Journal of Chemometrics

WILEY

Journal of CHEMOMETRICS

RESEARCH ARTICLE OPEN ACCESS

A Simulation Study of the Effects of Additive, Multiplicative, Correlated, and Uncorrelated Errors on Principal Component Analysis

Edoardo Saccenti[1] | Marieke E. Timmerman[2] | José Camacho[3]

[1]Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen, The Netherlands | [2]Department of Psychometrics and Statistics, University of Groningen, Groningen, The Netherlands | [3]Department of Signal Theory, Telematics and Communications, University of Granada, Granada, Spain

Correspondence: Edoardo Saccenti (edoardo.saccenti@wur.nl)

Received: 16 February 2024 | **Revised:** 17 July 2024 | **Accepted:** 18 July 2024

Funding: E.S. acknowledges the funding received from the Netherlands Organisation for Health Research and Development (ZonMW) through the PERMIT project (Personalized Medicine in Infections: From Systems Biomedicine and Immunometabolism to Precision Diagnosis and Stratification Permitting Individualized Therapies, project number 456008002) under the PerMed Joint Transnational call JTC 2018 (research projects on personalized medicine–smart combination of preclinical and clinical research with data and ICT solutions) and the funding from the European Union's Horizon 2020 research and innovation program through the DIAGONAL project (GA No. 953152). J.C. was supported by the Agencia Estatal de Investigación in Spain, MCIN/AEI/10.13039/501100011033, Grant PID2020-113462RB-I00.

Keywords: correlation | covariance | experimental noise

ABSTRACT

Measurement errors are ubiquitous in all experimental sciences. Depending on the particular experimental platform used to acquire data, different types of errors are introduced, amounting to an admixture of additive and multiplicative error components that can be uncorrelated or correlated. In this paper, we investigate the effect of different types of experimental error on the recovery of the subspace with principal component analysis (PCA) using numerical simulations. Specifically, we assessed how different error characteristics (variance, correlation, and correlation structure), loading structures, and data distributions influence the accuracy to estimate an error-free (true) subspace from sampled data with PCA. Quality was assessed in terms of the mean squared reconstruction error and the congruence to the error-free loadings, using the pseudorank and adjusting for rotational ambiguity. Analysis of variance reveals that the error variance, error correlation structure, and their interaction with the loading structure are the factors mostly affecting quality of loading estimation from sampled data. We advocate for the need to characterize and assess the nature of measurement error and the need to adapt formulations of PCA that can explicitly take into account error structures in the model fitting.

1 | Introduction

The aim of Principal Component Analysis [1–4] (PCA) is to reduce the dimensionality of a data set $\mathbf{X}$ ($N \times J$), while retaining as much of the variation of the data as possible. The PCA model can also provide information on which variables contribute, to what extent, to the observed data patterns, and it is therefore useful for data exploration [5].

For an $N \times J$ mean centered data set (it is assumed in this paper that $N \geq J$), the PCA model follows the expression

$$\mathbf{X} = \mathbf{F}\mathbf{A}^t, \tag{1}$$

where $\mathbf{F}(N \times J)$ is the score matrix and $\mathbf{A}(J \times J)$ is the loading matrix. The loading matrix $\mathbf{A}$ can be directly obtained from the eigendecomposition of the sample covariance matrix $\mathbf{C}$

$$\mathbf{C} = \frac{1}{N-1}\mathbf{X}^t\mathbf{X}, \tag{2}$$

which equals the sample correlation matrix if the data are scaled to unit variance before analysis.

Correlations among the J observed variables in $\mathbf{X}$ have a fundamental role in the construction and use of PCA for dimensionality reduction. The loadings, which together with the scores form the basis for interpreting the PCA model, are functions of such correlations. A meaningful dimensionality reduction with PCA is possible only if correlated variables are present in the original data.

All experimental data is affected by experimental errors. The errors in multivariate measurements (like those obtained from comprehensive *omics* platforms, such as liquid chromatography–mass spectrometry, gas chromatography–mass spectrometry, nuclear magnetic resonance, and RNA sequencing in transcriptomics) arise from different sources, depending on the type of instruments and the experimental protocols. Experimental errors may originate from sample work-up and preparation, from the use of internal standards, from signal deconvolution for identification and quantification [6, 7], or from specific machine characteristics, like carryover effects in mass spectrometry or thermal errors in detectors [8]. For instance, errors in the amount of internal standard added to a set of samples can affect all measured quantities in the same way, hence leading to positively correlated measurement errors.

Experimental errors play an important role in data analysis because they are a confounding factor in the study of meaningful chemical and biological variation [9]. They are even more relevant for multivariate analysis than for univariate analysis because experimental errors affect both the level of variables and the correlations among variables. Because correlations are the building blocks of virtually all multivariate methods, like PCA, partial least squares regression, and canonical correlation analysis [7], experimental errors affect their results and analysis thereof.

The effect of experimental errors on PCA loadings has been previously investigated, mostly in the context of additive errors [10, 11], which are more amenable of analytical investigation than, for example, multiplicative errors. Here, we present an investigation of the effect of diverse types of measurement errors, namely, additive correlated and uncorrelated errors and multiplicative correlated and uncorrelated errors, on the estimation and recovery of population true loadings (i.e., without error) using PCA with the true pseudorank (i.e., the mathematical rank of the population data in the absence of error [12]) and accounting for rotational ambiguity. The latter implies that we assess the recovery of the PCA subspace, rather than the principal components themselves, as we motivate in Section 3.5. We take into account error characteristics like variance, correlation magnitude and correlation structure, and their interplay with different loading structures and data distributions. We use simulations to explore a large array of different parameter configurations on several quality measures to quantify recovery and error of the sample loadings with respect to the error-free loadings, to which we refer to as the "true" loadings (or population loadings using statistical terminology).

The factors with the largest effects on the quality measures are identified through analysis of variance (ANOVA). Results are interpreted and discussed in the context of bias and distortion of the correlation coefficient according to experimental error characteristics.

The paper is organized as follows. Section 2 presents an overview of the problem of estimating correlation in the presence of experimental errors, starting from the bivariate case. Multivariate additive and multiplicative error models are introduced in Section 3, together with a description of the simulation strategy and analysis implemented to investigate the effect of error characteristics on PCA loading recovery. Results are presented in Section 4 and discussed in Section 5. The paper concludes with an overall discussion in Section 6.

2 | Background Theory

2.1 | Estimating Correlations

Because correlations are estimated pairwise, we present the estimation of correlations in a bivariate setting, building on [7]. The generalization to the multivariate case follows effortlessly by simple generalization to all pairwise combinations of the J variables in $\mathbf{X}$.

We consider two observable variables x_1 and x_2, which are randomly varying in a population and are correlated.

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim \mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0), \tag{3}$$

with population means

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_{x_1} \\ \mu_{x_2} \end{pmatrix}, \tag{4}$$

and population covariance matrix

$$\begin{aligned} \boldsymbol{\Sigma}_0 &= \begin{pmatrix} \sigma^2_{x_1} & \sigma_{x_1x_2} \\ \sigma_{x_1x_2} & \sigma^2_{x_2} \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{x_1} & 0 \\ 0 & \sigma_{x_2} \end{pmatrix} \begin{pmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{pmatrix} \begin{pmatrix} \sigma_{x_1} & 0 \\ 0 & \sigma_{x_2} \end{pmatrix}. \end{aligned} \tag{5}$$

The variance components $\sigma^2_{x_1}$ and $\sigma^2_{x_2}$ quantify the variability of x_1 and x_2, respectively. At the population level, the correlation ρ_{12} between x_1 and x_2 is given by

$$\rho_{12} = \frac{\sigma_{x_1x_2}}{\sqrt{\sigma^2_{x_1}\sigma^2_{x_2}}}, \tag{6}$$

where $\sigma_{x_1x_2}$ represents the covariance of x_1 and x_2.

The correlation coefficient ρ_{12} is estimated from the measured data. The sample correlation r_{12} is calculated as

$$r_{12} = \frac{s_{x_1 x_2}}{\sqrt{s^2_{x_1} s^2_{x_2}}}, \tag{7}$$

where $s^2_{x_1}$ and $s^2_{x_2}$ are the sample variances for variables x_1 and x_2, respectively, and $s_{x_1 x_2}$ is the sample covariance. Assuming that N observations (samples) are taken under random sampling, it holds that

$$\lim_{N \to \infty} r_{12} = \rho_{12}. \tag{8}$$

It is often assumed that $\rho_{12} \to \rho_{0_{12}}$, with $\rho_{0_{12}}$ being the true correlation between the variables without errors, but this is unfortunately not always true if the variables are measured in the presence of experimental errors.

2.2 | Estimating Correlations in the Presence of Experimental Errors

To set the scene, we consider a simple experimental error model including additive uncorrelated errors; under this model, two measured variables x_1 and x_2 are described as [13]

$$\begin{cases} x_1 = x_{0_1} + e_{au_1} \\ x_2 = x_{0_2} + e_{au_2} \end{cases}, \tag{9}$$

where x_{0_1} and x_{0_2} (true signals) have population means μ_1 and μ_2, variances $\sigma^2_{0_1}$ and $\sigma^2_{0_2}$, and correlation $\rho_{0_{12}}$. The error terms e_{au_1} and e_{au_1} are assumed to be independently normally distributed with zero mean and variances $\sigma^2_{au_1}$ and $\sigma^2_{au_2}$, respectively, and independent from the true signals, which is always assumed in what follows.

Under this model, the experimental error causes within-sample variability, which means that M measurement replicates $x_{i,1}, x_{i,2}, \ldots, x_{i,M}$ of observation x_i of variable x will have different values due to the random fluctuation of the error component e_{au}. Uncorrelated errors affect the observed correlation coefficient that is biased downwards (attenuated). The expected correlation coefficient ρ_{12} is given by [7, 14] (see the latter reference for a derivation):

$$\rho_{12} = \frac{\rho_{0_{12}}}{\sqrt{\left(1 + \frac{\sigma^2_{au_1}}{\sigma^2_{0_1}}\right)} \times \sqrt{\left(1 + \frac{\sigma^2_{au_2}}{\sigma^2_{0_2}}\right)}}, \tag{10}$$

with $\rho_{0_{12}}$ being the population correlation between x_{0_1} and x_{0_2} in the absence of error. From this, it follows that the correlation coefficient ρ_{12} is smaller than the true correlation $\rho_{0_{12}}$, and the attenuation is a function of the error variance, as illustrated in Figure 1.

In the case of additive uncorrelated errors, there is a clear monotonic relationship between error components and the correlation coefficients $\rho_{0_{12}}$ and ρ_{12} [7], as shown by Equation (10). If there is also a correlated error component, the correlation coefficient ρ_{12} becomes (see [7] eqs. 42–48 and 73–75 for a derivation)

$$\rho_{12} = \frac{\rho_{0_{12}} + \pi_{ac} \frac{\sigma_{ac_1}}{\sigma_{x_{0_1}}} \frac{\sigma_{ac_2}}{\sigma_{x_{0_2}}}}{\sqrt{1 + \frac{\sigma^2_{au_1}}{\sigma^2_{x_{0_1}}} + \frac{\sigma^2_{ac_1}}{\sigma^2_{x_{0_1}}}} \times \sqrt{1 + \frac{\sigma^2_{au_2}}{\sigma^2_{x_{0_2}}} + \frac{\sigma^2_{ac_2}}{\sigma^2_{x_{0_2}}}}}, \tag{11}$$

where π_{ac} is the correlation of the error components and $\sigma^2_{ac_1}$ and $\sigma^2_{ac_2}$ are the variances of the correlated additive error components acting on the variables x_1 and x_2, respectively.

In the presence of a correlated error component, the correlation ρ_{12} can be inflated or deflated with respect to $\rho_{0_{12}}$, depending on the error characteristics. This is illustrated in Figure 2A,B, where the relationship between the correlation ρ_{12} and the

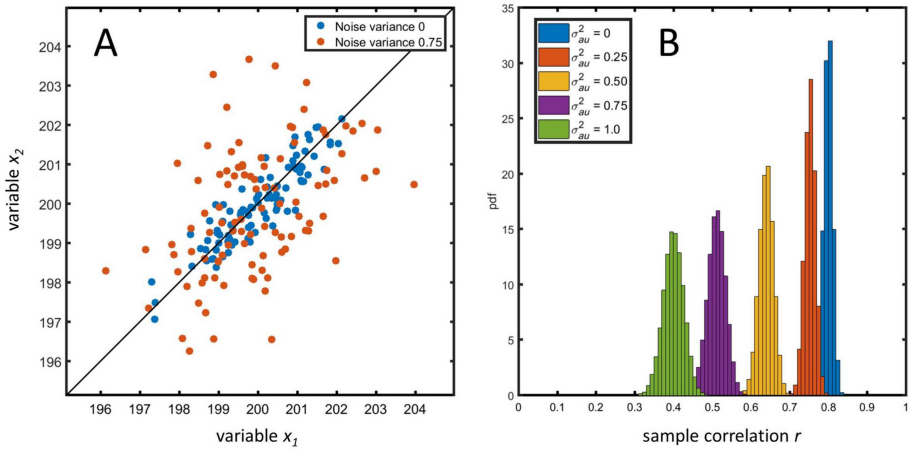

FIGURE 1 | (A) Scatter plot of two variables, x_1 and x_2 ($\sigma^2_{x_1} = \sigma^2_{x_2} = 1$), generated without ($\sigma^2_{au_1} = \sigma^2_{au_2} = 0$) and with uncorrelated additive errors ($\sigma^2_{au_1} = \sigma^2_{au_2} = 0.75$) with true correlation $\rho_{0_{12}} = 0.8$ (model in Equation 9). (B) Distribution of the sample correlation coefficient r_{12} for different levels of measurement errors ($\sigma^2_{au} = \sigma^2_{au_1} = \sigma^2_{au_2}$) for the true correlation $\rho_{0_{12}} = 0.8$. This example is adapted from Figure 1 of [7]; for more details on the simulations, see the Material and Methods section of [7].

