

# Errors-In-Variables Model Fitting for Partially Unpaired Data Utilizing Mixture Models

Wolfgang Högele<sup>1</sup> and Sarah Brockhaus<sup>1</sup>

<sup>1</sup> Munich University of Applied Sciences HM  
Department of Computer Science and Mathematics  
Lothstraße 64, 80335 München, Germany

[wolfgang.hoegele@hm.edu](mailto:wolfgang.hoegele@hm.edu)

November 19, 2024

## Abstract

We introduce a general framework for regression in the errors-in-variables regime, allowing for full flexibility about the dimensionality of the data, observational error probability density types, the (nonlinear) model type and the avoidance of ad-hoc definitions of loss functions. In this framework, we introduce model fitting for partially unpaired data, i.e. for given data groups the pairing information of input and output is lost (semi-supervised). This is achieved by constructing mixture model densities, which directly model the loss of pairing information allowing inference. In a numerical simulation study linear and nonlinear model fits are illustrated as well as a real data study is presented based on life expectancy data from the world bank utilizing a multiple linear regression model. These results show that high quality model fitting is possible with partially unpaired data, which opens the possibility for new applications with unfortunate or deliberate loss of pairing information in data.

**Keywords:** Errors-In-Variables; Mixture Models; Model Fitting; Semi-Supervised; Total Least Squares

## About the Authors

Dr. Högele is a Full Professor of Applied Mathematics and Computational Science at the Department of Computer Science and Mathematics at the Munich University of Applied Sciences HM, Germany.

Dr. Brockhaus is a Full Professor of Applied Mathematics and Statistics at the Department of Computer Science and Mathematics at the Munich University of Applied Sciences HM, Germany.

*This manuscript is accepted and will be published in STATISTICS (Taylor & Francis), 2024*

*This manuscript appears also on ArXiv.org*  
*arXiv:2406.18154 [stat.ME], <https://doi.org/10.48550/arXiv.2406.18154>*

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Methods</b>	<b>5</b>
2.1	General Nomenclature . . . . .	5
2.2	Model Fit with Completely Paired Data . . . . .	6
2.2.1	Example: Fitting a Line and Gaussian Disturbance . . . . .	7
2.2.2	Example: Fitting a Hyperplane and Gaussian Disturbance (Errors-In-Variables Multiple Linear Regression) . . . . .	7
2.2.3	Connection to Interval Data Regression . . . . .	8
2.2.4	Example: Linear Interval Data Regression . . . . .	9
2.3	Model Fit with Completely Unpaired Data . . . . .	9
2.4	Model Fit with Partially Unpaired Data . . . . .	10
2.4.1	Example: Fitting a Line and Gaussian Disturbance . . . . .	11
2.4.2	Example: Fitting a Hyperplane and Gaussian Disturbance (Errors-In-Variables Multiple Linear Regression) . . . . .	11
2.4.3	Example: Linear Interval Data Regression . . . . .	12
2.5	Extensions . . . . .	12
2.5.1	Numerical Implementation of the General Formula for Partially Unpaired Data	12
2.5.2	Evaluation of Unpaired Data Subgroups . . . . .	13
2.5.3	Simultaneous Estimation of the Underlying Density Functions . . . . .	13
2.5.4	Bayesian Extension . . . . .	14
<b>3</b>	<b>Simulation Study</b>	<b>14</b>
3.1	Demonstration for a Line Fit . . . . .	14
3.2	Nonlinear Model with Anisotropic Observation Errors . . . . .	16
<b>4</b>	<b>Real Data Study: Life Expectancy</b>	<b>17</b>
<b>5</b>	<b>Discussion and Conclusion</b>	<b>19</b>
<b>A</b>	<b>Completely Paired Data: Derivations for Errors in <math>y</math> only</b>	<b>20</b>
<b>B</b>	<b>Completely Paired Data: Relation to Deming Regression</b>	<b>21</b>
<b>C</b>	<b>Demonstration for a Plane Fit with Gaussian Disturbance for Partially Unpaired Data</b>	<b>22</b>
<b>D</b>	<b>Application of <math>R_{\delta}^2</math></b>	<b>24</b>

# 1 Introduction

Parametric model fitting is a standard task in many applications starting from problem specific models with a few meaningful parameters to huge, flexible models (such as in the training phase of ANNs) [Zhang, 1997, Bishop, 2006]. The general idea is that a defined parametric model family is fitted to given input / output data (i.e. supervised learning) with the purpose to estimate the *best fit* parameters of the model representing the data. Typically, loss functions are defined for this purpose, such as the squared or absolute losses [Wang et al., 2022a]. Practical difficulties of model fitting are i) finding the appropriate model family and their *parameters*, ii) defining an adequate loss function, iii) applying an efficient optimization algorithm and iv) dealing with data deficiencies (such as outliers, strong noise, incompleteness, partially lost pairing information, etc.).

A possibility to avoid the definition of an *ad-hoc* loss function is the modeling of known uncertainties in the data and applying Maximum Likelihood (ML) approaches, which, in consequence, lead inherently to data driven loss functions. For example, data uncertainties can appear only in the output data (such as for ordinary least squares) or in input and output data, also known as the *errors-in-variables* approaches. A well-known connection is between a normal distributed error in the output and the squared loss function. These approaches can always be extended to Bayesian estimations if prior distributions are assumed for the model parameters leading to Maximum A Posteriori (MAP) or Minimum Mean Squared Error (MMSE) algorithms based on the posterior distribution [Hoegele et al., 2013].

*Unpaired / unlabeled* data (also *broken sampling* in regression) can occur in many applications and is focus of current research, e.g. [Bai and Hsing, 2005, Liang et al., 2007, Wang et al., 2022b]. In this work, we understand by *partially unpaired* data (under the term *semi-supervised data*) specifically that for parts of the data the one-to-one pairing of input and output data is missing, but we still have paired subgroups in the data. *Semi-supervised* data often is understood as having additional input data without output data, e.g. [Kostopoulos et al., 2018, Qi and Luo, 2022]. This is different to the partially unpaired data in this paper and we regard this as a different flavor of *semi-supervised* data since we lose significant information compared to a fully *supervised* framework, but we still have (possibly weak) input/output relations in contrast to *unsupervised learning*. See Figure 1 for an illustration of different pairing information between input  $\mathbf{x}$  and output  $\mathbf{y}$ . It is demonstrated that mixtures of labeled and unlabeled data can improve the predictive performance in regression problems [Liang et al., 2007]. There are different strategies to deal with such deficiencies, such as altering the data set by data imputation approaches, e.g. see [Bennett, 2001, Sterne et al., 2009], or focusing on unordered subsets in hypothesis testing [Wang et al., 2022b]. We regard the work of Liang *et al.* [Liang et al., 2007] considering a general predictive Bayesian frameworks for mixed labeled and unlabeled data as closest to our goal. Although the work provides a very general framework for regression and classification, it misses the configurational complexity of partially unpaired data. In consequence, one task of this paper is to incorporate the largest possible variety of missing pairing information transparently to model fitting by avoiding data alteration (deletion or imputation) and actively constructing mixture model probability densities accurately representing the partial pairing.

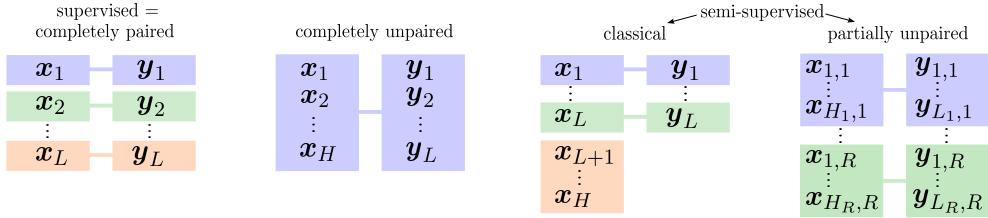


Figure 1: Nomenclature of data configurations: *completely paired*, *completely unpaired* and *semi-supervised* with the *classical* and the *partially unpaired* configuration.

Mixture Models are probability density functions, which are composed of the weighted sum of elementary probability densities. They are applied in a variety of applications, especially Gaussian Mixture Models (GMM) are very popular, which are the weighted sum of Gaussian densities, e.g. [Liang et al., 2007]. A main task is in the literature to find the GMM description of a given data set utilizing expectation maximization (EM) algorithms, also for the task of missing data deficiencies

[Michael et al., 2020, McCaw et al., 2022]. To the knowledge of the authors, a general framework for constructing mixture models for dealing with the high complexity of corrupt / incomplete pairing information of the data with the goal to support a model fitting problem is missing in the literature. Closely related approaches of utilizing mixture models in applied mathematics with lost pairing information are presented for computer vision in order to deal with unknown feature correspondences [Hoegele, 2024a] and for random equations with high combinatorial possibilities for the stochastic parameters [Hoegele, 2024b].

In this paper, we propose to consider the problem of model fitting in a new conclusive way. A general probabilistic framework for fitting models in data based purely on *observational error probability density functions* including *errors-in-variables* is presented, which has direct relations to well-known standard methods for completely paired data, such as ordinary least squares, Deming regression [Deming, 1964], total least squares[Markovsky and Van Huffel, 2007], interval data regression as well as multiple linear regression. Observational error densities can have different reasons, for example, errors only in the output can be classical Gaussian measurement errors, but if regression is performed additionally with measured input data then both, input and output errors (*errors-in-variables*), are typically described by Gaussians. Another example is *interval data*, which can be represented by uniform density functions and which can occur, e.g., in survey data, in particular, when asking for sensitive information like income. This framework will be generalized from *supervised* to *semi-supervised model fitting* by including (partially) unpaired data in one common line of stochastic argumentation. It is a key point that the presented derivations allow for full flexibility about i) the number and dimensions of the input / output data, ii) the type of individual error characteristics of each data point with *errors-in-variables* utilizing general density functions, iii) the type of (linear or nonlinear) models which should be fitted and iv) the pairing information level. In the schematic Figure 2 the concepts of this work are explained in an overview starting from the well-known ordinary least squares application of a line fit in *Subfigure A* to the most general concept of this paper in *Subfigure D* with a nonlinear fit in partially unpaired data. This presentation utilizes a one-dimensional input and output for illustrative purposes, which is generalized in the paper to arbitrary dimensions.

In summary, there are two main goals of this paper: I) presenting a general stochastic argumentation framework for model fitting (without *ad-hoc* loss functions) (see Figure 2A–C) and II) presenting an extension of model fitting within this framework to partially unpaired data utilizing mixture models (see Figure 2D).

In I) the main ideas of the stochastic argumentation are: a) Formulating the fitting problem as a Maximum Likelihood (ML) problem of difference random variables. b) Applying the law of total probability for densities wherever necessary to make sure that correct stochastic dependencies are utilized and identifying the density functions of the basic random variables in the ML problem. c) Presentation of the optimization problems by the resulting objective functions. Points a) to c) are presented repeatedly for different fitting scenarios and their generality is a *first main result* of this paper.

In II) the extension of model fitting to partially unpaired data is structured in the following three-step-approach: Presenting the cases for

- completely paired data sets as the standard case in model fitting (i.e., *supervised learning*) (Section 2.2).
- completely unpaired data sets are introduced mathematically utilizing mixture model random variables to model fitting (Section 2.3). In this extreme case it is impossible to model the relation between input and output.
- partially unpaired data sets, which lie between the two previous extremes and include different levels of pairing (i.e., *semi-supervised learning*) (Section 2.4).

This structure is chosen in order to allow clear and separated lines of argumentation which eventually conclude in the *second main result* of the paper.

In the results section, we demonstrate the applicability of this framework by simulation studies with Gaussian and uniform mixture models as observational error densities for a line fit (Section 3.1) and a fit of anisotropic noisy data with a cubic polynomial (Section 3.2). Further, in Section 4, we will demonstrate how this argumentation can be applied to multiple linear regression for *life*

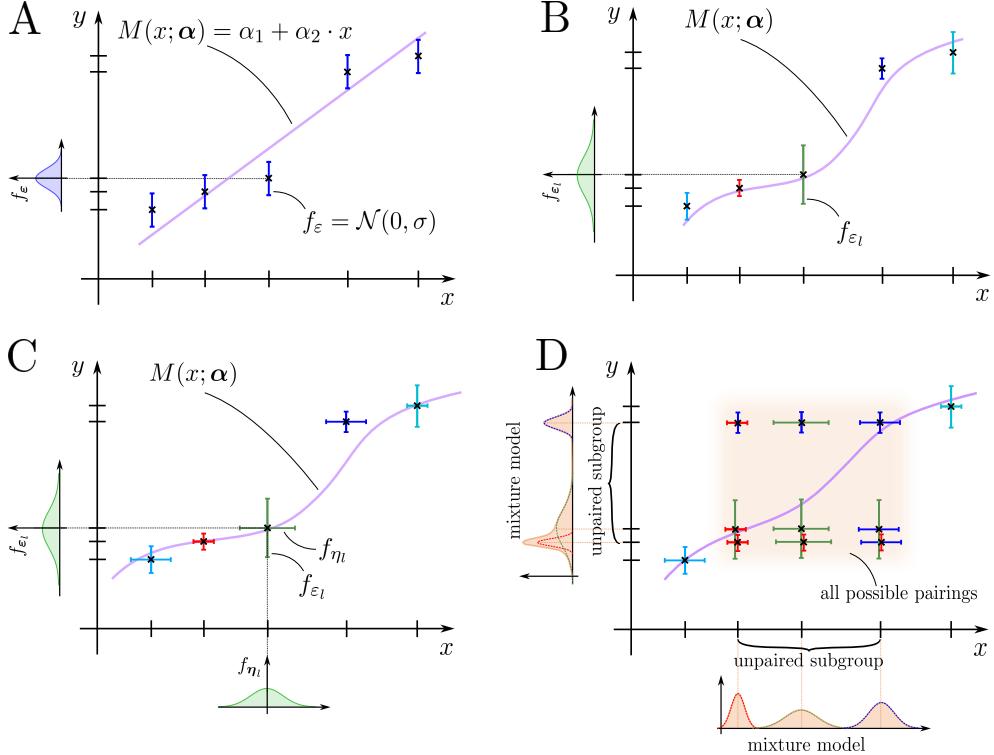


Figure 2: Schematic illustration of the model fitting concepts in this paper for an input  $x \in \mathbb{R}$  and output  $y \in \mathbb{R}$ , presented as steps of generalizations. A) Well-known *ordinary least squares for a line fit* model, which corresponds to *Maximum Likelihood* estimation with a constant Gaussian observational error density  $f_\epsilon$  in  $y$ . B) Generalization of case (A) to a *general nonlinear model*  $M(x; \alpha)$  and *general observational error density*  $f_{\epsilon_l}$  in  $y$  individual for each data point  $l = 1, \dots, L$ . This is presented in *Supplement A*. C) Generalization of (B) to the *errors-in-variables* approach including observational error density functions  $f_{\eta_l}$  in  $x$ . This is presented in *Section 2.2*. D) Generalization of (C) to also include partially unpaired data subgroups, which contains all possible pairings of data within a subgroup (in this case  $3 \cdot 3 = 9$  pairings). This is realized with mixture models, which are the weighted sum of the data point probability densities. This is presented in *Section 2.4*.

