chemengineering

Article

Parameter Estimation Strategies in Thermodynamics

Johannes Höller [1,*], Patricia Bickert [1], Patrick Schwartz [1], Martin von Kurnatowski [1], Joachim Kerber [2], Niklaus Künzle [2], Hilke-Marie Lorenz [2], Norbert Asprion [3], Sergej Blagov [3] and Michael Bortz [1]

[1] Fraunhofer Center for Machine Learning and ITWM, Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany; patricia.bickert@itwm.fraunhofer.de (P.B.); patrick.schwartz@itwm.fraunhofer.de (P.S.); martin.kurnatowski@itwm.fraunhofer.de (M.v.K.); michael.bortz@itwm.fraunhofer.de (M.B.)
[2] Lonza AG, Lonzastrasse, 3930 Visp, Switzerland; joachim.kerber@lonza.com (J.K.); niklaus.kuenzle@lonza.com (N.K.); hilke.lorenz@lonza.com (H.-M.L.)
[3] BASF SE, Carl-Bosch-Straße 38, 67056 Ludwigshafen, Germany; norbert.asprion@basf.com (N.A.); sergej.blagov@basf.com (S.B.)
[*] Correspondence: johannes.hoeller@itwm.fraunhofer.de; Tel.: +49-631-31600-4854

Received: 26 February 2019; Accepted: 26 May 2019; Published: 1 June 2019

Abstract: Many thermodynamic models used in practice are at least partially empirical and thus require the determination of certain parameters using experimental data. However, due to the complexity of the models involved as well as the inhomogeneity of available data, a straightforward application of basic methods often does not yield a satisfactory result. This work compares three different strategies for the numerical solution of parameter estimation problems, including errors both in the input and in the output variables. Additionally, the new idea to apply multi-criteria optimization techniques to parameter estimation problems is presented. Finally, strategies for the estimation and propagation of the model errors are discussed.

Keywords: regression; model adjustment; optimization

1. Introduction

Process engineering involves mechanistic models on different scales, ranging from pure-substance over mixture properties up to models for different apparatuses and flowsheets. On each scale, the corresponding model involves model parameters. Generally, these model parameters are not known a priori, but must be determined such that the model predictions match measurements from experiments. This procedure is designated as parameter estimation by model regression in the following. This work describes recent progress that has been made in implementing least-squares regression approaches in an industrially relevant context, including efficient error propagation from one scale (e.g., thermodynamic mixture properties) to another (e.g., flowsheets described by the MESH equations).

Least-squares regression probably belongs to the most popular methods of parameter estimation. The reason is that one can show that this method leads to the best parameter estimates assuming that the ground truth is known and the experimental measurements are normally and independently distributed around this ground truth (for more details, cf. the textbook [1]). However, these two assumptions are never strictly fulfilled in practice: Neither the ground truth is known nor are the measurements normally distributed. Thus, from a practical point of view, least squares can rather be considered as one possibility to measure the distance between model prediction and experiment. Other measures for the distance adapted to special situations, such as strong deviations of the measured data from normal distribution due to outliers, are known as well [2]. More recently, Bayesian parameter estimation has gained some attention [3,4]. Here, however, we restrict ourselves to the least-squares approach, which in countless situations has led to satisfactory parameter estimates (for a recent review, cf. [5]).

In the context of chemical engineering, three challenges are encountered when it comes to regression: The first is that measurement errors are present not only in the measured quantities corresponding to model predictions but also in those corresponding to the model inputs. The second challenge results from multiple model outputs and their respective weights in the loss function of the regression: Ideally, weights are obtained from the variances of the measurement errors [6], which are not known exactly. They can be estimated as well, resulting in tedious iterative loops [3,6]. For the practitioner, however, weights have only a limited physical meaning; the more important question is how well can the model describe a set of data with different output quantities. Finally, the third challenge consists in propagating the error estimates of the model parameters through different model scales, since these are usually implemented in different environments with limited interfaces. Thus, methods are needed that encode the parameter estimates and their errors obtained from the performed regression, e.g., for mixture properties from experiments in the lab, such that these can be easily used by flowsheet simulators, which describe the properties of entire processes. All three challenges are addressed in the present contribution.

