Corruption of the Pearson correlation coefficient by measurement error: estimation, bias, and correction under different error models

Edoardo Saccenti[1,+,*], **Margriet H. W. B. Hendriks**[2], **and Age K. Smilde**[3,+]

[1]Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen, the Netherlands.
[2]DSM Biotechnology Center, Delft, the Netherlands
[3]Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, the Netherlands.
[*]esaccenti@gmail.com
[+]These authors contributed equally to this work

ABSTRACT

Correlation coefficients are abundantly used in the life sciences. Their use can be limited to simple exploratory analysis or to construct association networks for visualization but they are also basic ingredients for sophisticated multivariate data analysis methods. It is therefore important to have reliable estimates for correlation coefficients. In modern life sciences, comprehensive measurement techniques are used to measure metabolites, proteins, gene-expressions and other types of data. All these measurement techniques have errors. Whereas in the old days, with simple measurements, the errors were also simple, that is not the case anymore. Errors are heterogeneous, non-constant and not independent. This hampers the quality of the estimated correlation coefficients seriously. We will discuss the different types of errors as present in modern comprehensive life science data and show with theory, simulations and real-life data how these affect the correlation coefficients. We will briefly discuss ways to improve the estimation of such coefficients. Submitted to *Scientific Reports* on June 14, 2019.

1 Introduction

The concept of correlation and correlation coefficient dates back to Bravais[1] and Galton[2] and found its modern formulation in the work of Fisher and Pearson[3,4], whose product–moment correlation coefficient ρ has become the most used measure to describe the linear dependence between two random variables. From the pioneering work of Galton on heredity, the use of correlation (or co-relation as is it was termed) spread virtually in all fields of research and results based on it pervade the scientific literature.

Correlations are generally used to quantify, visualize and interpret bivariate (linear) relationships among measured variables. They are the building blocks of virtually all multivariate methods such as Principal Component Analysis (PCA[5–7]), Partial Least Squares regression, Canonical Correlation Analysis (CCA[8]) which are used to reduce, analyze and interpret high-dimensional *omics* data sets and are often the starting point for the inference of biological networks such as metabolite-metabolite associations networks[9,10], gene regulatory networks[11,12] an co-expression networks[13,14].

Fundamentally, correlation and correlation analysis are pivotal for understanding biological systems and the physical world. With the increase of comprehensive measurements (liquid-chromatography mass-spectrometry, nuclear magnetic resonance, gas-chromatography mass-spectrometry in metabolomics and proteomics; RNA-sequencing in transcriptomics) in life sciences, correlations are used as a first tool for

visualization and interpretation, possibly after selection of a threshold to filter the correlations. However, the complexity and the difficulty of estimating correlation coefficients is not fully acknowledged.

Measurement error is intrinsic to every experimental technique and measurement platform, be it a simple ruler, a gene sequencer or a complicated array of detectors in a high-energy physics experiment, and in the early days of statistics it was known that measurement errors can bias the estimation of correlations[15]. This bias was called attenuation because it was found that under the error condition considered, the correlation was attenuated towards zero. The attenuation bias has been known and discussed in some research fields[16–19] but it seems to be totally neglected in modern *omics*-based science. Moreover, contemporary comprehensive *omics* measurement techniques have far more complex measurement error structures than the simple ones considered in the past and on which early results were based.

In this paper, we intend to show the impact of measurement errors on the quality of the calculated correlation coefficients and we do this for several reasons. First, to make the *omics* community aware of the problems. Secondly, to make the theory of correlation up to date with current *omics* measurements taking into account more realistic measurement error models in the calculation of the correlation coefficient and third, to proposes ways to alleviate the problem of distortion in the estimation of correlation induced by measurement error. We will do this by deriving analytical expressions supported by simulations and simple illustrations. We will also use real-life metabolomics data to illustrate our findings.

2 Measurement error models

We start with the simple case of having two correlated biological entities x_0 and y_0 which are randomly varying in a population. This may, e.g, be concentrations of two blood-metabolites in a cohort of persons or gene-expressions of two genes in cancer tissues. We will assume that those variables are normally distributed

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \sim N(\mu, \Sigma) \tag{1}$$

with underlying parameters

$$\mu = \begin{pmatrix} \mu_{x_0} \\ \mu_{y_0} \end{pmatrix} \tag{2}$$

and

$$\Sigma_0 = \begin{pmatrix} \sigma_{x_0}^2 & \rho_0 \sigma_{x_0 y_0} \\ \rho_0 \sigma_{x_0 y_0} & \sigma_{y_0}^2 \end{pmatrix} \tag{3}$$

Under this model the variance components $\sigma_{x_0}^2$ and $\sigma_{y_0}^2$ describe the biological variability for x_0 and y_0, respectively. The expected correlation coefficient is given by

$$\mathrm{E}[r] = \rho = \rho_0 = \frac{\sigma_{x_0 y_0}}{\sqrt{\sigma_{x_0}^2 \sigma_{y_0}^2}} \tag{4}$$

This is the correlation coefficient that is sought to be estimated from measured data, using the sample correlation coefficient r as a proxy. Assuming that N samples are taken, the sample correlation r is calculated as

$$r = \frac{\sum_{i=1}^{N} (x_{0i} - \bar{x}_0)(y_{0i} - \bar{y}_0)}{N s_x s_y} \tag{5}$$

where $(\bar{x}_0, \bar{y}_0)$ is the sample mean over N observations and s_x, s_y are the usual sample standard deviation estimators. We will use three different symbols to distinguish between three different correlation coefficients: r will indicate the sample correlation coefficient, $E(r) = \rho$ is the expected value of r and, finally, ρ_0 is the true biological correlation coefficient. In the errorless case, ρ equals ρ_0, but this does not hold anymore for measurements with error.

Whatever the nature of the variables x_0 and y_0 and whatever the experimental technique used to measure them there is always a random error component (also refereed to as noise or uncertainty) associated with the measurement procedure. This random error is by its own nature not reproducible (in contrast with systematic error which is reproducible and can be corrected for) but can be modeled, *i.e.* described, in a statistical fashion. Such models have been developed and applied in virtually every area of science and technology and can be used to adjust for measurement errors or to describe the bias introduced by it. In the following section we will introduce three error models and will show with both simulated and real data how measurement error impacts the estimation of the Pearson correlation coefficient.

2.1 Additive error

The most simple error model is the additive error model where the measured entities x and y are modeled as

$$\begin{cases} x = x_0 + \varepsilon_{au_x} \\ y = y_0 + \varepsilon_{au_y} \end{cases} \tag{6}$$

where it is assumed that the error components ε_{au_x} and ε_{au_y} are independently normally distributed around zero with variance σ^2_{au} and $\sigma^2_{au_x}$ and are also independent from x_0 and y_0. The subscripts au_x, au_y stand for *a*dditive *u*ncorrelated error (ε) on variable x and y.

Variable x and y represent measured quantities accessible to the experimenter. This error model describes the case in which the measurement error causes within-sample variability, which means that p measurement replicates $x_{i,1}, x_{i,2}, \ldots x_{i,p}$ of observation x_i of variable x will all have slightly different values due to the random fluctuation of the error component ε_{au_x}; the extent of the variability among the replicates depends on the magnitude of the error variance $\sigma^2_{au_x}$ (and similarly for the y variable). This can be seen in Figure 1A where it is shown that in the presence of measurement error (*i.e.* $\sigma^2_{au_x}, \sigma^2_{au_x} > 0$) the two variables x and y are more dispersed. Due to the measurement error, the expected correlation coefficient (see Equation (4)) is always biased downwards, *i.e.* $\rho < \rho_0$, as already shown by Spearman[15] (see Figure 1B) who also provided an analytical expression for the attenuation of the correlation coefficient as a function of the error components (a modern treatment can be found in reference[20]):

$$\rho = A\rho_0 \tag{7}$$

where

$$A = \frac{1}{\sqrt{\left(1 + \frac{\sigma^2_{au_x}}{\sigma^2_{x_0}}\right)\left(1 + \frac{\sigma^2_{au_y}}{\sigma^2_{y_0}}\right)}}. \tag{8}$$

Equation (7) implies that the expected sample correlation $\rho = \mathrm{E}[r]$ is different from the true correlation ρ_0 which is sought to be estimated. The attenuation A is always strictly smaller than 1 and it is a decreasing function of the size of the measurement error relative to the biological variation (see Figure 1C), as it

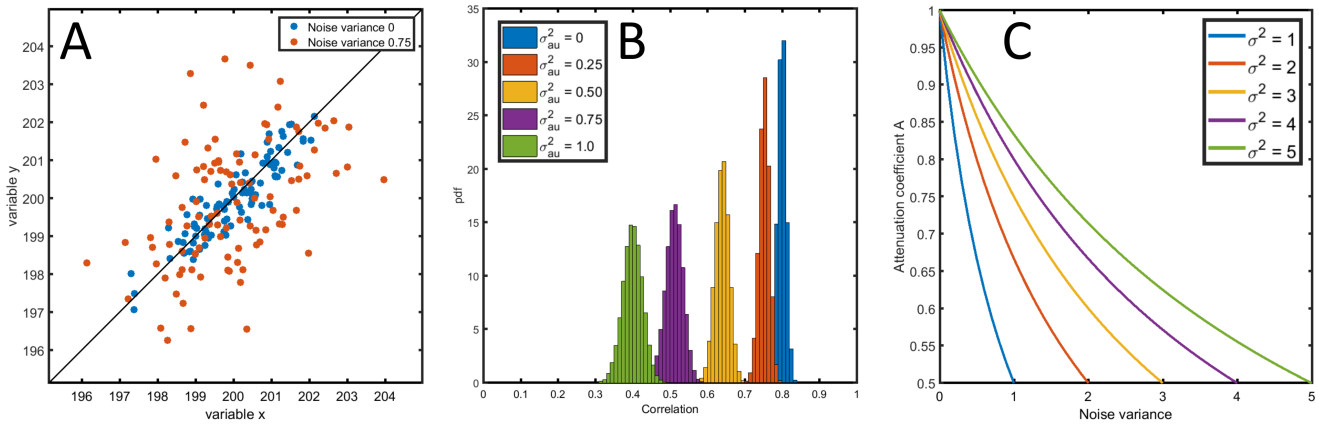

Figure 1. **A**: Correlation plot of two variables x and y ($\sigma^2_{x_0} = \sigma^2_{y_0} = 1$) generated without ($\sigma^2_{au_x} = \sigma^2_{au_y} = 0$) and with uncorrelated additive noise ($\sigma^2_{au_x} = \sigma^2_{au_y} = 0.75$) with underlying true correlation $\rho_0 = 0.8$ (model 6). **B**: Distribution of the sample correlation r for different levels of measurement error ($\sigma^2_{au} = \sigma^2_{au_x} = \sigma^2_{au_y}$) for a true correlation $\rho_0 = 0.8$.**C**: The attenuation coefficient A from Equation (8) as a function the measurement error for different level of the variance $\sigma^2 = \sigma^2_{x_0} = \sigma^2_{y_0}$ of the variables x_0 and y_0. See Material and Methods section 6.1.1 for details on the simulations.

can be seen from Equation (8). The attenuation of the correlation, despite being known since 1904, has sporadically resurfaced in the statistical literature in the psychological, epidemiology and behavioral sciences (where it is known as attenuation due to intra-person or intra-individual variability, see[19] and reference therein) but has been largely neglected in the life sciences, despite its relevance.