*expectancy* data from the world bank. These results demonstrate the importance of modeling the inherent uncertainties and the use of different levels of pairing information in data.

## 2 Methods

### 2.1 General Nomenclature

Throughout the paper, we utilize the following nomenclature:

- Observations are presented by input data  $\mathbf{x}_l \in \mathbb{R}^k$  and output data  $\mathbf{y}_l \in \mathbb{R}^m$  for  $l = 1, \dots, L$  as independent observations.
- We call the data set *completely paired* if for every  $l = 1, \dots, L$  there is a unique correspondence between  $\mathbf{x}_l$  and  $\mathbf{y}_l$ , typically written as tuples  $(\mathbf{x}_l, \mathbf{y}_l)$ .

We call the data *completely unpaired* if there is no pairing at all, i.e. there is a set of  $\mathbf{x}_h$  for  $h = 1, \dots, H$  and independently a set of  $\mathbf{y}_l$  for  $l = 1, \dots, L$  and there is no information which  $\mathbf{x}_h$  corresponds to which  $\mathbf{y}_l$ .

In consequence, *partially unpaired* data are a mix of both extremes, i.e. we have  $R$  subgroups of the data  $\mathbf{x}_h$  and  $\mathbf{y}_l$  and inside each subgroup there is no pairing information of the data (no correspondences) but it is guaranteed that no  $\mathbf{x}_h$ , or  $\mathbf{y}_l$  respectively, of one subgroup

corresponds to a  $\mathbf{x}_h$ , or  $\mathbf{y}_l$  respectively, of another subgroup. This means, we have pairing information only on the level of subgroups. These data configurations are illustrated in Figure 1.

- Probability density functions are denoted by  $\mathbf{X} \sim f_{\mathbf{X}}(x)$ , which are Lebesgue integrable  $f_{\mathbf{X}}(x) \in L^1$  and contain standard cases such as the normal or uniform distribution. Dirac distributions are used for theoretical discussions to show the connection between undisturbed and disturbed observations.
- We denote explicit models to be fitted from  $\mathbf{x}$  to  $\mathbf{y}$  by functions  $\mathbf{M}(\cdot; \boldsymbol{\alpha}) : \mathbb{R}^k \mapsto \mathbb{R}^m$  depending on the model parameters  $\boldsymbol{\alpha} \in \mathbb{R}^N$ . A toy example for an explicit model is the one-dimensional affine model of linear regression  $\mathbb{R} \mapsto \mathbb{R}$  ( $k = 1, m = 1, N = 2$ ):  $M(x; \boldsymbol{\alpha}) = \alpha_1 + \alpha_2 \cdot x$ .
- Following a flexible Bayesian view on random variables is essential in this work. In the classical perspective we have an undisturbed variable  $\mathbf{y}^*$  (the true value), which is disturbed by an error random variable  $\boldsymbol{\varepsilon}$  leading to the observation  $\mathbf{y}$ , in short:  $\mathbf{y} := \mathbf{y}^* + \boldsymbol{\varepsilon}$ . We interpret  $\mathbf{y}$  as a new random variable of observations which has a shifted density function of  $\boldsymbol{\varepsilon}$  (we interpret  $\mathbf{y}^*$  as a random variable with Dirac distribution at the true value). In this work, we consequently take an alternative point of view and introduce the definition by the reformulation:  $\mathbf{y}^* := \mathbf{y} - \boldsymbol{\varepsilon}$ . This time we interpret  $\mathbf{y}^*$  as a new random variable of the true values (as typical in Bayesian frameworks) which has a shifted density function of  $\boldsymbol{\varepsilon}$  (this time, we interpret  $\mathbf{y}$  as the random variable with Dirac distribution at the observed value). The meaning of  $\boldsymbol{\varepsilon}$  in these two perspectives is different but related, capturing the uncertainty of observation with a different center. We consequently utilize the latter notation in the rest of the paper.
- In the stochastic argumentation, we utilize the notation

$$f_{\bigcap_{l=1}^L \mathbf{Z}_l}(\mathbf{z}) := f_{\mathbf{Z}_1, \dots, \mathbf{Z}_L}(\mathbf{z})$$

with the meaning of the *common density function* of all individual random variables  $\mathbf{Z}_1, \dots, \mathbf{Z}_L$ .

- In the following derivations, we focus on the *argmax/argmin* of an expression. Since the *argmax/argmin* is independent of the application of strictly monotonic increasing functions, we neglect those in the course of argumentation, i.e. for  $c \in \mathbb{R}^+$  we will write

$$\operatorname{argmax}_{\boldsymbol{\alpha}} c \cdot f(\boldsymbol{\alpha}) = \operatorname{argmax}_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}) = \operatorname{argmax}_{\boldsymbol{\alpha}} \ln(f(\boldsymbol{\alpha})).$$

## 2.2 Model Fit with Completely Paired Data

The observations in this section are of type  $(\mathbf{x}_l, \mathbf{y}_l)$  as tuples for  $l = 1, \dots, L$ . Standard approaches, such as ordinary least squares (errors in  $\mathbf{y}$  only) in this argumentation framework are presented in *Supplement A*.

We introduce disturbances in the input and output data with the notation (which is referred in literature to *errors-in-variables* [Markovsky and Van Huffel, 2007]):

$$\mathbf{x}_l^* := \mathbf{x}_l - \boldsymbol{\eta}_l \quad (1)$$

$$\mathbf{y}_l^* := \mathbf{y}_l - \boldsymbol{\varepsilon}_l \quad (2)$$

with  $\mathbf{x}_l^* \in \mathbb{R}^k$  and  $\mathbf{y}_l^* \in \mathbb{R}^m$  the random variables of true values, and the uncertainty random variables  $\boldsymbol{\eta}_l \sim f_{\boldsymbol{\eta}_l}(\mathbf{s}) : \mathbb{R}^k \mapsto \mathbb{R}$  and  $\boldsymbol{\varepsilon}_l \sim f_{\boldsymbol{\varepsilon}_l}(\mathbf{s}) : \mathbb{R}^m \mapsto \mathbb{R}$  independent for all  $l = 1, \dots, L$ . We are interested in the case where the model is correctly chosen so that the true  $\mathbf{x}_l^*$  predicts the true  $\mathbf{y}_l^*$ :

$$\begin{aligned} \mathbf{M}(\mathbf{x}_l^*; \boldsymbol{\alpha}) &\stackrel{d}{=} \mathbf{y}_l^* \quad \forall l = 1, \dots, L \\ \mathbf{M}(\mathbf{x}_l - \boldsymbol{\eta}_l; \boldsymbol{\alpha}) &\stackrel{d}{=} \mathbf{y}_l - \boldsymbol{\varepsilon}_l \quad \forall l = 1, \dots, L, \end{aligned}$$

with  $\stackrel{d}{=}$  meaning equality *in distribution*. Due to  $\boldsymbol{\eta}_l$  and  $\boldsymbol{\varepsilon}_l$  being random variables, the differences of right and left side  $\mathbf{M}(\mathbf{x}_l - \boldsymbol{\eta}_l; \boldsymbol{\alpha}) - \mathbf{y}_l + \boldsymbol{\varepsilon}_l$  are interpreted as difference random variables for

all  $l = 1, \dots, L$  whose density function values should have the highest possible value at  $\mathbf{0} \in \mathbb{R}^m$  to achieve the most probable equality leading to the Maximum Likelihood approach:

$$\begin{aligned}
&\Rightarrow \operatorname{argmax}_{\boldsymbol{\alpha}} f_{\bigcap_{l=1}^L [M(\mathbf{x}_l - \boldsymbol{\eta}_l; \boldsymbol{\alpha}) - \mathbf{y}_l + \boldsymbol{\varepsilon}_l]}(\mathbf{0}) \\
&= \operatorname{argmax}_{\boldsymbol{\alpha}} \prod_{l=1}^L f_{M(\mathbf{x}_l - \boldsymbol{\eta}_l; \boldsymbol{\alpha}) - \mathbf{y}_l + \boldsymbol{\varepsilon}_l}(\mathbf{0}) \quad (\text{independency of } \boldsymbol{\eta}_l, \boldsymbol{\varepsilon}_l \forall l) \\
&= \operatorname{argmax}_{\boldsymbol{\alpha}} \prod_{l=1}^L \int_{\mathbb{R}^k} f_{M(\mathbf{x}_l - \mathbf{s}; \boldsymbol{\alpha}) - \mathbf{y}_l + \boldsymbol{\varepsilon}_l}(\mathbf{0}) \cdot f_{\boldsymbol{\eta}_l}(\mathbf{s}) d\mathbf{s} \quad (\text{law of total probability}) \\
&= \operatorname{argmax}_{\boldsymbol{\alpha}} \prod_{l=1}^L \int_{\mathbb{R}^k} f_{\boldsymbol{\varepsilon}_l}(\mathbf{y}_l - M(\mathbf{x}_l - \mathbf{s}; \boldsymbol{\alpha})) \cdot f_{\boldsymbol{\eta}_l}(\mathbf{s}) d\mathbf{s} \quad (\text{shifted } \boldsymbol{\varepsilon}_l) \tag{3}
\end{aligned}$$

$$= \operatorname{argmax}_{\boldsymbol{\alpha}} \prod_{l=1}^L \int_{\mathbb{R}^k} f_{\boldsymbol{\varepsilon}_l}(\mathbf{y}_l - M(\mathbf{s}; \boldsymbol{\alpha})) \cdot f_{\boldsymbol{\eta}_l}(\mathbf{x}_l - \mathbf{s}) d\mathbf{s} \quad (\text{integral shift}) \tag{4}$$

Applying the *law of total probability* allows recovering the observation density functions in the final ML expression. In the following, we present examples of this general formula (4) (or equivalently Equation (3) if beneficial).

*Remark:* By setting  $f_{\boldsymbol{\eta}_l}(\mathbf{s}) = \delta(\mathbf{s})$  (the Dirac distribution), we allow no variation of the  $\mathbf{x}_l^*$ -values and, in consequence, get the equation of ordinary least squares (*Supplement A*) by applying the sifting property. In consequence, this can be regarded as a true generalization of errors in  $\mathbf{y}$  only.

### 2.2.1 Example: Fitting a Line and Gaussian Disturbance

In the line of total least squares [Markovsky and Van Huffel, 2007], we are introducing Gaussian disturbances by  $\boldsymbol{\eta}_l \sim \mathcal{N}(0, \sigma_\eta^2)(s) \forall l = 1, \dots, L$  and  $\boldsymbol{\varepsilon}_l \sim \mathcal{N}(0, \sigma_\varepsilon^2)(s) \forall l = 1, \dots, L$ . Utilizing the one-dimensional affine model  $M(x; \boldsymbol{\alpha}) = \alpha_1 + \alpha_2 \cdot x$  and inserting it in Equation (4), we get

$$\Rightarrow \operatorname{argmax}_{\boldsymbol{\alpha}} \prod_{l=1}^L \int_{\mathbb{R}} e^{-\frac{1}{2\sigma_\varepsilon^2}(y_l - \alpha_1 - \alpha_2 \cdot s)^2 - \frac{1}{2\sigma_\eta^2}(x_l - s)^2} ds \tag{5}$$

$$\begin{aligned}
&= \operatorname{argmax}_{\boldsymbol{\alpha}} \prod_{l=1}^L \frac{e^{-\frac{(\alpha_1 + \alpha_2 \cdot x_l - y_l)^2}{2(\alpha_2^2 \sigma_\eta^2 + \sigma_\varepsilon^2)}}}{\sqrt{\alpha_2^2 \sigma_\eta^2 + \sigma_\varepsilon^2}} \\
&= \operatorname{argmin}_{\boldsymbol{\alpha}} \frac{L}{2} \ln(\alpha_2^2 \sigma_\eta^2 + \sigma_\varepsilon^2) + \sum_{l=1}^L \frac{(\alpha_1 + \alpha_2 \cdot x_l - y_l)^2}{2(\alpha_2^2 \sigma_\eta^2 + \sigma_\varepsilon^2)}. \tag{6}
\end{aligned}$$

Since this is a Deming regression type of problem, the solution to this minimization is also closely related to the classical Deming regression. The similarities and differences are presented in *Supplement B*.

