Periodica Polytechnica
Chemical Engineering

58(Sup), pp. 49-52, 2014
DOI:10.3311/PPch.7313
http://www.pp.bme.hu/ch/article/view/7313

RESEARCH ARTICLE

Reformulation of the Gaussian error propagation for a mixture of dependent and independent variables

Sándor Kristyán

RECEIVED 6 SEPTEMBER 2012; ACCEPTED 24 MARCH 2013

Abstract

The Gaussian error propagation is a state of the art expression in error analysis for estimating standard deviation for an expression $f(x_1,\ldots,x_n,z)$ via its variables. One of its basic assumptions is the independence of the measurable variables in its argument. However, in practice, measurable quantities are correlated somehow, and sometimes, z depends on some of the x_i's. We provide the generalized version of the Gaussian error propagation formula in this case. We will prove this with the formula for total derivative of a general multivariable function for which some of its variables are not independent from the others; a counterpart to the probability approach of this subject.

Keywords

Reformulation of Gaussian error propagation · dependent variables · physical/biological/economical measurements/calculations

Sándor Kristyán
Department of of Biological Nanochemistry, Institute of Molecular Pharmacology, Research Centre for Natural Sciences, Hungarian Academy of Sciences
Pusztaszeri út 59–67., H-1025 Budapest, Hungary
e-mail: kristyan.sandor@ttk.mta.hu

1 Introduction

Frequently, the final result of an experiment cannot be measured directly, rather, it is calculated from several measurable physical quantities, each of which has a mean value and an error, and we are interested in the resulting error in the final result of such an experiment. Often, the measurement protocol is very complex and the set of measured physical quantities is a mix of variables in which some are independent of others and some are not. More importantly, selecting only independent physical quantities to be measured is not always possible. These difficulties occur in data analysis after collecting the outcome of measurements, for example: in weather observation or meteorology, astro- or high-energy physics, physical-, chemical- or biological measurements, as well as economics.

Below we discuss a theorem, how the Gaussian error propagation reads if in its x_1, x_2, $x_3,\ldots,x_n$, $z_1,\ldots,z_m$ variables, the first n (the x_i's) are independent, but among the $z_1,\ldots,z_m$ each one depends on some of the x_i's. When all n variables are independent and no such z_j exists, the well known Eq.1 (written below) holds, commonly appearing in corresponding text and lab books. However, in many complex and/or large scale measurements, the variables may not be totally independent, and there may not be an alternative way to measure/choose purely independent variables. Statisticians use a procedure commonly called the delta method [1,2,3] to obtain an estimator of the variance when the estimator is not a simple sum of observations. The basic idea is to use a method from calculus called a Taylor series expansion to derive a linear function that approximates the more complicated function. To the best of our knowledge, although this case has been commonly formulated with algorithms using the concept of covariance via probability theory approach, still there is no compact expression formulated via calculus – here we do this.

2 Problem formulation

If the error in x_1 is Δx_1, then the error in f can be approximated as $(\partial f/\partial x_1)\Delta x_1$, and similarly for x_2, and finally $(\Delta f)^2 = (\partial f/\partial x_1)^2(\Delta x_1)^2 + (\partial f/\partial x_2)^2(\Delta x_2)^2$, leaving the cross term

$2(\partial f / \partial x_1)(\partial f / \partial x_2)(\Delta x_1)(\Delta x_2)$. More generally, one ends up with the famous Gaussian error propagation formula [4,5] which states that if $f = f(x_1,x_2,\ldots,x_n)$ and $x_1, x_2, \ldots,x_n$ are independent quantities, e.g. of measurement possessing Gaussian distribution, the standard deviation of f (denoted as s_f) is

$$(s_f)^2 = (\Delta f)^2 = \Sigma_{(i = 1\ldots n)}(\partial f / \partial x_i)^2(s_{xi})^2. \quad (1)$$

To complete Eq.1 for a measurement in practice, let u denote any of the independent variables among $x_1, x_2, \ldots,x_n$, and u_j is the j^{th} measured quantity for u, where $j = 1,2,\ldots,m(u)$. The mean of u is $u_{avrg} \equiv \Sigma_{(j=1\ldots m)}u_j/m$, and the standard deviation of u is $s_u \equiv [\Sigma_{(j=1\ldots m)}(u_j - u_{avrg})^2 / (m - 1)]^{1/2}$. On the other hand, to complete Eq.1 for probability variables u, one needs the corresponding unbiased estimate for expected value (E(u)) and its variance ($D^2(u)$).