It must be remarked that attenuation of the correlation is a population characteristic, that is, it is a property of the expected value of the sample correlation r as shown in Figure 1B. This holds true also for the distortion induced by other types of measurement error which will be discussed in Section 3.1, 3.2 and 3.3. The error model (6) can be extended to include a correlated error term ε_{ac}

$$\begin{cases} x = x_0 + \varepsilon_{au_x} + \varepsilon_{ac} \\ y = y_0 + \varepsilon_{au_y} \pm \varepsilon_{ac} \end{cases} \tag{9}$$

with ε_{ac} normally distributed around zero with variance σ^2_{ac}. The '$\pm$' models the sign of the error correlation. When ε_{ac} has a positive sign in both x and y the error is positively correlated; if the sign is discordant the error is negatively correlated. The subscript ac is used to indicate *a*dditive *c*orrelated error. The variance for x (and y) is given by

$$\sigma^2_x = \sigma^2_{x_0} + \sigma^2_{au_x} + \sigma^2_{ac} \tag{10}$$

and likewise for the variable y. In general, additive correlated noise can have different causes depending on the type of instruments and measurement protocols used. For example, in transcriptomics, metabolomics and proteomics, usually samples have to be pretreated (sample work-up) prior to the actual instrumental analysis. Any error in a sample work-up step may affects all measured entities in a similar way[21]. Another example is the use of internal standards for quantification: any error in the amount of internal standard added may also affect all measured entities in a similar way. Hence, in both cases this leads to (positively) correlated measurement error. In some cases in metabolomics and proteomics the data are preprocessed using deconvolution tools. In that case two co-eluting peaks are mathematically separated and quantified.

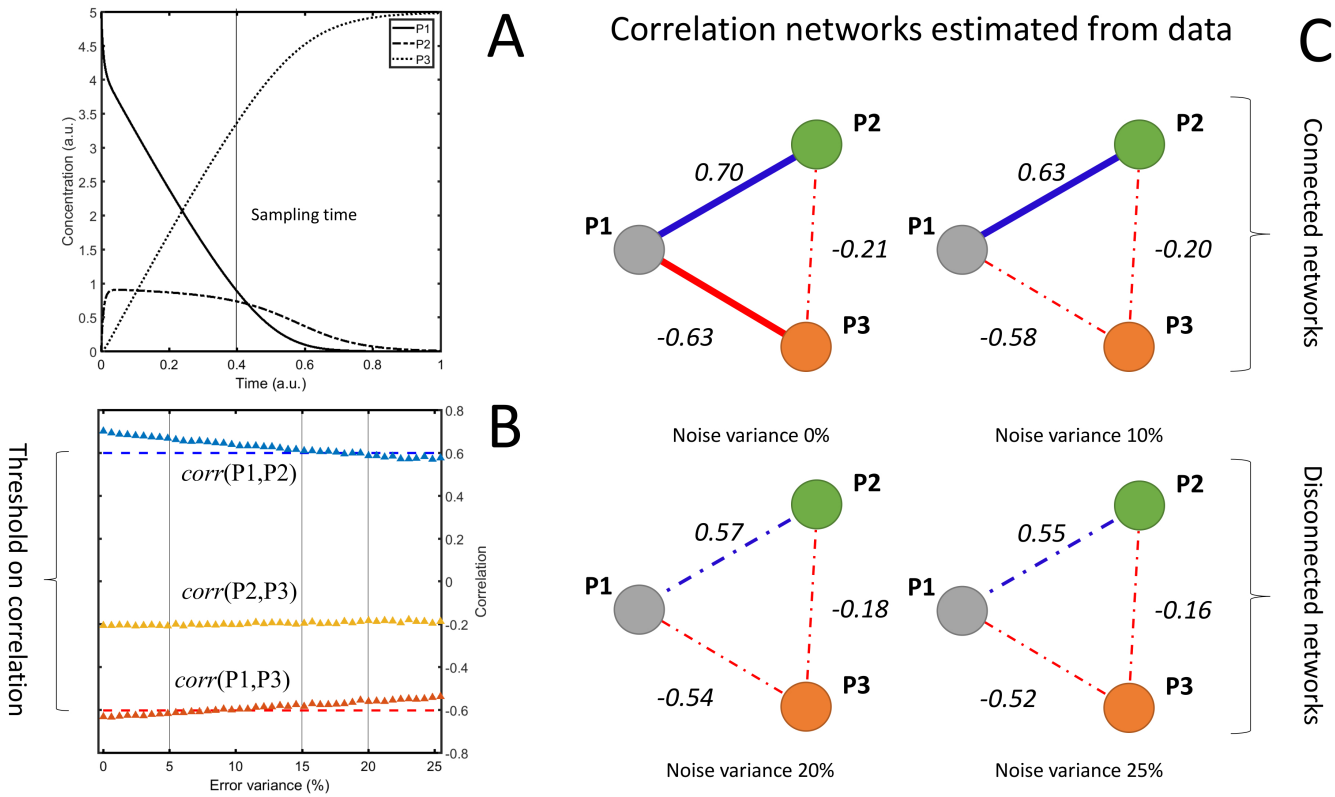

Figure 2. Consequences of measurement error when using correlation in systems biology. **A**: Time concentration profile of three metabolites P1, P2 and P3 generated through a simple enzymatic metabolic model; 100 profiles are generated by randomly varying the kinetic parameters defining the model and sampled at time 4 (a.u.). **B**: Average pairwise correlation of P1, P2 and P3 as a function of the variance of the additive uncorrelated error **C**: Inference of a metabolite-metabolite correlation network: two metabolites are associated if their correlation is above 0.6[22] (see threshold in **B**. The increasing level of measurement error hampers the network inference (compare the different panels). See Material and Methods section 6.1.2 for details on the simulations.

Since the total area under the curve is constant and (positive) error in one of the deconvoluted peak is compensated by a (negative) error in the second peak, this may give rise to negatively correlated measurement error.

To show the effect of additive uncorrelated measurement noise we consider the concentration profiles of three hypothetical metabolites P1, P2 and P3 simulated using a simple dynamic model (see Figure 2A and Section 6.1.2) where additive uncorrelated and correlated measurement noise is added before calculating the pairwise correlations among P1, P2 and P3: also in this case the magnitude of the correlation is attenuated, and the attenuation increases with the error variance (see Figure 2B). This has serious repercussions when correlations are used for the definition of association networks, as commonly done in systems biology and functional genomics[10,23]: measurement error drives correlation towards zero and this impacts network reconstruction. If a threshold of 0.6 is imposed to discriminate between correlated and non correlated variables as usually done in metabolomics[22], an error variance of around 15% (see Figure 2B, point where the correlation crosses the threshold) of the biological variation will attenuate the correlation to the point that metabolites will be deemed not to be associated even if they are biologically correlated leading to very different metabolite association networks (see Figure 2C.)

2.2 Multiplicative error

In many experimental situations it is observed that the measurement error is proportional to the magnitude of the measured signal; when this happens the measurement error is said to be multiplicative. The model for sampled variables in presence of multiplicative measurement error is

$$\begin{cases} x = x_0(1 + \varepsilon_{mu_x} + \varepsilon_{mc}) \\ y = y_0(1 + \varepsilon_{mu_y} \pm \varepsilon_{mc}) \end{cases} \tag{11}$$

where x_0, y_0, ε_{mu_x}, ε_{mu_y} and ε_{mc} have the same distributional properties as before in the additive noise case, and the last three terms represent the *m*ultiplicative *u*ncorrelated errors in x and y, respectively, and the *m*ultiplicative *c*orrelated error.

The characteristics of the multiplicative noise and the variance of the measured entities σ_x^2 depend on the level μ_{x0} of the signal to be measured (for a derivation of Equation (12) see Appendix A.1):

$$\sigma_x^2 = \sigma_{x_0}^2 + \left(\sigma_{x_0}^2 + \mu_{x_0}^2\right)\left(\sigma_{mu_x}^2 + \sigma_{mc}^2\right) \tag{12}$$

while in the additive case the standard deviation is similar for different concentrations and does not depend explicitly on the signal intensity, as shown in Equation (10).A similar equation holds for the variable y.

It has been observed that multiplicative errors often arise because of the different procedural steps like sample aliquoting[24]: this is the case of deep sequencing experiments where the multiplicative noise is possibly introduced by the pre-processing steps like, for example, linker ligation and PCR amplification which may vary from tag to tag and from sample to sample[25]. In other cases the multiplicative noise arise from the distributional properties of the signal, like in those experiments where the measurement come downs to counts like in the case of RNA fragments in an RNA-seq experiments or numbers of ions in a mass-spectrometer that are governed by Poisson distributions for which the standard deviation is equal to the mean. For another example, in NMR spectroscopy measured intensities are affected by the sample magnetization conditions: fluctuations in the external electromagnetic field or instability of the rf pulses affects the signal in a fashion that is proportional to the signal itself[26].

A multiplicative error distorts correlations and this affects the results of any data analysis approach which is based on correlations. To show the effect of multiplicative noise we consider the analysis of a simulated metabolomic data set starting from real MS data on which extra uncorrelated and correlated multiplicative measurement errors have been added. As can be seen in Figure 3A the addition of noise affects the underlying data structure: the noise free data is such that only a subset of the measured variables contributes to explain the pattern in a low dimensional projection of the data, *i.e.* have PCA loadings substantially different from zero (3B). The addition of extra multiplicative noise perturbs the loading structure to the point that all variables contribute equally to the model (3C), obscuring the real data structure and hampering the interpretation of the PCA results. This is not necessarily caused by the multiplicative nature of the noise, it is certainly caused by the correlated error part. Since the term ε_{mc} is common to all variables it introduces the same amount of correlation among all the variables and this leads to all the variables contributing to the latent vector (principal component). One may also observe that the variation explained by the first principal component increases when adding the correlated measurement error.