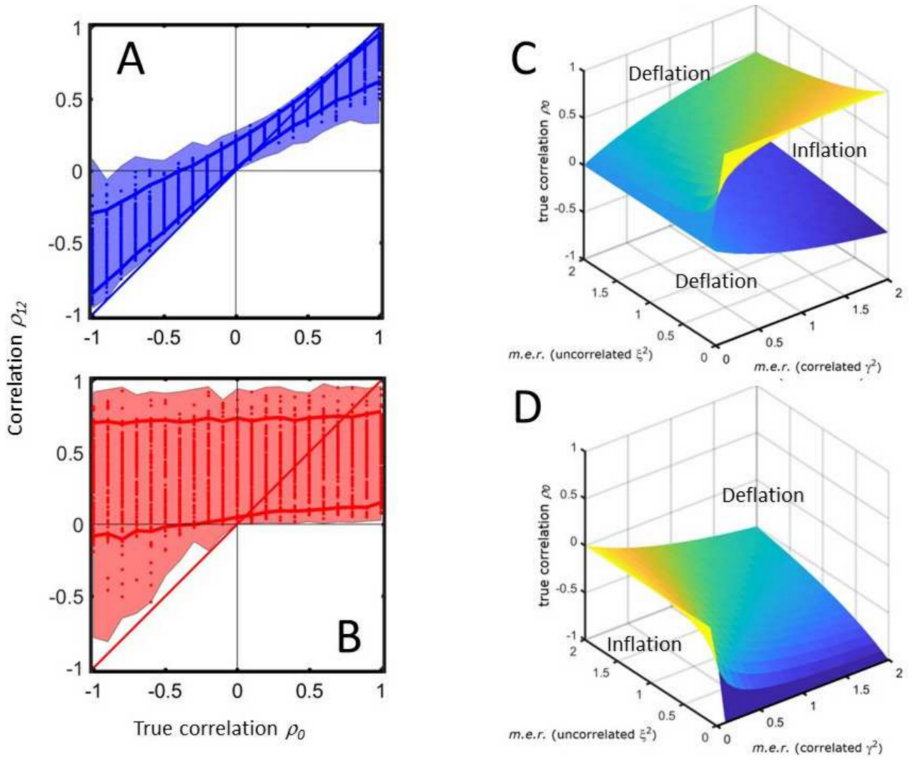

FIGURE 2 | (A, B) Correlation coefficient ρ_{12} between x_1 and x_2 as a function of the different realizations of error parameters for different values of the true correlation $\rho_{0_{12}}$. The shadowed area encloses the maximum and the minimum of the values of ρ calculated in the simulation using the different error parameters (100 of 10^5 Monte Carlo realizations are shown; sample size was 10^4). The dots represent the realized values of ρ_{12}, solid lines the fifth and the 95th percentiles of the observed values. (A) Additive measurement error with positive correlated error (11). (B) Multiplicative measurement error with positive correlated error (13). (C, D) Relationship between the population true correlation $\rho_{0_{12}}$ and the error model parameters expressed as the measurement error ratio (*m.e.r.*) $\xi^2 = \sigma^2_{au}/\sigma^2_x$ (uncorrelated error) and $\gamma^2 = \sigma^2_{ac}/\sigma^2_x$ (correlated error), taking for simplicity $\sigma^2_{x_{0_1}} = \sigma^2_{x_{0_2}} = \sigma^2_x$, $\sigma^2_{au_1} = \sigma^2_{au_2} = \sigma^2_{au}$, and $\sigma^2_{ac_1} = \sigma^2_{ac_2} = \sigma^2_{ac}$. The correlation ρ_{12} will be deflated or inflated depending on the relationship among $\rho_{0_{12}}$, ξ^2, and γ^2. (C) Positively correlated error ($\pi_{ac} = 1$). (D) Negatively correlated error ($\pi_{mc} = 1$). Analytical expression for the surfaces is given in eqs. (35) and (36) of [7]. (A–D) Adapted from [7]: For more details, see eqs. (35) to (38) therein. See the associated Material and Methods (in particular, the Section: Simulations in Figure 6) for the details on the Monte Carlo simulations used to generate the examples.

population true correlation $\rho_{0_{12}}$ is explored using Monte Carlo realizations of the error parameters. This shows how a positive true population correlation $\rho_{0_{12}}$ can result in a negative (sample) correlation ρ_{12}. Relationships have been established in [7] (see eqs. 35, 36, 71, and 72 therein) in the case of perfectly correlated error ($\pi_{ac} = 1$), describing inflation and deflation of the correlation coefficient ρ_{12} between x_1 and x_2 with respect to $\rho_{0_{12}}$, depending on error characteristics (variance and correlation). This is illustrated in Figure 2C,D.

If the error is multiplicative

$$\begin{cases} x_1 = x_{0_1}(1 + e_{mu_1} + e_{mc_1}) \\ x_2 = x_{0_2}(1 + e_{mu_2} + e_{mc_2}) \end{cases} \tag{12}$$

(see Equation 29 for extended definitions) with both correlated and uncorrelated components, the expected correlation coefficient ρ_{12} is (see [7], eqs. 42–48 and 76–78 for a derivation)

$$\rho_{12} = \frac{\rho_{0_{12}}(1 + \pi_{mc}\sigma_{mc_1}\sigma_{mc_2}) + \frac{\mu^2_{x_{0_1}}}{\sigma^2_{x_{0_1}}}\frac{\mu^2_{x_{0_2}}}{\sigma^2_{x_{0_2}}}\pi_{mc}\sigma_{mc_1}\sigma_{mc_2}}{\sqrt{1 + \left(1 + \frac{\mu^2_{x_{0_1}}}{\sigma^2_{x_{0_1}}}\right)\left(\sigma^2_{mu_1} + \sigma^2_{mc_1}\right)} \times \sqrt{1 + \left(1 + \frac{\mu^2_{x_{0_2}}}{\sigma^2_{x_{0_2}}}\right)\left(\sigma^2_{mu_2} + \sigma^2_{mc_2}\right)}}, \tag{13}$$

where π_{mc} is the correlation of the error components and $\sigma^2_{mc_1}$ and $\sigma^2_{mc_2}$ are the variance of the correlated multiplicative error components in the variables x_1 and x_2, respectively.

Generalizing x_1 and x_2 to x_i and x_j, Equations (11) and (13) give the asymptotic expression of the *ij*th element of the sample correlation matrix **C** when the data **X** contain experimental error (for an in-depth discussion, see Sections 5.1.1 and 5.1.2). Because the PCA loadings are a function of the sample correlation matrix elements, it follows that they are also a function of the error characteristics, on which they depend in a complicated, nonlinear fashion.

3 | Material and Methods

3.1 | Measurement Error Models

3.1.1 | Additive Error

Consider the $1 \times J$-dimensional vector $\mathbf{x}_0$ of the error-free variables (true signals) $x_{0_1}, x_{0_2}, \ldots, x_{0_j}, \ldots, x_{0_J}$ collected on a sample i (in the following, the index i is omitted for simplicity,

and we will use the subscript "0" to indicate error-free variable[s]):

$$\mathbf{x}_0 = \left(x_{0_1}, x_{0_2}, \ldots, x_{0_j}, \ldots, x_{0_J}\right). \tag{14}$$

We assume data to be sampled from a J-variate distribution (whose nature will be introduced in Section 3.3.3) with populations means $\boldsymbol{\mu}_0$ $(1 \times J)$,

$$\boldsymbol{\mu}_0 = \left(\mu_{0_1}, \mu_{0_2}, \ldots, \mu_{0_J}\right), \tag{15}$$

and population covariance–correlation matrix $\boldsymbol{\Sigma}_0$,

$$\boldsymbol{\Sigma}_0 = \boldsymbol{\Lambda}_0 \mathbf{R}_0 \boldsymbol{\Lambda}_0, \tag{16}$$

where $\boldsymbol{\Lambda}_0$ is a $J \times J$ diagonal matrix with the population standard deviation σ_{0_j} of variable jth on its diagonal,

$$\boldsymbol{\Lambda}_0 = \begin{pmatrix} \sigma_{0_1} & 0 & \ldots & 0 \\ 0 & \sigma_{0_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \ldots & \sigma_{0_J} \end{pmatrix}, \tag{17}$$

and $\mathbf{R}_0$ is the population correlation matrix

$$\mathbf{R}_0 = \begin{pmatrix} 1 & \rho_{0_{12}} & \ldots & \rho_{0_{1J}} \\ \rho_{0_{12}} & 1 & \ldots & \rho_{0_{2J}} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{0_{1J}} & \ldots & \ldots & 1 \end{pmatrix}. \tag{18}$$

The additive experimental error model (which is a generalization to the multivariate case of the bivariate model presented in Equation 9) consists of both *u*ncorrelated ($\mathbf{e}_{au}$) and *c*orrelated ($\mathbf{e}_{ac}$) errors and is formulated as

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{e}_{au} + \mathbf{e}_{ac}, \tag{19}$$

where $\mathbf{x}$ is the $1 \times J$ vector of measured variables. The $1 \times J$ vectors

$$\mathbf{e}_{au} = \left(e_{au_1}, e_{au_2}, \ldots, e_{au_J}\right), \tag{20}$$

$$\mathbf{e}_{ac} = \left(e_{ac_1}, e_{ac_2}, \ldots, e_{ac_J}\right) \tag{21}$$

are realizations of the additive error components. We will consider only Gaussian distributed errors with zero population mean $\mathbf{0}_J = (0,0, \ldots, 0)_J$:

$$\mathbf{e}_{au} \sim \mathcal{N}(\mathbf{0}_J, \boldsymbol{\Sigma}_{au}), \tag{22}$$

$$\mathbf{e}_{ac} \sim \mathcal{N}(\mathbf{0}_J, \boldsymbol{\Sigma}_{ac}). \tag{23}$$

In (22), $\boldsymbol{\Sigma}_{au}$ is the covariance matrix of the uncorrelated error component:

$$\boldsymbol{\Sigma}_{au} = \begin{pmatrix} \sigma^2_{au_1} & 0 & \ldots & 0 \\ 0 & \sigma^2_{au_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \ldots & \sigma^2_{au_J} \end{pmatrix}, \tag{24}$$

where $\sigma^2_{au_j}$ is the variance of the additive uncorrelated error on the jth variable.

In (23), $\boldsymbol{\Sigma}_{ac}$ is the covariance matrix of the correlated error component, with

$$\boldsymbol{\Sigma}_{\mathrm{ac}} = \boldsymbol{\Gamma}_{\mathrm{ac}} \boldsymbol{\Pi}_{\mathrm{ac}} \boldsymbol{\Gamma}_{\mathrm{ac}} \tag{25}$$

and

$$\boldsymbol{\Gamma}_{ac} = \begin{pmatrix} \sigma_{ac_1} & 0 & \ldots & 0 \\ 0 & \sigma_{ac_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \ldots & \sigma_{ac_J} \end{pmatrix}, \tag{26}$$

and $\boldsymbol{\Pi}_{ac}$ is the correlation matrix of the additive error components

$$\boldsymbol{\Pi}_{ac} = \begin{pmatrix} 1 & \pi_{ac_{12}} & \ldots & \pi_{ac_{1J}} \\ \pi_{ac_{12}} & 1 & \ldots & \pi_{ac_{2J}} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{ac_{1J}} & \ldots & \ldots & 1 \end{pmatrix}. \tag{27}$$

Different shapes of correlation matrices $\boldsymbol{\Pi}_{ac}$ (error correlation structures [ECSs]) are given in Section 3.2.

Note that Equations (24)–(27) give a general case, where error components may differ between variables, and the correlation between errors may differ between pairs of variables. In the simulations that will be realized and analyzed in the remainder of the paper, we will simplify the models by setting variance components to be the uniform across the variables. In particular, we set

$$\left.\begin{array}{r} \sigma^2_{0_j} = 1 \\ \sigma^2_{au_j} = \sigma^2_{au} \\ \sigma^2_{ac_j} = \sigma^2_{ac} \end{array}\right\} \text{for each variable } j = 1,2, \ldots, J; \tag{28}$$

that is, we consider the case of homogeneous additive errors. For simplicity, we will also set $\sigma^2_{au} = \sigma^2_{ac}$.

3.1.2 | Multiplicative Error

The multiplicative error model consists of uncorrelated ($\mathbf{e}_{mu}$) and correlated ($\mathbf{e}_{mc}$) errors:

$$\mathbf{x} = \mathbf{x}_0 \odot \left(1 + \mathbf{e}_{mu} + \mathbf{e}_{mc}\right), \tag{29}$$

where $\mathbf{x}$ is the vector of observed variables, $\odot$ is the Hadamard element-wise matrix product, and the $1 \times J$ vectors $\mathbf{e}_{mu}$ and $\mathbf{e}_{mc}$,

$$\mathbf{e}_{mu} = \left(e_{mu_1}, e_{mu_2}, \ldots, e_{mu_J}\right), \tag{30}$$

$$\mathbf{e}_{mc} = \left(e_{mc_1}, e_{mc_2}, \ldots, e_{mc_J}\right), \tag{31}$$

are the multiplicative error components. As in the additive case, we will consider only Gaussian distributed errors with zero population mean $\mathbf{0}_J = (0,0, \ldots, 0)_J$:

$$\mathbf{e}_{mu} \sim \mathcal{N}(\mathbf{0}_J, \mathbf{\Sigma}_{mu}), \tag{32}$$

$$\mathbf{e}_{mc} \sim \mathcal{N}(\mathbf{0}_J, \mathbf{\Sigma}_{mc}). \tag{33}$$

In (32), $\mathbf{\Sigma}_{\text{mu}}$ is the covariance matrix of the uncorrelated error component:

$$\mathbf{\Sigma}_{mu} = \begin{pmatrix} \sigma^2_{mu_1} & 0 & \ldots & 0 \\ 0 & \sigma^2_{mu_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \ldots & \sigma^2_{mu_J} \end{pmatrix}, \tag{34}$$

where $\sigma^2_{mu_j}$ is the variance of the additive uncorrelated error on the jth variable.