In Section 2, we compare different methods to numerically solve parameter estimation problems for highly nonlinear thermodynamic models for pure-substance and mixture properties. These could be, for example, gE-models [7,8] or equation-of-state approaches such as PC-SAFT [9,10]. Further references can be found in the textbook [11]. In conventional least-squares regression, it is assumed that measurement errors occur only in the model outputs. Measurement errors in both model inputs and outputs are accounted for by data reconciliation or errors-in-variables methods [12,13]. Data reconciliation aims to correct the measurement errors by the incorporation of models for the process. Models at different scales can be employed, ranging from simple models describing only parts of the process, e.g., the conservation of mass, to complex models, which are designed to describe the process completely. Data reconciliation does not differentiate between input and output quantities. Therefore, data reconciliation combined with parameter estimation can serve as an errors-in-variables model for parameter estimation. Errors-in-variables methods lead to restricted nonlinear optimization problems, where the number of equality constraints increases with the number of measured points. This can become a challenging optimization problem, requiring advanced solvers which are capable of dealing with sparsity. In industrial environments, such solvers are not always available. We therefore show how the constrained least-squares optimization problem can be recast as an unrestricted problem. This idea originates in the work by Patino–Leal [14,15]. Our contribution consists in an alternative derivation of this formulation and in a comparison of the regression results from conventional least squares, error-in-variables least squares with constraints, and the reformulation without constraints. Section 3 provides the comparison of the different parameter estimation strategies for two specific thermodynamic examples.

The second challenge, related to a suitable choice of weights in a multivariate regression problem with unknown variances of the measurements, is tackled in a multi-criteria framework in Section 4. An adaptive scalarization scheme is used to choose the weights such that, in the case of conflicting regression objectives, the resulting Pareto boundary is computed with the least numerical effort [16–18]. This approach has been successfully applied to parameter estimation in the context of molecular simulations [19] and the parametrization of equations of state [20]. In this work, the multi-criteria framework will be, to the best of the authors' knowledge for the first time, extended to the determination of process parameters of a flowsheet simulation.

Finally, in Section 5, we discuss the third challenge dealing with error propagation. It is well known that an estimate of the covariance matrix for the model parameters can be obtained from an analysis of the regression results [21,22]. In the present paper, we review different measures for the errors resulting from this matrix. Furthermore, we propose a scheme allowing for efficient error propagation in the case where the regression task has been performed in a model environment different from the one eventually employing the parameter estimates. This situation frequently occurs in process engineering, since the properties of pure substances and their mixtures are estimated from regression of thermodynamic models to lab experiments and are then used by flowsheet simulators.

We would like to mention that all of the described algorithms were implemented in in-house tools of the industrial partners with very specific interfaces. Therefore, employing existing commercial solutions was not an option.

2. Parameter Estimation Approaches

The physical process involved in measuring an input–output relation can be described by the regression model

$$y_i = f(x_i; \beta) + \epsilon_i. \tag{1}$$

Here, the measured value of the output y_i is obtained by evaluating the model $f(x; \beta)$ at the input x_i and adding an error ϵ_i, which is assumed to be normally distributed with zero mean and standard deviation σ^{y_i}. The model depends on the vector of parameters $\beta \in \mathbb{R}^{N_\beta}$, which are to be estimated from a set of measured data $\{(x_i, y_i)|i = 1, ..., N\}$.

At this point, the input quantities x_i may be considered as vectors of dimension N_x and y_i, f, ϵ_i as vectors of dimension N_y. If $N_y > 1$, we assume that the measurement errors ϵ_i are uncorrelated, i.e., σ^{y_i} is a diagonal $N_y \times N_y$-matrix. This assumption is sufficient for many situations in practice, because the measurements are performed by independent sensors not being correlated with each other. Since, in most cases, estimates for the correlations of measurement errors are not available, we discard them for the sake of a simpler notation.