2.3 Realistic error

The measurement process usually consists of different procedural steps and each step can be viewed as a different source of measurement error with its own characteristics, which sum to both additive and

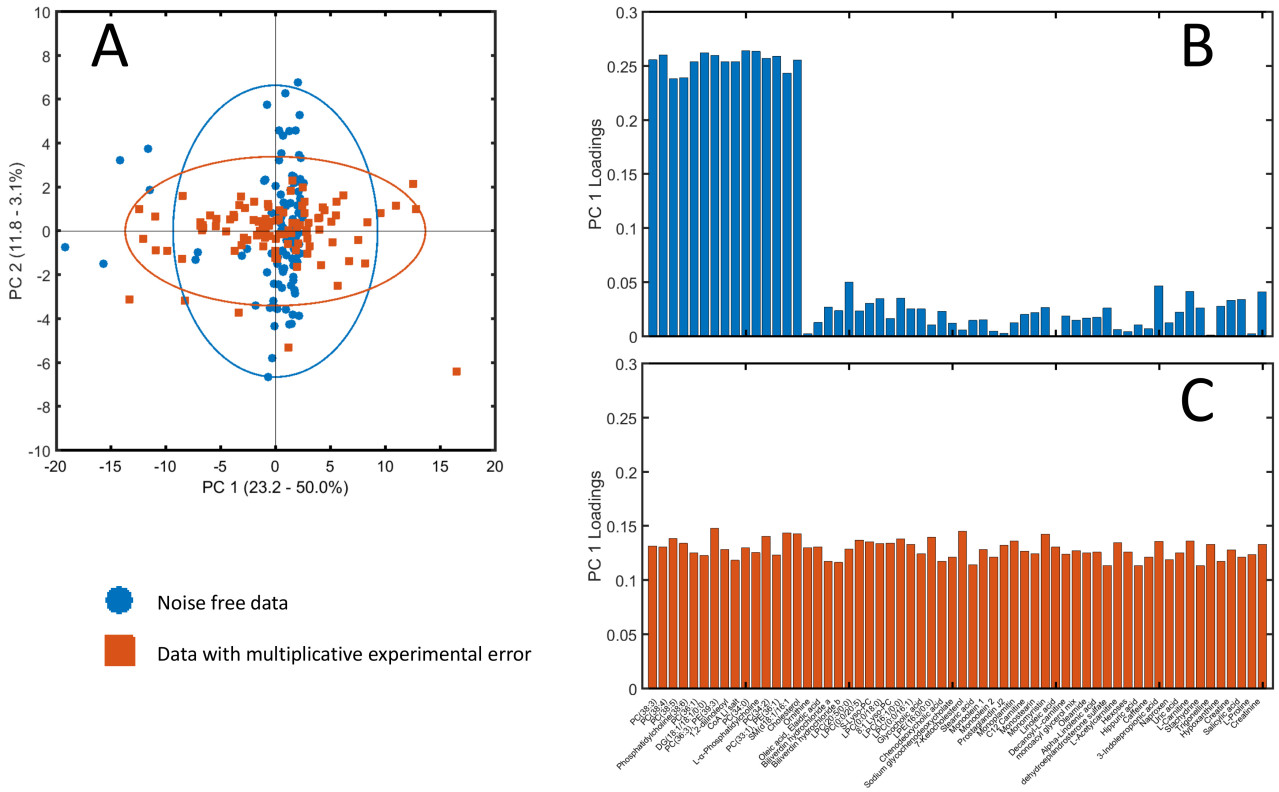

Figure 3. Consequences of correlated measurement error for data analysis. **A** Scatter plot of the overlayed view of the first two components of two PCA models of simulated data set one without and one with multiplicative error. For visualization purposes are plotted in the same graph, but the subspaces spanned by the first two principal components for the two data sets are of course different. The labels on both axis also present the percentages explained variation for the two analyses. **B** Loading plot for the error free data. **C** Loading plot for the data with multiplicative error. See Material and Methods section 6.1.3 for details on the simulations.

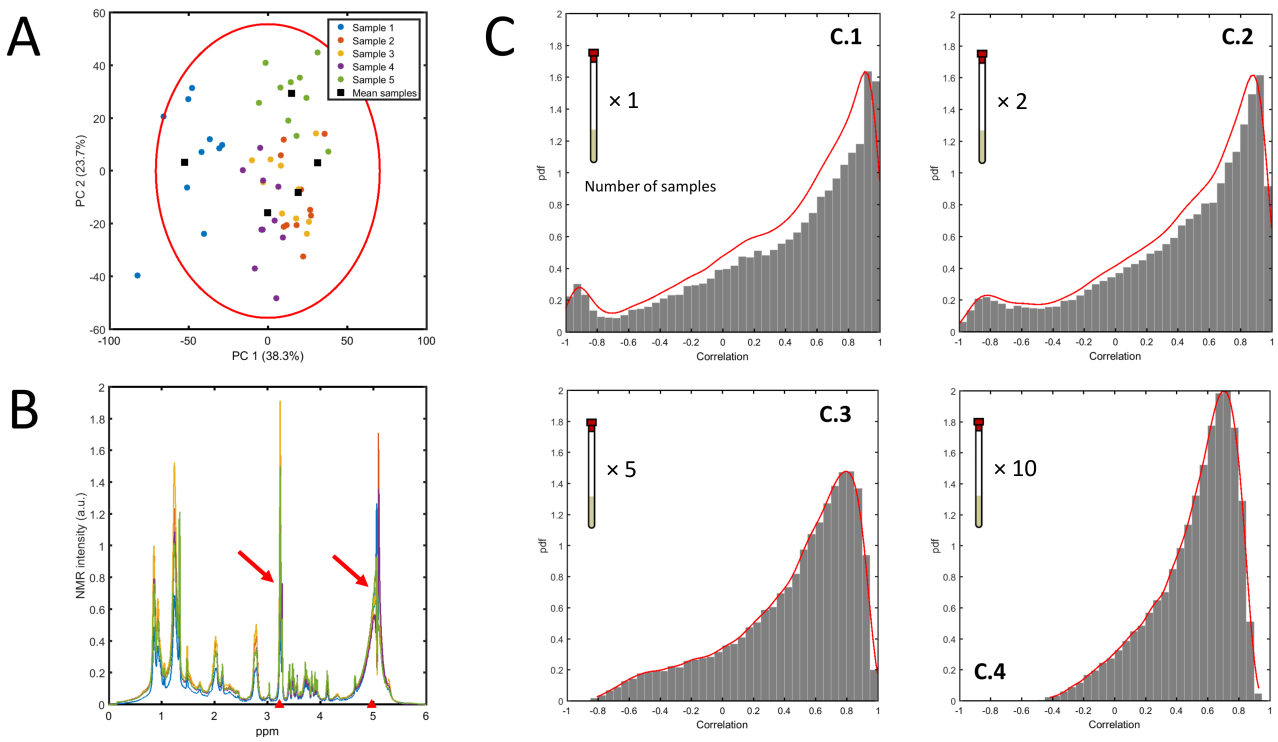

Figure 4. **A**: PCA plot of 5 different samples of fish extracts measured with technical replicates (15×) using NMR[29]. **B**: Overlap of the average binned NMR spectra of the 5 samples: the two resonances whose correlation is investigated are highlighted (3.23 and 4.98 ppm). **C**: Distribution of the sample correlation between the two resonances calculated, taking as input the average over different numbers of technical replicates (see inserts). See Material and Methods section 6.1.4 for more details on the estimation procedure.

multiplicative error components as is the case of comprehensive *omics* measurements[27]. The model for this case is:

$$\begin{cases} x = x_0(1+\varepsilon_{mu_x}+\varepsilon_{mc})+\varepsilon_{au_x}+\varepsilon_{ac} \\ y = y_0(1+\varepsilon_{mu_y}\pm\varepsilon_{mc})+\varepsilon_{au_y}\pm\varepsilon_{ac} \end{cases} \tag{13}$$

where all errors have been introduced before and are all assumed to be independent of each other and independent of the true (biological) signals (x_0 and y_0).
This realistic error model has a multiplicative as well as an additive component and also accommodates correlated and uncorrelated noise. It is an extension of a much-used error model for analytical chemical data which only contains uncorrelated error[28]. From model (13) it follows that the error changes not only quantitatively but also qualitatively with changing signal intensity: the importance of the multiplicative component increases when the signal intensity increases, whereas the relative contribution of the additive error component increases when the signal decreases.

Since most of the measurements do not usually fall at the extremity of the dynamic range of the instruments used, the situation in which both additive and multiplicative error are important is realistic. For example, this is surely the case of comprehensive NMR and Mass Spectrometry measurements, where multiplicative errors are due to sample preparation and carry-over effect (in the case of MS) and the additive noise is due to thermal noise in the detectors[29]. To illustrate this we consider an NMR experiment where a different number of technical replicates are measured for five samples (Figure 4A and 4B). We are

interested in establishing the correlation patterns across the (binned) resonances. For sake of simplicity we focus on two resonances, binned at 3.22 and 4.98 ppm. If one calculates the correlation using only one (randomly chosen) replicate per sample the resulting correlation can be anywhere between -1 and 1 (see Figure 4C.1). The variability reduces considerably if more replicates are taken and averaged before calculating the correlation ((see Figure 4C), but there is still a rather large variation, induced by the limited sample size.Averaging across the technical replicates reduces variability among the sample means: however this not accompanied by an equal reduction in the variability of the correlation estimation. This is because the error is not taken into account in the calculation of the correlation coefficient given by Equation (4)) which is derived under a zero error model (see Equation (1)).

3 Estimation of Pearson's correlation coefficient in presence of measurement error

Any two measured variables x and y can be considered as random variables and the expected correlation $\mathrm{E}[r] = \rho$ is defined using the Pearson's moment formula:

$$\mathrm{E}[r] = \rho = \frac{\mathrm{E}[xy] - \mathrm{E}[x]\,\mathrm{E}[y]}{\sqrt{\mathrm{E}[x^2] - \mathrm{E}[x]^2}\sqrt{\mathrm{E}[y^2] - \mathrm{E}[y]^2}} \tag{14}$$

where r is given by Equation (5) by taking x and y in place of x_0 and y_0.

In the ideal case of an error free measurement where the only variability is due to intrinsic biological variation, the expected ρ coincides with the true correlation ρ_0 and is given by Equation (4) and estimated using Equation (5). If additive uncorrelated error is present the expected correlation is given by Equations (7) and (8) which explicitly take into account the error component. In the next Section we will derive analytical expressions, akin to Equations (7) and (8), for the expected correlation for variables sampled with measurement error (additive, multiplicative and realistic) as introduced in Section 2.

The sample correlation r tends to ρ_0 for indefinitely sample size, $N \to \infty$, but it is is biased estimator for finite sample size of ρ_0 since[30]:

$$\mathrm{E}[r] = \rho = \rho_0 \left(1 + \frac{1-\rho_0^2}{2N} + O\left(\frac{1}{N^2}\right)\right), \tag{15}$$

with $\rho < \rho_0$. Although this is relevant for practical applications and a correction formula can be derived by taking $\rho = r\left(1 + \frac{1-r^2}{2N}\right)$ (for more on this see reference[30,31] and references therein), in the following we will not consider distortions in the correlation arising from this estimation bias (15) or from finite sample size ($N < \infty$) (these will be topics of a follow-up paper). Instead, we will focus exclusively on distortions arising because of the presence of measurement error and this is the only bias we refer to.