In (34), $\mathbf{\Sigma}_{mc}$ is the covariance matrix of the correlated error component

$$\mathbf{\Sigma}_{\text{mc}} = \mathbf{\Gamma}_{\text{mc}} \mathbf{\Pi}_{\text{mc}} \mathbf{\Gamma}_{\text{mc}}, \tag{35}$$

with

$$\mathbf{\Gamma}_{\text{mc}} = \begin{pmatrix} \sigma_{\text{mc}_1} & 0 & \ldots & 0 \\ 0 & \sigma_{\text{mc}_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \ldots & \sigma_{\text{mc}_J} \end{pmatrix}, \tag{36}$$

and $\mathbf{\Pi}_{mc}$ is the additive error correlation matrix

$$\mathbf{\Pi}_{\text{mc}} = \begin{pmatrix} 1 & \pi_{\text{mc}_{12}} & \ldots & \pi_{\text{mc}_{1J}} \\ \pi_{\text{mc}_{12}} & 1 & \ldots & \pi_{\text{mc}_{2J}} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{\text{mc}_{1J}} & \ldots & \ldots & 1 \end{pmatrix}. \tag{37}$$

As in the case of additive error, we set, for the sake of simplicity,

$$\left.\begin{array}{c} \sigma^2_{0_j} = 1 \\ \sigma^2_{mu_j} = \sigma^2_{mu} \\ \sigma^2_{mc_j} = \sigma^2_{mc} \end{array}\right\} \text{for each variable } j = 1,2, \tag{38}$$

that is, we consider the case of homogeneous multiplicative errors. For simplicity, we will also set $\sigma^2_{mu} = \sigma^2_{mc}$.

3.2 | Error Correlation Structures in the Simulations

In our simulation for experimental errors, we used four different types of error correlation structures (ECS). The additive and multiplicative error correlation matrices, $\mathbf{\Pi}_{ac}$ (27) and $\mathbf{\Pi}_{mc}$ (37), respectively, will take the form of one of the following correlation types.

3.2.1 | Average Correlation Structure

The average correlation matrix $\mathbf{\Pi}_{\text{average}}$ has elements π_{ij} satisfying the condition

$$\frac{2}{J^2 - J} \sum_{i>j} \pi_{ij} = \pi. \tag{39}$$

The average correlation in $\mathbf{\Pi}$ is π (with the average taken over all possible correlation pairs, excluding the diagonal), where all variables may have a different degree of correlation. We generated five random correlation matrices, satisfying the property (39) with $\pi = 0.2, 0.4, 0.6, 0.8$, and 1, using the algorithmic procedure proposed by [15], which is based on the linear transformation of normal random variables satisfying a given set of constraints that are determined algorithmically. We refer the reader to the original publication for more details.

3.2.2 | Hub Correlation Structure

The hub correlation structure $\mathbf{\Pi}_{Hub}$ describes a general correlation structure using k groups of variables: The observed variables (which are also termed, a bit confusingly, observations, hence the name hub-observation correlation model) within each group are correlated with a single observation (the so-called *hub*) in the group with decreasing strength [16]. The k groups are independent, implying zero correlation among variables belonging to different groups. We set the first variable in each group to be the hub variable, and the correlation $\pi_{1,i}$ (with $i > 1$) between variable $i = 1,2, \ldots, g$ and the hub variable is defined as

$$\pi_{1,i} = \pi - \left(\frac{i-2}{g-2}\right)^{\gamma} (\pi - \pi_{min}). \tag{40}$$

We simulated a random hub correlation structure with two groups of equal size $J/2$ using a quadratic attenuation ($\gamma = 2$). The π_{min} is the minimum correlation in the hub, and it is determined algorithmically; given π, correlation matrices were generated using the generative algorithm by Hardin et al. [16].

3.2.3 | Toeplitz Correlation Structure

A Toeplitz correlation structure (also called autoregressive model) [17] has the matricial form $\mathbf{\Pi}_{Toeplitz}$:

$$\mathbf{\Pi}_{\text{Toeplitz}} = \begin{pmatrix} 1 & \pi & \pi^2 & \pi^3 & \ldots & \pi^{J-1} \\ \pi & 1 & \pi & \pi^2 & \ldots & \pi^{J-2} \\ \pi^2 & \pi & 1 & \pi & \ldots & \pi^{J-3} \\ \pi^3 & \pi^2 & \pi & 1 & \ldots & \pi^{J-4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \pi^{J-1} & \pi^{J-2} & \pi^{J-3} & \pi^{J-4} & \ldots & 1 \end{pmatrix}. \tag{41}$$

Under this correlation structure, the error realizations e_{ac_j} (additive error, see 21) and e_{mc_j} (multiplicative error, see 31) corresponding to adjacent variables are highly correlated at the population level, while those further away are less correlated. The magnitude of the correlation decays exponentially with respect to $|i-j|$.

3.2.4 | Uniform Error Correlation

The uniform correlation matrix $\mathbf{\Pi}_{uniform}$ is defined as

$$\mathbf{\Pi}_{\text{uniform}} = \begin{pmatrix} 1 & \pi & \dots & \pi \\ \pi & 1 & \dots & \pi \\ \vdots & \vdots & \ddots & \vdots \\ \pi & \dots & \dots & 1 \end{pmatrix}. \tag{42}$$

Under this correlation model, the error realizations e_{ac_j} (additive error, see 21) and e_{mc_j} (multiplicative error, see 31) are all identically pairwisely correlated with population correlation π; that is, $\mathbb{E}[\text{corr}(e_{ac_j}, e_{ac_i})] = \pi$ (same for e_{mc_j}).

In the absence of established ECSs for omics data, these correlation structures were chosen to be representative of different scenarios, ranging from the well-established Toeplitz model, with clear relationships between adjacent variables, to the limiting case of uniform correlations. They can be considered to be extensions of the model proposed in Figure 1 of [18], which also provides a multivariate framework for the modeling of ECSs. A graphical depiction of the four types of correlation structures is shown in Figure 3.

Given that actual error realizations e_{ac_i} and e_{mc_i} are drawn at random using models (33) and (21), it holds that $\mathbb{E}[\text{corr}(e_{ac_i}, e_{ac_j})] = \pi_{ij}$ and $\mathbb{E}[\text{corr}(e_{mc_j}, e_{mc_i})] = \pi_{ij}$, where π_{ij} is the i,j element of any of the four error population correlation matrices.

A description of the modeling of the ECS through experimental characterization of several sources of error to the total variance NMR measurements is given in [8], but these results are not directly applicable to our simulation scheme.

3.3 | Data Simulation

The error-free data $\mathbf{X}_0^{sim}$ was simulated according to the PCA model:

$$\mathbf{X}_0^{sim} = \mathbf{F}_0^{sim}\mathbf{A}_0^{sim^t}, \tag{43}$$

where $\mathbf{F}_0^{sim}$ is an $N \times J$ matrix of simulated components (scores) and $\mathbf{A}_0^{sim}$ is the $J \times J$ loading matrix, where the first $K = 3$ nonzero loadings are the structural loadings to be defined in Equations (50)–(54). In the next sections, we will explain how the loadings and scores were constructed. When N independent observations of $\mathbf{x}$ are collected, they can be combined in an $N \times J$ data matrix $\mathbf{X}$ so that for additive error, it holds that

$$\mathbf{X} = \mathbf{X}_0^{sim} + \mathbf{E}_{au} + \mathbf{E}_{ac}, \tag{44}$$

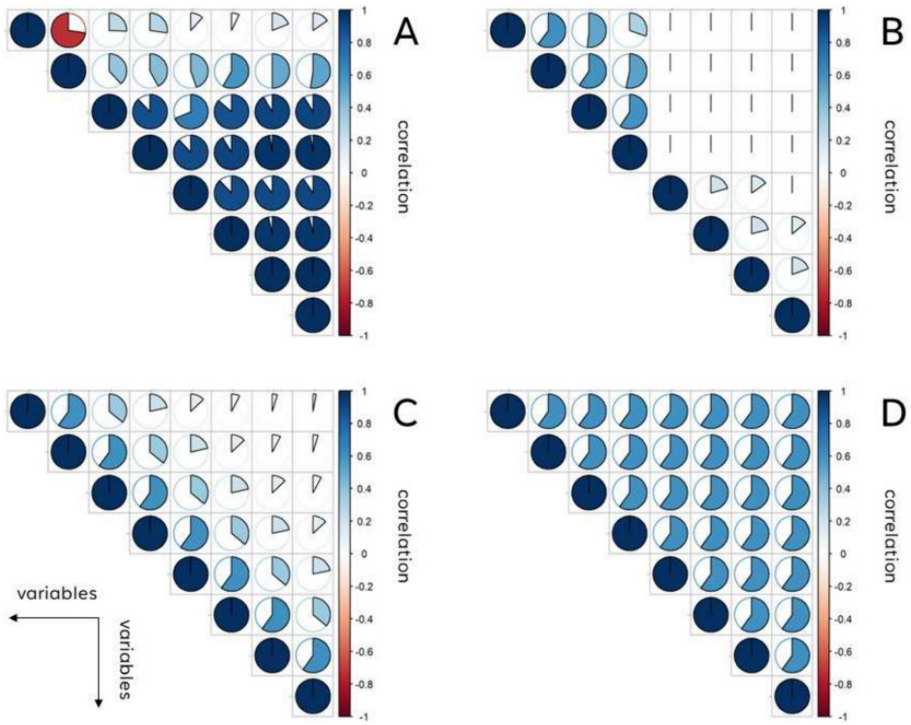

FIGURE 3 | Graphical illustration of the four error correlation structures described in Section 3.2. (A) Average correlation, Equation (39). (B) Hub correlation, Equation (40). (C) Toeplitz correlation, Equation (41). (D) Uniform correlation, Equation (42). Correlation matrices are shown for the case $\pi = 0.6$.

where $\mathbf{E}_{au}$ and $\mathbf{E}_{ac}$ are $N \times J$ data matrices collected from N realizations of the uncorrelated and correlated errors $\mathbf{e}_{au}$ and $\mathbf{e}_{ac}$, respectively. For multiplicative error, it holds that

$$\mathbf{X} = \mathbf{X}_0^{sim} \odot (\mathbf{1} + \mathbf{E}_{mu} + \mathbf{E}_{mc}), \tag{45}$$

where $\mathbf{E}_{mu}$ and $\mathbf{E}_{mc}$ are $N \times J$ data matrices collected from N realizations of the uncorrelated and correlated errors $\mathbf{e}_{mu}$ and $\mathbf{e}_{mc}$ and $\mathbf{1}$ is an $N \times J$ matrix of 1s ($\mathbf{1}_{ij} = 1$).

Note that Equations (44) and (45) give the most general case where uncorrelated and correlated error components coexist. In the simulation and analysis that will follow, we will treat the correlated and uncorrelated cases independently. In particular, we will consider the four models:

1. Additive uncorrelated error:

$$\mathbf{X} = \mathbf{X}_0^{sim} + \mathbf{E}_{au}. \tag{46}$$

2. Multiplicative uncorrelated error:

$$\mathbf{X} = \mathbf{X}_0^{sim} \odot (\mathbf{1} + \mathbf{E}_{mu}). \tag{47}$$

3. Additive correlated error:

$$\mathbf{X} = \mathbf{X}_0^{sim} + \mathbf{E}_{ac}. \tag{48}$$

4. Multiplicative correlated error:

$$\mathbf{X} = \mathbf{X}_0^{sim} \odot (\mathbf{1} + \mathbf{E}_{mc}). \tag{49}$$

3.3.1 | Loading Structures

We considered four population loading matrix structures $J \times J$, where the first K loadings are structural (i.e., describe data structure) and the last $J - K$ are set to 0.

Three population loading structures are inspired by the simulation scheme devised in [19], where it was observed that the quality of estimated loadings depends on the degree of simplicity of the loading structure and on the rotation criterion applied. In our study, we expect these three loading structures to have limited effect on the recovery because we assess the recovery of the subspace and thus account for rotational ambiguity. The simple and complex structures are defined as (only the first $K = 3$ loadings shown, with the remaining being equal to 0)

$$\mathbf{A}^S_{simple}(c) = \begin{pmatrix} c & 0 & 0 \\ c & 0 & 0 \\ c & 0 & 0 \\ c & 0 & 0 \\ 0 & c & 0 \\ 0 & c & 0 \\ 0 & c & 0 \\ 0 & 0 & c \end{pmatrix}, \tag{50}$$

$$\mathbf{A}^S_{complex}(c) = \begin{pmatrix} c & c & c \\ -c & c & c \\ c & -c & c \\ c & c & -c \\ -c & -c & c \\ -c & c & -c \\ c & -c & -c \\ -c & -c & -c \end{pmatrix}. \tag{51}$$

For the simple case, we set $c = 1$, while for the complex case, we set $c = \sqrt{1/K}$, to keep the proportion of structural variance per variable the same across conditions (as the error variance was manipulated in the same way across conditions).

The simple loading structure $\mathbf{A}^S_{simple}$ is sparse, and it is well aligned with the varimax criterion [20]. It is stable after varimax rotation over different samples [19]. We have slightly modified this structure to have four, three, and one nonzero loadings in components in the first, second, and third components, rather than three, three, and two, respectively, as in [19].

The complex loading structure $\mathbf{A}^S_{complex}$ has a circumplex structure, implying that it has a nonunique axis position in terms of the varimax criterion. Therefore, its varimax rotated loadings are highly unstable over samples [19].