If the model is a linear function of the parameters to be estimated, a linear parameter-estimation problem has to be solved. This leads to a quadratic optimization problem and can, in principle, be solved analytically. On the other hand, if the model depends nonlinearly on the parameters, the parameter-estimation problem becomes nonlinear as well. In this case, a nonlinear optimization method is required. In thermodynamics, many highly-nonlinear effects occur and, therefore, this work focuses on nonlinear parameter estimation problems. The methods used can also be applied to linear problems.

2.1. Least Squares

The simplest and most common strategy for the estimation of model parameters from measured data is the method of least squares. The approach consists in minimizing the sum of squared residuals, which are defined as the differences between the measured values y_i and the ones predicted by the model for the corresponding measured inputs x_i. It can be derived from the maximum-likelihood estimator under the condition that all measurement errors are normally distributed and independent. The detailed derivation can be found in numerous textbooks (e.g., [23]). Therefore, only a short summary is given here. The best values for the model parameters β are calculated by maximization of the probability to obtain the measured dataset $\{(x_i, y_i)|i = 1, ..., N\}$, leading to the optimization problem

$$\min_{\beta} s^A, \tag{2}$$

$$s^A = \sum_{i=1}^{N} \left(\frac{f(x_i; \beta) - y_i}{\sigma^{y_i}} \right)^2. \tag{3}$$

For the multiple output case, the single terms of s^A should formally be written as $(f(x_i) - y_i)^T(\sigma^{y_i})^{-2}(f(x_i) - y_i)$, whereas, for diagonal matrices, Equation (3) can be interpreted component-wise. In the following, we employ the simpler notation. In Equation (1), the measurement error is assumed to appear only in the model outputs, the errors of the input quantities σ^{x_i} are not considered. In thermodynamic applications (and many others), the differentiation between input and output quantities is often only based on the formulation of the model. However, both quantities are subject to measurement errors.

2.2. Reconciliation Formulation with Constraints

To include errors of input variables, a simultaneous data reconciliation and parameter estimation problem is proposed [12,13], which is also known as the errors-in-variables model [14].

$$\min_{\beta, x'} s^B, \tag{4}$$

$$s^B = \sum_{i=1}^{N} \left[\left(\frac{x'_i - x_i}{\sigma^{x_i}} \right)^2 + \left(\frac{f(x'_i; \beta) - y_i}{\sigma^{y_i}} \right)^2 \right], \tag{5}$$

where σ^{x_i} denotes the errors of the input variables. However, the price for including the errors of the input variables is high: Optimization solvers specialized for least-squares functions usually require derivatives for every residual separately. Here, the introduction of $N \times N_x$ additional optimization variables increases the size of the gradient from N_β to $(N_\beta + N \times N_x)$ and the number of residuals from $N \times N_y$ to $N \times (N_x + N_y)$. The size of the Hessian matrix increases by the square of the factors. Most of the additional entries are zeros, and it is possible to calculate the matrices efficiently. However, storing them in memory efficiently is more difficult and requires to use solvers with implementations, which are adapted to these structures.

As shown in Section 2.3, the problem can be reformulated and approximated to a form without these additional variables but still including the errors of the input variables (see Equation (15)). Before addressing this, we reformulate Equation (5) using constraints, which leads to a constrained optimization problem

$$\min_{\beta, x', y'} s^{B*},$$

$$s^{B*} = \sum_{i=1}^{N} \left[\left(\frac{x'_i - x_i}{\sigma^{x_i}} \right)^2 + \left(\frac{y'_i - y_i}{\sigma^{y_i}} \right)^2 \right] \tag{6}$$

$$s.t. \quad 0 = f(x'_i; \beta) - y'_i \quad \forall\, i = 1, ..., N. \tag{7}$$

This version can directly be applied to implicit models and represents the first step in reducing the number of optimization variables.