Before moving on, we define more specifically the error components and some properties which will be useful for the calculation of the expected correlation coefficient. The error terms comparing in models (9), (11) and (13) are assumed to have the following distributional properties

$$\begin{pmatrix} \varepsilon_{au_x} \\ \varepsilon_{au_y} \end{pmatrix} \sim N(0, \Sigma_A) \quad \text{and} \begin{pmatrix} \varepsilon_{mu_x} \\ \varepsilon_{mu_y} \end{pmatrix} \sim N(0, \Sigma_M) \tag{16}$$

with

$$\Sigma_A = \begin{pmatrix} \sigma^2_{au_x} & 0 \\ 0 & \sigma^2_{au_y} \end{pmatrix} \quad \text{and} \Sigma_M = \begin{pmatrix} \sigma^2_{mu_x} & 0 \\ 0 & \sigma^2_{mu_y} \end{pmatrix} \tag{17}$$

and

$$\varepsilon_{mc} \sim N(0, \sigma^2_{mc}) \quad \text{and} \varepsilon_{ac} \sim N(0, \sigma^2_{ac}) \tag{18}$$

From definitions (16), (17) and (18) the following properties can be derived

p1) The expected value of the errors $\mathrm{E}[\varepsilon_\alpha]$ is zero:

$$\mathrm{E}[\varepsilon_\alpha] = 0 \quad \forall\ \alpha \quad \text{in } \{\mathrm{au}, \mathrm{ac}, \mathrm{mu}, \mathrm{mc}\}. \tag{19}$$

p2) The covariance between x_0 (y_0) and the error terms is zero because x_0 (y_0) and errors are independent,

$$\mathrm{E}[x_0 \varepsilon_\alpha] - \mathrm{E}[x_0]\,\mathrm{E}[\varepsilon_\alpha] = 0 \quad \forall\ \alpha \text{ in } \{\mathrm{au}, \mathrm{ac}, \mathrm{mu}, \mathrm{mc}\} \tag{20}$$

p3) The covariance between the different error components is zero because the errors are independent from each other.

$$\mathrm{E}[\varepsilon_\alpha \varepsilon_{\alpha'}] - \mathrm{E}[\varepsilon_\alpha]\,\mathrm{E}[\varepsilon_{\alpha'}] = 0 \quad \forall\ \alpha, \neq \alpha' \text{ in } \{\mathrm{au}, \mathrm{ac}, \mathrm{mu}, \mathrm{mc}\}. \tag{21}$$

3.1 The Pearson correlation in the presence of additive noise

We show here a detailed derivation of the expected correlation among two variables x and y sampled under the additive error model (9). The variance for variable x (similar considerations hold for y) is given by

$$\mathrm{var}(x) = \mathrm{E}[x^2] - \mathrm{E}[x]^2 \tag{22}$$

where

$$\mathrm{E}[x] = \mathrm{E}[x_0 + \varepsilon_{au_x} + \varepsilon_{ac}] = \mu_{x_0}. \tag{23}$$

and

$$\begin{aligned} \mathrm{E}[x^2] &= \mathrm{E}[x_0^2 + \varepsilon^2_{au_x} + \varepsilon^2_{ac} + 2x_0\varepsilon_{au_x} + 2x_0\varepsilon_{ac} + 2\varepsilon_{au_x}\varepsilon_{ac}] \\ &= \sigma^2_{x_0} + \mu^2_{x_0} + \sigma^2_{au_x} + \sigma^2_{ac} \end{aligned} \tag{24}$$

It follows that

$$\mathrm{var}(x) = \sigma^2_{x_0} + \sigma^2_{au_x} + \sigma^2_{ac} \tag{25}$$

The covariance of x and y is

$$\mathrm{cov}(x, y) = \mathrm{E}[xy] - \mathrm{E}[x]\,\mathrm{E}[y] \tag{26}$$

with

$$\mathrm{E}[xy] = \mathrm{E}[x_0 y_0 + x_0\varepsilon_{au_y} + x_0\varepsilon_{ac} + \varepsilon_{au_x} y_0 + \varepsilon_{au_x}\varepsilon_{au_y} + \varepsilon_{au_x}\varepsilon_{ac} + \varepsilon_{ac} y_0 + \varepsilon_{ac}\varepsilon_{au_y} + \varepsilon^2_{ac}] \tag{27}$$

From considerations (20) and (21), Equation (27) reduces to

$$\mathrm{E}[xy] = \mathrm{E}[x_0 y_0] + \mathrm{E}[\varepsilon^2_{ac}] \tag{28}$$

with

$$\mathrm{E}[x_0 y_0] = \mathrm{cov}(x_0, y_0) + \mathrm{E}[x_0]\,\mathrm{E}[y_0] \tag{29}$$
$$= \sigma_{x_0 y_0} + \mu_{x_0}\mu_{y_0} \tag{30}$$

and $\mathrm{E}[\varepsilon_{ac}^2] = \pm\sigma_{ac}^2$ ($\pm$ depending on the sign of the measurement error correlation. From Equations (23), (28) and (29) it follows

$$\mathrm{cov}(x, y) = \sigma_{x_0 y_0} \pm \sigma_{ac}^2 \tag{31}$$

Plugging (25) and (31) into (14) and defining the attenuation coefficient A^a

$$A^a = \frac{1}{\sqrt{1 + \frac{\sigma_{au_x}^2}{\sigma_{x_0}^2} + \frac{\sigma_{ac}^2}{\sigma_{x_0}^2}}\sqrt{1 + \frac{\sigma_{au_y}^2}{\sigma_{y_0}^2} + \frac{\sigma_{ac}^2}{\sigma_{y_0}^2}}} = \frac{1}{\sqrt{1 + \xi_x^2 + \gamma_x^2}\sqrt{1 + \xi_y^2 + \gamma_y^2}}, \tag{32}$$

where $\xi_x^2 = \sigma_{au_x}^2/\sigma_{x_0}^2$, $\xi_y^2 = \sigma_{au_y}^2/\sigma_{y_0}^2$, $\gamma_x^2 = \sigma_{ac}^2/\sigma_{x_0}^2$ and $\gamma_y^2 = \sigma_{ac}^2/\sigma_{y_0}^2$. The expected Pearson correlation in presence of additive measurement error is obtained as:

$$\rho = \mathrm{E}[r] = A^a(\rho_0 \pm \gamma_x\gamma_y) \tag{33}$$

where the sign $\pm$ signifies positively and negatively correlated error.

The attenuation coefficient A^a is a decreasing function of the measurement error ratios, that is, the ratio between the variance of the uncorrelated and the correlated error to the variance of the true signal. Compared to Equation (7), in formula (33) there is an extra additive term related to the correlated measurement error expressing the impact of the correlated measurement error relative to the original variation. In the presence of only uncorrelated noise (*i.e.* $\sigma_{ac}^2 = 0$), Equation (33) reduces to the Spearman's formula for the correlation attenuation given by (7) and (8). As previously discussed, in this case the correlation coefficient is always biased towards zero (attenuated).

Given the true correlation ρ_0, the Pearson correlation coefficient (33) is completely determined by the measurement error ratios. Assuming the errors on x and y to be the same ($\sigma_{au_x}^2 = \sigma_{au_y}^2 =$, $\sigma_{mu_x}^2 = \sigma_{mu_y}^2$, an assumption not unrealistic if x and y are measured with the same instrument and under the same experimental conditions during an *omics* comprehensive experiment), we can set $\xi_x = \sigma_{au_x}/\sigma_{x_0}$ and $\xi_y = \sigma_{au_y}/\sigma_{y_0}$ and taking for simplicity $\sigma_{x_0}^2 = \sigma_{y_0}^2$ (thus $\xi_x = \xi_y = \xi$ and $\gamma_x = \gamma_y = \gamma$), Equation (33) can be simplified to

$$\rho = \frac{\rho_0 \pm \gamma^2}{1 + \xi^2 + \gamma^2} \tag{34}$$

and ρ can be visualized graphically as a function of the uncorrelated and correlated measurement error ratios ξ and γ as shown in Figure 5.

In the presence of positively correlated noise, the expected correlation ρ is attenuated towards 0 if the correlated noise increases and inflated if the additive uncorrelated noise increases (Figure 5A) when $\rho_0 > 0$. If $\rho_0 < 0$ the distortion introduced by the uncorrelated noise can be so severe that the expected correlation ρ can become positive. When the error is negatively correlated (Figure 5B), the correlation ρ is biased towards 0 when $\rho_0 > 0$ (and can change sign), while it can be attenuated or inflated if $\rho_0 < 0$.

A set of rules can be derived to describe quantitatively the bias of ρ. For positive correlated measurement error (for negatively correlated measurement error see Appendix A.3) if the true correlation ρ_0

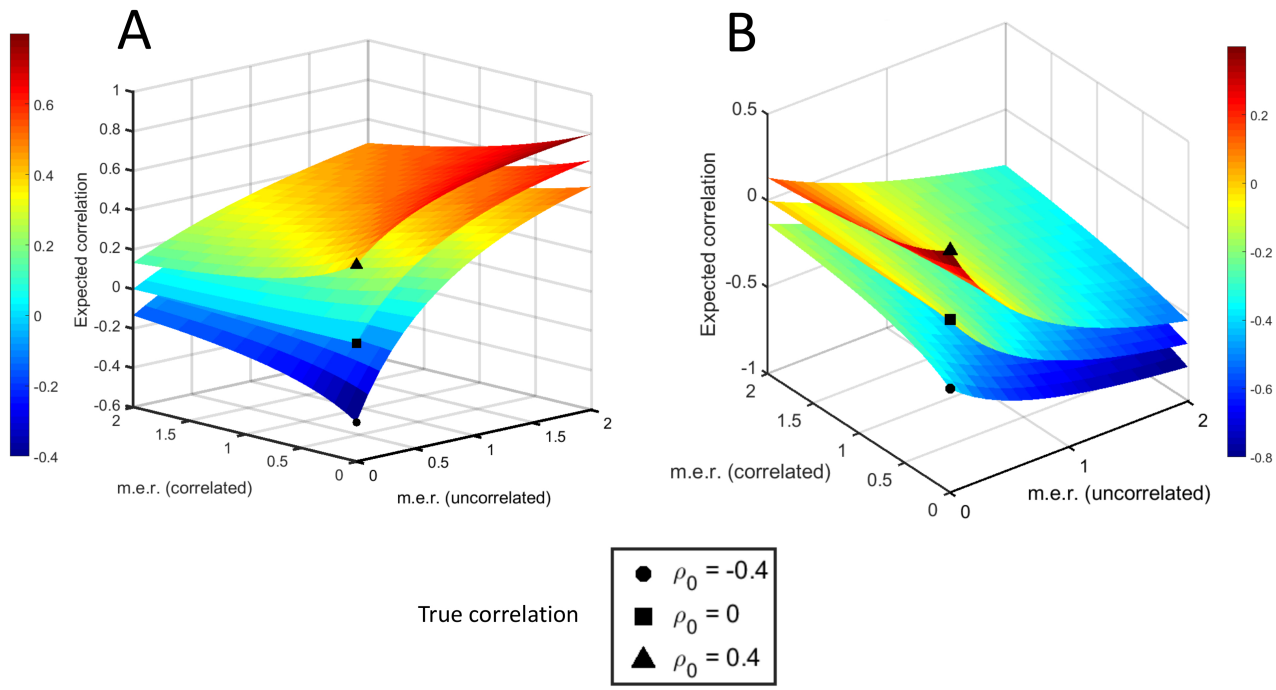

Figure 5. Expected correlation coefficient $\rho = \mathrm{E}[r]$ in the presence of additive measurement error as a function of the uncorrelated (ξ^2) and correlated (γ^2) measurement error ratios (*m.e.r.*) for different values of the true correlation ρ_0; $\mathrm{E}[r]$ is given by Equation (34). **A**: Positively correlated error. **B**: Negatively correlated error.

is positive the expected correlation ρ is always strictly positive: this is shown on Figure 6A where the relationship between ρ and ρ_0 is shown by means of Monte Carlo simulation (see Figure caption for more details). The magnitude of ρ ($\|\rho\|$) depends on how A^a (for readability in the following equations we will use A) and the additive term $\gamma_x\gamma_y > 0$ compensate each other. In particular when $\rho_0 > 0$

$$\rho \to \begin{cases} 0 < \rho < \rho_0 & \text{if } \rho_0 > \frac{A}{1-A}\gamma_x\gamma_y \\ \rho_0 & \text{if } \rho_0 = \frac{A}{1-A}\gamma_x\gamma_y \\ > \rho_0 & \text{if } \rho_0 < \frac{A}{1-A}\gamma_x\gamma_y \end{cases} \tag{35}$$

This means that ρ is always a biased estimator of the true correlation ρ_0, with the exception of the second case which happens only for specific values of γ and ρ_0. This is unlikely to happen in practice.