A simple example can illustrate what the misapplication of independency or dependency can cause. Let $f = x_1+x_2$, and $x_2 = x_1$ with the obvious $\Delta x_1 = \Delta x_2$. Assuming them to be independent variables (although they are not), $(\Delta f)^2 = (\partial f/\partial x_1)^2(\Delta x_1)^2 + (\partial f/\partial x_2)^2(\Delta x_2)^2 = (\Delta x_1)^2 + (\Delta x_2)^2 = 2(\Delta x_1)^2$ or $\Delta f = \sqrt{2}\ (\Delta x_1)$. The correct expression is $f = x_1 + x_2 = 2x_1$ and $(\Delta f)^2 = (\partial f / \partial x_1)^2(\Delta x_1)^2 = (df / dx_1)^2(\Delta x_1)^2$, or more simply $\Delta f = (df / dx_1)(\Delta x_1) = 2(\Delta x_1)$, i.e. the misapplication underestimates it ($\sqrt{2} < 2$). Note: the equation of Gaussian error propagation degrades to the simple estimation of derivatives with the elementary numerical device $\Delta f / (\Delta x_1) \approx df / dx_1$ for one variable ($n = 1$), given that in numerical analysis the Δx_1 is a small step while in error analysis the Δx_1 is the standard deviation. Similarly, if $f = x_1 - x_2$, then the misapplication yields $\Delta f = \sqrt{2}(\Delta x_1)$ again, but the correct expression yields $\Delta f = 0$ (since $f = x_1 - x_2 = x_1 - x_1 = 0$), i.e. the misapplication overestimates it ($\sqrt{2} > 0$). The latter is a warning for a general perspective: in a statistical test for a hypothesis, predicting small positive value instead of zero may mistakenly suggest a statement to be true or false.

Now we outline how a measurement can come up with a mix of dependent and independent variables. Let us suppose that one has to calculate a quantity of which dependence is $f(x_1,x_2,x_3,x_4(x_2,x_3,x_5))$, where x_1,x_2,x_3 and x_5 are independent variables, and x_4 is not, i.e. dependent as it is indicated. However, x_1,x_2,x_3 and x_4 can be measured directly, but not so in the case of x_5. Algebraically it means $f(x_1,x_2,x_3,x_4(x_2,x_3,x_5)) \equiv g(x_1,x_2,x_3,x_4,x_5) \equiv h(x_1,x_2,x_3,x_5)$ with the proper relationship among f, g and h. In other words, x_5 does not show up alone in the argument, but with x_2 and x_3 via x_4. In this work we call these f-forms and h-forms. In the example mentioned above $f = x_1 + x_2$ with $x_1 = x_2$, so $h = 2x_1$. In this way, the general definition of f- and h-forms is obvious. The h-form may have fewer variables than the f-form, but not necessarily. In the particular case above, both have four variables, but in the case of $f = x_1 + x_2$ with $x_1 = x_2$, f has two variables, as opposed to h which has only one. Below, we will need their partial derivatives, and e.g. in the case above $\partial f / \partial x_5 = 0$, despite that, $(\partial f / \partial x_4)(\partial x_4 / \partial x_5)$ is generally not zero. This is because x_5 does not appear in the argument of f, but otherwise it is possible. In other words, one has to be careful with the partial derivatives. It is obvious, that the f-form has a mixture of dependent and independent variables in its argument, while the h-form has only independent variables, but both have the same graph. Below, we will consider the general function $f(x_1,\ldots,x_n,z_1,\ldots,z_m)$, where $x_1,\ldots,x_n$ are independent variables, and $z_1,\ldots,z_m$ are dependent variables. The latter means that these depend on at least one of $x_1,\ldots,x_n$, e.g. $z_1 = z_1(x_1, x_2)$, $z_2 = z_2(x_2, x_3, x_5)$ with $n \geq 5$, $m \geq 2$ and so on. Algebraically the f-form can be reduced to h-form, because sometimes the relationship is indeed known, and the latter has only independent variables in its argument. However, sometimes even the exact analytical relationship is unknown, or in practice only the f-form can be used to evaluate that particular measurement and the h-form cannot. We try to enumerate that the effect of "mixture variables" can be positive or negative alike. It clearly shows that the unknown biases committed might be compensated by each other. The correlation of variables has a paradoxical outcome, e.g. the probability of chance correlation is diminished if the variables selected from a large pool are correlated [6].