2.3. Reconciliation Formulation without Constraints

In Equation (6), the input and output variables are treated on equal footing, which is expressed by collecting them in a single vector $z_i = (x_i, y_i)$, and the corresponding errors σ^{z_i} into one diagonal matrix. Furthermore, the quantity g is introduced as

$$g(z; \beta) = g(x, y; \beta) = f(x; \beta) - y. \tag{8}$$

This form has the additional benefit that it can be directly applied to models given only implicitly (i.e., $g(x,y;\beta)=0$). Equations (6) and (7) with the substitution Equation (8) constitute a restricted least-squares optimization problem, where the number of constraints scales with the number of measurements N. For large N, advanced optimization solvers are necessary to solve this problem. In this section, a method is presented that leads to an unrestricted optimization problem, which is equivalent to Equations (6)–(8). This method is due to Patino–Leal [14,15]. In the original work, it was derived in a Bayesian framework. Here, we present an alternative derivation based solely on the restricted problem (Equations (6)–(8)). This reformulation relies on a linearization of the constraints in Equations (7) and (8), and is sketched below. To remove the additional optimization variables from Equation (6), the constraints (i.e., the actual model equations) are linearized with respect to the variables z_i' around some point ξ_i:

$$\min_{\beta,z'} s,$$

$$s = \sum_{i=1}^{N} \left(\frac{z_i' - z_i}{\sigma^{z_i}}\right)^2 = \sum_{i=1}^{N} \left(\frac{z_i' - z_i}{\sigma^{z_i}}\right)^T \left(\frac{z_i' - z_i}{\sigma^{z_i}}\right) \tag{9}$$

$$s.t. \quad 0 = g(\xi_i;\beta) + (z_i' - \xi_i)\frac{\partial g}{\partial z}(\xi_i;\beta) \quad \forall\, i = 1,...,N. \tag{10}$$

Note that, for linear models, the approximation is exact, and, for explicit models, the y_i dependence is already linear by definition of g. The impact of this approximation on the results strongly depends on the accuracy of the approximation in the vicinity of the actual result, which in turn relies on the choice of the point, which is used for the linearization. Now, while the number of optimization variables has not been reduced yet, the minimization problem with respect to the z' variables, for fixed parameter values β, has a much simpler structure-quadratic objective function and linear constraints—and can be solved analytically. It is very similar to the problems appearing in data reconciliation dealing with unmeasured variables described, e.g., in [12]. The analytical solution is given by

$$z_i' = z_i - V_i B_i^T (B_i V_i B_i^T)^{-1} (g_i + B_i^T (z_i - \xi_i)), \tag{11}$$

with the abbreviations

$$B_i = \frac{\partial g(z;\beta)}{\partial z}|_{z=\xi_i}, \tag{12}$$

$$g_i = g(\xi_i;\beta), \tag{13}$$

$$V_i = \mathrm{diag}((\sigma^{z_i})^2). \tag{14}$$

Inserting these results in the optimization problem in Equation (9) yields

$$\min_{\beta} s^C,$$

$$s^C = \sum_{i=1}^{N} (g_i + B_i(z_i - \xi_i))^T (B_i V_i B_i^T)^{-1} (g_i + B_i(z_i - \xi_i)). \tag{15}$$

This strategy can readily be applied to models with multiple inputs and outputs and even works directly for implicit models. The values $\xi_i = (x_i, y_i)$ are called linearization points and play a similar role as the additional variables x' in the reconciliation strategy given in Equation (5). However, in contrast to Equation (5), the ξ_i are calculated within the evaluation of the residuals. For large problems, this is an advantage compared to Equation (5), where, due to the introduction of a high number of extra optimization variables, the optimization takes substantially longer. On the other hand, in Equation (15), the additional calculations are performed within the evaluation of the objective, and thus the numerical effort grows only linearly with the number of data points.

Thus far, the reformulation leading to Equation (15) is based on an initial linearization of the constraints. The optimal point for linearization of the constraints is the true value of the measured quantities, because there the nonlinearity of the model does not play a role any longer. For reliable measurements, the measured values are a reasonable approximation for the linearization points $\xi_i = z_i$. For different linearization points, the linearization has to be adapted in a fixed-point iteration after each solution of the unrestricted problem Equation (15) in the following way [14,15]:

$$\xi_i^{k+1} = z_i - V_i B_i^T (B_i V_i B_i^T)^{-1} (g(\xi_i^k; \beta) + B_i (z_i - \xi_i^k)), \tag{16}$$

which is expected to converge reliably also in the presence of non-linearities.