### 2.2.2 Example: Fitting a Hyperplane and Gaussian Disturbance (Errors-In-Variables Multiple Linear Regression)

Further extending this argumentation to hyperplanes for  $\mathbf{x}_l \in \mathbb{R}^k$  and  $y_l \in \mathbb{R}$ , and  $\boldsymbol{\eta}_l \sim \mathcal{N}(\mathbf{0}, \operatorname{diag}(\sigma_{\eta,1}^2, \dots, \sigma_{\eta,k}^2))(\mathbf{s}) \forall l = 1, \dots, L$  and  $\boldsymbol{\varepsilon}_l \sim \mathcal{N}(0, \sigma_\varepsilon^2)(s) \forall l = 1, \dots, L$  for fitting an affine hyperplane model  $M(\mathbf{x}; \boldsymbol{\alpha}) =$

$\alpha_1 + \sum_{n=1}^k \alpha_{n+1} \cdot x_n$  leads (utilizing Equation (4)) to the general optimization problem

$$\begin{aligned} & \Rightarrow \operatorname{argmax}_{\boldsymbol{\alpha}} \prod_{l=1}^L \int_{\mathbb{R}^k} e^{-\frac{1}{2 \sigma_{\varepsilon}^2}\left(y_l-\alpha_1-\sum_{n=1}^k \alpha_{n+1} s_n\right)^2-\frac{1}{2}\left(\sum_{n=1}^k \frac{\left(x_{l,n}-s_n\right)^2}{\sigma_{\eta,n}^2}\right)} d \mathbf{s} \\ & =\operatorname{argmax}_{\boldsymbol{\alpha}} \prod_{l=1}^L \frac{e^{-\frac{\left(\alpha_1+\sum_{n=1}^k \alpha_{n+1} \cdot x_{l,n}-y_l\right)^2}{2\left(\left(\sum_{n=1}^k \alpha_{n+1}^2 \sigma_{\eta,n}^2\right)+\sigma_{\varepsilon}^2\right)}}}{\sqrt{\left(\sum_{n=1}^k \alpha_{n+1}^2 \sigma_{\eta,n}^2\right)+\sigma_{\varepsilon}^2}} \\ & =\operatorname{argmin}_{\boldsymbol{\alpha}} \frac{L}{2} \ln \left(\left(\sum_{n=1}^k \alpha_{n+1}^2 \sigma_{\eta,n}^2\right)+\sigma_{\varepsilon}^2\right)+\sum_{l=1}^L \frac{\left(\alpha_1+\sum_{n=1}^k \alpha_{n+1} \cdot x_{l,n}-y_l\right)^2}{2\left(\left(\sum_{n=1}^k \alpha_{n+1}^2 \sigma_{\eta,n}^2\right)+\sigma_{\varepsilon}^2\right)} . \end{aligned}$$

### 2.2.3 Connection to Interval Data Regression

Interval data is defined as data for which only the borders of an interval in which the true data point lies are observed. Performing model fitting for such data is an active field of research, e.g. for multilinear linear regression models [Lima Neto and De Carvalho, 2008, Souza et al., 2017]. First, interval data for regression is defined the following way: For each data point coordinate  $x_{l,i}$  ( $i = 1, \dots, k$ ) and  $y_{l,j}$  ( $j = 1, \dots, m$ ), respectively, we only know the interval borders, i.e.  $x_{l,i} \in [\underline{x}_{l,i}, \overline{x}_{l,i}]$  and  $y_{l,i} \in [\underline{y}_{l,i}, \overline{y}_{l,i}]$ , which are independently measured. This information can be interpreted as a uniform distribution with probability mass inside the interval and zero outside. By introducing  $\underline{\mathbf{x}}_l := \frac{1}{2} (\underline{\mathbf{x}}_l + \overline{\mathbf{x}}_l)$  and  $\underline{\mathbf{y}}_l := \frac{1}{2} (\underline{\mathbf{y}}_l + \overline{\mathbf{y}}_l)$ , and  $v_{l,i} := \frac{1}{2} (\overline{x}_{l,i} - \underline{x}_{l,i})$  and  $w_{l,j} := \frac{1}{2} (\overline{y}_{l,j} - \underline{y}_{l,j})$ , this is equivalent to the general description of Equations (1) and (2) with

$$\begin{aligned} \boldsymbol{\eta}_l & \sim \prod_{i=1}^k U_{[-v_{l,i}, v_{l,i}]}(s_i) \\ \boldsymbol{\varepsilon}_l & \sim \prod_{j=1}^m U_{[-w_{l,j}, w_{l,j}]}(s_j) , \end{aligned}$$

with  $U_{[a,b]}(s)$  the density function of the uniform distribution on the interval  $[a, b]$ . This is obvious, since we can define an interval by either the two interval borders or the midpoint and its half width.

Second, this means we can apply Equation (3) for fitting a model  $\mathbf{M}$  into that interval data, leading to

$$\begin{aligned} & \Rightarrow \operatorname{argmax}_{\boldsymbol{\alpha}} \prod_{l=1}^L \int_{\mathbb{R}^k} \prod_{j=1}^m U_{[-w_{l,j}, w_{l,j}]} \left( \frac{1}{2} \left( \underline{y}_{l,j} + \overline{y}_{l,j} \right) - M_j \left( \frac{1}{2} (\underline{\mathbf{x}}_l + \overline{\mathbf{x}}_l) - \mathbf{s}; \boldsymbol{\alpha} \right) \right) \\ & \quad \cdot \prod_{i=1}^k U_{[-v_{l,i}, v_{l,i}]}(s_i) d \mathbf{s} . \end{aligned}$$

This can be further simplified to the  $\operatorname{argmax}_{\boldsymbol{\alpha}}$  of

$$\prod_{l=1}^L \frac{1}{\prod_{i=1}^k v_{l,i}} \cdot \int_{\substack{[-v_{l,1}, v_{l,1}] \times \\ \cdots \times [-v_{l,k}, v_{l,k}]}} \prod_{j=1}^m U_{[-w_{l,j}, w_{l,j}]} \left( \frac{1}{2} \left( \underline{y}_{l,j} + \overline{y}_{l,j} \right) - M_j \left( \frac{1}{2} (\underline{\mathbf{x}}_l + \overline{\mathbf{x}}_l) - \mathbf{s}; \boldsymbol{\alpha} \right) \right) d \mathbf{s} .$$

For each  $\mathbf{s}$  the integrand is either zero or the positive normalization constant of the density of  $\boldsymbol{\varepsilon}_l$ , leading to a  $k$ -dimensional constant region for the integrand, whose volume is integrated over

the  $k$ -dimensional box  $[-v_{l,1}, v_{l,1}] \times \cdots \times [-v_{l,k}, v_{l,k}]$ . This means, the resulting optimization is searching for  $\boldsymbol{\alpha}$  which maximizes the overlapping volume of the  $k$ -dimensional region with the  $k$ -dimensional box for all data points  $l = 1, \dots, L$  under consideration of the weights  $v_{l,i}$  and  $w_{l,i}$ . This is an intuitive general understanding of model fitting with interval data.

#### 2.2.4 Example: Linear Interval Data Regression

Utilizing the one-dimensional affine model  $M(x; \boldsymbol{\alpha}) = \alpha_1 + \alpha_2 \cdot x$ , we can further derive

$$\begin{aligned} &\Rightarrow \operatorname{argmax}_{\boldsymbol{\alpha}} \prod_{l=1}^L \frac{1}{v_l} \cdot \int_{[-v_l, v_l]} U_{[-w_l, w_l]} \left( \frac{1}{2} (\underline{y}_l + \bar{y}_l) - \left( \alpha_1 + \alpha_2 \left( \frac{1}{2} (\underline{x}_l + \bar{x}_l) - s \right) \right) \right) ds \\ &\stackrel{\alpha_2 \neq 0}{\Rightarrow} \operatorname{argmax}_{\boldsymbol{\alpha}} \prod_{l=1}^L \frac{1}{v_l} \cdot \int_{[-v_l, v_l]} \frac{1}{2w_l} \chi_{[c_{l,\min}(\boldsymbol{\alpha}), c_{l,\max}(\boldsymbol{\alpha})]}(s) ds \end{aligned}$$

with the abbreviations  $c_{l,\pm}(\boldsymbol{\alpha}) := \frac{1}{2} (\underline{x}_l + \bar{x}_l) + \frac{1}{\alpha_2} (\alpha_1 - \frac{1}{2} (\underline{y}_l + \bar{y}_l) \pm w_l)$ ,  $c_{l,\min} = \min(c_{l,\pm})$ ,  $c_{l,\max} = \max(c_{l,\pm})$  and  $\chi_{[a,b]}(s)$  the *characteristic function* (1 if  $s \in [a, b]$ , else 0). This can further be simplified to

$$\Rightarrow \operatorname{argmax}_{\boldsymbol{\alpha}} \prod_{l=1}^L \frac{1}{v_l \cdot w_l} \cdot \max [\min [v_l, c_{l,\max}(\boldsymbol{\alpha})] - \max [-v_l, c_{l,\min}(\boldsymbol{\alpha})], 0].$$

For an example fit according to this formula, see the results Section 3.1

*Remark:* With this framework of argumentation, we can also introduce uncertainty about the knowledge of the integral borders in a transparent way by not assuming a strict uniform distribution, but a distribution blurred at the borders.

### 2.3 Model Fit with Completely Unpaired Data

In this section, we are losing the property of tuples, i.e. observations of type  $\mathbf{x}_h$  for  $h = 1, \dots, H$  and  $\mathbf{y}_l$  for  $l = 1, \dots, L$  are unpaired. Only a set of  $\mathbf{x}_h$  and a set of  $\mathbf{y}_l$  observations are available. Please note, even the sizes  $H$  and  $L$  can be different. Although there is theoretical research about the usability of such *broken sampling* data sets, e.g. [Bai and Hsing, 2005], obviously such data will only lead to very limited regression results if there are no further assumptions about the involved probability densities since we only have marginal distributions. Nonetheless, we want to introduce a formulation by mixture models for this case, which will later be utilized for partially unpaired data directly. We start this argumentation with possible disturbances in input and output data:

$$\begin{aligned} \mathbf{x}_h^* &:= \mathbf{x}_h - \boldsymbol{\eta}_h \\ \mathbf{y}_l^* &:= \mathbf{y}_l - \boldsymbol{\varepsilon}_l \end{aligned}$$

with  $\mathbf{x}_h^* \in \mathbb{R}^k$  and  $\mathbf{y}_l^* \in \mathbb{R}^m$  the random variables of the true values, and the uncertainty random variables  $\boldsymbol{\eta}_h \sim f_{\boldsymbol{\eta}_h}(\mathbf{s}) : \mathbb{R}^k \mapsto \mathbb{R}$  and  $\boldsymbol{\varepsilon}_l \sim f_{\boldsymbol{\varepsilon}_l}(\mathbf{s}) : \mathbb{R}^m \mapsto \mathbb{R}$  independent for all  $h = 1, \dots, H$  and  $l = 1, \dots, L$ . A new step is now to introduce the two mixture model random variables

$$\begin{aligned} \mathbf{X}^* &\sim f_{\mathbf{X}^*}(\mathbf{s}) = \frac{1}{H} \sum_{h=1}^H f_{\mathbf{x}_h - \boldsymbol{\eta}_h}(\mathbf{s}) = \frac{1}{H} \sum_{h=1}^H f_{\boldsymbol{\eta}_h}(\mathbf{x}_h - \mathbf{s}) \\ \mathbf{Y}^* &\sim f_{\mathbf{Y}^*}(\mathbf{s}) = \frac{1}{L} \sum_{l=1}^L f_{\mathbf{y}_l - \boldsymbol{\varepsilon}_l}(\mathbf{s}) = \frac{1}{L} \sum_{l=1}^L f_{\boldsymbol{\varepsilon}_l}(\mathbf{y}_l - \mathbf{s}), \end{aligned}$$

which exactly contain the ignorance of the pairing, i.e. all  $\mathbf{x}_h$  and  $\mathbf{y}_l$  observations are present in these mixture models at once. We follow the same technical argumentation as in the previous section:

$$M(\mathbf{X}^*; \boldsymbol{\alpha}) \stackrel{d}{=} \mathbf{Y}^*.$$

Due to  $\mathbf{X}^*$  and  $\mathbf{Y}^*$  being random variables, the difference of right and left side  $\mathbf{M}(\mathbf{X}^*; \boldsymbol{\alpha}) - \mathbf{Y}^*$  is again interpreted as a difference random variable, whose density function value should have highest value at  $\mathbf{0} \in \mathbb{R}^m$ . This leads to the ML approach:

$$\begin{aligned}
&\Rightarrow \operatorname{argmax}_{\boldsymbol{\alpha}} f_{\mathbf{M}(\mathbf{X}^*; \boldsymbol{\alpha}) - \mathbf{Y}^*}(\mathbf{0}) \\
&= \operatorname{argmax}_{\boldsymbol{\alpha}} \int_{\mathbb{R}^k} f_{\mathbf{M}(\mathbf{s}; \boldsymbol{\alpha}) - \mathbf{Y}^*}(\mathbf{0}) \cdot f_{\mathbf{X}^*}(\mathbf{s}) d\mathbf{s} \quad (\text{law of total probability}) \\
&= \operatorname{argmax}_{\boldsymbol{\alpha}} \int_{\mathbb{R}^k} f_{\mathbf{Y}^*}(\mathbf{M}(\mathbf{s}; \boldsymbol{\alpha})) \cdot f_{\mathbf{X}^*}(\mathbf{s}) d\mathbf{s} \quad (\text{shifted } \mathbf{Y}^*) \\
&= \operatorname{argmax}_{\boldsymbol{\alpha}} \int_{\mathbb{R}^k} \left( \frac{1}{L} \sum_{l=1}^L f_{\boldsymbol{\varepsilon}_l}(\mathbf{y}_l - \mathbf{M}(\mathbf{s}; \boldsymbol{\alpha})) \right) \cdot \left( \frac{1}{H} \sum_{h=1}^H f_{\boldsymbol{\eta}_h}(\mathbf{x}_h - \mathbf{s}) \right) d\mathbf{s} \quad (\text{def. of } \mathbf{X}^*, \mathbf{Y}^*) \\
&= \operatorname{argmax}_{\boldsymbol{\alpha}} \frac{1}{LH} \sum_{l=1}^L \sum_{h=1}^H \int_{\mathbb{R}^k} f_{\boldsymbol{\varepsilon}_l}(\mathbf{y}_l - \mathbf{M}(\mathbf{s}; \boldsymbol{\alpha})) \cdot f_{\boldsymbol{\eta}_h}(\mathbf{x}_h - \mathbf{s}) d\mathbf{s}. \tag{7}
\end{aligned}$$

*Remark:* The double sum in Equation (8) takes care of all combinations of  $\mathbf{x}_h$  and  $\mathbf{y}_l$  coming directly from a strict stochastic derivation with these mixture model random variables. In the completely paired case with independent observations, a product appears in Equation (4) which corresponds to this double sum for the completely unpaired case.

*Remark:* This fit with completely unpaired data is practically useless. This means, there will be broad, probably non-distinct or multiple maxima in this objective function. Still, this argumentation helps in a theoretical perspective since it is applied directly to the partially unpaired data where we have a range of different levels of pairing information.