If $\rho_0 < 0$ it holds that

$$\rho \to \begin{cases} < \rho_0 & \text{if } -\frac{A}{A+1}\gamma_x\gamma_y < \rho_0 < 0 \\ = \rho_0 & \text{if } \rho_0 = -\frac{A}{A+1}\gamma_x\gamma_y \\ > \rho_0 & \text{if } \rho_0 < -\frac{A}{A+1}\gamma_x\gamma_y \end{cases} \tag{36}$$

The interpretation of Equation (36) is similar to that of Equation (35) but additionally, the correlation coefficient can even change sign. In particular, this happens when

$$|\rho_0| > \sqrt{\gamma_x\gamma_y} \tag{37}$$

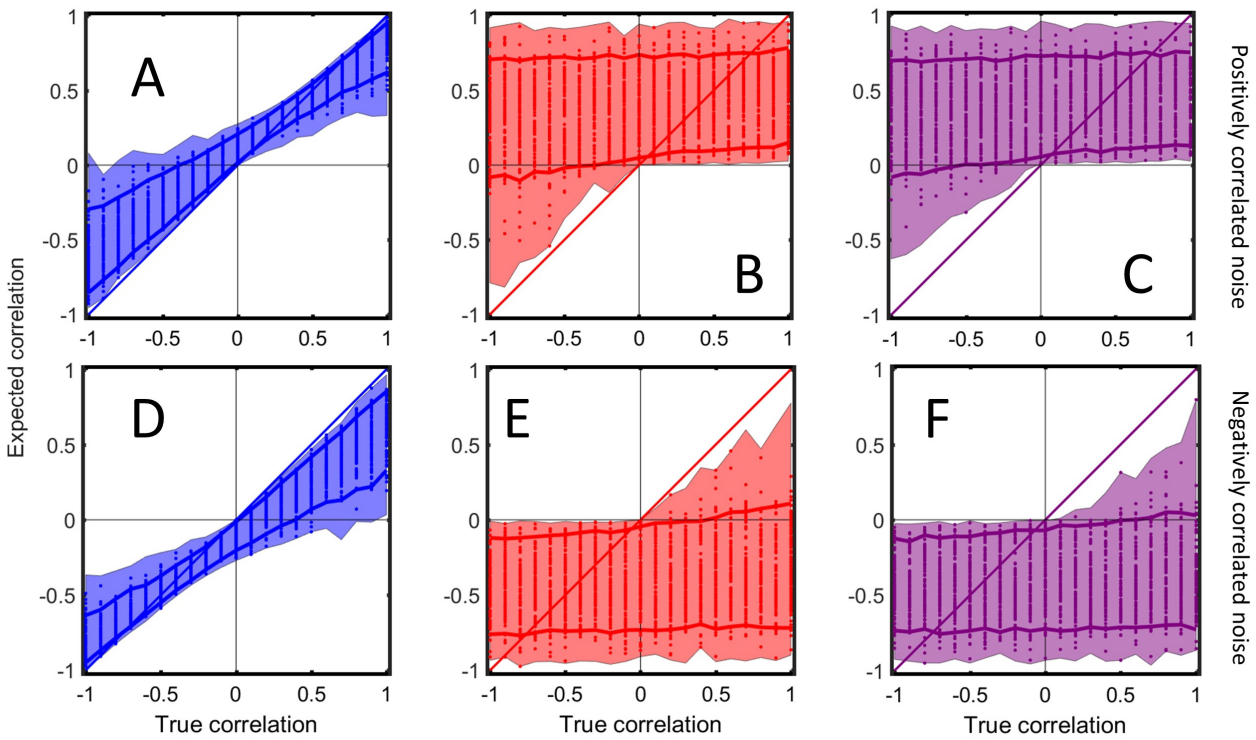

Figure 6. Calculations of the expected value of the correlation coefficient ρ as a function of the different realizations of the signal means and the size of theerror components for different values of the true correlation ρ_0. The shadowed area encloses the maximum and the minimum of the values of ρ calculated in the simulation using the different error models. The dots represent the realized values (only 100 of 10^5 Monte Carlo realizations are shown). The solid lines represent the 5-th and the 95-th percentiles of the observed values. **A**: Additive measurement error with positive correlated error. **B**: Multiplicative measurement error with positive correlated error. **C**) General case with both additive and multiplicative measurement error with positive correlated error. **D**: Additive measurement error with negative correlated error. **E**: Multiplicative measurement error with negative correlated error. **F**: Realistic case with both additive and multiplicative measurement error with negative correlated error. For more details on the simulations see Material and Methods section 6.1.5.

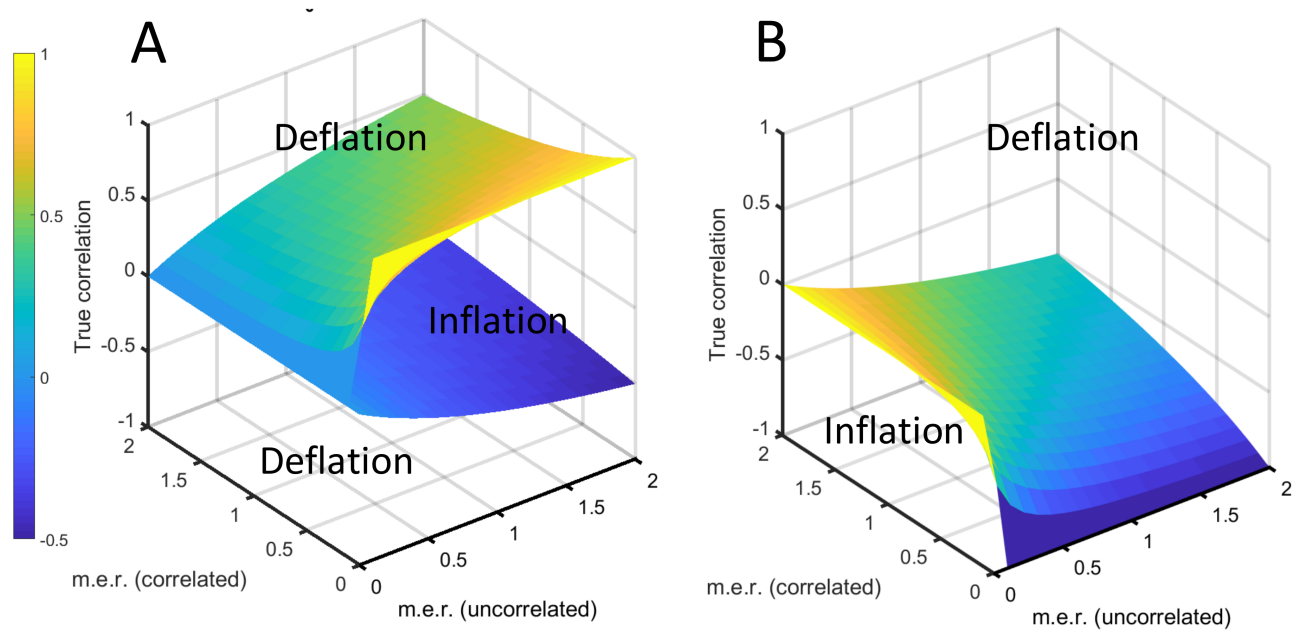

Figure 7. Limiting surfaces S for the inflation and deflation region of the expected correlation coefficient in presence of additive measurement error. The surfaces are a function of the uncorrelated (ξ^2) and correlated (γ^2) measurement error ratios (*m.e.r.*). **A**: S in the case of positively correlated error. **B**: S for negatively correlated error. The plot refers to ρ defined by Equation (34) with $\xi_x^2 = \xi_y^2 = \xi^2$ and $\gamma_x^2 = \gamma_y^2 = \gamma^2$.

The terms $S = \frac{A}{1-A}\gamma_x\gamma_y$ and $S = \frac{A}{A+1}\gamma_x\gamma_y$ in Equations (35), (36), (66) and (67) describe limiting surfaces S of ρ_0 values delineating the regions of attenuation and inflation of the expected correlation coefficient ρ. As can be seen from Figure 7, these surfaces are not symmetric with respect to zero correlation, indicating that the behavior of ρ is not symmetric around 0 with respect to the sign of ρ_0 and of the correlated error.

3.2 The Pearson correlation in presence of multiplicative noise

The correlation in the presence of multiplicative noise can be derived using similar arguments and detailed calculations can be found in the Appendix A.1. Here we only state the main result:

$$\rho = \rho_0(1 \pm \sigma_{mc}^2)A^m \pm \delta_x\delta_y\sigma_{mc}^2 A^m \tag{38}$$

with $\delta_x = \mu_{x_0}/\sigma_{x_0}$, $\delta_y = \mu_{y_0}/\sigma_{y0}$ (biological signal to biological variation ratios) and A is the attenuation coefficient:

$$A^m = \frac{1}{\sqrt{1+\left(1+\frac{\mu_{x_0}^2}{\sigma_{x_0}^2}\right)\left(\frac{\sigma_{mu_x}^2}{\sigma_{x_0}^2}+\frac{\sigma_{mc}^2}{\sigma_{x_0}^2}\right)}\sqrt{1+\left(1+\frac{\mu_{y_0}^2}{\sigma_{y_0}^2}\right)\left(\frac{\sigma_{mu_y}^2}{\sigma_{y_0}^2}+\frac{\sigma_{mc}^2}{\sigma_{y_0}^2}\right)}} \tag{39}$$

In this case, the expected sample correlation depends explicitly on the mean of the variables, as an effect of the multiplicative nature of the error component.

Expressions similar to Equation (35) and Equation (67) in Appendix A.3 can be derived but they are cumbersome to interpret. Our simulations show that if the signal intensity is not too large, the sample

correlation can change sign as shown see Figure 6B); if the signal intensity is very large the multiplicative error will have a very large effect and if the correlated error is positive the sample correlation ρ will be also positive, and will be negative if the error are negatively correlated. but simulations cannot be exhaustive as shown in Figure 6B.)