2.3.1. Computational Aspects

Optimization algorithms specialized to least-squares problems [24] have been developed which offer a significant gain in performance compared to conventional solvers which do not take advantage of the least-squares structure. The performance is enhanced by deriving an approximation of the Hessian matrix from the gradients of the individual terms instead of from the gradient of the sum of squares. The Hessian of a quadratic expression contains the first and the second derivatives of the argument. Since the gradients of the individuals terms are, in most cases, already calculated to determine the gradients of $s^{A,B,C}$, this part of the Hessian can be evaluated exactly without additional costs and only the part involving second derivatives has to be approximated. Different methods to incorporate this information exist: Spedicato and Vespucci [24] proposed specialized Hessian update strategies, whereas Schittkowski [25] reformulated the individual objective terms as constraints and approximates the Hessian of the constraints individually.

Those specialized solvers usually rely on a sum of squares form of the objective function. The strategy given by Equation (15) does not have this structure. It is, however, possible to reformulate the objective function s^C into a form consistent with standard least-squares algorithms. To this end, the QR decomposition of the matrix $D_i^T B_i^T = Q_i R_i$, with D_i defined by $V_i = D_i D_i^T$, is introduced. Then, the inverse matrix in Equation (15) can be obtained and the objective function can be equivalently written as

$$s^C = \sum_i v_i^T v_i, \tag{17}$$

where the vectors v_i are calculated by solving the linear system of equations

$$\begin{aligned} R_i^T R_i \tilde{v}_i &= g(\xi_i; \beta) + B_i(\xi_i)(z_i + \xi_i), & (18) \\ v_i &= Q_i R_i \tilde{v}_i. & (19) \end{aligned}$$

The computational costs are only slightly increased, since Equation (15) requires a matrix inversion as well. A detailed derivation is contained in Appendix A.

2.3.2. Linearization Points

A simple version of Equation (15) is to use the measured values z_i as an approximation of ξ_i. However, if one wants to use the optimal values for the linearization points ξ_i given in Equation (16), an additional loop inside the residual evaluation is required, which leads to significantly more model evaluations. One solution to this problem is to only use $\xi_i^0 = z_i$ as starting value in the first iteration of the optimization (or whenever a misfit is detected) and the final value of the previous iteration otherwise. Then, Equation (16) may require only a small numbers of iterations, in the extreme case just one. For strongly nonlinear models and parameter values far from optimal (or even physically reasonable) ranges, however, the convergence may be limited. In such cases, it is preferable to start the optimization with the approximate formulation, until reasonable parameter values are obtained.

An alternative formulation for the iteration in Equation (16) is

$$\begin{aligned} \xi_i^{k+1} &= \xi_i^k + u_i^k, && (20) \\ B_i(\xi_i^k;\beta)u_i^k &= -g(\xi_i^k;\beta), && (21) \\ ||D_i^{-1}(z_i - \xi_i^k - u_i^k)|| &\rightarrow \text{min}, && (22) \end{aligned}$$

with $V_i = D_i D_i^T$ (as in the previous section). This requires an algorithm for the linear least-squares problem with equality constraints given in Equations (20)–(22). QR decomposition is a typical component of such algorithms, but it is no longer necessary to use it explicitly. The objective function s^C of Equation (17) can be calculated directly as $v_i = \frac{z_i - \xi_i}{\sigma^{z_i}}$. The derivation is provided in Appendix B.

3. Comparison of Parameter Estimation Strategies

To demonstrate the behavior of the different regression strategies presented in Section 2, two examples from thermodynamic practice are discussed.