The intermediate loading structure $\mathbf{A}^S_{inter}$ is defined as

$$\mathbf{A}^S_{inter} = \mathbf{D}^{-1}\mathbf{A}, \tag{52}$$

where $\mathbf{A}$ is a weighted sum of the two extreme cases:

$$\mathbf{A} = \mathbf{A}^S_{simple}(c_1) + \mathbf{A}^S_{complex}(c_2), \tag{53}$$

where $c_1 = \sqrt{0.4}$ and $c_2 = \sqrt{(1 - 0.4) \times 1/q)}$ and $\mathbf{D}$ is a $J \times J$ diagonal matrix with the diagonal elements being equal to the square root of the diagonal elements of the matrix $\mathbf{A}\mathbf{A}^t$.

The fourth loading structure $\mathbf{A}^S_{real}$ that we termed "experimental" is taken equal to the loadings of the first three principal components of a 60×8 experimental metabolomic data set (see Section 3.8 for more details).

$$\mathbf{A}^S_{real} = \begin{pmatrix} 0.04 & 0.01 & -1.00 \\ -0.13 & 0.95 & -0.05 \\ -0.77 & 0.41 & 0.08 \\ -0.91 & -0.35 & -0.05 \\ -0.96 & 0.17 & 0.00 \\ -0.89 & -0.30 & -0.04 \\ 0.87 & -0.48 & 0.02 \\ -0.50 & -0.86 & 0.00 \end{pmatrix}. \tag{54}$$

3.3.2 | Construction of the Scores

The elements $f_{0_{jl}}$ of the score matrix $\mathbf{F}_0^{sim}$ were generated following different data distributions, which will be described in the next section. We take the element of $\mathbf{F}_0^{sim}$ to be uncorrelated.

3.3.3 | Distributions of Error-Free Scores

We considered two different data distributions to generate the uncorrelated elements $f_{0_{jl}}$ of the $N \times J$ score matrix $\mathbf{F}_0^{sim}$.

Normal distribution:

$$f_{0_{jl}} \sim \mathcal{N}(0,1). \tag{55}$$

Log-normal distribution:

$$f_{0_{jl}} = e^{y_{0_{jl}}}, \tag{56}$$

with

$$y_{0_{jl}} \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right), \tag{57}$$

where $\mu_0 = -0.3466$ and $\sigma_0 = 0.8326$ so that the generated scores have expected mean and variance equal to 1.

These score distributions were chosen to represent the statistical idealized case (Gaussian distribution) and more realistic situations like distribution of biological entities (log-normal distribution); they are shown in Figure 4.

3.3.4 | Correlation Structure of the Error-Free Data

Equation (43) and the population loading structures $\mathbf{A}_0^{sim}$ (see Equations 50–52 and 54) implicitly define the population covariance–correlation structure (18) underlying the error-free data generation. If the data are mean centered, it holds that

$$\begin{aligned}\boldsymbol{\Sigma}_0 &= \mathbb{E}\left[\frac{1}{N-1}\left(\mathbf{X}_0^{\text{sim}}\mathrm{O}_n\right)^{\text{t}}\mathbf{O}_n\mathbf{X}_0^{\text{sim}}\right] \\ &= \mathbf{A}_0^{\text{sim}}\mathbb{E}\left[\frac{1}{N-1}(\mathbf{F}_0^{\text{sim}})^{\text{t}}\mathbf{O}_n\mathbf{F}_0^{\text{sim}}\right]\left(\mathbf{A}_0^{\text{sim}}\right)^{\text{t}} \\ &= \mathbf{A}_0^{\text{sim}}\mathbb{E}[\mathbf{I}]\left(\mathbf{A}_0^{\text{sim}}\right)^{\text{t}} \\ &= \mathbf{A}_0^{\text{sim}}\left(\mathbf{A}_0^{\text{sim}}\right)^{\text{t}},\end{aligned} \tag{58}$$

with $\mathbf{O}_n = \left(\mathbf{I}_n - \frac{1}{N}\mathbf{J}_n\right)$ being the centering matrix and $\mathbf{J}_n$ an n-by-n matrix of all 1s; the third step follows from the score $\mathbf{F}_0^{sim}$ being uncorrelated after mean centering and having variances of 1 at the population level.

3.4 | Experimental Design for the Simulated Data

In the analysis of the effect of uncorrelated error (for both additive and multiplicative components), the factors that were manipulated are the loading matrix type, the data distribution, and the error variance. In the case of correlated error (for both additive and multiplicative error components), additional factors are the error correlation π and the ECS. The factor levels are given in Table 1.

The simulation design was full factorial. For the uncorrelated error, there are 4(loading matrix types) × 2(data distribution) × 6(error variance) = 48 conditions. For the correlated error, there are 4(loading matrix types) × 2(data distributions) × 7(error variances) × 5(error correlations) × 4(ECSs) = 1120 conditions.

The same $\mathbf{X}_0$ and error realizations were used to generate simulated data matrices $\mathbf{X}$ with additive and multiplicative errors, namely, $\mathbf{E}_{au} = \mathbf{E}_{mu}$ in models (46) and (47) and $\mathbf{E}_{ac} = \mathbf{E}_{mc}$ in models (48) and (49).

In all cases, $N = 2^{16}$ observations were used. Such a high number of observations was chosen to reduce finite sample size

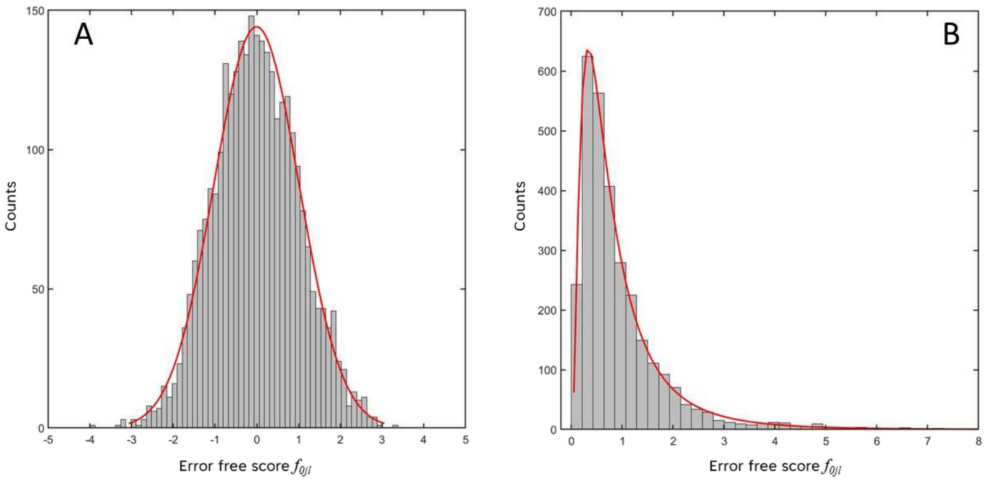

FIGURE 4 | Data distributions used to generate the uncorrelated elements $f_{0_{jl}}$ of the $N \times J$ score matrix $\mathbf{F}_0^{sim}$. (A) Normal distribution, Equation (55). (B) Log-normal distribution, Equations (56)–(57).

TABLE 1 | Overview of experimental factors and associated levels that were varied in the simulations.

Design factor	Levels	Number of levels
Loading type	simple, intermediate, complex, experimental	4
Distribution	Normal, Log-normal	2
Error variance	0.5, 1, 2, 4, 8, 16	6
Error correlation	0.2, 0.4, 0.6, 0.8, 1	5
Error correlation structure	average, hub, Toeplitz, uniform	4

Note: Loading type indicates the different population structural loading matrices $\mathbf{A}^{S}_{simple}$ simple (50), complex (51), intermediate (52), and experimental (54). Distribution refers to the distribution of the error-free scores $\mathbf{F}_0^{sim}$ used to generate error-free data $\mathbf{X}_0$ (see Section 3.3.3). Error variance refers to $\sigma^2_{error} = \sigma^2_{au} = \sigma^2_{ac} = \sigma^2_{mu} = \sigma^2_{mc}$. Error correlation structures Π are Toeplitz (41), hub (40), and average (39); the error correlation is π in the error correlation structures Π.

effects on loading estimations. For each condition, 250 different realizations of $\mathbf{X}_0^{sim}$ and error matrices $\mathbf{E}$ were (randomly) generated.

3.5 | Rotation of the PCA Solutions

When considering the recovery of a PCA solution, we need to distinguish whether we are interested in the principal components and associated loadings themselves or in the principal subspaces they define. These characteristics of PCA are well known and have been discussed and addressed elsewhere: See, for instance, [19, 21, 22].

For the principal components, there is a sign indeterminacy, meaning that per principal component, the loadings and its accompanying scores can be multiplied by −1 without changing the principal components. Further, if any components have exactly the same variance, then the order of these components (or of any linear combination of them) is arbitrary. The closer the variance of the components, the more sensitive to noise the model is in terms of rotation or reordering. However, if we consider the whole low-dimensional space rather than each component on an individual basis, this problem can be avoided.

The choice of working with the subspace is based on the fact that subspaces are meaningful, regardless of whether PCA is used to find an abstract representation of the data (i.e., for data compression or exploration as done in chemometrics and data science [23]) or to identify interpretable dimensions (as done in areas like psychometrics and econometrics [24]).

If the interest is in the projection of the data onto a low-dimensional space, then there is an indeterminacy in the orientation of the axes. This implies that one can rotate the loadings and counter-rotate the principal component scores without changing the projected data. Herewith, one could retain the orthogonality of the axes, if one would use an orthogonal rotation, or also allow for oblique axes, if one would use an oblique rotation.

In the current case, we are interested in the recovery of the projection of the data in low-dimensional space rather than principal components themselves. For this reason, it is necessary to adjust for the rotational ambiguities. This can be achieved by rotating the loadings of one model towards the loadings of the other model or towards a common target set of loadings. A common approach is to apply a so-called Procrustes rotation, which rotates a given matrix to attain maximum similarity with a target matrix by minimizing a given criterion [25].

When we set $\mathbf{B}$ to be a matrix of rotated loadings and $\mathbf{T}$ as the target matrix, the criterion h to be minimized is

$$h(\mathbf{B}) = \frac{1}{2}\mathrm{tr}\left((\mathbf{B}-\mathbf{T})^{\mathrm{T}}(\mathbf{B}-\mathbf{T})\right), \tag{59}$$

where tr(*) is the trace operator. Here, we applied the oblique Procrustes rotation: Each sample loading matrix calculated from the simulated data $\mathbf{X}$ under the different error models (Equations 4 and 45) was rotated towards the known error-free loadings (which in the present case is one of the structural loading matrices defined in Equations 50–52 and 54) used to generate the error-free data (Equation 43), before calculation of the quality metrics defined in Section 3.6. The choice of an oblique rotation (instead of an orthogonal) is motivated by the fact that oblique rotation maximizes the recovery, in this way taking into account that sample components can be oblique, also when the error-free components are orthogonal.

3.6 | Quality Criteria for Loading Recovery

The agreement between the known structural population loadings $\mathbf{A}^S$ and the sample loadings $\mathbf{A}^{samp}$, as estimated with PCA retaining K components, from the generated data $\mathbf{X}$, was evaluated using two quality criteria, indicative of the estimation error of and the proportionality to the population loadings. That is, for all conditions, we used the mean squared error (MSE) and Tucker's ϕ congruence index. For the multiplicative noise conditions, we also considered the fold change ratio *FC* of the MSE.

3.6.1 | MSE of the Reconstruction

The MSE of the estimated loadings of component k is defined as

$$\mathrm{MSE}_k = \frac{1}{J}\sum_j \left(a^{samp}_{jk} - a^{S}_{jk}\right)^2, \tag{60}$$

where a_{jk}^{samp} and a_{jk}^{S} are the loadings for variable j in component k in the sample loading matrix $\mathbf{A}^{samp}$ and error-free loading matrix $\mathbf{A}^{S}$, respectively. The MSE is a measure for the recovery in an absolute sense. For each loading matrix, we calculated the MSE as the mean of MSE_k over the three components:

$$\text{MSE} = \frac{1}{3}(\text{MSE}_1 + \text{MSE}_2 + \text{MSE}_3). \tag{61}$$

3.6.2 | Tucker's ϕ

The congruence of the sample loadings to population structural loadings was quantified per component k as using Tucker's ϕ congruence index [26, 27]:

$$\phi_k = \frac{\sum_j a_{jk}^{samp} a_{jk}^{S}}{\sqrt{\left(\sum_j \left(a_{jk}^{samp}\right)^2\right)\left(\sum_j \left(a_{jk}^{S}\right)^2\right)}}. \tag{62}$$

For each loading matrix, we calculated the ϕ as the mean ϕ_k across the three components:

$$\phi = \frac{1}{3}(\phi_1 + \phi_2 + \phi_3). \tag{63}$$

Tucker's ϕ_k takes values from -1 to 1, with 1 (and -1) indicating perfect proportionality of columns (though opposite in sign). Perfect proportionality of loadings implies an identical interpretation of the loading structure. ϕ values larger than 0.95 indicate practically equivalent loading vectors, and fair equivalence is established for $0.85 < \phi_k < 0.95$ [28].

3.6.3 | Fold Change Ratio of MSE

We defined the fold change ratio FC of the mean squared reconstruction error MSE as

$$FC = \log_2\left(\frac{\text{MSE}_{additive}}{\text{MSE}_{multiplicative}}\right). \tag{64}$$

The FC is taken over MSE values (61) computed on data generated under the same conditions, namely, the same error-free data $\mathbf{X}_0$ on which an error of a different nature (additive and multiplicative) with the same characteristics (variance, correlation structure, and correlation) is added (see Equations 44 and 45).