3.3 The Pearson correlation in presence of realistic noise

When both additive and multiplicative error are present, the expected correlation coefficient is a combination of formula (33) and (38) (see Appendix A.2 for detailed derivation):

$$\rho = \rho_0(1 \pm \sigma_{mc}^2)A^c \pm (\gamma_x\gamma_y + \delta_x\delta_y\sigma_{mc}^2)A^c \tag{40}$$

Where the γ and δ parameters have been previously defined for the additive and multiplicative case. A^c is the attenuation coefficient:

$$A^c = \frac{1}{\sqrt{1+\left(1+\frac{\mu_{x_0}^2}{\sigma_{x_0}^2}\right)\left(\frac{\sigma_{mux}^2}{\sigma_{x_0}^2}+\frac{\sigma_{mc}^2}{\sigma_{x_0}^2}\right)+\frac{\sigma_{aux}^2}{\sigma_{x_0}^2}+\frac{\sigma_{ac}^2}{\sigma_{x_0}^2}}\sqrt{1+\left(1+\frac{\mu_{y_0}^2}{\sigma_{y_0}^2}\right)\left(\frac{\sigma_{muy}^2}{\sigma_{y_0}^2}+\frac{\sigma_{mc}^2}{\sigma_{y_0}^2}\right)+\frac{\sigma_{auy}^2}{\sigma_{y_0}^2}+\frac{\sigma_{ac}^2}{\sigma_{y_0}^2}}}. \tag{41}$$

General rules governing the sign of the numerator and denominator in Equation (40) cannot be determined since it depends on the interplay of the six error components, the true mean and product thereof. However, the simulation results presented in Figures 6C show that the behavior of ρ in presence of realistic measurement error is qualitatively similar to that in presence of only multiplicative error.

4 Correction for correlation bias

Because virtually all kinds of measurement are affected by measurement error, the correlation calculated from sampled data is distorted to some degree depending on the level of the measurement error and on its nature. We have seen that experimental error can inflate or deflate the correlation and that ρ (and hence its sample realization r) is almost always a biased estimation of the true correlation ρ_0. An estimator that gives a theoretically unbiased estimate of the correlation coefficient between two variables x and y taking into account the measurement error model can be derived. For simple uncorrelated additive noise this is given by the Spearman's formula (42): this is a known results which in the past has been presented and discussed in many different fields[16–19]. To obtain similar correction formulas for the error models considered here it is sufficient to solve for ρ_0 from the defining Equations (33), (38) and (40). The correction formulas are as follows (the $\pm$ indicates positive and negatively correlated error):

1. Correction for simple additive error (only uncorrelated error):

$$\rho_0 = A^{-1}\rho \tag{42}$$

2. Correction for additive error:

$$\rho_{\pm}^{corrected} = \frac{1}{A^a}\rho \mp \gamma_x\gamma_y \tag{43}$$

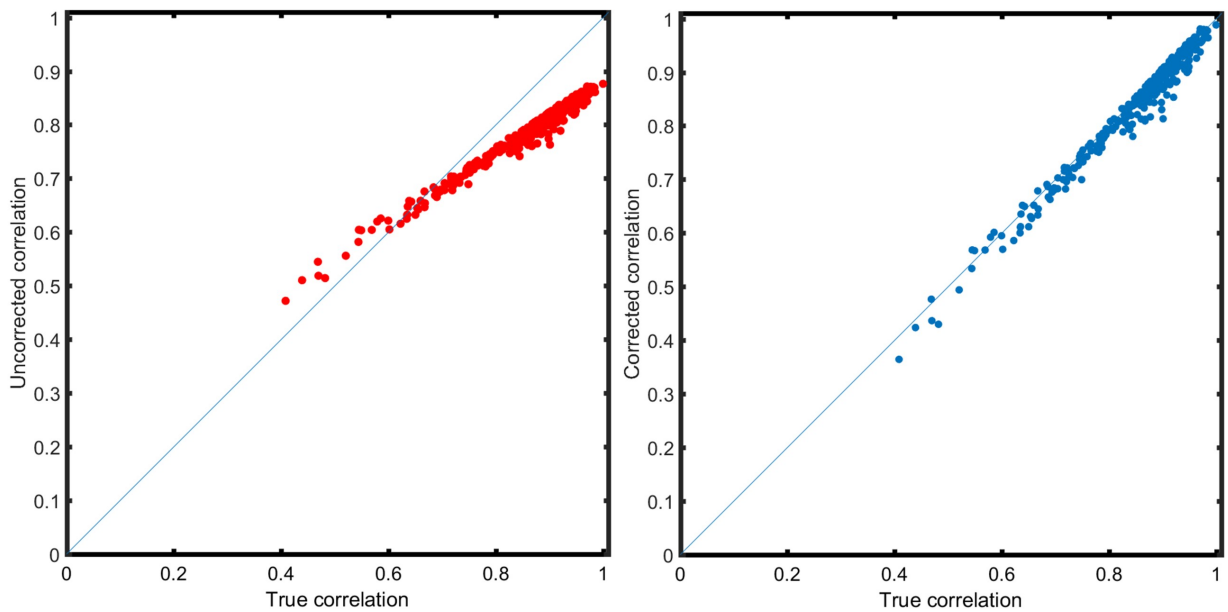

Figure 8. Correction of the distortion induced by the measurement error. **A**: Expected (pairwise) correlations ρ among 25 metabolites calculated from simulated data with additive and multiplicative measurement error *vs* the true correlation ρ_0. **B**: Corrected correlation coefficients using Equation 45 and using the known error variance components. See Section 6.1.6 for details on the data simulation.

3. Correction for multiplicative error:

$$\rho_{\pm}^{corrected} = \frac{1}{A^m(1 \pm \sigma_{ac}^2)}\rho \mp \frac{\sigma_{ac}^2}{1 \pm \sigma_{ac}^2}\delta_x\delta_y \tag{44}$$

4. Correction for realistic error:

$$\rho_{\pm}^{corrected} = \frac{1}{A^c(1 \pm \sigma_{mc}^2)}\rho \mp \frac{\gamma_x\gamma_y + \delta_x\delta_y\sigma_{mc}^2}{1 \pm \sigma_{mc}^2} \tag{45}$$

In practice, to obtain a corrected estimation of the correlation coefficient ρ_0, the ρ is substituted by r in (43), (44) and (45), which is the sample correlation calculated from the data. The effect of the correction is shown, for the realistic error model (13), in Figure 8 where the true know error variance components have been used. It should be noted that it is possible that the corrected correlation exceeds ± 1.0. This phenomenon has already been observed and discussed[16,32]: it is due to the fact that the sampling error of a correlation coefficient corrected for distortion is greater than would be that of an uncorrected coefficient of the same size (at least for the uncorrelated additive noise[4,18,33]). When this happens the corrected correlation can be rounded to ± 1.0[19,33].

5 Discussion

Since measurement error cannot be avoided, correlation coefficients calculated from experimental data are distorted to a degree which is not known and that has been neglected in life sciences applications but can expected to be considerable when comprehensive omics measurement are taken.

As previously discussed, the attenuation of the correlation coefficient in the presence of additive (uncorrelated) error has been know since more than one century. The analytical description of the distortion of the correlation coefficient in presence of more complex measurement error structures (Equations (33), (38) and (40)) has been presented here for the first time to the best of our knowledge.

The inflation or attenuation of the correlation coefficient depends on the relationship between the value of true correlation ρ_0 and the error component. In most cases in practice, ρ is a biased estimator for ρ_0. In absence of correlated error, there is always attenuation; in the presence of correlated error there can also be increase (in absolute value) of the correlation coefficient. This has also been observed in regression analysis applied to nutritional epidemiology and it has been suggested that correlated error can, in principle, be used to compensate for the attenuation[34]. Moreover, the distortion of the correlation coefficient has also implications for hypothesis testing to assess the significance of the measured correlation r.

To illustrate the counterintuitive consequences of correlated measurement error consider the following. Suppose that the true correlation is null. In that case, Equations (33), (38) and (40) reduce to

$$\rho = A^a \gamma_x \gamma_y \tag{46}$$

$$\rho = A^m \delta_x \delta_y \sigma_{mc}^2 \tag{47}$$

$$\rho = \pm(\gamma_x \gamma_y + \delta_x \delta_y \sigma_{mc}^2) A^c \tag{48}$$

which implies that the expected value of the correlation coefficient is not zero. Moreover, in real-life situations there is also sampling variability superimposed on this which may in the end result in estimated correlations of the size as found in several omics applications (in metabolomics observed correlation are usually lower than 0.6[10,22]; similar patterns are also observed in transcriptomics[35,36]) while the true biological correlation is zero.

The correction equations presented need the input of estimated variances. Such estimates also carry uncertainty and the quality of these estimates will influence the quality of the corrections. This will be the topic of a follow-up paper. Prior information regarding the sizes of the variance components would be valuable and this points to new requirements for system suitability tests of comprehensive measurements. In metabolomics, for example, it would be worthwhile to characterize an analytical measurement platform in terms of such error variances including sizes of correlated error using advanced (and to be developed) measurement protocols.

Distortion of the correlation coefficient has implications also for experimental planning. In the case of additive uncorrelated noise, the correction depends explicitly on the sample size N used to calculate r and on the number of replicates n_x, n_y used to estimate the intraclass correlation (*i.e.* the error variance components): since in real life the total sample size $N \times (n_x + n_y)$ is fixed, there is a trade off between the sample size and the number of replicates that can be measured and the experimenter has to decide whether to increase N or n_x.

The results presented here are derived under the assumption of normality of both measurement and measurement errors. If x_0 and y_0 are normally distributed, then x and y will be, in presence of additive

measurement error, normally distributed, with variance given by (10). For multiplicative and realistic error the distribution of x and y will be far from normality since it involves the distribution of the product of normally distributed quantities which is usually not normal[37]. It is known that departure from normality can result in the inflation of the correlation coefficient[38,39] and this will add to the distortion induced by the measurement error.

We think that in general correlation coefficients are trusted too much on face value and we hope to have triggered some doubts and pointed to precautions in this paper.

6 Material and Methods

6.1 Simulations

We provide here details on the simulation performed and shown in Figures 1 - 4, 6 and 8.

6.1.1 Simulations in Figure 1

$N = 100$ realizations of two variables x and y were generated under model with additive uncorrelated measurement error (9), with $\rho_0 = 0.8$, $\sigma^2_{x_0} = \sigma^2_{y_0} = 1$ and $\mu = (100, 100)$. Error variance components were set to $\sigma^2_{au_x} = \sigma^2_{au_y} = 0$ and to $\sigma^2_{au_x} = \sigma^2_{au_y} = 0.75$ (Panel A).