3.1. Example: Vapor Pressure

The standard DIPPR model for the vapor pressure of methanol was chosen as a first test model to compare the different regression strategies. The measured dataset was provided by BASF SE and consists of 40 data points for the (logarithmic) vapor pressure and the temperature. The model describes the temperature dependence of the logarithmic vapor pressure according to

$$\ln(p) = a + \frac{b}{T} + c\ln(T). \tag{23}$$

Herein, a, b, and c denote the model parameters to be estimated, and pressure p and temperature T are given in units of Pa and K, respectively. The model parameters were determined using four different regression strategies with three different initial parameter values:

1. the ordinary least-squares strategy s^A (Equation (3))
2. the reconciliation strategy s^B (Equation (5))
3. the approximate error-in-variables model s^C (Equation (17)) without linearization point optimization
4. the full error-in-variables model s^C together with Equation (20) (s^{C*})

The results are shown in Table 1. They include the observed run time of each regression run. These values were obtained by taking the mean value of repeated micro-benchmark runs on a Windows machine with an Intel Core i7-7700 CPU and 2 × 8 GB DDR4-2400 main memory. The parameter values agree with each other within a few percent, except for the strategies s^C and s^{C*} for the third initial value. The error was defined as the minimum value of the objective function $s^{A,B,C,C*}$ and therefore expected to differ between the methods. Note that the highest error was obtained by s^A, because all other methods split the error between input and output. The reconciliation strategy s^B required many iterations. However, the approximate version s^C yielded, except in one case, almost the same result with much fewer iterations. The last strategy s^{C*} agreed with the third one, but took the highest time in total, because of the additional loop for the linearization points. For the third initial value, the last two strategies converged to a different local minimum; however, the effect of the varying parameter sets on the model prediction was very small.

The model predictions for the corresponding cases are shown in Figure 1. The difference between any two predictions evaluated at the data points was smaller than 0.04 (i.e., below 0.5% of the absolute value). If only the different strategies starting from the same initial point were compared, this bound reduced to 0.0003. The fact that different parameter sets led to similar model predictions was related to the correlations between the model parameters. The variation of one parameter was compensated by adjusting a correlated parameter, and the parameters could not be independently estimated from the data. For this example, the ordinary least-squares strategy (s^A) already worked quite well, and we compared it with respect to the results and the performance of the other methods. The strategy s^B works well for sufficiently small problems, whereas the error-in-variables strategy s^C is also suited for larger problems. The last strategy s^{C*} is useful for situations where the approximation in s^C does not work well and the problem size is too large for s^B.

Table 1. Regression results for different methods and initial points.

Method	Start	*a*	*b*	*c*	Iterations	Error	Time [ms]
Initial point	I	100.896	−7210.917	−12.44128	-	-	-
	II	1	1	1	-	-	-
	III	50	−5000	0	-	-	-
s^A	I	44.5796	−5511.94	−2.87739	24	0.23211	2.35
	II	44.581	−5512.01	−2.8776	19	0.23211	1.87
	III	44.581	−5512.01	−2.8776	15	0.23211	1.49
s^B	I	43.7994	−5471.03	−2.76444	192	0.107258	156.01
	II	43.7992	−5471.03	−2.76441	179	0.107258	145.33
	III	43.7991	−5471.02	−2.7644	169	0.107258	137.41
s^C	I	43.7851	−5470.32	−2.76237	21	0.107503	11.93
	II	43.7833	−5470.23	−2.76211	31	0.107503	17.37
	III	34.8895	−5000.28	−1.47651	9	0.133713	5.37
s^{C*}	I	43.7985	−5470.98	−2.76431	29	0.107258	228.6
	II	43.799	−5471.02	−2.76438	31	0.107258	239.06
	III	34.8928	−5000.36	−1.477	10	0.133517	913.83

