which is much smaller.

4. Practical rule

Assume that the parameters are written in decreasing order of importance. Then the best strategy is to use a division by 2, as long as possible. This means that the number of balls should be chosen in the sequence $1, 2, 2^2, 2^3, \ldots$. If we use $N = 2^k$ balls, we divide the interval of variation of k parameters into 2, which means that the centers are of the form $\frac{2j+1}{2}$ $(j = 0,1)$ for the first k and of the form $\frac{1}{2}$ for the last $K - k$. The value of the radius is:

$$r_{2^k}^2 = \frac{1}{4}\left(\frac{k}{4} + K - k\right) \tag{3}$$

5. Comparison between two binary steps

Let us see what happens if we use 2^{k+1} balls instead of 2^k. Then, by the paragraph above:

$$r_{2^{k+1}}^2 = \frac{1}{4}\left(\frac{k+1}{4} + K - (k+1)\right)$$

and therefore :

$$r_{2^{k+1}}^2 - r_{2^k}^2 = -\frac{3}{16}$$

So, the multiplication by 2 of the number of balls gives a decrease on r^2 of a constant quantity, namely $\frac{3}{16}$. Using any intermediate number of balls is rather useless.

6. Reaching a given threshold

Assume we want to find the number of balls necessary to have $r^2 \le A$, for a given A. Then, by (3),

$$K - \frac{3k}{4} \le 4A$$

which gives :

$$k \ge \frac{4}{3}(K - 4A) \tag{4}$$

This conclusion is very strong. It shows that, unless we accept a threshold which is proportional to K, the number of necessary balls is exponential : $N \approx 2^{4K/3}$.

IV. Comparison with the theoretical bounds

We now compare the radius obtained in this construction with the radius indicated by Theorem 1.

A. Numerical comparison

It is easy to do when numerical values are chosen. So let us take $K = 40,\ N = 256 = 2^8$. We find $j_0 = 11$ in the definition given in Theorem 1 ; therefore, the theoretical radius satisfies :

$$r_{th}^2 = 12$$

For the practical radius, we use formula (2) above, and we get:

$$r_{obs}^2 = \frac{1}{4}\sum_{k=1}^{K}\frac{1}{n_k^2} = \frac{1}{4}\left(\frac{8}{4}+32\right) = \frac{34}{4} = 8.5$$

So the order of magnitude is correct, once again.

B. Theoretical comparison

A general comparison between the estimates given by Theorem 1 and the practical construction is harder to obtain, because it requires the evaluation of partial sums of binomial coefficients. We use Corollary 3 above, with $N = 2^k$ balls, and we want to find the largest j such that :

$$\sum_{i=0}^{j}\binom{K}{i} \le \frac{2^K}{2^k}$$

If we take k proportional to K, that is:

$$k = \alpha K$$

Corollary 3 gives:

$$r_{th}^2 = \frac{K}{2}\left(1-\sqrt{2Log2\,\alpha}\right)$$

On the other hand, the practical construction gives:

$$r^2 = \frac{1}{4}\left(\frac{k}{4}+K-k\right) = \frac{1}{4}\left(K-\frac{3k}{4}\right)$$

that is:

$$r_{obs}^2 = \frac{K}{4}\left(1-\frac{3\alpha}{4}\right)$$

So there is roughly a factor 2 between the square of the radii, which is a correct order of magnitude.

V. References

[BB1] Bernard Beauzamy : Introduction to Banach Spaces and their Geometry. North Holland, Collection "Notas de Matematica", vol. 68. Première édition : 1982, seconde édition : 1985.

[Chernoff] http://en.wikipedia.org/wiki/Chernoff_bounds

[PIT] Olga Zeydina and Bernard Beauzamy : Probabilistic Information Transfer. SCM SA, 2013. ISBN: 978-2-9521458-6-2, ISSN : 1767-1175.