Next, for the sake of brevity, we will call and use the errors Δf and Δx_i, i.e. the standard deviation belonging to their mean or exact values. The measured variables (x_i) obey the Gaussian distribution, so their actual error is smaller than these threshold (Δx_i) values at a certain significance level. Even if f is not known analytically, via the measured or non-explicitly (e.g. recursively, etc.) calculated f(x) at x and $x + \Delta x$, the derivative of f can be approximated numerically. On the other hand, if the measured x suffers an error of size as the standard deviation (that is $x \pm \Delta x$, i.e. the maximal expected deviation on a certain significance level), the error made in f, the Δf (which is also a standard deviation), can be estimated as $(\partial f / \partial x)\Delta x$, if $(\partial f / \partial x)$ is known – that is $(\Delta f)^2 \approx (\partial f / \partial x)^2\Delta x^2$, which is Eq.1 for one variable.

3 The way to the reformulation via calculus

Without losing generality, let us suppose that there is only one dependent z, and we consider the $f(x_1,\ldots,x_n,z)$, where $x_1,\ldots,x_n$ are independent variables and $z = z(x_1,\ldots,x_n)$. The latter includes two distinct cases: 1.: z depends on at least one (there exist i s.t. $\partial f / \partial x_i \neq 0$, $i=1,\ldots,n$), more, or all (for all i, $\partial f / \partial x_i \neq 0$) variables, 2.: z does not depend on any of the x_i (for all i, $\partial f / \partial x_i = 0$). If z does not depend on any x_i, that is, the set $\{x_1,\ldots,x_n, z\}$ contains only independent variables, the total derivative is

$$df = \Sigma_{(i = 1\ldots n)}(\partial f / \partial x_i)dx_i+(\partial f / \partial z)dz, \quad (2)$$

and the Gaussian error propagation comes from applying Eq.1 with the extension for one more variable

$$(\Delta f)^2 = \Sigma_{(i = 1\ldots n)}(\partial f / \partial x_i)^2(\Delta x_i)^2 + (\partial f / \partial z)^2(\Delta z)^2. \quad (3)$$

Again, the independence is strictly necessary for both, Eqs. 2,3. The close relationship between the two algebraic structures between Eqs. 2 and 3 is visible. (Again, the way to Eq. 3, which is used for estimating standard deviation, the square of the exact expression in Eq. 2 was taken, along with replacing the derivative (d) with standard deviation (Δ) and leaving all cross terms.) If z depends on at least one x_i, generally $h(x_1,\ldots,x_n) = f(x_1,\ldots,x_n, z(x_1,\ldots,x_n))$ holds with a proper h, then Eqs. 2,3 are false.