3.7 | Simulation Analysis

N-way ANOVA was used to explore the simulation results and to identify which data and error characteristics have a large(r) effect on the quality of loading recovery, as expressed by MSE (61) and Tucker's congruence coefficient ϕ (63), which were used as the response in the ANOVA models. Because in most analysis residuals deviated from the normality assumption, we used the rank transformation on the responses [29]. This transformation was recently observed to improve the statistical power of ANOVA [30] using simulated power curves [31].

In the case of uncorrelated error, the ANOVA factors were the loading type, the data distribution, the error variance, and all pairwise interactions. In the case of correlated error, the factors were the loading type, the data distribution, the error variance, the ECS, the error correlation magnitude, and all pairwise interactions. The levels of each factor are given in Table 1. The results for the fold change FC (64) were visually investigated. Note that because the two distributions used to generate error-free scores have different population means (0 for the normal and 1 for the log-normal), the two distributions factor levels can also be read, implicitly, as factor "population means" with two levels, 0 and 1.

To quantify the amount of variance of the response explained by the different factors and combinations thereof, we used the partial eta squared η^2_{part} as a measure of the effect size. Partial eta squared can be used, at the ordinal level, to compare the effects of different factors in the same design [32].

3.8 | Experimental Data

The experimental loading structure given in Equation (54) was obtained by considering the first three principal components of a (standardized to unit variance) data set containing abundances of 465 metabolites measured with mass spectrometry in the liver biopsy of 315 mice fed two different diets. Of the 465 variables, eight were selected (variables 4, 9, 49, 99, 149, 199, 249, and 299, corresponding to 14-hydroxy-E4-neuroprostane, 24,25-dihydrolanosterol, 5-L-glutamyl-L-alanine, anserine, coutaric acid, glucosylceramide [d18:1/16:0], 1-stearoyl-2-hydroxy-sn-glycero-3-phosphate, and Ne,Ne dimethyllysine). For more details about the experimental designs, experimental protocols, and technical details, we refer to the original publication [33].

3.9 | Software

Calculations were performed using Matlab R2023a 9.14 (Natick, Massachusetts, The MathWorks Inc.) for Windows and using R [34] version 4.3.3 and RStudio version 2023.06.1. Multivariate data were generated with the in-built Matlab function mvnrnd; singular-value decomposition was performed using the Matlab function mvnrnd; N-way ANOVA was performed using the anovan function. All other operations were performed using in-house scripting routines. Hub correlation matrices (40) were generated using the R function `simcor.H` provided by Hardin et al. [16]. The parameters used were $k = 2$, $\epsilon = 0.01$, $\gamma = 2$, size $= (5,2)$, and edim $= 2$. R code to generate random correlation matrices with given average correlation was obtained from the authors of the original publication [15]. Correlation plots were made using the corrplot function from the R corrplot package [35]. Code for analysis is freely available at github.com/esaccenti/PCAnoiseLoadings.

4 | Results

In analyzing our simulation results, we focus on identifying which factors and interactions, and levels in them, affect the quality of the loading estimates the most, as indicated by our quality measures. Therefore, we do not focus on the actual values, that is, we do not attempt to establish numerical relationships between the error model parameters and the values of the quality measures. We also do not consider the statistical significance of the factors explicitly. Because the sample size is very large (we used 250 replicates per simulation condition) and each manipulated factor is expected to have an effect, every statistical test (ANOVA and post hoc) is expected to show significance.

4.1 | Effect on Experimental Error on the Loading Reconstruction Error (MSE)

In this section, we study the loading MSE in the presence of different types of error.

4.1.1 | MSE for Additive Uncorrelated Error

Results of the ANOVA of the MSE (61) in the case of uncorrelated error (model 46) are given in Table 2, together with loading congruence results (discussed later) to facilitate comparison. For the MSE for the additive error, the factor with the largest η^2_{part} is the loading type (i.e., complexity of the loading patterns) and its interaction with the error variance (σ^2_{au} in error models 24) followed by the main effect of the latter. The average MSEs for the levels of these factors and their interactions are depicted in Figure 5. These plots provide a more detailed interpretation of the results and given a similar information, but without statistical inference, as the traditional visualizations associated to post hoc tests.

TABLE 2 | ANOVA results for the reconstruction MSE (61) and Tucker's ϕ congruence (63) simulations for sample loadings estimated from data generated with uncorrelated additive (Equations 19, 20, and 24) and multiplicative (Equations 29, 30, and 34) uncorrelated experimental errors.

	Uncorrelated error			
	Additive		Multiplicative	
	η^2_{part} MSE	η^2_{part} ϕ	η^2_{part} MSE	η^2_{part} ϕ
Loading type	**0.98**	**0.95**	**0.91**	**0.92**
Distribution	0.00	0.00	0.20	**0.81**
Error variance	**0.74**	**0.91**	0.26	**0.90**
Loading type×distribution	0.00	0.01	0.33	0.40
Loading type×error variance	**0.98**	**0.63**	**0.93**	**0.64**
Distribution×error variance	0.00	0.00	0.16	0.11

Note: See caption of Table 1 for an explanation of factor names. The effects with effect size > 0.55 are in bold.

Given the large MSS of the interaction shown in Table 2, we base our interpretation of the results only on Figure 5C and do not attempt to interpret the main factors [29]. We see a differential behavior of the MSE for simple (50) and intermediate (52) loading structures, which increases substantially with the error variance (σ^2_{au}), in comparison to the MSE for complex (51) and experimental (54) loading structures, which decrease first and then increase with increasing error variance. The latter is an unexpected behavior, which interestingly illustrates that a low quality of estimation in the loadings can indeed happen for very low levels of error.

4.1.2 | MSE for Multiplicative Uncorrelated Error

Table 2 shows the η^2_{part} for multiplicative uncorrelated error (model 47). The most important effects are those due to the loading type and its interaction with the error variance σ^2_{mu}. Similar considerations to the case of additive error hold (the figure with average MSE for the levels not shown): For simple (50) and intermediate (52) loading structures, the MSE increases with the error variance, (σ^2_{mu}), while for complex (51) and experimental (54) loading structures, the MSE decreases and then remains stable.

4.1.3 | MSE for Additive Correlated Error

Results for the MSE for the case of correlated error (model 48) are given in Table 3. In the case of additive error, the most relevant effects are, with practically equivalent η^2_{part}, the loading type, the error variance (σ^2_{ac} in error models 26), the ECS, and the interaction between the loading type and the ECS. The average MSE for the levels of the previous factors and the interaction is shown in Figure 6.

The effect of the error variance on the MSE, in Figure 6A, shows that the MSE increases with the increasing error variance. This result, although intuitive, contrasts with what we found for uncorrelated noise. To interpret the effect of the loading type and the ECS in the MSE, we inspect the interaction in Figure 6D, given its high η^2_{part} in the ANOVA. We observe that the interaction shows a different behavior, with high MSE in the case of uniform (ECS = 1) error correlation, similar for all loading types, and lower MSE in the case of Toeplitz (ECS = 2), hub (ECS = 3), and average (ECS = 4) error correlations, with exception for the complex loading type (LT = 3).

The uniform structure has stronger correlation with respect to the other structures (for instance, in the hub structure, only subsets of error components are correlated; in the average structure, both positive and negative correlations can coexist, especially for low average correlation; see Equation 39). This could explain why the effect on MSE is stronger because a stronger error correlation is imposed on top of that defined by the population loadings.

4.1.4 | MSE for Multiplicative Correlated Error

The η^2_{part} values for the MSE for multiplicative correlated error (model 49), Table 3, indicate that the main effects are for the

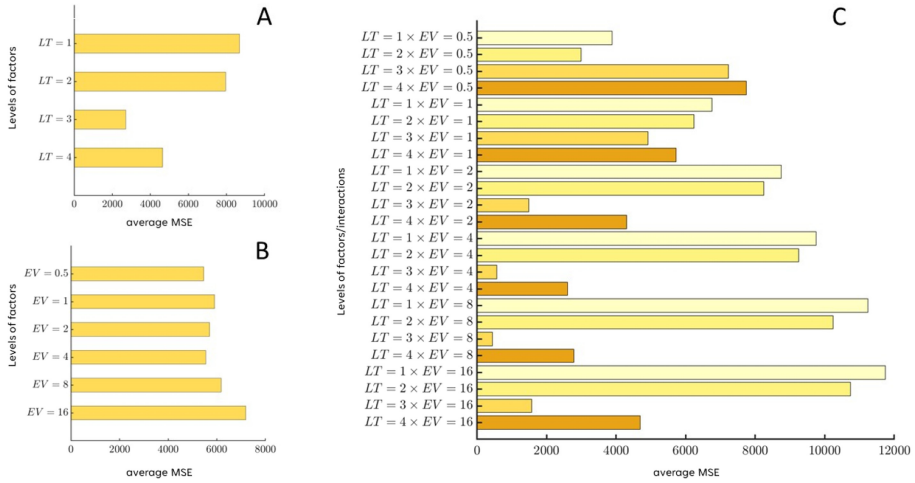

FIGURE 5 | Additive uncorrelated error: averaged (per factor level/interaction) mean square error (MSE, 61) for the loading reconstruction for different factors/interactions of the ANOVA model. (A) Loading type (LT): 1 (simple), 2 (intermediate), 3 (complex), 4 (experimental); see Equations (50)–(52) and (54). (B) Uncorrelated error variance (EV) σ^2_{au}. (C) MSE for the interaction ($LT \times EV$) between EV and LT. For the multiplicative uncorrelated error, similar patterns are observed.

TABLE 3 | ANOVA results for the reconstruction MSE (61) and Tucker's ϕ congruence (63) simulations for sample loadings estimated from data generated with correlated additive (Equations 19, 21, and 26) and multiplicative (Equations 29, 31, and 35) experimental errors.

	Correlated error			
	Additive		**Multiplicative**	
Factor	η^2_{part} **MSE**	η^2_{part} ϕ	η^2_{part} **MSE**	η^2_{part} ϕ
Loading type	**0.68**	**0.71**	0.11	**0.56**
Distribution	0.00	0.00	0.19	**0.59**
Error variance	**0.70**	**0.79**	0.36	**0.72**
Error corr struct	**0.60**	**0.83**	**0.69**	**0.85**
Error corr level	0.22	0.25	0.01	0.03
Loading type×distribution	0.00	0.00	0.07	0.08
Loading type×error variance	0.35	0.29	**0.62**	0.17
Loading type×error corr struct	0.49	0.44	0.21	0.49
Loading type×error corr level	0.14	0.08	0.22	0.02
Distribution×error variance	0.00	0.00	0.01	0.01
Distribution×error corr struct	0.00	0.00	0.03	0.05
Distribution×error corr level	0.00	0.00	0.00	0.00
Error variance×error corr struct	0.19	0.35	0.23	0.37
Error variance×error corr level	0.06	0.06	0.00	0.02
Error corr struct×error corr level	0.04	0.09	0.01	0.07

Note: See caption of Table 1 for an explanation of factor names. The effects with effect size > 0.55 are in bold.

interaction of the error variance σ^2_{mc} and the loading type and for the main effects of the ECS. The levels of these factors are shown in Figure 7A,C. Like in the case of additive correlated error, the MSE grows with the error variance, and we see higher MSE in the case of uniform (ECS = 1) ECSs.

4.1.5 | MSE Ratio Between Additive and Multiplicative Error

A legitimate question is whether, overall, the reconstruction error on loadings (as expressed by the MSE 61) is larger in the

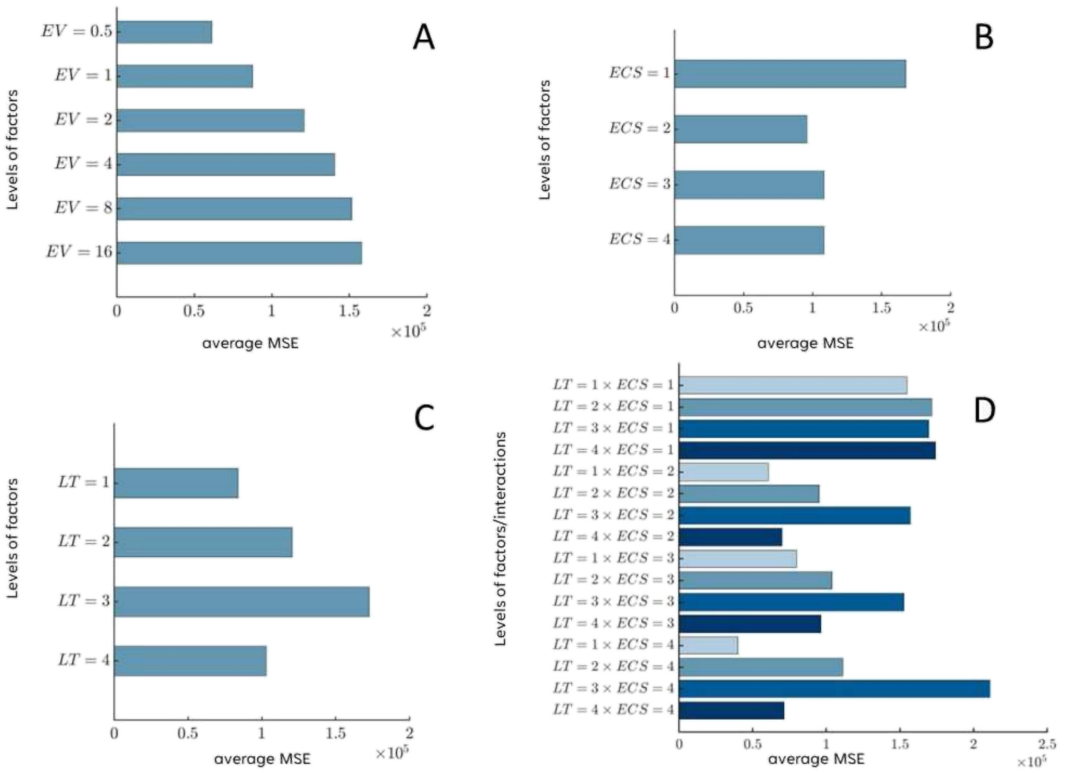

FIGURE 6 | Additive correlated error: averaged (per factor/interaction level) mean square error (MSE, 61) for the loading reconstruction for different factors/interactions of the ANOVA model. (A) Additive correlated error variance (*EV*) (σ^2_{ac}). (B) Error correlation structure (*ECS*): 1 (average), 2 (hub), 3 (Toeplitz), 4 (uniform); see Equations (39)–(42) for definitions. (C) Loading type (*LT*): 1 (simple), 2 (intermediate), 3 (complex), 4 (experimental); see Equations (50)–(52) and (54). (D) MSE for the interaction (*LT* × *ECS*) between *LT* and *ECS*.