## 2.4 Model Fit with Partially Unpaired Data

This section introduces the argumentation for *partially unpaired data*, which is a main result of this paper. For this case, we partition the  $H$  observations  $\mathbf{x}_h$  and  $L$  observations  $\mathbf{y}_l$  into  $r = 1, \dots, R$  disjoint independent groups, i.e. observations of group  $r$  of type  $\mathbf{x}_{h,r}$  for  $h = 1, \dots, H_r$  and  $\mathbf{y}_{l,r}$  for  $l = 1, \dots, L_r$  are unpaired with  $H_r$  and  $L_r$  representing the number of elements in subgroup  $r$ . This means, we have a set of  $\mathbf{x}_h$ -values and a set of  $\mathbf{y}_l$ -values for each subgroup and we have pairing information on the group level. The number of input and output elements in each subgroup  $H_r$  and  $L_r$  are not necessarily the same. Again, we allow disturbances in  $\mathbf{x}_h$  and  $\mathbf{y}_l$ :

$$\begin{aligned}
\mathbf{x}_h^* &:= \mathbf{x}_h - \boldsymbol{\eta}_h \\
\mathbf{y}_l^* &:= \mathbf{y}_l - \boldsymbol{\varepsilon}_l
\end{aligned}$$

with  $\mathbf{x}_h^* \in \mathbb{R}^k$  and  $\mathbf{y}_l^* \in \mathbb{R}^m$  the random variables of the true values, and the independent uncertainty random variables  $\boldsymbol{\eta}_h \sim f_{\boldsymbol{\eta}_h}(\mathbf{s}) : \mathbb{R}^k \mapsto \mathbb{R}$  and  $\boldsymbol{\varepsilon}_l \sim f_{\boldsymbol{\varepsilon}_l}(\mathbf{s}) : \mathbb{R}^m \mapsto \mathbb{R}$  ( $l = 1, \dots, L$ ). The main argument for dealing with unpaired data is presented by mixture models, i.e. we define ( $r = 1, \dots, R$ ):

$$\begin{aligned}
\mathbf{X}_r^* &\sim f_{\mathbf{X}_r^*}(\mathbf{s}) = \frac{1}{H_r} \sum_{h=1}^{H_r} f_{\boldsymbol{\eta}_{h,r}}(\mathbf{x}_{h,r} - \mathbf{s}) \\
\mathbf{Y}_r^* &\sim f_{\mathbf{Y}_r^*}(\mathbf{s}) = \frac{1}{L_r} \sum_{l=1}^{L_r} f_{\boldsymbol{\varepsilon}_{l,r}}(\mathbf{y}_{l,r} - \mathbf{s}).
\end{aligned}$$

Following our standard line of argumentation, we get

$$\mathbf{M}(\mathbf{X}_r^*; \boldsymbol{\alpha}) \stackrel{d}{=} \mathbf{Y}_r^* \quad \forall r = 1, \dots, R, \tag{9}$$

and again focus on the difference random variables  $\mathbf{M}(\mathbf{X}_r^*; \boldsymbol{\alpha}) - \mathbf{Y}_r^*$  for all  $r = 1, \dots, R$  at  $\mathbf{0}$ . Following the ML approach, we arrive at

$$\begin{aligned} &\Rightarrow \operatorname{argmax}_{\boldsymbol{\alpha}} f_{\bigcap_{r=1}^R [\mathbf{M}(\mathbf{X}_r^*; \boldsymbol{\alpha}) - \mathbf{Y}_r^*]}(\mathbf{0}) \\ &= \operatorname{argmax}_{\boldsymbol{\alpha}} \prod_{r=1}^R f_{\mathbf{M}(\mathbf{X}_r^*; \boldsymbol{\alpha}) - \mathbf{Y}_r^*}(\mathbf{0}) \quad (\text{group independency}) \\ &= \operatorname{argmax}_{\boldsymbol{\alpha}} \prod_{r=1}^R \int_{\mathbb{R}^k} f_{\mathbf{Y}_r^*}(\mathbf{M}(\mathbf{s}; \boldsymbol{\alpha})) \cdot f_{\mathbf{X}_r^*}(\mathbf{s}) \, d\mathbf{s} \quad (\text{cp. equ. (7)}) \end{aligned} \quad (10)$$

$$= \operatorname{argmax}_{\boldsymbol{\alpha}} \prod_{r=1}^R \left[ \frac{1}{L_r H_r} \sum_{l=1}^{L_r} \sum_{h=1}^{H_r} \int_{\mathbb{R}^k} f_{\boldsymbol{\varepsilon}_{l,r}}(\mathbf{y}_{l,r} - \mathbf{M}(\mathbf{s}; \boldsymbol{\alpha})) \cdot f_{\boldsymbol{\eta}_{h,r}}(\mathbf{x}_{h,r} - \mathbf{s}) \, d\mathbf{s} \right] \quad (\text{cp. equ. (8)})$$

(11)

Equation (11) is the most general formula we derive in this paper, since it contains the previous cases (completely paired data  $R = H = L$  and completely unpaired data  $R = 1$ ). Most importantly, all other possibilities of partial pairing are contained in this equation. For example, ordinary least squares for paired data (a standard regression approach) is achieved by setting  $R = L = H$  (equals group size 1) and setting  $f_{\boldsymbol{\eta}_h}(\mathbf{s}) = \delta(\mathbf{s})$ .

*Remark:* In the partially unpaired setup, we always work with input / output correspondences, only on a subgroup basis. A completely paired data subset ( $=$  *supervised data*) is represented by subgroups of size one. An additional pure input data subset ( $=$  *unsupervised*) can be approximated by neglecting the information about  $\mathbf{Y}_r^*$ , which corresponds to extremely flat  $f_{\boldsymbol{\varepsilon}_{l,r}}$ , arriving at the classical definition of *semi-supervised*. This means, in Equation (11) the first term in the integral  $f_{\boldsymbol{\varepsilon}_{l,r}}$  gets essentially constant (independent of the prediction  $\mathbf{M}(\mathbf{s}; \boldsymbol{\alpha})$ ) and, therefore, this part becomes practically noninformative with respect to the optimization on  $\boldsymbol{\alpha}$ . Thus, as expected, the completely unsupervised part of the data only on the input side can be neglected since it contains no information about the model parameters  $\boldsymbol{\alpha}$ .

*Remark:* An interesting point is that the same cannot be said about having unsupervised data on the output side, e.g. neglecting information about  $\mathbf{X}_r^*$ . This time  $f_{\boldsymbol{\eta}_{h,r}}$  becomes a flat distribution, essentially leaving  $f_{\boldsymbol{\varepsilon}_{l,r}}(\mathbf{y}_{l,r} - \mathbf{M}(\mathbf{s}; \boldsymbol{\alpha}))$  in the integral of Equation (11) evaluated for all possible  $\mathbf{s}$ . This time the change of  $\boldsymbol{\alpha}$  can have direct influence on the optimization, essentially taking care that the output values  $\mathbf{y}_{l,r}$  are plausible / possible (i.e. in the probabilistically blurred image of  $\mathbf{M}(\mathbf{s}; \boldsymbol{\alpha})$ ) for a given parameter set  $\boldsymbol{\alpha}$ . This shows an asymmetry with respect to the classical semi-supervised setup [Liang et al., 2007].

#### 2.4.1 Example: Fitting a Line and Gaussian Disturbance

Gaussian disturbances in input and output variables with  $\eta_h \sim \mathcal{N}(0, \sigma_\eta^2)(s) \forall h = 1, \dots, H$  and  $\varepsilon_l \sim \mathcal{N}(0, \sigma_\varepsilon^2)(s) \forall l = 1, \dots, L$ , and utilizing the one-dimensional affine model  $M(x; \boldsymbol{\alpha}) = \alpha_1 + \alpha_2 \cdot x$  and inserting it, we get by applying Equation (11)

$$\Rightarrow \operatorname{argmax}_{\boldsymbol{\alpha}} \prod_{r=1}^R \left[ \frac{1}{L_r H_r} \sum_{l=1}^{L_r} \sum_{h=1}^{H_r} \frac{e^{-\frac{(\alpha_1 + \alpha_2 \cdot x_{h,r} - y_{l,r})^2}{2(\alpha_2^2 \sigma_\eta^2 + \sigma_\varepsilon^2)}}}{\sqrt{\alpha_2^2 \sigma_\eta^2 + \sigma_\varepsilon^2}} \right],$$

which represents a solution to the Deming type problem for partially unpaired data.

#### 2.4.2 Example: Fitting a Hyperplane and Gaussian Disturbance (Errors-In-Variables Multiple Linear Regression)

Gaussian disturbances in input and output variables with  $\eta_h \sim \mathcal{N}(\mathbf{0}, \operatorname{diag}(\sigma_{\eta,1}^2, \dots, \sigma_{\eta,k}^2))(\mathbf{s}) \forall h = 1, \dots, H$  and  $\varepsilon_l \sim \mathcal{N}(0, \sigma_\varepsilon^2)(s) \forall l = 1, \dots, L$ , and utilizing the  $k$ -dimensional affine model  $M(\mathbf{x}; \boldsymbol{\alpha}) = \alpha_1 + \sum_{n=1}^k \alpha_{n+1} \cdot x_n$  and inserting it, we get by applying

Equation (11)

$$\Rightarrow \operatorname{argmax}_{\alpha} \prod_{r=1}^R \left[ \frac{1}{L_r H_r} \sum_{l=1}^{L_r} \sum_{h=1}^{H_r} \frac{e^{-\frac{\left(\alpha_1 + \sum_{n=1}^k \alpha_{n+1} \cdot x_{h,r,n} - y_{l,r}\right)^2}{2\left(\left(\sum_{n=1}^k \alpha_{n+1}^2 \sigma_{\eta,n}^2\right) + \sigma_{\varepsilon}^2\right)}}}{\sqrt{\left(\sum_{n=1}^k \alpha_{n+1}^2 \sigma_{\eta,n}^2\right) + \sigma_{\varepsilon}^2}} \right],$$

representing errors-in-variables multiple linear regression for partially unpaired data.

#### 2.4.3 Example: Linear Interval Data Regression

As final example, we present interval data that are given with  $\underline{x}_h := \frac{1}{2}(\underline{x}_h + \bar{x}_h)$  and  $\underline{y}_l := \frac{1}{2}(\underline{y}_l + \bar{y}_l)$ , and  $v_{h,i} := \frac{1}{2}(\bar{x}_{h,i} - \underline{x}_{h,i})$  and  $w_{l,j} := \frac{1}{2}(\bar{y}_{l,j} - \underline{y}_{l,j})$ , and

$$\begin{aligned} \boldsymbol{\eta}_h &\sim \prod_{i=1}^k U_{[-v_{h,i}, v_{h,i}]}(s_i) \\ \boldsymbol{\varepsilon}_l &\sim \prod_{j=1}^m U_{[-w_{l,j}, w_{l,j}]}(s_j). \end{aligned}$$

The one-dimensional affine model  $M(x; \alpha) = \alpha_1 + \alpha_2 \cdot x$ , leads to the  $\operatorname{argmax}_{\alpha}$  of

$$\prod_{r=1}^R \left[ \frac{1}{L_r H_r} \sum_{l=1}^{L_r} \sum_{h=1}^{H_r} \frac{1}{v_{h,r} \cdot w_{l,r}} \cdot \max [ \min [v_{h,r}, c_{l,h,r,\max}(\alpha)] - \max [-v_{h,r}, c_{l,h,r,\min}(\alpha)], 0 ] \right],$$

with  $c_{l,h,r,\pm}(\alpha) := \frac{1}{2}(\underline{x}_{h,r} + \bar{x}_{h,r}) + \frac{1}{\alpha_2} \left( \alpha_1 - \frac{1}{2}(\underline{y}_{l,r} + \bar{y}_{l,r}) \pm w_{l,r} \right)$ ,  $c_{l,h,r,\min} = \min(c_{l,h,r,\pm})$  and  $c_{l,h,r,\max} = \max(c_{l,h,r,\pm})$ .

## 2.5 Extensions

### 2.5.1 Numerical Implementation of the General Formula for Partially Unpaired Data

We want to stress that the implementation of the general formula (11) is not recommended if avoidable, due to typically high computational costs. A typical way to avoid this, is to work with specific probability density types or model families, such as presented in the examples following Equation (11). For the general case, we provide the following implementation recommendations: First, a beneficial numerical implementation strategy is to avoid the (possibly massive) multiplication in the general Equation (11). In consequence, we rewrite this by applying the natural logarithm and multiplying it by  $-1$  in order to generate a practically useful minimization problem

$$\operatorname{argmin}_{\alpha} - \sum_{r=1}^R \ln \left( \frac{1}{L_r H_r} \left[ \sum_{l=1}^{L_r} \sum_{h=1}^{H_r} \int_{\mathbb{R}^k} f_{\boldsymbol{\varepsilon}_{l,r}}(\boldsymbol{y}_{l,r} - \mathbf{M}(\mathbf{s}; \alpha)) \cdot f_{\boldsymbol{\eta}_{h,r}}(\boldsymbol{x}_{h,r} - \mathbf{s}) d\mathbf{s} \right] \right).$$

Second, although the formulation of Equation (11) shows the combinatorics of the possible correspondences in unpaired data subsets, this is not an efficient way for implementation since it involves the approximation of  $L_r \cdot H_r$  integrals for each subgroup  $r$ . It is recommended to utilize Equation (10), by first evaluating the mixture models  $f_{\boldsymbol{Y}_r^*}$  and  $f_{\boldsymbol{X}_r^*}$  for appropriate  $\mathbf{s}$  for the numerical integration and then solving only one integral numerically for each subgroup, leading to the formula

$$\operatorname{argmin}_{\alpha} - \sum_{r=1}^R \ln \left( \int_{\mathbb{R}^k} f_{\boldsymbol{Y}_r^*}(\mathbf{M}(\mathbf{s}; \alpha)) \cdot f_{\boldsymbol{X}_r^*}(\mathbf{s}) d\mathbf{s} \right).$$

Third, numerical approximation of the integral is necessary. For high dimensional input data dimensions  $\boldsymbol{x}_h \in \mathbb{R}^k$  it is preferable to apply advanced Monte Carlo integration. For example, if

we utilize  $p = 1, \dots, P$  samples  $\mathbf{s}_{r,p}$  drawn from the mixture model density  $f_{\mathbf{X}_r^*}$ , then we can apply Monte Carlo integration with the formula

$$\operatorname{argmin}_{\boldsymbol{\alpha}} - \sum_{r=1}^R \ln \left( \frac{1}{P} \sum_{p=1}^P f_{\mathbf{Y}_r^*}(\mathbf{M}(\mathbf{s}_{r,p}; \boldsymbol{\alpha})) \right) , \quad (12)$$

which increases computational efficiency significantly.

Fourth, the choice of optimization algorithm depends strongly on the dimensionality of the parameters  $\boldsymbol{\alpha} \in \mathbb{R}^N$ . For low dimensions such as  $N < 10$  standard minimization routines such as Quasi-Newton optimization are recommended. For high dimension optimization problems stochastic gradient descent or simulated annealing are certainly preferable approaches. As starting values of these iterative optimization routines the ordinary least squares solutions can be utilized, if applicable.