6.1.2 Simulations in Figure 2

The time concentrations profiles $P_1(t)$, $P_2(t)$ and $P_3(t)$ of three hypothetical metabolites P1, P2 and P3 are simulated using the following dynamic model

$$\begin{cases} \frac{d}{dt}P_1(t) = -k_1P-1(t)(E_T - P_2(t)) + k_{-1}P_2(t) \\ \frac{d}{dt}P_2(t) = -k_{-1}P_1(t) + k_1P_1(t)(E_T - P_2(t)) - k_2P_2(t) \\ \frac{d}{dt}P_3(t) = +k_2P_2(t) \end{cases}$$

which is the model of an irreversible enzyme-catalyzed reaction described by Michaelis-Menten kinetics. Using this model, $N = 100$ concentration time profiles for P1, P2 and P3 were generated by solving the system of differential equations after varying the kinetic parameters k_1, k_{-1} and k_2 by sampling them from a uniform distribution. For the realization of the jth concentration profile

$$k_1^j \approx U(0.9\times k_1, 1.1\times k_1) k_{-1}^j \approx U(0.9\times k_{-1}, 1.1\times k_{-1}) k_2^j \approx U(0.9\times k_2, 1.1\times k_2) E_T^j \approx U(0.9\times E_T, 1.1\times E_T)$$

with population values $k_1 = 30, k_{-1} = 20, k_2 = 10$, and $E_T = 1$. Initial conditions were set to $(P_{10}, P_{2_0}, P_{3_0}) = (P^j_{1_0}, 0, 0)$ with $P^j_{1_0} \approx U(0.9\times P_{10}, 1.1\times P_{1_0})$ and $P_{1_0} = 5$. All quantities in arbitrary units. Time profiles were sampled at $t = 0.4$ $a.u$ and collected in a data matrix $\mathbf{X}_0$ of size 100×3. The variability in data matrix $\mathbf{X}_0$ is given by biological variation. The concentration time profiles of P1, P2 and P3 shown in Panel A are obtained using the population values for the kinetic parameters and for the initial conditions.

Additive uncorrelated and correlated measurement error is added on $\mathbf{X}_0$ following model (9) where P1, P2 and P3 in $\mathbf{X}_0$ play the role of x_0, y_0 and of an additional third variable z_0 which follows a similar model. The variance of the error component was varied in 50 steps between 0 and 25% of the sample variance $s^2_{x_0}, s^2_{y_0}$ and $s^2_{z_0}$ calculated from $\mathbf{X}_0$. The variance of the correlated error was set to $\sigma^2_a c = 0.05$ in all simulations. Pairwise Pearson sample correlations $r_{i,j}$ with $i, j = \{P1, P2, P3\}$ were calculated for the noise free case $\mathbf{X}_0$ and for data with measurement error added. 100 noise realization were simulated for each noise values and the average correlation across the 100 realization is calculated and it is shown in Panel B.

The mini metabolite-metabolite association networks shown in Panel C are defined by first taking the sample Pearson correlation r_{ij} among P1, P2 and P3 and then imposing a threshold on r to define the Connectivity matrix A_{ij}

$$A_{ij} = \begin{cases} 1 \text{ if } |r_{ij}| > 0.6 \\ 0 \text{ otherwise} \end{cases}. \tag{49}$$

For more details see reference[10].

6.1.3 Simulations in Figure 3

Principal component analysis was performed on a 100×133 experimental metabolomic data set (see Section 6.2 for a description). The 15 variables with the highest loading (in absolute value) and the 45 variables with the smallest loading (in absolute value) on the first principal component where selected to form a 100×60 data set $\mathbf{X}_0$ (we call this now the noise free data, as if it only contained biological variation). On this subset anew a principal component analysis was performed. Then multiplicative correlated and uncorrelated measurement error was added on $\mathbf{X}_0$. The variance of the additive error was set $\sigma^2_{mu_j} = 0.05 \times s^2_{j_0}$ with $j = 1, 2, \ldots, 60$ where $s^2_{j_0}$ is the variance calculated for the jth column of $\mathbf{X}_0$, *i.e.*, the biological variance. The variance of the correlated error was fixed to 5% of the average variance observed in the noise free data ($\sigma^2_{mc} = 0.045$.

6.1.4 Simulations in Figure 4

Let x_{ij} and y_{ij} the intensities of the resonances measured at 3.23 and 4.98 in randomly drawn replicate j of sample Fi ($i = 1, 2, \ldots, 5$) and define the 5×1 vectors of means

$$\mathbf{x}_J = \frac{1}{J} \begin{pmatrix} \sum_j x_{1j} \\ \vdots \\ \sum_j x_{5j} \end{pmatrix} \text{ and } \mathbf{y}_J = \frac{1}{J} \begin{pmatrix} \sum_j y_{1j} \\ \vdots \\ \sum_j y_{5j} \end{pmatrix} \tag{50}$$

The sample correlation $r_J = \text{corr}(\mathbf{x}_J, \mathbf{y}_J)$ is calculated for $J = 1, 2, 5$, and 10; for each J the replicates used to calculate $\mathbf{x}_J$ and $\mathbf{y}_J$ are randomly and independently sampled, for each sample separately, from the total set of the 12 to 15 replicates available per sample. The procedure is repeated 10^5 times to construct the distributions of the Pearson sample correlation shown in Figure 4C.

6.1.5 Simulations in Figure 6

Simulation results presented in Figure 6 show the results from calculations of the sample correlation coefficient as a function of the true correlation ρ_0 and of the true means (μ_{x_0} and μ_{y_0}), the variances ($\sigma^2_{x_0}$ and $\sigma^2_{y_0}$ of the signals x_0 and y_0 and the measurement error variances as they appear in the definitions of ρ under the different error models (Equations (33), (38) and (40)). The calculations were done multiple times for varying values for μ_{x_0} and μ_{y_0}, which were randomly and independently sampled from a uniform distribution $U(0, \mu_0)$, where μ_0 was set to be equal to 23.4, which was the maximum values observed in Data set 1 (see Section 6.2.)Values for $\sigma^2_{x_0}$ and $\sigma^2_{y_0}$ were randomly and independently sampled from a uniform distribution $U(0, \sigma^2_0)$, where σ^2_0 was set to be equal to the average variance observed in the experimental Data set 1. The values of the variance of all error components are randomly and independently sampled from $U(0, \frac{1}{4}\sigma^2_0)$. The overall procedure was repeated 10^4 for each value of ρ_0 in the range $[-1, 1]$ in steps of 0.1.

6.1.6 Simulations in Figure 8

The first 25 variables from Data set 1 have been selected and used to compute the means μ_0 and the correlation/covariance matrix Σ_0 used to generate noise-free data $\mathbf{X}_0 \sim N(\mu_0, \Sigma_0)$ of size $10^4 \times 25$ on which additive and multiplicative measurement error (correlated and uncorrelated) is added (error model (13)) to obtain $\mathbf{X}$. All error variances are set to 0.1 which is approximately equal to 5% of the average variance observed in $\mathbf{X}_0$. Pairwise sample correlations among the 25 metabolites are calculated from $\mathbf{X}$. The correlations are corrected using Equation (45) using the known distributional and error parameters (μ_0, Σ_0) used to generate the data. The data generation is repeated 10^3 times and correlations (uncorrected and corrected) are averaged over the repetitions.

6.2 Data sets

Data set 1: A publicly available data set containing measurements of 133 blood metabolites from 2139 subjects was used as a base for the simulation to obtain realistic distributional and correlation patterns among measured features. The data comes from a designed case-cohort and a matched sub-cohort (controls) stratified on age and sex from the TwinGene project[40]. The first 100 observation were used in the simulation described in Section 6.1.3 and shown in Figure 3.

Data were downloaded from the Metabolights public repository[41] (www.ebi.ac.uk/metabolights) with accession number MTBLS93. For full details on the study protocol, sample collection, chromatography, GC-MS experiments and metabolites identification and quantification see the original publication[42] and the Metabolights accession page.

Data set 2: This data set was acquired in the framework of a study aiming to the ”Characterization of the measurement error structure in Nuclear Magnetic Resonance (NMR) data for metabolomic studies”[29]. Five biological replicates of fish extract F1 - F5 were originally pretreated in replicates (12 to 15) and acquired using ^{1}H NMR. The replicates account for variability in sample preparation and instrumental variability. For details on the sample preparation and NMR experiments we refer to the original publication.

6.3 Software

All calculations were performed in Matlab (version 2017a 9.2). Code to generate data under measurement error models (9), (11) and (13) is available at systemsbiology.nl under the SOFTWARE tab.

References

1. Bravais, A. *Analyse mathématique sur les probabilités des erreurs de situation d'un point* (Impr. Royale, 1844).
2. Galton, F. Co-relations and their measurement, chiefly from anthropometric data. *Proc. Royal Soc. Lond.* **45**, 135–145 (1889).
3. Pearson, K. Note on regression and inheritance in the case of two parents. *Proc. Royal Soc. Lond.* **58**, 240–242 (1895).
4. Spearman, C. Demonstration of formulae for true measurement of correlation. *The Am. J. Psychol.* 161–169 (1907).
5. Pearson, K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, Dublin Philos. Mag. J. Sci.* **2**, 559–572 (1901).
6. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. educational psychology* **24**, 417 (1933).

7. Jolliffe, I. *Principal component analysis* (Springer, 2011).

8. Härdle, W. & Simar, L. *Applied multivariate statistical analysis*, vol. 22007 (Springer, 2007).

9. Müller-Linow, M., Weckwerth, W. & Hütt, M.-T. Consistency analysis of metabolic correlation networks. *BMC Syst. Biol.* **1**, 44 (2007).

10. Jahagirdar, S., Suarez-Diez, M. & Saccenti, E. Simulation and reconstruction of metabolite-metabolite association networks using a metabolic dynamic model and correlation based-algorithms. *J. proteome research* (2019).

11. Dunlop, M. J., Cox III, R. S., Levine, J. H., Murray, R. M. & Elowitz, M. B. Regulatory activity revealed by dynamic correlations in gene expression noise. *Nat. genetics* **40**, 1493 (2008).

12. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804 (2012). DOI 10.1038/nmeth.2016.

13. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *science* **302**, 249–255 (2003).

14. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. applications genetics molecular biology* **4** (2005).

15. Spearman, C. The proof and measurement of association between two things. *The Am. journal psychology* **15**, 72–101 (1904).

16. Thouless, R. H. The effects of errors of measurement on correlation coefficients. *Br. J. Psychol.* **29**, 383 (1939).

17. Beaton, G. H. *et al.* Sources of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation. *The Am. journal clinical nutrition* **32**, 2546–2559 (1979).

18. Rosner, B. & Willett, W. Interval estimates for correlation coefficients corrected for within-person variation: implications for study design and hypothesis testing. *Am. journal epidemiology* **127**, 377–386 (1988).

19. Adolph, S. C. & Hardin, J. S. Estimating phenotypic correlations: correcting for bias due to intraindividual variability. *Funct. Ecol.* **21**, 178–184 (2007).

20. Fuller, W. A. *Measurement error models*, vol. 305 (John Wiley & Sons, 2009).

21. Moseley, H. N. Error analysis and propagation in metabolomics data analysis. *Comput. structural biotechnology journal* **4**, e201301006 (2013).

22. Camacho, D., de la Fuente, A. & Mendes, P. The origin of correlations in metabolomics data. *Metabolomics* **1**, 53–63 (2005). URL http://link.springer.com/10.1007/s11306-005-1107-3. DOI 10.1007/s11306-005-1107-3.

23. Rosato, A. *et al.* From correlation to causation: analysis of metabolomics data using systems biology approaches. *Metabolomics* **14**, 37 (2018).

24. Werner, M., Brooks, S. H. & Knott, L. B. Additive, multiplicative, and mixed analytical errors. *Clin. chemistry* **24**, 1895–1898 (1978).

25. Balwierz, P. J. *et al.* Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepcage data. *Genome biology* **10**, R79 (2009).

26. Mehlkopf, A., Korbee, D., Tiggelman, T. & Freeman, R. Sources of t1 noise in two-dimensional nmr. *J. Magn. Reson. (1969)* **58**, 315–323 (1984).