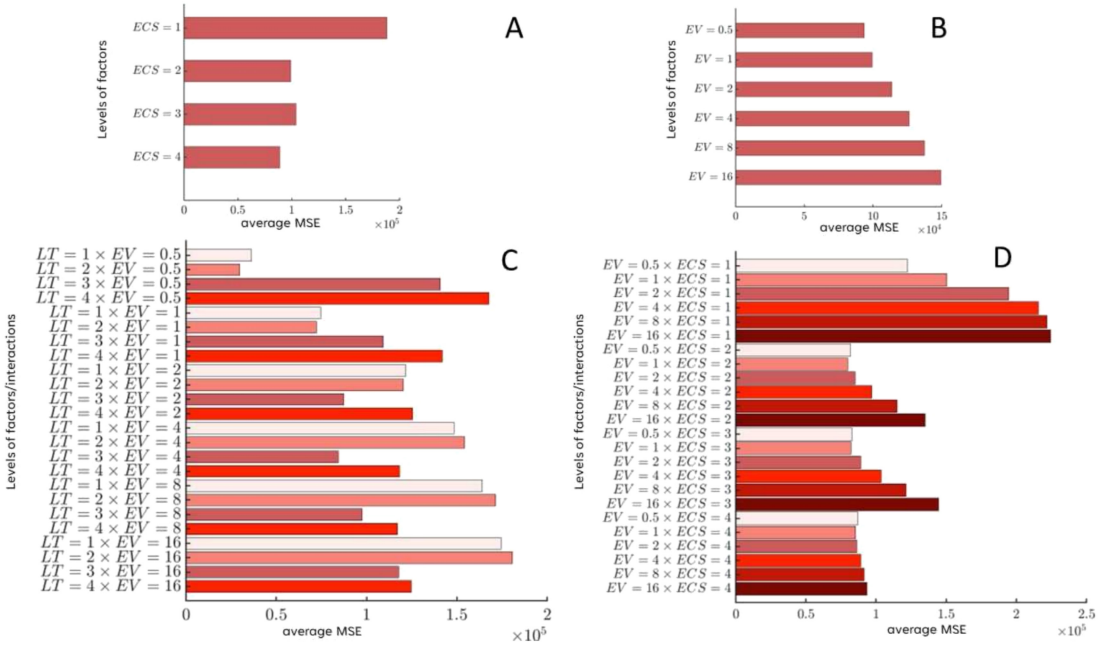

FIGURE 7 | Multiplicative correlated error: averaged (per factor/interaction level) mean square error (MSE, 61) for the loading reconstruction for different factors/interactions of the ANOVA model. (A) Error correlation structure (*ECS*): 1 (average), 2 (hub), 3 (Toeplitz), 4 (uniform); see Equations (39)–(42) for definitions. (B) Multiplicative correlated error variance (*EV*) (σ^2_{mc}). (C) MSE for the interaction (*LT* × *EV*) between *LT* and *EV*. (D) MSE for the interaction (*EV* × *ECS*) between *EV* and *ECS*.

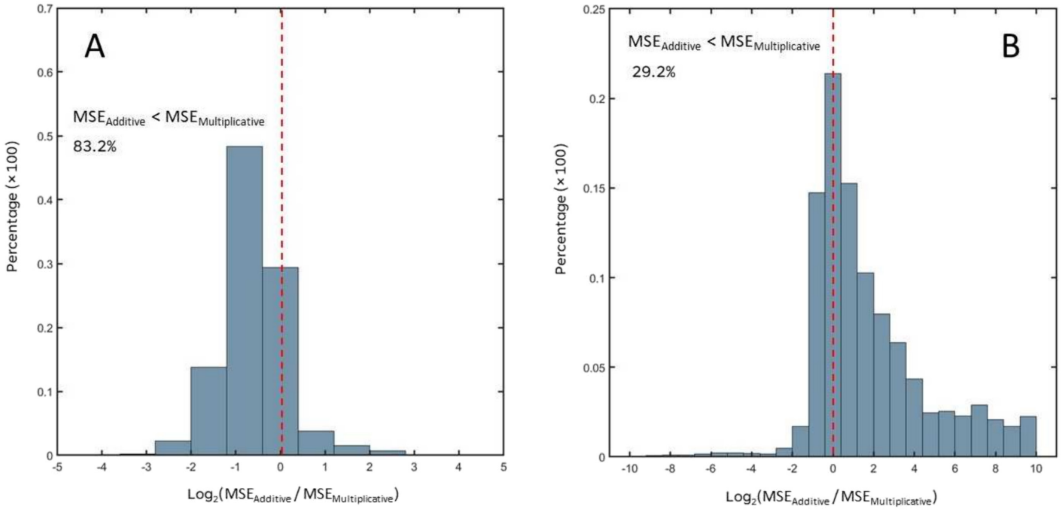

FIGURE 8 | Distribution of the fold change *FC* (Equation 64) of the mean squared reconstruction error (MSE) for loading reconstructed in the presence of additive and multiplicative error. (A) *FC* calculated in the presence of uncorrelated error. (B) *FC* calculated in the presence of correlated error.

case of additive or multiplicative error. To answer this question, we considered the ratio between the reconstruction error in the two cases (see Equation 64): Figure 8 shows the distribution of the log2 ratio of the MSE error calculated in the case of additive error to the MSE calculated for multiplicative case. For uncorrelated error (Figure 8A), the reconstruction error is almost always smaller when the error is additive (83.2% over all simulation realized for all factors and factor combinations). For correlated error, the reconstruction error is often larger when the error is additive (70.8%).

Because of the complex interaction between the factors, it is complicated to define situations where the two conditions ($FC > 0$ or $FC < 0$) apply, but for zero or (very) low error variance, the data realized under the additive and multiplicative error models are similar (see Equations 19 and 29, respectively) and so are the reconstruction errors. Note that the error-free data $\mathbf{X}_0$ are the same for both additive and multiplicative simulations (so that simulations are paired for optimal comparability).

4.2 | Effect of Experimental Error on Loading Congruence

In this section, we study the congruence between sample and population loadings in the presence of error. Overall, patterns of variation for the congruence are similar to those of MSE.

We do observe a major difference in Figure 9 between the distribution of congruence values for correlated error and for uncorrelated error. Congruence values are lower for the former, in some cases much lower, indicating serious distortion of loading patterns. Correlated error distorts correlation in a complicated fashion, with correlation being attenuated or enhanced depending on the interplay between the error correlation level and the variance (see Figures 6 and 7 in [7] and associated text). That is, given the structure of the error correlation matrix used in our simulations where in the average (39), Toeplitz (41), and hub (40) correlation matrices, different pairs of variables have different correlations, we expect the loading patterns also to be affected in a somehow unpredictable fashion, resulting in the observed degradation of congruence values in the case of correlated error.

4.2.1 | Congruence in the Presence of Uncorrelated Error

The ANOVA results for Tucker's ϕ congruence (63) in the case of uncorrelated error are given in Table 2. For additive error (model 46), these results are consistent with those for the MSE (also in Table 2). For multiplicative error (model 47), we see a clear effect of the data distribution in the congruence, whose effect was much smaller for the MSE. We also see that error variance has a much larger effect than on MSE.

Note that Tucker's ϕ values > 0.95 indicate equivalent loadings and $0.85 < \phi < 0.95$ fair equivalence. The distribution of congruence values is given in Figure 9. We observe that full or fair equivalence could be establish in almost all conditions of uncorrelated error, indicating that the shape of loading patterns is maintained.

4.2.2 | Congruence in the Presence of Correlated Error

The ANOVA results of Tucker's ϕ congruence (63) in the case of correlated error (models 48 and 49) are given in Table 3. Like in the previous case, generally speaking, these results are consistent with what was observed for the MSE for the case of additive errors. In the case of multiplicative errors, we observe a stronger effect for the distribution, as expected, and for the error variance.

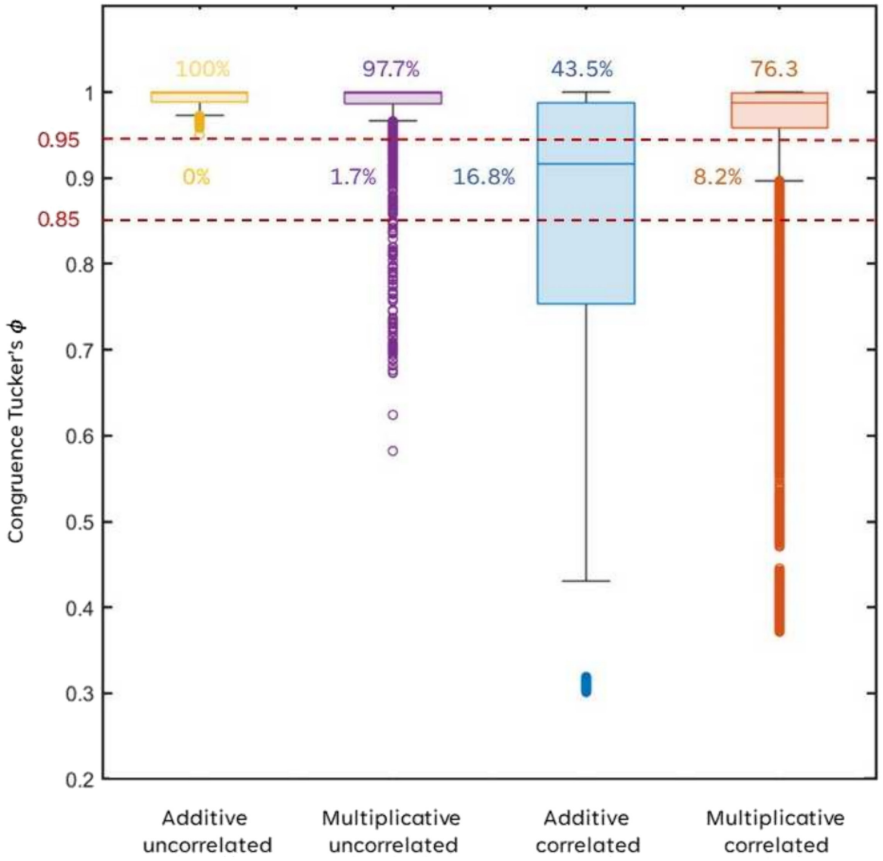

FIGURE 9 | Distribution of Tucker's ϕ congruence index values (Equation 63) between sample loadings estimated in the presence of error and the population loadings. (A) Congruence in the presence of additive uncorrelated error. (B) Congruence in the presence of multiplicative uncorrelated error. (C) Congruence in the presence of additive correlated error. (D) Congruence in the presence of multiplicative correlated error.

5 | Discussion

5.1 | Making Sense of the Results Considering the Sample Correlations

The simulations results previously discussed can be interpreted by considering the relationships between the correlation coefficient ρ_{ij}, the error-free correlation $\rho_{0_{ij}}$, and the error characteristics defined by Equations (10), (11), and (13).

In particular, we will consider the limiting behavior of the sample correlation coefficient, the sample correlation matrix $\mathbf{C}$, and the PCA loadings. The loadings, being the eigenvectors of the sample correlation matrix, are a function of the correlation matrix.

5.1.1 | Effects of Additive Error

We begin by considering the case of purely uncorrelated error: Its multivariate formulation can be obtained from Equation (19) by dropping the correlated term $\mathbf{e}_{ac}$ and considering only the relevant covariance–correlation matrices, namely, $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_{\text{au}}$ given by Equations (16) and (17). Under this model, the expected correlation ρ_{ij} between any two variables x_i and x_j is (this and other following formulas are intended to be $\rho_{ij} = 1$ when $i = j$)

$$\rho_{ij} = \frac{\rho_{0_{ij}}}{\sqrt{\left(1 + \frac{\sigma^2_{au_i}}{\sigma^2_{0_i}}\right)} \times \sqrt{\left(1 + \frac{\sigma^2_{au_j}}{\sigma^2_{0_j}}\right)}}, \tag{65}$$

which is a generalization of (10). The correlation ρ_{ij} is element i, j of the correlation matrix estimated from the data $\mathbf{X}$, which will be, as previously discussed, different from the population correlation $\mathbf{R}_0$ (18).

Equation (65) establishes that in the asymptotic regime of infinite error variance or signal variance tending to zero, the expected correlation is zero:

$$\lim_{\left(\sigma^2_{au_i}, \sigma^2_{au_j}\right) \to (\infty, \infty)} \rho_{ij} = 0, \tag{66}$$

$$\lim_{\left(\sigma^2_{0_i}, \sigma^2_{0_j}\right) \to (0,0)} \rho_{ij} = 0. \tag{67}$$

From this (and from Equation 8 for a sufficiently large sample size), it follows that the sample correlation matrix $\mathbf{C}$ (2) will tend to the $J \times J$ identity matrix $\mathbf{I}$ (which is also the limit of the population correlation matrix $\mathbf{R}$), where J is the number of variables:

$$\lim_{\substack{\sigma^2_{au_i}\to\infty \\ \forall i}} \mathbf{C} = \mathbf{I}, \tag{68}$$

$$\lim_{\substack{\sigma^2_{0_i}\to 0 \\ \forall i}} \mathbf{C} = \mathbf{I}. \tag{69}$$

The $\mathbf{I}$ matrix has J unit eigenvalues ($\lambda_1 = \lambda_2 = \ldots = \lambda_J = 1$) and $J \times 1$ eigenvectors with 1 on the jth position (for the eigenvector associated to the jth eigenvalue) and zero otherwise. Thus, in the limiting situation of high error variance (or low signal variance), the loadings become identical (in absolute value) for all variables. Overall, by deflating the true correlation existing between the variables to zero, the presence of additive uncorrelated error renders the use of PCA suboptimal because PCA is an efficient dimensionality and exploratory tool only when variables are correlated.