Fifth, due to the choice of  $f_{\varepsilon_{l,r}}$  and  $f_{\eta_{h,r}}$  (in the best case representing the true data errors), the optimization problem can be more or less difficult. For example, selecting these densities with too small standard deviations, the objective function might contain a large number of non-distinct local extrema next to each other, which is difficult for local optimization algorithms. On the other side, selecting these densities with too large standard deviations may lead to very broad extrema, which can be helpful for the optimization algorithm but strongly reduces the information value of the observed data.

### 2.5.2 Evaluation of Unpaired Data Subgroups

Presenting the mathematical argumentation framework does not mean, that a practical model fitting problem at hand is well stated. Let us focus on the input set  $D = \{\mathbf{x}_h, h = 1, \dots, H\}$ . We are choosing  $R$  unpaired subgroups  $S_r$  ( $r = 1, \dots, R$ ) which in total represents a partition of these input values

$$\bigcup_{r=1}^R S_r = D \quad \wedge \quad S_r \cap S_m = \emptyset \quad \forall r, m \in \{1, \dots, R\}, r \neq m .$$

The question arises which partitioning is beneficial for the fit and which is not. At this point, we only want to discuss this problem by exploring the extremes:

- A) If all subgroups  $S_r$  (approximately) contain a representative sample of the whole data set  $D$ , then the model fitting is qualitatively the same as if we would use the completely unpaired case, which can be regarded as useless for a practical model fit, since no useful pairing information is contained in such a partitioning.
- B) If all subgroups  $S_r$  are presenting different, separated areas of the input data set, i.e. each subgroup is very dissimilar to  $D$ .

In consequence, one way to judge about the practical usefulness of the partitioning is to look for dissimilarity of each  $S_r$  to  $D$  and between subgroups  $S_r$ . The question arises: What is a good measure to determine the *dissimilarity* between  $S_r$  and  $D$  and between subgroups  $S_r$  for all  $r = 1, \dots, R$ ? Only for high dissimilarity, the pairing inside the subgroups will not degrade the model fitting result strongly. In *Supplement C* an illustrative example for this effect is presented. For designing data observation processes with deliberately partially unpaired data (maybe due to privacy protection, or observational costs etc.) it could be helpful to measure such dissimilarities directly and we regard this as future work.

### 2.5.3 Simultaneous Estimation of the Underlying Density Functions

In Equation (11) we assumed known and fixed input and output density functions  $f_{\eta_{h,r}}$  and  $f_{\varepsilon_{l,r}}$  for data groups  $r = 1, \dots, R$  with their element indices  $h = 1, \dots, H_r$  and  $l = 1, \dots, L_r$ . An extension of the proposed estimation concept is that the density functions are depending on *unknown parameters*, i.e.  $\beta_{h,r}$  for  $x$ - and  $\gamma_{l,r}$  for  $y$ -values, which we want to estimate simultaneously with the model

parameters  $\boldsymbol{\alpha}$ . We denote this by density functions  $f_{\boldsymbol{\eta}_{h,r}}(\cdot; \boldsymbol{\beta}_{h,r})$  and  $f_{\boldsymbol{\varepsilon}_{l,r}}(\cdot; \boldsymbol{\gamma}_{l,r})$ . This leads to the extended Maximum Likelihood problem:

$$\operatorname{argmax}_{(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})} \prod_{r=1}^R \left[ \frac{1}{L_r H_r} \sum_{l=1}^{L_r} \sum_{h=1}^{H_r} \int_{\mathbb{R}^k} f_{\boldsymbol{\varepsilon}_{l,r}}(\mathbf{y}_{l,r} - \mathbf{M}(\mathbf{s}; \boldsymbol{\alpha}); \boldsymbol{\gamma}_{l,r}) \cdot f_{\boldsymbol{\eta}_{h,r}}(\mathbf{x}_{h,r} - \mathbf{s}; \boldsymbol{\beta}_{h,r}) d\mathbf{s} \right].$$

*Remark:* One challenge of this extension is that the optimization gets high-dimensional with possibly many local maxima, which might occur due to the flexible interplay of densities with large standard deviations and the matching of the unpaired data groups. Further, too many parameters scaling with the data size may lead to problems of identifiability. In consequence, we assume that there will be the need to combine parameters and force the densities to smaller standard deviations, e.g. by penalizing parameters  $(\boldsymbol{\beta}, \boldsymbol{\gamma})$  which correspond to large standard deviations. There are many standard approaches for additive penalizing and we regard this out of the scope of the current presentation. A pragmatic approach for combining parameters could be to assume that for each input and output coordinate the error densities are known and identical with zero mean and unknown standard deviations, i.e.  $\boldsymbol{\beta} := \boldsymbol{\sigma}_\eta \in \mathbb{R}^k$  and  $\boldsymbol{\gamma} := \boldsymbol{\sigma}_\varepsilon \in \mathbb{R}^m$  (independent of  $h, l$  and  $r$ ). This allows a global estimation of the coordinate errors simultaneously to the model parameters and increases the degrees of freedom of the optimization problem only by  $k + m$ . For example, this corresponds for the *errors-in-variables multiple linear regression* in Section 2.4.2 to additionally maximize all  $\beta_n := \sigma_{\eta,n}$  ( $n = 1, \dots, k$ ) and  $\gamma := \sigma_\varepsilon$ , increasing the degrees of freedom from  $k + 1$  to  $2 \cdot (k + 1)$ .

#### 2.5.4 Bayesian Extension

In this derivation, we defined the likelihood function, which we need to maximize in the previous sections by

$$\mathcal{L}(\mathbf{0} | \boldsymbol{\alpha}) := \prod_{r=1}^R f_{\mathbf{M}(\mathbf{X}_r^*; \boldsymbol{\alpha}) - \mathbf{Y}_r^*}(\mathbf{0}).$$

The unusual perspective in this likelihood derivation is that our *observation* is  $\mathbf{0}$  since we must find the parameters  $\boldsymbol{\alpha}$  of the difference random variable  $\mathbf{M}(\mathbf{X}_r^*; \boldsymbol{\alpha}) - \mathbf{Y}_r^*$  to make  $\mathbf{0}$  most likely. This can be extended to a classical *Bayesian* perspective by introducing a prior for the random variable  $\boldsymbol{\alpha} \sim \pi(\boldsymbol{\alpha})$ . In consequence, the posterior density (utilizing Bayes' rule) gets

$$\begin{aligned} \pi(\boldsymbol{\alpha} | \mathbf{0}) &= c \cdot \mathcal{L}(\mathbf{0} | \boldsymbol{\alpha}) \cdot \pi(\boldsymbol{\alpha}) \\ &= c \cdot \left( \prod_{r=1}^R \int_{\mathbb{R}^k} f_{\mathbf{Y}_r^*}(\mathbf{M}(\mathbf{s}; \boldsymbol{\alpha})) \cdot f_{\mathbf{X}_r^*}(\mathbf{s}) d\mathbf{s} \right) \cdot \pi(\boldsymbol{\alpha}), \end{aligned}$$

with  $c$  a normalization constant. By utilizing a non- or weakly informative prior  $\pi(\boldsymbol{\alpha})$  (such as  $\pi(\boldsymbol{\alpha}) = \text{const.}$  on a large enough domain) the previously presented optimization problem is identical to the maximization of the posterior, leading to a *Maximum A Posteriori* (MAP) estimate. This allows to interpret the plotted objective functions in the results Section 3.1 as presentations of the density functions of  $\boldsymbol{\alpha}$ , *containing directly* the inherent estimation uncertainties about  $\boldsymbol{\alpha}$  graphically as intensity maps.

## 3 Simulation Study

The purpose of the results section is to illustrate the presented general fitting approach by examples to improve understanding of the derived formulas.

### 3.1 Demonstration for a Line Fit

At first, a simple line fit is illustrated. For each of the following scenarios we vary the number of subgroups  $R$ . The following scenarios are investigated:

- *Base scenario A*: The correct parameter values are  $\alpha_1 = 0$  and  $\alpha_2 = 0.5$ . We utilize  $L = H = 300$  data points and the Gaussian data point disturbances are drawn with  $\sigma_\eta = \sigma_\varepsilon = 0.2$  and expectation 0.
- *Scenario B*: same as A, but with increased Gaussian disturbances  $\sigma_\eta = \sigma_\varepsilon = 0.6$ .
- *Scenario C*: same as A, but with  $L = H = 36$ .
- *Scenario D*:  $L = H = 100$  data points in Figure 3 and  $L = H = 300$  in Figure 4. Interval data regression with uniform disturbances with standard deviations  $\sigma_\eta = \sigma_\varepsilon = 0.2$  and their corresponding interval boxes.

See Figure 3 for presentation of example objective function and resulting line fits of scenario A and D. The columns correspond to different numbers of unpaired subgroups  $R \in \{1, 3, 12, L\}$ . For

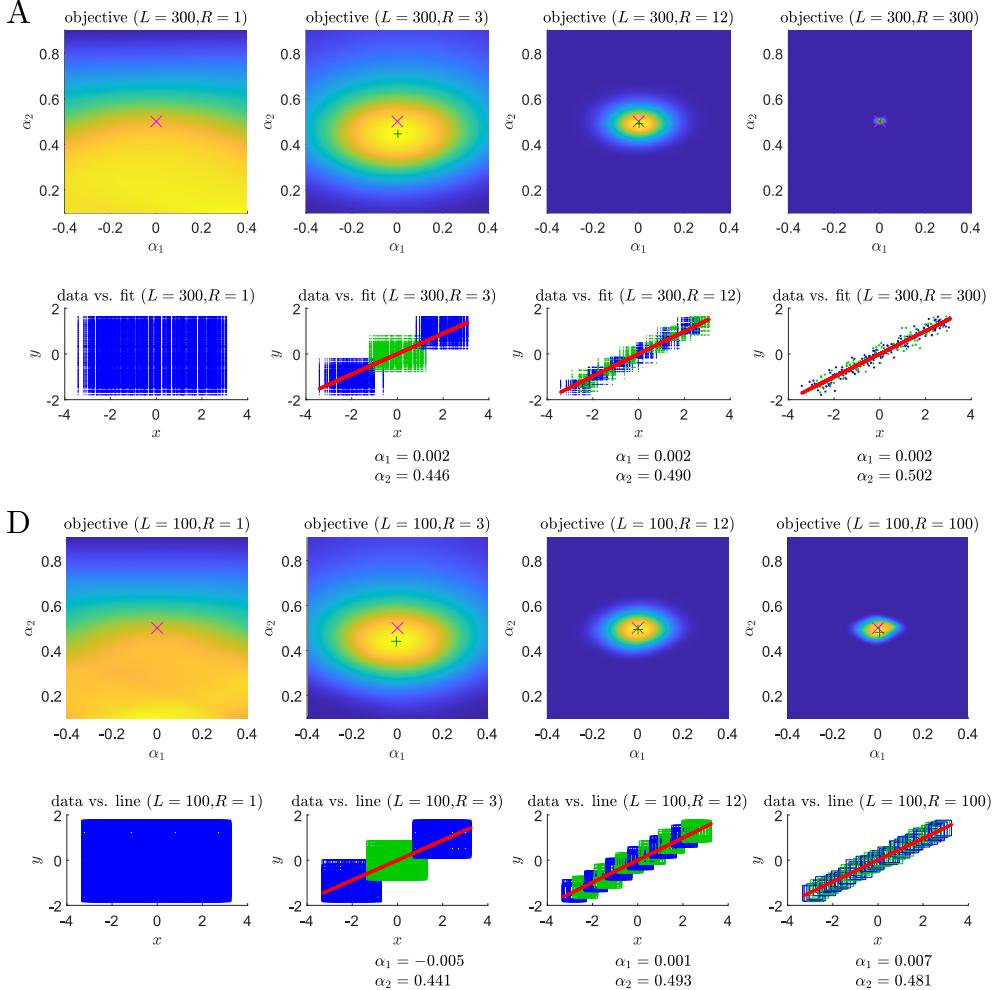


Figure 3: Illustration of example results for different line fit scenarios according to scenarios A (Gaussian error) and D (interval data). For each case: top row: objective functions with true parameters (red crosses) and maximum (green crosses), bottom row: data presentation and line fit results. Inside the color-coded unpaired subgroups (green and blue) all possible correspondences are plotted. Columns: Four different scenarios of partial pairing with  $R = 1$ ,  $R = 3$ ,  $R = 12$  and  $R = L = H$  groups.

the *completely unpaired* case ( $R = 1$ ) only the objective function is presented.