An elementary example can demonstrate how Eq. 2 breaks or survives if dependence arises among the variables. If $f(x_1,x_2,z) = x_1^2x_2^3z$ then $\partial f / \partial x_1 = 2x_1x_2^3z$ and misapplying the partial derivatives for dependent variables for a case like $z = x_1x_2$: $\partial f / \partial x_1 = 2x_1x_2^3z \sim 2x_1^2x_2^4$, tilde means an "equality by mistake". Given the h-form $h(x_1,x_2) = x_1^2x_2^3z = x_1^3x_2^4$, and $\partial h / \partial x_1 = 3x_1^2x_2^4 \neq 2x_1^2x_2^4$. In fact, the substitution with $z = x_1x_2$ was used at the wrong point, because $df = (\partial f / \partial x_1)dx_1 + (\partial f / \partial x_2)dx_2 + (\partial f / \partial z)dz = 2x_1x_2^3zdx_1 + 3x_1^2x_2^2zdx_2 + x_1^2x_2^3dz = 2x_1^2x_2^4dx_1 + 3x_1^3x_2^3dx_2 + x_1^2x_2^3(x_2dx_1 + x_1dx_2) = 3x_1^2x_2^4dx_1 + 4x_1^3x_2^3dx_2$, where $dz = x_2dx_1 + x_1dx_2$ was used in the second step, i.e. at a proper point. Eq. 2 can be applied directly in the h-form, because it only contains the independent x_1 and x_2, giving the same $dh = 3x_1^2x_2^4dx_1+4x_1^3x_2^3dx_2$. (We note as a finer detail, calculation of dh needed fewer algebraic operations than df.) The critical point was that the $\partial f / \partial x_1= 2x_1x_2^3z$ and $\partial h/\partial x_1 = 3x_1^2x_2^4$, and the similar ones for index 2, are not equivalent for substitution of z into the former, although f and h have exactly the same graph. In a more general case, f has n + 1 variables, while h has n, and if z depended on at least one of x_i's, the total derivative in Eq. 2 has to be reformulated as

$$\begin{aligned} df &= \Sigma_{(i = 1\ldots n)}(\partial f / \partial x_i)dx_i + (\partial f / \partial z)dz = \\ &\Sigma_{(i = 1\ldots n)}(\partial f / \partial x_i)dx_i + (\partial f / \partial z)(\Sigma_{(i = 1\ldots n)}(\partial z / \partial x_i)dx_i) = \\ &\Sigma_{(i = 1\ldots n)}[(\partial f / \partial x_i) + (\partial f / \partial z)(\partial z / \partial x_i)]dx_i. \end{aligned} \quad (4)$$

For Eq. 4 we have used the chain rule only. If z does not depend on some x_i's for those $\partial z / \partial x_i = 0$. If z does not depend on any of x_i's, all $(\partial f / \partial z)(\partial z / \partial x_i) = 0$, and with an abstract composition, in fact z becomes an element of the independent set $\{x_1,\ldots,x_n\}$, so Eq. 4 reduces to Eq. 2 or to the general expression of total derivative for independent variables, as expected. We note that Eq. 2 is a fundamentally known and listed equation in corresponding mathematical textbooks and tables, but Eq. 4 is not, although it is an almost immediate consequence.

Eq. 3 is not accurate if z depends on at least one x_1, … or x_n. In this case, Eq. 2 is also inaccurate, in fact it is false. While Eq. 2 is used for manipulating exact expressions, Eq. 3 is used for estimating standard deviations. In other words, not using Eq. 4 as opposed to Eq. 2, for dependent variables is a mistake. While not developing Eq. 3, as Eq. 2 has been developed to Eq. 4, would yield a weaker estimation only for Gaussian error propagation. If $x_1,\ldots, x_n$ are independent and z depends on at least one of $x_1, \ldots, x_n$, the trivial

$$(\Delta h)^2 = \Sigma_{(i = 1\ldots n)}(\partial h / \partial x_i)^2(\Delta x_i)^2 \quad (5)$$

still holds for the h-form. However, not the h-form but the f-form is known or to be used by some conditions/restrictions of the measurement [7]. For this reason, a more useful and accurate expression is developed here for practice. It is by employing the algebraic relationship between Eqs. 2 and 3, but starting from Eq. 4. The Gaussian error propagation in this case is

$$(\Delta f)^2 = \Sigma_{(i = 1\ldots n)}[(\partial f / \partial x_i) + (\partial f / \partial z)(\partial z / \partial x_i)]^2(\Delta x_i)^2. \quad (6)$$

More generally, if $y = f(x_1,\ldots,x_n,z_1,\ldots,z_m)$ with dependent variables $z_j = z_j(x_1,\ldots,x_n)$ for j = 1,…,m, then

$$\begin{aligned} (\Delta f)^2 = \Sigma_{(i = 1\ldots n)}[&(\partial f / \partial x_i) + \\ &\Sigma_{(j = 1\ldots m)}(\partial f / \partial z_j)(\partial z_j / \partial x_i)]^2(\Delta x_i)^2. \end{aligned} \quad (7)$$