Concerning the congruence to population loadings, we note that in our simulation, we used the same variance for all error components ($\sigma^2_{au_j} = \sigma^2_{au}$ in model 24 and $\sigma^2_{mu_j} = \sigma^2_{mu}$ in model 34). Therefore, we expect all variables to be affected equally, thus preserving the overall loading patterns and resulting in large congruence values.

The analysis of simulations presented in Table 2 shows that the most important factors affecting loading quality in the presence of error are the error variance and the population loading type $\mathbf{A}^S_{simple}$ (see Equations 50–52). The magnitude of attenuation of the observed correlations ρ_{ij} (65) also depends on $\rho_{0_{ij}}$ (for the same error variance, the larger the $\rho_{0_{ij}}$, the smaller the attenuation). This can explain why the loading structure and its interaction with error variance are important factors contributing to the variability observed in the quality measures for the loading reconstruction.

If the error is additive and correlated, the ρ_{ij} element of the expected sample correlation matrix will be (this is obtained by dropping the uncorrelated terms in (11) and generalizing the formula to x_i and x_j and suppressing x to simplify the notation)

$$\rho_{ij} = \frac{\rho_{0_{ij}} + \pi_{ac}\frac{\sigma_{ac_i}}{\sigma_{0_i}}\frac{\sigma_{ac_j}}{\sigma_{0_j}}}{\sqrt{1+\frac{\sigma^2_{ac_i}}{\sigma^2_{0_i}}}\times\sqrt{1+\frac{\sigma^2_{ac_j}}{\sigma^2_{0_j}}}}. \tag{70}$$

In the limit of infinite error variance, the behavior of ρ_{ij} is totally determined by the variance of the correlated error:

$$\lim_{\left(\sigma^2_{ac_i},\sigma^2_{ac_j}\right)\to(\infty,\infty)} \rho_{ij} = \pi_{ac_{ij}}, \tag{71}$$

$$\lim_{\left(\sigma^2_{0_i},\sigma^2_{0_j}\right)\to(0,0)} \rho_{ij} = \pi_{ac_{ij}}, \tag{72}$$

from which it follows that the sample correlation matrix tends to the error correlation matrix (which in the multivariate error model specifications, we have indicated with $\mathbf{\Pi}_{ac}$; see Equation 27):

$$\lim_{\substack{\sigma^2_{ac_i}\to\infty \\ \forall i}} \mathbf{C} = \mathbf{\Pi}_{ac}, \tag{73}$$

$$\lim_{\substack{\sigma^2_{0_i}\to 0 \\ \forall i}} \mathbf{C} = \mathbf{\Pi}_{ac}. \tag{74}$$

In this extreme case, the PCA will just describe the error structure, rather than the underlying relationships between the observed variables that are of interest, and this is consistent with the ECS (which is defined by $\mathbf{\Pi}_{ac}$) being an important factor in the ANOVA model results given in Table 3.

The loading structure depends on the correlation structure $\mathbf{\Pi}_{ac}$ of the correlated error: As this is in general not known, it is not possible to derive a general rule. If the error is perfectly correlated, with $\pi_{ac_{ij}} = 1$ in $\mathbf{\Pi}_{ac}$ (27) for every variable, it is $\mathbf{\Pi}_{ac} = \mathbf{1}\mathbf{1}^t$, which leads to a degenerate PCA model with unity loadings for the first (and only) component. Also in this case, the results of simulation given in Table 3 indicate that the most important factors are the error variance and the loading structure: Similar considerations hold for the results obtained for the case of purely additive uncorrelated error.

The loading structure and its interactions with error variance and ECS appear as important factors contributing to the error reconstruction MSE and congruence ϕ (Tables 2 and 3). Although the loading structure appears explicitly in the data generation procedure, these interactions may not be immediately evident from the explanation offered in terms of the (asymptotic) behavior of the sample correlation matrix.

In the case of additive error, the interaction between the loading structure and the error variance and ECS can be made more explicit by calculating the (expected) error (bias) on the loadings. This has been derived by Hellton and Thoresen [10] using perturbation theory. For error variance $\sigma^2 \to 0$ conditionally on the noise-free data $\mathbf{X}_0$ as (Theorem 2 in [10]: Note that index notation has been changed to be consistent with our notation)

$$\begin{aligned}\mathbb{E}(\Delta\mathbf{v}_k|\mathbf{X}_0) = \sigma^2 \sum_{l\neq k}\frac{\mathbf{v}_l^T\mathbf{\Sigma}_{au}\mathbf{v}_k}{\lambda_k-\lambda_l}\mathbf{v}_l - 2\frac{\sigma^2}{N}\sum_{l\neq k}\frac{\mathbf{v}_l^T\mathbf{\Sigma}_{au}\mathbf{v}_k}{\lambda_k-\lambda_l}\mathbf{v}_l + \\ + \frac{\sigma^2}{N}\sum_{l\neq k}\sum_{m\neq k}\lambda_l\frac{\mathbf{v}_k^T\mathbf{\Sigma}_{au}\mathbf{v}_k}{(\lambda_k-\lambda_l)(\lambda_k-\lambda_m)}\mathbf{v}_l + O(\sigma^3),\end{aligned} \tag{75}$$

where $\mathbf{v}_k$ is the kth eigenvector (loading vector) of the covariance–correlation matrix of the error-free data $\mathbf{X}_0$ associated with the λ_k eigenvalue and $\mathbf{\Sigma}_{au}$ is the error correlation matrix; k runs from 1 to J when $J \leq N$ and to N when $J > N$.

Equation (75) shows how the error in the loadings depends on the loadings themselves and thus depends on the loading structure, the error variance, and the ECS. In particular, the interaction between the loading structure and ECS is expressed by the term $\mathbf{v}_l^T\mathbf{\Sigma}_{au}\mathbf{v}_k$: for $k = l$, which is the projection of the error covariance–correlation matrix onto the space spanned by the bv_k eigenvector (loading).

In the case of uncorrelated and homogeneous error (the case addressed in our study, see 28), the expected bias $\Delta\mathbf{v}_k$ on the

loadings is zero, because $\mathbf{v}_T^l \boldsymbol{\Sigma}_{au} \mathbf{v}_k = 0$ for $k \neq l$. However, note that in our simulation strategy, the error variance σ^2 is not negligible with respect to the signal variance and, as such, we observed MSE > 0, as shown in Figure 3B. From Equation (75), it follows that errors in the estimation are introduced if the errors are not homogeneous, a case that we have not addressed in our simulations.

Nonhomogeneous additive error increases loading variability (given by eq. 9 in [10]), and this results in variables with a larger error being given higher importance in the PCA model. The extent of the bias on the loadings depends on whether the error variance on a particular variable is larger or smaller than the average error on all other variables (Section 3.1 in [10]).

5.1.2 | Effects of Multiplicative Error

Using the same arguments used in the case of additive error, we can derive the limiting behavior of the (sample) correlation matrix in the presence of multiplicative error. If the errors are uncorrelated, the ρ_{ij} element of the correlation matrix is obtained by setting with $\pi_{mc_{ij}} = 0$ (and taking $\sigma_{mc_i} = \sigma_{mc_i} = 0$ for convenience) and generalizing to the multivariate case, which yields

$$\rho_{ij} = \frac{\rho_{0_{ij}}}{\sqrt{1 + \left(1 + \frac{\mu_{0_i}^2}{\sigma_{0_i}^2}\right)\sigma_{mu_i}^2} \times \sqrt{1 + \left(1 + \frac{\mu_{0_j}^2}{\sigma_{0_j}^2}\right)\sigma_{mu_j}^2}}. \tag{76}$$

Similar considerations apply for what concerns the limiting behavior of $\mathbf{C}$ with respect to error variance and, in addition, signal mean, which can help to explain results of simulations in Table 3.

If the errors are correlated, the matter is even more intricate, because the effect of the error depends on the interplay between error characteristics (correlation and variance) and the data itself, as per the nature of multiplicative error:

$$\rho_{ij} = \frac{\rho_{0_{ij}}(1 + \pi_{mc}\sigma_{mc_i}\sigma_{mc_j}) + \frac{\mu_{0_i}}{\sigma_{x_{0_i}}}\frac{\mu_{0_j}}{\sigma_{0_j}}\pi_{mc}\sigma_{mc_i}\sigma_{mc_j}}{\sqrt{1 + \left(1 + \frac{\mu_{0_i}}{\sigma_{0_i}^2}\right)\sigma_{mc_i}^2} \times \sqrt{1 + \left(1 + \frac{\mu_{0_i}^2}{\sigma_{0_i}^2}\right)\sigma_{mc_j}^2}}. \tag{77}$$

For very large error variance or for very small signal variance, the correlated part of the error will dominate, and we again have the limiting conditions

$$\lim_{\substack{\sigma_{mc_i}^2 \to \infty \\ \forall\, i}} \mathbf{C} = \boldsymbol{\Pi}_{\text{mc}}, \tag{78}$$

$$\lim_{\substack{\mu_{x_{0_i}} \to \infty \\ \forall\, i}} \mathbf{C} = \boldsymbol{\Pi}_{\text{mc}}. \tag{79}$$

Again with reference to the ANOVA model results in both Tables 2 and 3, it should be noted that for multiplicative errors, the distribution played a larger role than in the case of the additive error. For the additive error case, the distribution had little to no effect, while it had some effect for the multiplicative error. The latter is explained by the explicit appearance of term $\mu_{0_i}^2$ in (77), which can be directly related to the data distribution. In fact for the normal distribution (55), the population mean is 0 ($\mu_{0_j} = 0$ for all variables j), while for the log-normally distributed scores (56), the population mean is 1 ($\mu_{0_j} = 1$ for all variables). This difference in means has its repercussions for the expected correlations and thus for the estimated loadings. The case of multiplicative error is by far more complicated to address analytically: Vaswani and Guo [36] addressed a related error model, but their theoretical framework and results are not easily transferable to our setting.

5.2 | Extending to a Generalized Error Model

In this study, we have considered the two error models independently: This was dictated by the necessity of keeping the simulations within reasonable complexity and to be able to compare directly the effect of additive and multiplicative errors. Moreover, when considering high-throughput measurement technologies, it is not unreasonable to assume that the multiplicative error will dominate for high signal intensity (this corresponds in our data models to have large $\mu_{x_{0_i}}$ in the error-free data), while the additive error will contribute mostly at low signal intensity ($\mu_{x_{0_i}} \to 0$). However, there are situations when neither of the two limiting conditions is realized, or it is realized only for a subset of variables, and both additive and multiplicative errors coexist, producing data and error models that are rather complicated, given by the combination of models (19) and (29):

$$\mathbf{x} = \mathbf{x}_0 \odot \left(1 + \mathbf{e}_{mu} + \mathbf{e}_{mc}\right) + \mathbf{e}_{au} + \mathbf{e}_{ac}. \tag{80}$$

Under this model, the (expected) variance $\text{var}(x_j)$ of the x_j variable is (see [7], eqs. 61–69 for a derivation)

$$\text{var}(x_j) = \sigma_j^2 + \left(\sigma_j^2 + \mu_{0_j}^2\right)\left(\sigma_{mu_j}^2 + \sigma_{mc_j}^2\right) + \sigma_{au_j}^2 + \sigma_{ac_j}^2. \tag{81}$$

This is fundamentally a two-component error model (visualized in Figure 10). It can be seen as a generalization to the Rocke–Lorenzato model [37], which assumes constant measurement error at low signal intensity and constant coefficient of variation at high intensity (i.e., multiplicative error component). Also, the correlation coefficient and hence the i,jth element of the expected sample correlation matrix can be generalized under the complete model (80) (for a derivation, see [7], eqs. 61–69). Namely, it holds

$$\rho_{ij} = \frac{\rho_{0_{ij}}(1 + \pi_{mc}\sigma_{mc_i}\sigma_{mc_j}) + \pi_{ac}\frac{\sigma_{ac_i}}{\sigma_{0_i}} + \frac{\mu_{0_i}}{\sigma_{0_i}}\frac{\mu_{0_j}}{\sigma_{0_j}}\pi_{mc}\sigma_{mc_i}\sigma_{mc_j}}{\sqrt{1 + \left(1 + \frac{\mu_{0_i}^2}{\sigma_{0_i}^2}\right)\left(\sigma_{mu_i}^2 + \sigma_{mc_i}^2\right) + \frac{\sigma_{au_i}^2}{\sigma_{i_0}^2} + \frac{\sigma_{ac_i}^2}{\sigma_{i_0}^2}} \times \\ \times \sqrt{1 + \left(1 + \frac{\mu_{0_j}^2}{\sigma_{j_0}^2}\right)\left(\sigma_{mu_j}^2 + \sigma_{mc_j}^2\right) + \frac{\sigma_{au_j}^2}{\sigma_{j_0}^2} + \frac{\sigma_{ac_j}^2}{\sigma_{j_0}^2}}}. \tag{82}$$