See Figure 4 for a systematic evaluation of the line fits for scenarios A to D utilizing 1000 simulated fits with random data errors. Presented are the box plots of the residual errors for  $\alpha_1$  and  $\alpha_2$ . Qualitative conclusions from Figures 3 and 4: First, utilizing completely unpaired data leads to arbitrary insufficient results which can be observed by the non-distinct maxima of the objective

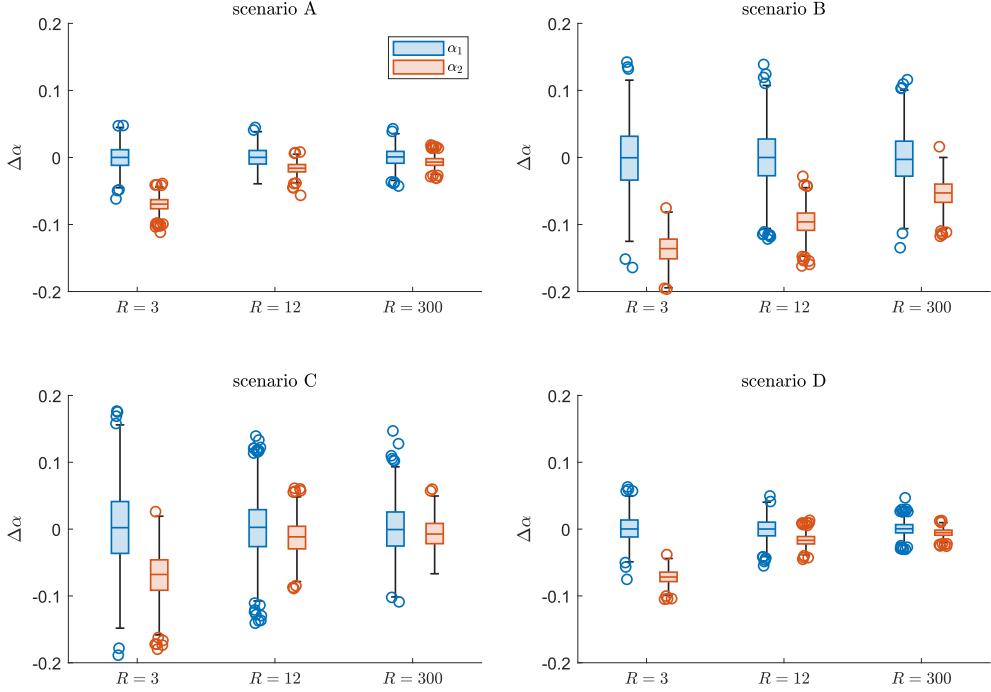


Figure 4: Evaluation of scenarios  $A$  to  $D$  (with  $L = 300$ ) for 1000 fits. Presented are the box plots of the residual errors  $\Delta\alpha = \alpha_{\text{fit}} - \alpha_{\text{truth}}$  of the intercept  $\alpha_1$  and slope  $\alpha_2$ . For each scenario different pairings are presented with  $R = 3$ ,  $R = 12$  and  $R = L = H = 300$ .

functions. Second, the fewer subgroups  $R$  are utilized, the broader (and more uncertain) gets the maximum in the objective function in Figure 3. Further, for very few subgroups, such as  $R = 3$ , a bias on the slope  $\alpha_2$  is introduced for all scenarios as presented in Figure 4. On the other side, it is obvious that reasonable estimation of the parameters is absolutely possible even if only partially unpaired data is available (comparing  $R = 12$  and  $R = 300$ ). Third, comparing scenarios  $A$  to  $B$ : The uncertainty of estimation increases with an increased noise level of the data for all cases of  $R$ . Fourth, comparing scenarios  $B$  to  $C$ : The higher noise level in the data leads to similar uncertainties in comparison to fewer data. Fifth, comparing scenarios  $A$  to  $D$ : The main difference utilizing interval data compared to Gaussian disturbances is that also for the completely paired case no distinct maximum appears but a *plateau* of high intensity values are observable in the objective function in Figure 3. In Figure 4 it can be seen that interval data leads to similar results with a stronger bias on the slope for  $R = 3$ .

In *Supplement C* a plane fit is presented as an example for 2D input variables.

### 3.2 Nonlinear Model with Anisotropic Observation Errors

In order to demonstrate the flexibility of this framework, the fitting of a nonlinear model  $\mathbb{R} \mapsto \mathbb{R}$  ( $k = 1, m = 1, N = 4$ )

$$M(x; \boldsymbol{\alpha}) = \alpha_1 + \alpha_2 \cdot x + \alpha_3 \cdot x^2 + \alpha_4 \cdot x^3,$$

is presented with anisotropic Gaussian disturbances  $\sigma_\eta = 0.2$  and  $\sigma_\varepsilon = 0.1$  for  $L = H = 300$  data points. In this model, nonlinearity holds with respect to  $x$  and not with respect to  $\boldsymbol{\alpha}$ , which is deliberate and not necessary. In Figure 5 the results are presented for a completely paired case ( $R = L = H$ ) (left) and a partially unpaired case (right) with two areas of lost pairing information. The general algorithm was implemented according to section 2.5.1 with the simple trapezoidal rule for integration and the *Nelder-Mead*-optimization in *Matlab*. For comparison of fitting results, Figure 5 shows the generating cubic function (black dashed line), the model fitting result according to Equation (11) (red line) and two simple comparison model fits (violet continuous and dotted lines). This comparison model fit is the ordinary least squares fit application of the cubic model (neglecting the noise in  $x$ -direction) with two different simple but intuitive

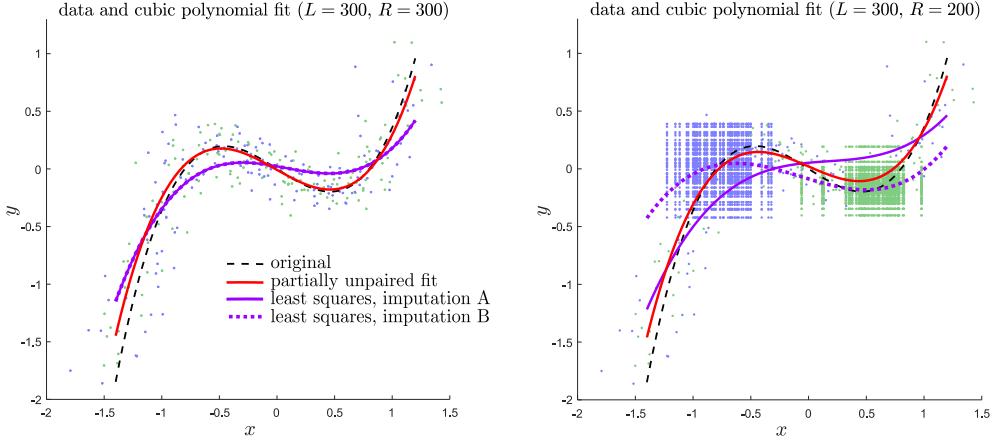


Figure 5: Illustration of example results for cubic model fitting scenarios on data with Gaussian disturbances. The generating cubic function (black dashed line), the partially unpaired data model fitting (red line) and two simple comparison model fits (violet continuous and dotted lines) are presented. Left: completely paired case ( $R = L = H$ ). Right: partially unpaired case with two unpaired areas with 51 lost data correspondences each ( $R = 200$ ). Inside the color-coded unpaired groups (green or blue dots) all possible correspondences are plotted as dots.

*imputation* treatments of the unpaired data: A) (= violet continuous line) Taking the average of the  $x$ - and  $y$ -values as a new artificial data point in these two areas and else neglect the unpaired data. B) (= violet dotted line) Including all possible combinations of the unpaired data directly in the least squares fit. Both comparison approaches are regarded as suboptimal, but intuitive data *imputations* for an unexperienced practitioner utilizing *ordinary least squares*, and therefore, presented for demonstration.

In Figure 5 (left) the benefit of including an error model in  $x$  additionally to  $y$  is presented (*ordinary least squares* does not contain an error model in  $x$ ), showing clearly superior results of the fit (red line) compared to the overlapping violet lines. In Figure 5 (right) the performance of the both simple comparison model fits (violet lines) decrease significantly compared to the case on the left while the model fit (red line) stays robust, dealing in a stable way with the lost pairing information.

## 4 Real Data Study: Life Expectancy

It will be demonstrated how this framework can be utilized for a *errors-in-variables multiple linear regression* problem with observational errors in  $\mathbf{x}$  and  $y$  on real data. This means, the model utilized is

$$M(\mathbf{x}; \boldsymbol{\alpha}) = \alpha_1 + \sum_{n=1}^k \alpha_{n+1} \cdot x_n .$$

The task will be to fit this model to life expectancy data for most countries in the world. The specific data set is taken from the *world bank databank*<sup>1</sup> and utilizes  $k = 4$  input variables  $x_1 = \text{Birth rate, crude (per 1,000 people)} [\text{SP.DYN.CBRT.IN}]$ ,  $x_2 = \text{Urban population (percent of total population)} [\text{SP.URB.TOTL.IN.ZS}]$ ,  $x_3 = \text{Political Stability and Absence of Violence/Terrorism: Estimate} [\text{PV.EST}]$ ,  $x_4 = \text{logarithm of Incidence of tuberculosis (per 100,000 people)} [\text{SH.TBS.INCD}]$  and the output variable  $y = \text{Life expectancy at birth, total (years)} [\text{SP.DYN.LE00.IN}]$ . The corresponding plot matrix is presented in Figure 6 for all 192 countries for which the variables were available.

As we want to demonstrate the *errors-in-variables* approach, we need to define error densities for each variable. Since in the world bank data there are no error margins provided, we assume normally distributed errors with mean zero and standard deviations of 15% of the standard deviation of the full data set for each variable  $x_1$  (1.49),  $x_2$  (3.52),  $x_3$  (0.14),  $x_4$  (0.24) and  $y$  (1.12).

<sup>1</sup>Data taken from <https://databank.worldbank.org/source/world-development-indicators> (June 2024), Database: World Development Indicators, data year 2020.

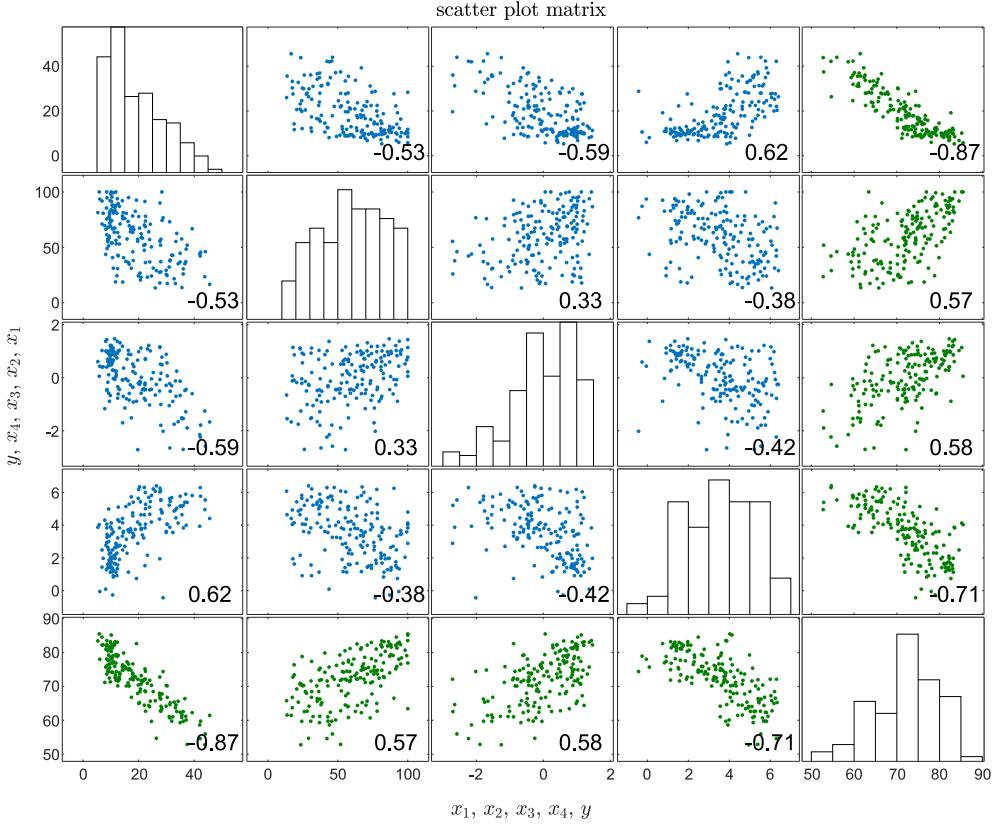


Figure 6: Pair plot of the input data  $x_1 = \text{Birth rate, crude (per 1,000 people)}$ ,  $x_2 = \text{Urban population (percent of total population)}$ ,  $x_3 = \text{Political Stability and Absence of Violence/Terrorism: Estimate}$ ,  $x_4 = \text{logarithm of Incidence of tuberculosis (per 100,000 people)}$  (blue plots) and the output variable  $y = \text{Life expectancy at birth, total (years)}$  (green plots). The Pearson correlation coefficient is presented in each correlation plot. On the diagonal are the histograms of each variable.

For evaluation purposes, we perform a train-test-split with 172 training countries and 20 test countries in order to judge if we can learn from the training countries the life expectancy for the test countries based on the input variables. A major challenge is that the goodness of fit cannot be measured with the classical  $R^2$  since it is only valid for cases with no errors in the input variables. In consequence, we utilize the extension of  $R^2$  to the *errors-in-variables* approach  $R_\delta^2$  [Cheng et al., 2014] for multiple linear regression (see *Supplement D* for more details).

First, we consider the case of the model fit with perfect pairing ( $R = L = H$ ). Since the proposed algorithm was implemented with Monte Carlo methods as described in Section 2.5.1, it was verified with an explicit solution for the multiple linear regression model [Cheng et al., 2014] and showed equivalent results. The results are presented in Figure 7A as correlation and error plots of predicted and real output values for the train and test data sets. They show that the training allows a high quality prediction for the test countries with the multiple linear regression model for life expectancy. It is noted that the scatter plots themselves are defective, since the prediction is performed by taking the input values  $\mathbf{x}$  assuming no errors. In order to illustrate also the errors in the input values, other plot types would need to be established.

Second, one major approach in this work is to investigate explanatory power of the model fit if the train data is partially unpaired, i.e. if we only consider data of groups of countries rather than individual countries for model training. To achieve this, we introduce the country grouping along an additional criteria which is not part of the input: *GDP per capita (current USD) [NY.GDP.PCAP.CD]* with increasing GDP per capita for each group. Please note, although the grouping is performed by GDP, the GDP itself is not part of the predictors and the overall information level for the predictors is reduced by this grouping compared to the completely paired data set. We considered 3 different groupings with group sizes  $L_r = H_r$  (the last group always consists of residual countries) of 4 ( $R = 44$ ), 8 ( $R = 22$ ) and 16 ( $R = 11$ ). The results of the model fit based

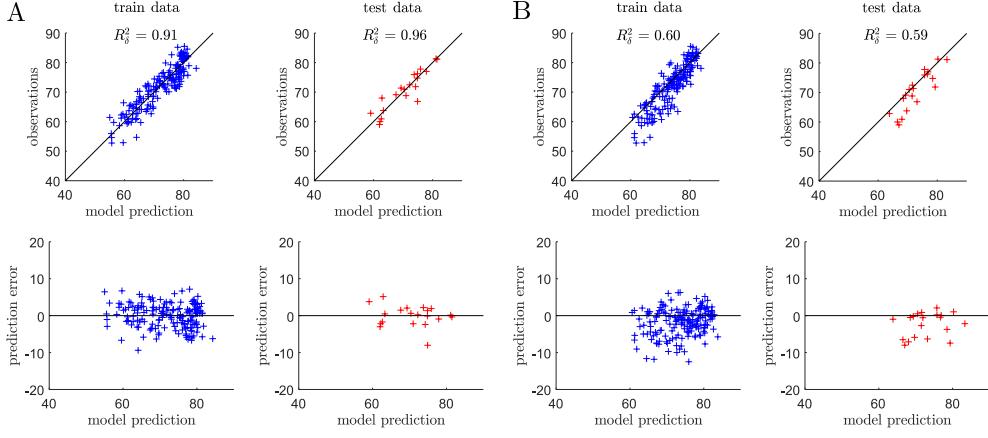


Figure 7: Plot results for real and predicted values of the model fit. A) Utilizing all 172 countries (completely paired) and B) utilizing only  $R = 11$  groups of countries (partially unpaired) with the  $R_\delta^2$  values as goodness of fit measure. Top: Correlation plots, Bottom: Residual error plots. Left: Train data, Right: Test data.

on this country groups are presented in Table 1. The  $R_\delta^2$  for train and test data is presented which

number of groups	group sizes	$R_\delta^2$ (train)	$R_\delta^2$ (test)
172	1	0.91	0.96
44	4	0.74	0.74
22	8	0.70	0.68
11	16	0.60	0.59

Table 1: Coefficient of determination  $R_\delta^2$  in train and test data for different group sizes.

show overall very high goodness-of-fit values but with decreasing values for increasing group sizes, as one might expect. Please note, the  $R_\delta^2$  for testing contains a rather strong noise component since it is calculated by only 20 data points in the test set. In addition, in Figure 7B (for group size 16) it is demonstrated that the fit results still align well for training and testing even for 11 subgroups. It is noted, that the grouping is part of the training to get  $\alpha$ , and not part of these scatter plots. This indicates that the model fit and the prediction for the test countries work quite well based only on the country groups, confirming that the pairing information can be reduced and this still provides a valuable model fit.