27. Van Batenburg, M. F., Coulier, L., van Eeuwijk, F., Smilde, A. K. & Westerhuis, J. A. New figures of merit for comprehensive functional genomics data: the metabolomics case. *Anal. chemistry* **83**, 3267–3274 (2011).

28. Rocke, D. M. & Lorenzato, S. A two-component model for measurement error in analytical chemistry. *Technometrics* **37**, 176–184 (1995).

29. Karakach, T. K., Wentzell, P. D. & Walter, J. A. Characterization of the measurement error structure in 1d 1h nmr data for metabolomics studies. *Anal. Chimica Acta* **636**, 163–174 (2009).

30. Fisher, R. A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biom.* **10**, 507–521 (1915).

31. Olkin, I., Pratt, J. W. *et al.* Unbiased estimation of certain correlation coefficients. *The Annals Math. Stat.* **29**, 201–211 (1958).

32. Pearson, K. & Lee, A. On the laws of inheritance in man: I. inheritance of physical characters. *Biom.* **2**, 357–462 (1903).

33. Winne, P. H. & Belfry, M. J. Interpretive problems when correcting for attenuation. *J. Educ. Meas.* 125–134 (1982).

34. Day, N. *et al.* Correlated measurement error—implications for nutritional epidemiology. *Int. J. Epidemiol.* **33**, 1373–1381 (2004).

35. Pereira, V., Waxman, D. & Eyre-Walker, A. A problem with the correlation coefficient as a measure of gene expression divergence. *Genet.* **183**, 1597–1600 (2009).

36. Reynier, F. *et al.* Importance of correlation between gene expression levels: application to the type i interferon signature in rheumatoid arthritis. *PloS one* **6**, e24828 (2011).

37. Springer, M. D. *The algebra of random variables* (Wiley and Sons, 1979).

38. Kowalski, C. J. On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. *J. Royal Stat. Soc. Ser. C (Applied Stat.* **21**, 1–12 (1972).

39. Bishara, A. J. & Hittner, J. B. Reducing bias and error in the correlation coefficient due to nonnormality. *Educ. psychological measurement* **75**, 785–804 (2015).

40. Magnusson, P. K. *et al.* The swedish twin registry: establishment of a biobank and other recent developments. *Twin Res. Hum. Genet.* **16**, 317–329 (2013).

41. Haug, K. *et al.* Metabolights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic acids research* **41**, D781–D786 (2012).

42. Ganna, A. *et al.* Large-scale non-targeted metabolomic profiling in three human population-based studies. *Metabolomics* **12**, 4 (2016).

A Mathematical Appendix

A.1 Derivation of ρ in presence of multiplicative measurement error

In presence of purely multiplicative error it holds

$$\mathrm{E}[x] = \mathrm{E}[x_0(1+\varepsilon_{mu_x} \pm \varepsilon_{mc})] = \mu_{x_0} \tag{51}$$

and

$$\begin{aligned}\mathrm{E}[x^2] &= \mathrm{E}[x_0^2 + x_0^2(\varepsilon_{mu_x}^2 + \varepsilon_{mc}^2 + 2\varepsilon_{mu_x} + \varepsilon_{mc} \pm \varepsilon_{mc} \pm 2\varepsilon_{mu_x}\varepsilon_{mc})] \\ &= \sigma_{x_0}^2 + \mu_{x_0}^2 + \sigma_{mu_x}^2\left(\sigma_{x_0}^2 + \mu_{x_0}^2\right) + \sigma_{mc}^2\left(\sigma_{x_0}^2 + \mu_{x_0}^2\right)\end{aligned} \tag{52}$$

where we used again properties p1 to p3 in the main text to calculate the expectation of the cross terms. For $\mathrm{E}[xy]$ it holds

$$\mathrm{E}[xy] = \mathrm{E}[x_0y_0 + x_0y_0(\varepsilon_{mc} \pm \varepsilon_{mc} \pm \varepsilon_{mc}^2 \pm \varepsilon_{mc}\varepsilon_{mu_y} + \varepsilon_{mc}\varepsilon_{mu_x} + \varepsilon_{mu_x} + \varepsilon_{mu_y} + \varepsilon_{mu_x}\varepsilon_{mu_y})] \tag{53}$$

Because of the derived properties (independence of x_0, y_0 and the error terms) the expectations of all cross terms is null except

$$\begin{aligned}\pm\mathrm{E}[x_0y_0\varepsilon_{mc}^2] &= \pm\mathrm{E}[x_0y_0]\,\mathrm{E}[\varepsilon_{mc}^2] \\ &= \pm\sigma_{mc}^2\left(\sigma_{x_0y_0}^2 + \mu_{x_0}\mu_{y_0}\right)\end{aligned} \tag{54}$$

where $\mathrm{E}[x_0y_0]$ is given by Equation (29). Plugging (51), (52) and (53) in (14), the expected correlation coefficient is

$$\mathrm{E}[r] = \rho = \frac{\sigma_{x_0y_0} \pm \sigma_{mc}^2(\sigma_{x_0y_0} + \mu_{x_0}\mu_{y_0})}{\sqrt{\sigma_{x_0}^2 + \left(\sigma_{x_0}^2 + \mu_{x_0}^2\right)\left(\sigma_{mu_x}^2 + \sigma_{mc}^2\right)}\sqrt{\sigma_{y_0}^2 + \left(\sigma_{y_0}^2 + \mu_{y_0}^2\right)\left(\sigma_{mu_y}^2 + \sigma_{mc}^2\right)}} \tag{55}$$

that can re-written as (38) by setting $\gamma_x = \sigma_{ac}^2/\sigma_{x_0}^2$ and $\gamma_y = \sigma_{ac}^2/\sigma_{y_0}^2$ $\delta_x = \mu_{x_0}/\sigma_{x_0}$, $\delta_y = \mu_{y_0}/\sigma_{y0}$ and defining the attenuation coefficient A^m (39).

A.2 Derivation of ρ in presence of realistic measurement error

To simplify calculations we set

$$\begin{cases} M_x = x_0(1 + \varepsilon_{mu_x} \pm \varepsilon_{mc}) \\ A_x = \varepsilon_{au_x} \pm \varepsilon_{ac} \end{cases} \tag{56}$$

and similarly we define M_y and A_y for variable y. It holds

$$\mathrm{E}[A_x] = 0 \text{ and } \mathrm{E}[M_x] = \mu_{x_0} \tag{57}$$

and

$$\mathrm{E}[A_x^2] = \sigma_{au_x}^2 + \sigma_{ac}^2 \tag{58}$$

$\mathrm{E}[M_x^2]$ is given by Equation (52). Because error components are independent and with zero expectation (properties p1 to p3) it holds

$$\mathrm{E}[M_xA_x] = \mathrm{E}[M_xA_y] = \mathrm{E}[M_yA_x] = 0 \tag{59}$$

and

$$\mathrm{E}[M_xM_y] = \sigma_{x_0y_0} + \mu_{x_0}\mu_{y_0} \pm (\sigma_{x_0y_0} + \mu_{x_0}\mu_{y_0})\sigma_{mc}^2 \tag{60}$$

$$\mathrm{E}[A_x A_y] = \pm\sigma_{ac}^2 \tag{61}$$

It follows that

$$\mathrm{E}[x] = \mu_{x_0} \tag{62}$$

$$\begin{aligned}\mathrm{E}[x^2] &= \mathrm{E}[M_x^2] + \mathrm{E}[A_x^2] + 2\,\mathrm{E}[M_x A_x] \\ &\sigma_{x_0}^2 + \mu_{x_0}^2 + \sigma_{mu_x}^2\left(\sigma_{x_0}^2 + \mu_{x_0}^2\right) + \sigma_c^2\left(\sigma_{x_0}^2 + \mu_{x_0}^2\right) + \sigma_{au_x}^2 + \sigma_{ac}^2\end{aligned} \tag{63}$$

$$\begin{aligned}\mathrm{E}[xy] &= \mathrm{E}[M_x M_y] + \mathrm{E}[A_x A_y] + \mathrm{E}[M_x A_y] + \mathrm{E}[M_y A_x] \\ &= \sigma_{x_0 y_0} + \mu_{x_0}\mu_{y_0} \pm (\sigma_{x_0 y_0} + \mu_{x_0}\mu_{y_0})\sigma_c^2 \pm \sigma_{ac}^2\end{aligned} \tag{64}$$

Plugging (62), (63), and (64) into (14) one gets the expression for the correlation coefficient in presence of additive and multiplicative measurement error:

$$\begin{aligned}\mathrm{E}[r] = \rho = \frac{\sigma_{x_0 y_0} \pm (\sigma_{x_0 y_0} + \mu_{x_0}\mu_{y_0})\sigma_{mc}^2 \pm \sigma_{ac}^2}{\sqrt{\sigma_{x_0}^2 + \left(\sigma_{x_0}^2 + \mu_{x_0}^2\right)\left(\sigma_{mu_x}^2 + \sigma_{mc}^2\right) + \sigma_{au_x}^2 + \sigma_{ac}^2} \times} \\ \times\sqrt{\sigma_{y_0}^2 + \left(\sigma_{y_0}^2 + \mu_{y_0}^2\right)\left(\sigma_{mu_y}^2 + \sigma_{mc}^2\right) + \sigma_{au_y}^2 + \sigma_{ac}^2}\end{aligned} \tag{65}$$

that can re-written as (40) by using previously defined $\gamma_x, \gamma_y, \delta_x$ and δ_y and defining the attenuation coefficient A^c (41).

A.3 Behavior of ρ in the case of additive negatively correlated error

For negative correlated error, when the true correlation is positive

$$\rho \to \begin{cases} 0 < \rho < \rho_0 & \text{if } \rho_0 < \frac{A}{A-1}\gamma_x\gamma_y \\ \rho_0 & \text{if } \rho_0 = \frac{A}{A-1}\gamma_x\gamma_y \\ > \rho_0 & \text{if } \rho_0 > \frac{A}{A-1}\gamma_x\gamma_y \end{cases} \tag{66}$$

Since $\frac{A}{A-1}\gamma_x\gamma_y < 0$, ρ is always smaller than the true correlation. When the true correlation is negative ($\rho_0 < 0$) the expected correlation is always negative, but it can be, in absolute value, smaller or larger than the absolute value of the true correlation:

$$\rho \to \begin{cases} < \rho_0 & \text{if } \frac{A}{A-1}\gamma_x\gamma_y < \rho_0 < 0 \\ = \rho_0 & \text{if } \rho_0 = \frac{A}{A-1}\gamma_x\gamma_y \\ > \rho_0 & \text{if } \rho_0 > \frac{A}{A-1}\gamma_x\gamma_y \end{cases} \tag{67}$$

Acknowledgements

This work was partially supported by the European Commission funded FP7 project INFECT (Contract No. 305280). The authors acknowledge Peter Wentzell (Halifax, Canada) for kindly making available the NMR data set.

Author contributions statement

E.S. and A.S conceived the study and performed theoretical calculations. E.S., M.H and A.S analysed and interpreted the the results. E.S. and M.H. performed simulations. E.S., M.H and A.S wrote, reviewed and approved the manuscript in its final form.

Competing interests

The authors declare no competing financial interests.