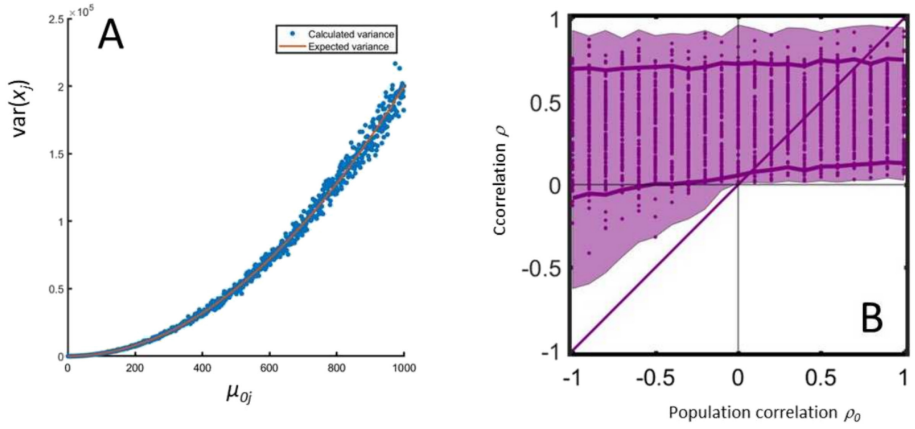

FIGURE 10 | (A) Expected variance (red curve) of x_j (81) generated in the presence of multiplicative and additive errors (80) as a function of the population mean μ_{0_j}. The realized variance over 100 replicates of x_j is represented by the blue dots. Simulation realized with $\sigma^2_{x_j} = \sigma^2_{au_j} = \sigma^2_{ac_j} = \sigma^2_{mu_j} = \sigma^2_{mc_j} = 1$. For small μ_{0_j} ($\rightarrow 0$), additive measurement can be considered to be present; for large values $\mu_{0_j} \rightarrow \infty$, the multiplicative error dominates. (B) Expected correlation coefficient ρ_{12} between x_1 and x_2 as a function of the different realizations of error parameters for different values of the true correlation $\rho_{0_{12}}$. The shadowed area encloses the maximum and the minimum of the values of ρ_{12} calculated in the simulation using the different error parameters. The dots represent the realized values of ρ, solid lines the fifth and the 95th percentiles of the observed values. (B) is adapted from Figure 6 in [7].

The relationship between the correlation ρ_{ij} (82) and the true (population) correlation ρ_0 is explored using Monte Carlo realizations of the error parameters underlying model (80) (viz., σ^2_j, $\sigma^2_{au_j}$, $\sigma^2_{ac_j}$, $\sigma^2_{mu_j}$, and $\sigma^2_{mc_j}$ plus population means μ_{0_j}). The behavior of ρ_{ij} under the general model is markedly similar to that observed under the multiplicative error model shown in Figure 2D, which suggests that the multiplicative error component may be dominant in this regime. This is also consistent with Figure 8, showing that, in most of the simulations, the MSE error on the loading population is larger in the presence of multiplicative errors than in the presence of additive errors.

6 | Conclusions

Our simulation study goes beyond classical results like those discussed by [10], by extending our analysis to additive correlated error and to multiplicative error, both correlated and uncorrelated. Our investigation highlights the important role and some particular characteristics of all error types.

Several approaches have been proposed to incorporate information about the measurement errors in the PCA model, ranging from a modified version of PCA [38, 39] to a framework for maximum likelihood PCA that includes an assumed known covariance matrix for the errors [39–42], which relates to exploratory common factor analysis [43, 44].

Another possible approach could be to exploit the fact that the correlation can be corrected if the error parameters are known, in all three cases (additive, multiplicative, and general models), as shown by Saccenti et al. [7] (see eqs. 79–81). This correction of course requires the (full) knowledge of the variance–covariance matrix of the errors. Approaches have been proposed for systematically characterizing the measurement error covariance matrix for a given experimental platform [8, 18, 45], together with methods based on fitting the Wishart distribution [9].

In the absence of knowledge of the error structure and relative error size, the interpretation of a PCA model is on an uncertain ground. Therefore, we advocate, when possible, to assess explicitly the nature and size of measurement error involved in the variables that are to be subjected to PCA. Ideally, the resulting knowledge can be incorporated in an adapted PCA version. We believe it would be of interest in future research to develop methods to express the uncertainty in the model estimates that is due to the estimated (rather than known) error properties.

Acknowledgments

E.S. acknowledges the funding received from the Netherlands Organisation for Health Research and Development (ZonMW) through the PERMIT project (Personalized Medicine in Infections: From Systems Biomedicine and Immunometabolism to Precision Diagnosis and Stratification Permitting Individualized Therapies, project number 456008002) under the PerMed Joint Transnational call JTC 2018 (research projects on personalized medicine–smart combination of preclinical and clinical research with data and ICT solutions) and the funding from the European Union's Horizon 2020 research and innovation program through the DIAGONAL project (GA No. 953152). J.C. was supported by the Agencia Estatal de Investigación in Spain, MCIN/AEI/10.13039/501100011033, Grant PID2020-113462RB-I00.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Peer Review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/cem.3595.

References

1. K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, no. 11 (1901): 559–572.

2. H. Hotelling, "Analysis of a Complex of Statistical Variables Into Principal Components," *Journal of Educational Psychology* 24, no. 6 (1933): 417.

3. E. R. Malinowski, *Factor Analysis in Chemistry* (New York: Wiley, 1991).

4. R. Bro and A. K. Smilde, "Principal Component Analysis," *Analytical Methods* 6, no. 9 (2014): 2812–2831.

5. I. T. Jolliffe and J. Cadima, "Principal Component Analysis: A Review and Recent Developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, no. 2065 (2016): 20150202.

6. H. N. B. Moseley, "Error Analysis and Propagation in Metabolomics Data Analysis," *Computational and Structural Biotechnology Journal* 4, no. 5 (2013): e201301006.

7. E. Saccenti, M. H. W. B. Hendriks, and A. K. Smilde, "Corruption of the Pearson Correlation Coefficient by Measurement Error and Its Estimation, Bias, and Correction Under Different Error Models," *Scientific Reports* 10, no. 1 (2020): 438.

8. T. K. Karakach, P. D. Wentzell, and J. A. Walter, "Characterization of the Measurement Error Structure in 1D ^{1}H NMR Data for Metabolomics Studies," *Analytica Chimica Acta* 636, no. 2 (2009): 163–174.

9. P. D. Wentzell, C. S. Cleary, and M. Kompany-Zareh, "Improved Modeling of Multivariate Measurement Errors Based on the Wishart Distribution," *Analytica Chimica Acta* 959 (2017): 1–14.

10. K. H. Hellton and M. Thoresen, "The Impact of Measurement Error on Principal Component Analysis," *Scandinavian Journal of Statistics* 41, no. 4 (2014): 1051–1063.

11. T. Raykov, G. A. Marcoulides, and T. Li, "On the Fallibility of Principal Components in Research," *Educational and Psychological Measurement* 77, no. 1 (2017): 165–178.

12. N. M. Faber, L. M. C. Buydens, and G. Kateman, "Aspects of Pseudorank Estimation Methods Based on the Eigenvalues of Principal Component Analysis of Random Matrices," *Chemometrics and Intelligent Laboratory Systems* 25, no. 2 (1994): 203–226, https://www.sciencedirect.com/science/article/pii/0169743994850437.

13. W. A. Fuller, *Measurement Error Models*, Vol. 305, (New York: John Wiley & Sons, 2009).

14. C. Spearman, "The Proof and Measurement of Association Between Two Things," *The American Journal of Psychology* 15, no. 1 (1904): 72–101.

15. J. Tuitman, S. Vanduffel, and J. Yao, "Correlation Matrices With Average Constraints," *Statistics & Probability Letters* 165 (2020): 108868.

16. J. Hardin, S. R. Garcia, and D. Golan, "A Method for Generating Realistic Correlation Matrices," *Annals of Applied Statistics* 7, no. 3 (2013): 1733–1762.

17. U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications* (Berkeley and Los Angeles: University of California Press, 1958).

18. M. N. Leger, L. Vega-Montoto, and P. D. Wentzell, "Methods for Systematic Investigation of Measurement Error Covariance Matrices," *Chemometrics and Intelligent Laboratory Systems* 77, no. 1–2 (2005): 181–205.

19. M. E. Timmerman, H. A. L. Kiers, and A. K. Smilde, "Estimating Confidence Intervals for Principal Component Loadings: A Comparison Between the Bootstrap and Asymptotic Results," *British Journal of Mathematical and Statistical Psychology* 60, no. 2 (2007): 295–314.

20. H. F. Kaiser, "The Varimax Criterion for Analytic Rotation in Factor Analysis," *Psychometrika* 23, no. 3 (1958): 187–200.

21. S. Chatterjee, "Variance Estimation in Factor Analysis: An Application of the Bootstrap," *British Journal of Mathematical and Statistical Psychology* 37, no. 2 (1984): 252–262.

22. L. Milan and J. Whittaker, "Application of the Parametric Bootstrap to Models That Incorporate a Singular Value Decomposition," *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 44, no. 1 (1995): 31–49.

23. J. Camacho, J. Picó, and A. Ferrer, "Data Understanding With PCA: Structural and Variance Information Plots," *Chemometrics and Intelligent Laboratory Systems* 100, no. 1 (2010): 48–56.

24. A. C. Visser-Keizer, A. Hogenkamp, H. J. Westerhof-Evers, I. J. L. Egberink, and J. M. Spikman, "Dutch Multifactor Fatigue Scale: A New Scale to Measure the Different Aspects of Fatigue After Acquired Brain Injury," *Archives of Physical Medicine and Rehabilitation* 96, no. 6 (2015): 1056–1063, https://www.sciencedirect.com/science/article/pii/S0003999314013501.

25. B. F. Green, "The Orthogonal Approximation of an Oblique Structure in Factor Analysis," *Psychometrika* 17, no. 4 (1952): 429–440.

26. C. Burt, "Factor Analysis and Canonical Correlations," *British Journal of Psychology* 1 (1948): 95–106.

27. L. R. Tucker, *A Method for Synthesis of Factor Analysis Studies*, Vol. 984, (Princeton, NJ: Educational Testing Service, 1951).

28. U. Lorenzo-Seva and J. M. F. Ten Berge, "Tucker's Congruence Coefficient as a Meaningful Index of Factor Similarity," *Methodology* 2, no. 2 (2006): 57–64.

29. D. C. Montgomery, *Design and Analysis of Experiments* (New York: Wiley, 2020).

30. O. P. Merchanskaya, M. D. S. Armstrong, C. G. Llorente, et al. "Considerations for Missing Data, Outliers and Transformations in Permutation Testing for ANOVA, ASCA(+) and Related Factorizations." Preprint, submitted 2024. https://arxiv.org/abs/2408.06739.

31. J. Camacho, C. Díaz, and P. Sánchez-Rovira, "Permutation Tests for ASCA in Multivariate Longitudinal Intervention Studies," *Journal of Chemometrics* 37, no. 7 (2023): e3398.

32. J. T. E. Richardson, "Eta Squared and Partial Eta Squared as Measures of Effect Size in Educational Research," *Educational Research Review* 6, no. 2 (2011): 135–147.

33. E. G. Williams, N. Pfister, S. Roy, et al., "Multiomic Profiling of the Liver Across Diets and Age in a Diverse Mouse Population," *Cell Systems* 13, no. 1 (2022): 43–57.

34. R Core Team, *R: A Language and Environment for Statistical Computing* (Vienna, Austria: R Foundation for Statistical Computing, 2024), https://www.R-project.org/.

35. T. Wei and V. Simko, "R Package 'corrplot': Visualization of a Correlation Matrix (Version 0.92)," (2021).

36. N. Vaswani and H. Guo, "Correlated-PCA: Principal Components' Analysis When Data and Noise Are Correlated," *Advances in Neural Information Processing Systems* 29 (2016).

37. D. M. Rocke and S. Lorenzato, "A Two-Component Model for Measurement Error in Analytical Chemistry," *Technometrics* 37, no. 2 (1995): 176–184.

38. G. Sanguinetti, M. Milo, M. Rattray, and N. D. Lawrence, "Accounting for Probe-Level Noise in Principal Component Analysis of Microarray Data," *Bioinformatics* 21, no. 19 (2005): 3748–3754.

39. P. D. Wentzell and S. Hou, "Exploratory Data Analysis With Noisy Measurements," *Journal of Chemometrics* 26, no. 6 (2012): 264–281.

40. P. D. Wentzell, D. T. Andrews, D. C. Hamilton, K. Faber, and B. R. Kowalski, "Maximum Likelihood Principal Component Analysis," *Journal of Chemometrics: A Journal of the Chemometrics Society* 11, no. 4 (1997): 339–366.

41. P. D. Wentzell and M. T. Lohnes, "Maximum Likelihood Principal Component Analysis With Correlated Measurement Errors: Theoretical and Practical Considerations," *Chemometrics and Intelligent Laboratory Systems* 45, no. 1–2 (1999): 65–85.

42. M. E. Tipping and C. M. Bishop, "Probabilistic Principal Component Analysis," *Journal of the Royal Statistical Society Series B: Statistical Methodology* 61, no. 3 (1999): 611–622.

43. R. Cudeck, "10—Exploratory Factor Analysis," in *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, eds. H. E. A. Tinsley and S. D. Brown (San Diego: Academic Press, 2000): 265–296, https://www.sciencedirect.com/science/article/pii/B9780126913606500112.

44. L. R. Goldberg and W. F. Velicer, "Principles of Exploratory Factor Analysis," *Differentiating Normal and Abnormal Personality* 2 (2006): 209–337.

45. L. Blanchet, J. Réhault, C. Ruckebusch, J. P. Huvenne, R. Tauler, and A. de Juan, "Chemometrics Description of Measurement Error Structure: Study of an Ultrafast Absorption Spectroscopy Experiment," *Analytica Chimica Acta* 642, no. 1–2 (2009): 19–26.