## 5 Discussion and Conclusion

In this work, we presented a general framework for model fitting scenarios with stochastic uncertainties for completely paired and partially unpaired data utilizing mixture models. The main advantage of this approach is its generality allowing for full flexibility about i) the number and dimensions of the data points, ii) the (possibly) individual error characteristics of each data point in the *errors-in-variables* framework, iii) the type of (linear or nonlinear) model to be fitted, iv) the specific level of pairing information, and v) completely avoiding *ad-hoc* loss functions. The presented loss function is derived from the data's pairing structure and data specific error characteristics, making it the most suitable fit for these problems. We present our framework as a generalization of *total least squares* [Markovsky and Van Huffel, 2007], extending it to a broader *errors-in-variables* context. By employing Gaussian errors with a line model, classical results are reproduced, but our approach also accommodates other scenarios, such as interval data through uniform distributions. The primary random variables in our study are the uncertainty variables  $\eta$  and  $\epsilon$ , with all stochastic derivations stemming directly from their definitions. This rigorous foundation is a key advantage of our approach.

The results in simulations and the real data study indicate by examples that there can be a trade-off between the level of pairing information (number and shape of unpaired subgroups) and estimation

accuracy, leading to a problem specific practical saturation in the accuracy level one can achieve by utilizing partially paired data. This means, the information about the full pairing of data is not as important for the fitting process as one might think, and consequently, accurate results can be performed also with reduced pairing information. Reduced pairing information can be useful, for example, in cases where the data is partially corrupt, or by deliberately leaving out pairing information due to data privacy policies (e.g. anonymizing data by building unpaired subgroups). We advocate to broaden the meaning of *semi-supervised* learning, as we did in this paper for *partially unpaired data*, in order to capture different scenarios of loss of pairing information which are practically relevant.

Although the presented framework might be general, the derived formulas lead only for specific selections of density function and model types to closed form solutions. The practical implementation can still be challenging, especially in cases with a high number of parameters  $\alpha$ , leading to a high dimensional optimization problem of an objective function with possibly non-distinct or non-unique extrema. Further, if the involved probability densities are not leading to expressions where the integral in Equation (11) can be exactly solved, numerical approximations of these integrals can be challenging especially for high-dimensional input data.

In the paper, it is only briefly presented how the ML approach can be directly extended applying MAP approaches. This was done to directly interpret the plotted likelihood functions in the results section as posterior densities by utilizing a non- or weakly informative prior. With this interpretation, we are able to directly quantify the uncertainties of the parameters  $\alpha$  of the model fit, allowing the calculation of credibility intervals or regions.

In general, Maximum Likelihood estimators are asymptotically consistent under some conditions, like identifiability. This extends to our framework, for fully paired data. However, for partially unpaired data, one would first need a meaningful definition of how the data and the data subgroups grow towards infinity. This is strongly related to the question of how the *dissimilarity* of unpaired subgroups of the data can be measured, as discussed in Section 2.5.2. This remains an interesting open question with certainly a differentiated answer which we direct to future work on this topic. Although the presentation of model fitting in this paper had regression problems in mind, the same argumentation can be applied to classification tasks. The adaption is that the output data  $\mathbf{y}$  and the image of  $\mathbf{M}(\cdot; \alpha)$  is discrete and finite. Further discretization, such as discrete density functions  $f_{\varepsilon_l}$  can be modeled by Dirac distributions in order to directly apply the presented equations, e.g. applying the *sifting property* for the obtained integral in Equation (11).

In total, this is a general argumentation framework for model fitting with many possible applications and an introduction to the specific treatment for partially unpaired data. The focus of this presentation is on the applied researcher, explaining all derivations and results in detail as well as providing numerical implementation strategies and interpretations of numerical examples. Further work is encouraged in order to extend this framework or provide further examples (e.g., benchmarking compared to alternative fitting methods) of expressive applications.

## A Completely Paired Data: Derivations for Errors in $\mathbf{y}$ only

In this case, the stochastic disturbances are only present in the output data:

$$\mathbf{y}_l^* := \mathbf{y}_l - \varepsilon_l \quad (13)$$

with  $\mathbf{y}_l^* \in \mathbb{R}^m$  the random variable of the true value, and the uncertainty random variable  $\varepsilon_l \sim f_{\varepsilon_l}(\mathbf{s}) : \mathbb{R}^m \mapsto \mathbb{R}$  independent for all  $l = 1, \dots, L$ . In a Bayesian context, the observed and true input values are the same  $\mathbf{x}_l = \mathbf{x}_l^*$ . With this, we introduce the technical argumentation of model fitting by

$$\mathbf{M}(\mathbf{x}_l; \alpha) = \mathbf{y}_l^* \quad \forall l = 1, \dots, L \quad (14)$$

$$\mathbf{M}(\mathbf{x}_l; \alpha) = \mathbf{y}_l - \varepsilon_l \quad \forall l = 1, \dots, L, \quad (15)$$

i.e. for given (undisturbed)  $\mathbf{x}_l$  we want to predict the true value  $\mathbf{y}_l^*$ . The first step in this technical presentation is to bring all basic random variables to the left side and equal this to  $\mathbf{0}$ :

$$\mathbf{M}(\mathbf{x}_l; \alpha) - \mathbf{y}_l + \varepsilon_l = \mathbf{0} \quad \forall l = 1, \dots, L. \quad (16)$$

We follow the interpretation: due to  $\varepsilon_l$  being a random variable, the left side is interpreted as a shifted random variable which density function value should have highest value at  $\mathbf{0} \in \mathbb{R}^m$ , following the idea of Maximum Likelihood for the parameters  $\alpha$ .

$$\Rightarrow \operatorname{argmax}_{\alpha} f_{\bigcap_{l=1}^L [M(\mathbf{x}_l; \alpha) - \mathbf{y}_l + \varepsilon_l]}(\mathbf{0}) \quad (17)$$

$$= \operatorname{argmax}_{\alpha} \prod_{l=1}^L f_{M(\mathbf{x}_l; \alpha) - \mathbf{y}_l + \varepsilon_l}(\mathbf{0}) \quad (\text{independence of } \varepsilon_l) \quad (18)$$

$$= \operatorname{argmax}_{\alpha} \prod_{l=1}^L f_{\varepsilon_l}(\mathbf{y}_l - M(\mathbf{x}_l; \alpha)). \quad (\text{shifted } \varepsilon_l) \quad (19)$$

We recognize this as the common standard result of Maximum Likelihood (ML) in this new way of technical argumentation and we present standard examples in the following.

### Example: Gaussian Disturbance

Introducing Gaussian disturbances, we get  $\varepsilon_l \sim \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 \cdot I_{m \times m})(s) \forall l = 1, \dots, L$  this results in

$$\Rightarrow \operatorname{argmax}_{\alpha} \prod_{l=1}^L e^{-\frac{1}{2\sigma_\varepsilon^2} \| \mathbf{y}_l - M(\mathbf{x}_l; \alpha) \|^2} \quad (\text{inserting pdf}) \quad (20)$$

$$= \operatorname{argmax}_{\alpha} e^{-\frac{1}{2\sigma_\varepsilon^2} \sum_{l=1}^L \| \mathbf{y}_l - M(\mathbf{x}_l; \alpha) \|^2} \quad (21)$$

$$= \operatorname{argmin}_{\alpha} \sum_{l=1}^L \| \mathbf{y}_l - M(\mathbf{x}_l; \alpha) \|^2 \quad (22)$$

which is the case of multivariate (nonlinear) ordinary least squares.

### Example: Fitting a Line and Gaussian Disturbance (Linear Regression)

Further utilizing the one-dimensional affine model  $M(x; \alpha) = \alpha_1 + \alpha_2 \cdot x$  and inserting it, we get

$$\Rightarrow \operatorname{argmin}_{\alpha} \sum_{l=1}^L (y_l - \alpha_1 - \alpha_2 \cdot x_l)^2 \quad (23)$$

which has the classical unique solution of the normal equations of ordinary least squares leading to a fitted line with parameters

$$\alpha_1 = \frac{\bar{x}^2 \cdot \bar{y} - \bar{x} \cdot \bar{x}\bar{y}}{\bar{x}^2 - \bar{x}^2}, \quad \alpha_2 = \frac{\bar{x}\bar{y} - \bar{x} \cdot \bar{y}}{\bar{x}^2 - \bar{x}^2} \quad (24)$$

with

$$\bar{x} := \frac{1}{L} \sum_{l=1}^L x_l, \quad \bar{y} := \frac{1}{L} \sum_{l=1}^L y_l \quad (25)$$

$$\bar{x}^2 := \frac{1}{L} \sum_{l=1}^L x_l^2, \quad \bar{x}\bar{y} := \frac{1}{L} \sum_{l=1}^L x_l \cdot y_l. \quad (26)$$

## B Completely Paired Data: Relation to Deming Regression

Deming regression is equivalent to the maximum likelihood for independent normally distributed observation errors in  $x_l$  and  $y_l$  ( $l = 1, \dots, L$ ), i.e.  $\eta_l \sim \mathcal{N}(0, \sigma_\eta^2)(s) \forall l = 1, \dots, L$  and  $\varepsilon_l \sim \mathcal{N}(0, \sigma_\varepsilon^2)(s) \forall l = 1, \dots, L$  with a line model. Since the true value  $s_l$  for  $x_l$  is unknown just like the parameters  $\alpha$  they are estimated by maximizing them simultaneously with the parameters  $\alpha$ ,

which leads to the effect that the estimation of the true values influences the estimation of the parameters  $\alpha$ :

$$\operatorname{argmax}_{\alpha, s_1, \dots, s_L} \prod_{l=1}^L e^{-\frac{1}{2\sigma_\varepsilon^2}(y_l - \alpha_1 - \alpha_2 \cdot s_l)^2 - \frac{1}{2\sigma_\eta^2}(x_l - s_l)^2}$$

In the case of this paper, we are also estimating the best parameters  $\alpha$  but independently of any specific true value  $s_l$ : We are averaging over all possible true values by the use of the law of total probability, compare Equation (5), which can be interpreted as an *integrated Deming regression*. This is a valid alternative perspective leading to a slightly more defensive estimation of  $\alpha$  which is not influenced by the estimation of  $s_l$ . An interesting observation is that the second part of derived objective function in Equation (6)

$$\operatorname{argmin}_{\alpha} \sum_{l=1}^L \frac{(\alpha_1 + \alpha_2 \cdot x_l - y_l)^2}{2(\alpha_2^2 \sigma_\eta^2 + \sigma_\varepsilon^2)}, \quad (27)$$

actually leads to the *classical Deming equations* and the first part of Equation (6)  $\frac{L}{2} \ln(\alpha_2^2 \sigma_\eta^2 + \sigma_\varepsilon^2)$  can be interpreted as a *penalty* added to the classical Deming regression, showing the more defensive estimation approach. Especially for large  $\sigma_\eta$  the parameter  $\alpha_2$  will tend slightly more to zero. The derivation of the classical Deming regression in this context is directly the minimization of Equation (27) by setting the gradient to zero

$$\nabla_{\alpha} \sum_{l=1}^L \frac{(\alpha_1 + \alpha_2 \cdot x_l - y_l)^2}{2(\alpha_2^2 \sigma_\eta^2 + \sigma_\varepsilon^2)} = \mathbf{0}, \quad (28)$$

whose solution results in the classical Deming regression coefficients

$$\alpha_1 = \frac{1}{L} \sum_{l=1}^L y_l - \alpha_2 \cdot x_l \quad (29)$$

$$\alpha_2 = \frac{s_{yy} - \frac{\sigma_\varepsilon^2}{\sigma_\eta^2} \cdot s_{xx} + \sqrt{\left(s_{yy} - \frac{\sigma_\varepsilon^2}{\sigma_\eta^2} \cdot s_{xx}\right)^2 + 4 \frac{\sigma_\varepsilon^2}{\sigma_\eta^2} s_{xy}^2}}{2 s_{xy}} \quad (30)$$

with

$$s_{xx} = \frac{1}{L} \sum_{l=1}^L (x_l - \bar{x})^2, \quad s_{yy} = \frac{1}{L} \sum_{l=1}^L (y_l - \bar{y})^2, \quad s_{xy} = \frac{1}{L} \sum_{l=1}^L (x_l - \bar{x}) \cdot (y_l - \bar{y}). \quad (31)$$

In conclusion, the presented approach leads for *Deming type problems* to formulas, which we call *integrated Deming regression*. These formulas can be interpreted as penalized *classical Deming regression* showing the more defensive approach by averaging over the true value during estimation of  $\alpha$  compared to estimating them simultaneously in *classical Deming regression*.

## C Demonstration for a Plane Fit with Gaussian Disturbance for Partially Unpaired Data

One possible part of demonstrating the flexibility of this framework is to show how it works in higher dimensions, which will be indicated by the previously introduced plane fit model. We are using  $L = H = 1600$  data points  $\mathbf{x}_h \in \mathbb{R}^2$  and  $y_l \in \mathbb{R}$  and numbers of subgroups are  $R \in \{6, 18, 100, L = H = 1600\}$ . The data generation parameters are the same as for the base scenario A in the line fit section but with the true values  $\alpha_1 = 0$ ,  $\alpha_2 = 0.2$  and  $\alpha_3 = 0.4$ .