Furthermore, if z_2 depends on z_1 too, as $z_2 = z_2(x_1,..,x_n,z_1)$ and so on, even Eq. 7 can be developed further with the chain rule for the derivatives of embed functions. Moreover, if z_2 depends on $(x_1,..,x_n,z_1,z_2)$, i.e. an implicit expression is given, the derivation rule for implicit function helps. (That is, if $w(x,z(x)) = 0$ or z, then $(\partial w / \partial x) + (\partial w / \partial z)(dz / dx) = 0$ or (dz / dx), and dz / dx can be expressed.) If $\partial z_j / \partial x_i = 0$ for all i = 1,…,n and all j = 1,...,m in Eqs. 6,7, all z_j fall into the independent set of $\{x_1,.., x_n\}$, and Eqs. 6,7 reduce to Eq. 3 or Eq. 1, i.e. to the general expression of Gaussian error propagation for independent variables, as expected.

4 Reformulation of the Gaussian error propagation

Explaining the title of this work, one must recall the known form for the standard deviation of f (denoted as s_f) when its variables are not independent, that is

$$\begin{aligned} &(s_f)^2 = (\Delta f)^2 = \\ &\Sigma_{(i = 1\ldots n + m)}\Sigma_{(j = 1\ldots n + m)}(\partial f / \partial \xi_i)(\partial f / \partial \xi_j)\mathrm{cov}(\xi_i,\xi_j) \end{aligned} \quad (8)$$

with the terminology of probability to compare with Eqs. 6 and 7. The $\mathrm{cov}(\xi_i,\xi_j)$ is the covariance of probability variables ξ_i and ξ_j as well as if i = j then $\mathrm{cov}(\xi_i,\xi_i) = (\Delta\xi_i)^2$, more, if $\mathrm{cov}(\xi_i,\xi_j) = \delta_{ij}(\Delta\xi_i)^2$ with δ_{ij} the Kronecker-delta, it reduces to the form as in Eq. 1 (along with correspondence n + m → n). Notice that in Eq. 8 the variables $(\xi_1,\xi_2,\ldots,\xi_{n+m})$ correspond to the $(x_1,\ldots,x_n,z_1,\ldots,z_m)$ as grouped in Eqs. 6,7, and in Eqs. 6,7 the $\mathrm{cov}(x_i,x_j) = \delta_{ij}(\Delta x_i)^2$ for i,j=1,…n, but generally $\mathrm{cov}(z_i,z_j) \neq 0$ if $i \neq j$ and generally $\mathrm{cov}(x_i,z_j) \neq 0$. The products (terms) in the double sum for $(s_f)^2$ in Eq. 8 (belonging to the terminology of probability theory) can be identified one by one with the products (terms) from the expansion of Eqs. 6 or 7, but the latter contains partial derivatives only (as entities from calculus).

Acknowledgements

Financial support for this research from LIPOMEDICINA and NANOSEN9 at RCNS-HAS as well as from Hungarian national found OTKA-104195 and 112312 (2012) is kindly acknowledged.

References

1 Hosmer D. W., Lemeshow S., May S., *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. John Wiley & Sons, Inc., Hoboken, New Jersey, USA (2008).

2 Oehlert G. W., *A Note on the Delta Method*. American Statistician, 46(1), 27-29 (1992). DOI: 10.1080/00031305.1992.10475842

3 Rice J., *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, CA, USA (1994).

4 Taylor J. R., *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books, Sausalito, CA, USA (1997).

5 Bork P. V., Grote H., Notz D., Regler M., *Data Analysis Techniques in High Energy Physics Experiments.* Cambridge University Press, Cambridge (1993).

6 Rencher A. C., Pun F. C., *Inflation of R^2 in Best Subset Regression*. Technometrics, 22(1), 49-53 (1980). DOI: 10.1080/00401706.1980.10486100

7 Kristyán S., *A Least-Square Computation Method for Smoothing and Differentiation of Two-Dimensional Data*. Periodica Polytechnica Electrical Engineering, 33 (1-2), 63-70 (1989).