Obviously the partitioning of the data has much more possibilities due to a much richer neighboring information for  $\mathbf{x}_h$  in 2D. In consequence, the discussed dissimilarity of the groups to the total data set gets more difficult to study. Nonetheless, it can be demonstrated that the utilization with more pairing information does increase accuracy and meaningful estimation can be performed even with a low number of groups.

In Figure 8 the fitting results are presented for separated groups (no overlapping of data groups in the  $\mathbf{x}$ -plane). In addition, in Figure 9 the same data is utilized but a random relabeling/switching of group labels is performed for approximately 30% of the data, which makes all groups slightly more similar to each other. As proposed in *Section 2.5.2*, the plane fitting results get worse the more similar the grouping gets.

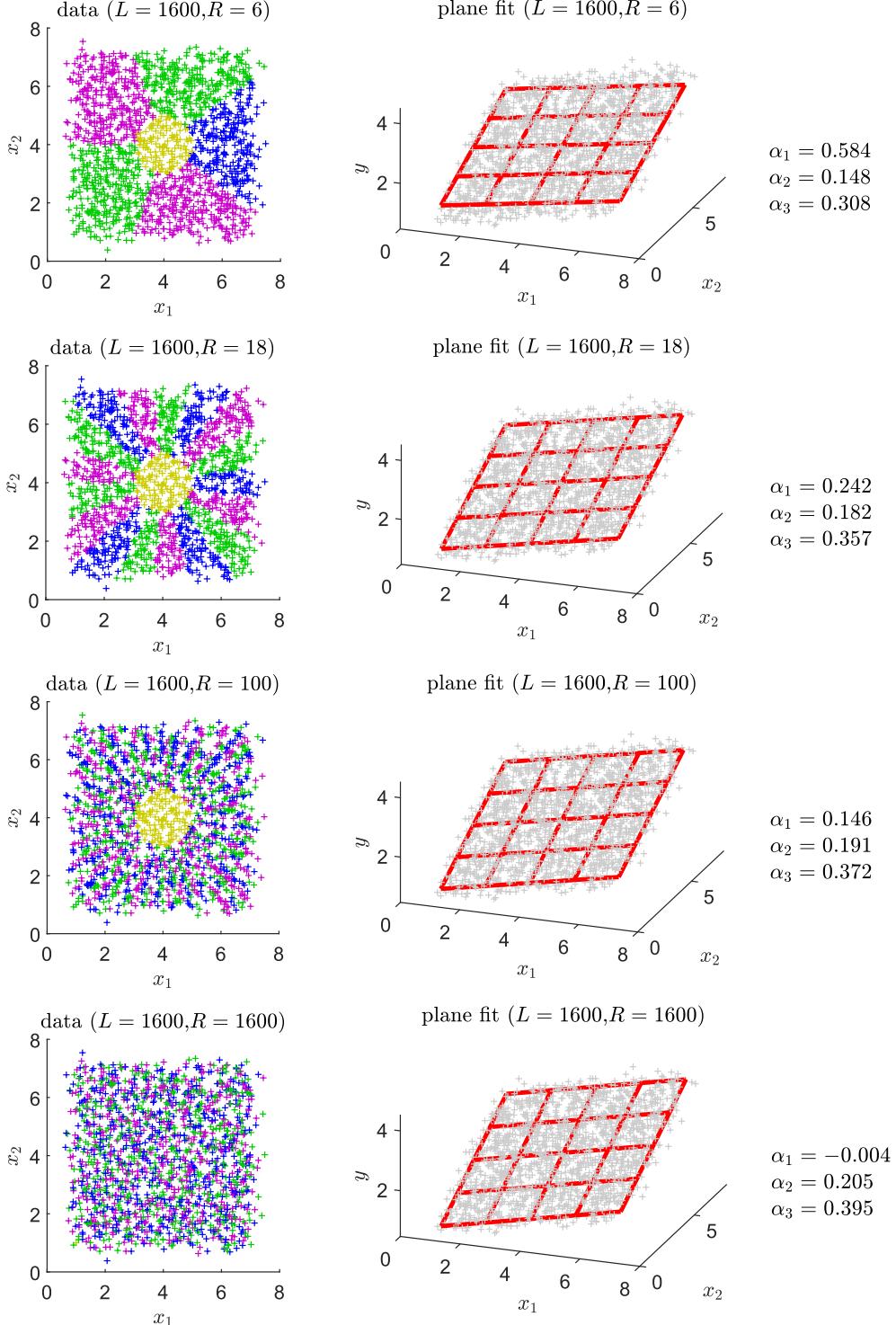


Figure 8: Illustration of example results for different plane fitting scenarios utilizing a partitioning of separated groups. Four different scenarios of partial pairing with  $R = 6$ ,  $R = 18$ ,  $R = 100$  and  $R = L = H$  unpaired subgroups. For each case: Left plot: representing the unpaired data subgroups by color-coding. Right plot: presenting the fitted plane (red) in the full data (gray).

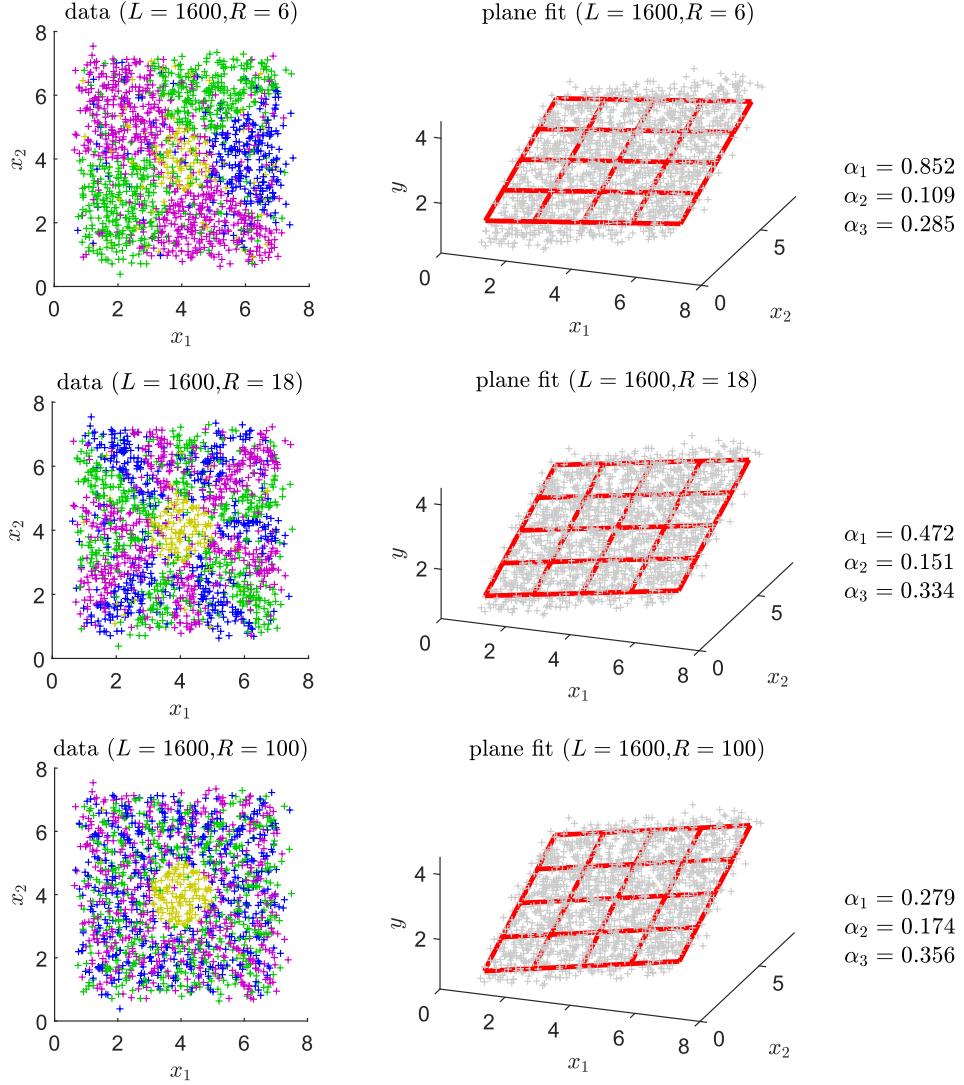


Figure 9: Illustration of example results for different plane fitting scenarios utilizing a partitioning of overlapping groups (i.e., randomly switching the group label for approximately 30% of the data points compared to Figure 8). Three different scenarios of partial pairing with  $R = 6$ ,  $R = 18$  and  $R = 100$  unpaired subgroups are demonstrated in the same type of presentation as in Figure 8.

## D Application of $R_\delta^2$

In [Cheng et al., 2014] a consistent goodness of fit measure for *errors-in-variables* multiple linear regression is presented if the standard deviations of the input variables are known. We show its direct application in this appendix. With  $X$  the  $L \times k$ -Matrix of  $L$  observations and  $k$  predictor variables,  $\mathbf{y}$  the  $L \times 1$  vector of output observations,  $\Sigma_\delta$  the  $k \times k$  covariance matrix of predictor variables,  $S = \frac{1}{L} X^T P X$  and  $P = I_{L \times L} - \frac{1}{L} \mathbf{1}_{L \times L}$  ( $\mathbf{1}$  being the matrix consisting of 1s) the goodness of fit is defined by

$$R_\delta^2 = \min \left( \frac{\mathbf{b}^T S \mathbf{b}}{\frac{1}{L} \mathbf{y}^T P \mathbf{y} + \mathbf{b}^T \Sigma_\delta \mathbf{b}}, 1 \right),$$

where  $\mathbf{b}$  is the vector of fitted slopes, i.e., in the notation of the multiple linear regression model of this paper  $\mathbf{b} = (\alpha_2, \alpha_3, \dots, \alpha_k)^T$  being independent of the intercept  $\alpha_1$ .

## References

- Zhengyou Zhang. Parameter estimation techniques: a tutorial with application to conic fitting. *Image and Vision Computing*, 15(1):59–76, January 1997. doi: 10.1016/S0262-8856(96)01112-2.
- Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, 2006. ISBN 978-0-387-31073-2.
- Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. A Comprehensive Survey of Loss Functions in Machine Learning. *Annals of Data Science*, 9(2):187–212, April 2022a. doi: 10.1007/s40745-020-00253-5.
- Wolfgang Hoegele, Rainer Loeschel, Barbara Dobler, Oliver Koelbl, and Piotr Zygmanski. Bayesian Estimation Applied to Stochastic Localization with Constraints due to Interfaces and Boundaries. *Mathematical Problems in Engineering*, 2013:1–17, 2013. doi: 10.1155/2013/960421.
- Zhidong Bai and Tailen Hsing. The broken sample problem. *Probability Theory and Related Fields*, 131(4):528–552, April 2005. doi: 10.1007/s00440-004-0384-5.
- Feng Liang, Sayan Mukherjee, and Mike West. The Use of Unlabeled Data in Predictive Modeling. *Statistical Science*, 22(2), May 2007. doi: 10.1214/088342307000000032.
- Yudong Wang, Yanlin Tang, and Zhi-Sheng Ye. Paired or Partially Paired Two-sample Tests With Unordered Samples. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1503–1525, September 2022b. doi: 10.1111/rssb.12541.
- Georgios Kostopoulos, Stamatis Karlos, Sotiris Kotsiantis, and Omiros Ragos. Semi-supervised regression: A recent review. *Journal of Intelligent & Fuzzy Systems*, 35(2):1483–1500, August 2018. doi: 10.3233/JIFS-169689.
- Guo-Jun Qi and Jiebo Luo. Small Data Challenges in Big Data Era: A Survey of Recent Progress on Unsupervised and Semi-Supervised Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2168–2187, April 2022. doi: 10.1109/TPAMI.2020.3031898.
- Derrick A. Bennett. How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, 25(5):464–469, October 2001. doi: 10.1111/j.1467-842X.2001.tb00294.x.
- J. A C Sterne, I. R White, J. B Carlin, M. Spratt, P. Royston, M. G Kenward, A. M Wood, and J. R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338(jun29 1):b2393–b2393, September 2009. doi: 10.1136/bmj.b2393.
- Semhar Michael, Tatjana Miljkovic, and Volodymyr Melnykov. Mixture modeling of data with multiple partial right-censoring levels. *Advances in Data Analysis and Classification*, 14(2):355–378, June 2020. doi: 10.1007/s11634-020-00391-x.
- Zachary R. McCaw, Hugues Aschard, and Hanna Julienne. Fitting Gaussian mixture models on incomplete data. *BMC Bioinformatics*, 23(1):208, December 2022. doi: 10.1186/s12859-022-04740-9.
- Wolfgang Hoegele. A Stochastic-Geometrical Framework for Object Pose Estimation Based on Mixture Models Avoiding the Correspondence Problem. *Journal of Mathematical Imaging and Vision*, June 2024a. doi: 10.1007/s10851-024-01200-2.
- Wolfgang Hoegele. Combinatorial potential of random equations with mixture models: Modeling and simulation, March 2024b. arXiv:2403.20152 [cs, math, stat].
- W. Edwards Deming. *Statistical adjustment of data*. Dover publ, New York, unabridged and corr. republication edition, 1964. ISBN 978-0-486-64685-5.
- Ivan Markovsky and Sabine Van Huffel. Overview of total least-squares methods. *Signal Processing*, 87(10):2283–2302, October 2007. doi: 10.1016/j.sigpro.2007.04.004.

Eufrásio De A. Lima Neto and Francisco De A.T. De Carvalho. Centre and Range method for fitting a linear regression model to symbolic interval data. *Computational Statistics & Data Analysis*, 52(3):1500–1515, January 2008. doi: 10.1016/j.csda.2007.04.014.

Leandro C. Souza, Renata M.C.R. Souza, Getúlio J.A. Amaral, and Telmo M. Silva Filho. A parametrized approach for linear regression of interval data. *Knowledge-Based Systems*, 131: 149–159, September 2017. doi: 10.1016/j.knosys.2017.06.012.

C.-L. Cheng, Shalabh, and G. Garg. Coefficient of determination for multiple measurement error models. *Journal of Multivariate Analysis*, 126:137–152, April 2014. doi: 10.1016/j.jmva.2014.01.006.