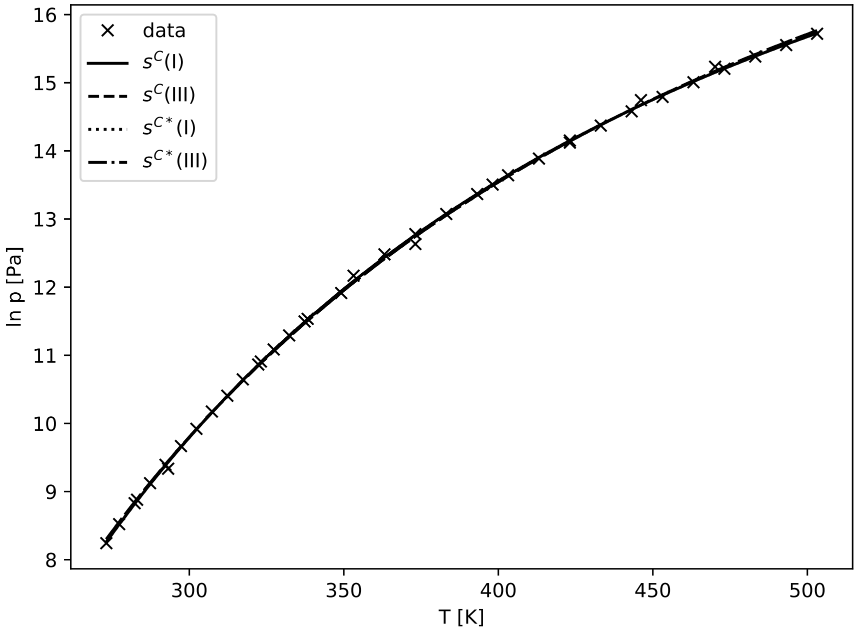


Figure 1. Regression results compared to data for regression runs with different initial parameters.

3.2. Example: NRTL

For a second comparison of the different regression problem formulations, the NRTL model for a water–methanol binary mixture was chosen. For this test, the simplest parameterization of the NRTL model was used

$$\frac{G^E}{RT} = \sum_{j=1}^{2} x_j \frac{S_{1j}}{S_{2j}}, \tag{24}$$

$$S_{1i} = \sum_{j=1}^{2} x_j P_{ji} \tau_{ji}, \tag{25}$$

$$S_{2i} = \sum_{j=1}^{2} x_j P_{ji}, \tag{26}$$

$$P_{ji} = \exp(-\alpha \tau_{ji}), \tag{27}$$

which has three parameters ($\alpha, \tau_{12}, \tau_{21}$). Here, G^E denotes the excess Gibbs free energy, which is normalized by temperature T and gas constant R. The model has a single input x_1, which is the mole fraction of methanol in the liquid phase, and $x_2 = 1 - x_1$ being the corresponding mole fraction of water is not independent. The dataset, which already has been published in [17], contains the temperatures and the molar fractions of methanol in the liquid and the vapor phase. Based on the extended Raoult's law, from these data activity coefficients $\ln \gamma_{1,2} = \frac{\partial G^E/RT}{\partial x_i}$ for both phases can be estimated. The regression was performed using the differences between the logarithmic activity coefficients $\ln(\gamma_{1,2})$ estimated from the experimental data and the model. Since the data contained no information about the measurement errors, a standard deviation of $\sigma = 4\%$ was assumed for the activity coefficients. For the mole fractions, the standard deviation was set to 4% for values above 0.1. Below this threshold, a constant standard deviation of 0.004 was used. The parameter α was not optimized, but set to a fixed value of 0.3, since it could not be reliably estimated from this dataset.

The results are summarized in Table 2, which contains the parameter values, the numbers of needed iterations, and timing estimates obtained with the same settings as in Section 3.1. For all regression calculations, the NLPLSQ solver [25] was used with a residual guess of 10 and accuracy settings of 10^{-6} and 10^{-9} (QP).

Table 2. Regression results and statistics for the NRTL example.

Method	α	τ_{12}	τ_{21}	Iterations	Error	Time [ms]
Initial	-	1	1	-	-	-
s^A	0.3 *	−0.254835	1.10159	6	1290.34	174
s^B	0.3 *	−0.232939	1.07197	8	1132.79	58,870
s^C	0.3 *	−0.233823	1.07287	9	1127.48	820
s^{C*}	0.3 *	−0.234391	1.07364	9	1127.59	1050

*: The parameter α was not optimized.

Figure 2 shows the data together with the corresponding model predictions for all four strategies. The model curves are very similar. The maximum of the absolute differences between the model predictions evaluated at the data points was 0.013 (i.e., about 1% of the value of γ_2 close to $x_1 = 1$) and reduced to 0.003, if strategy s^A was excluded.

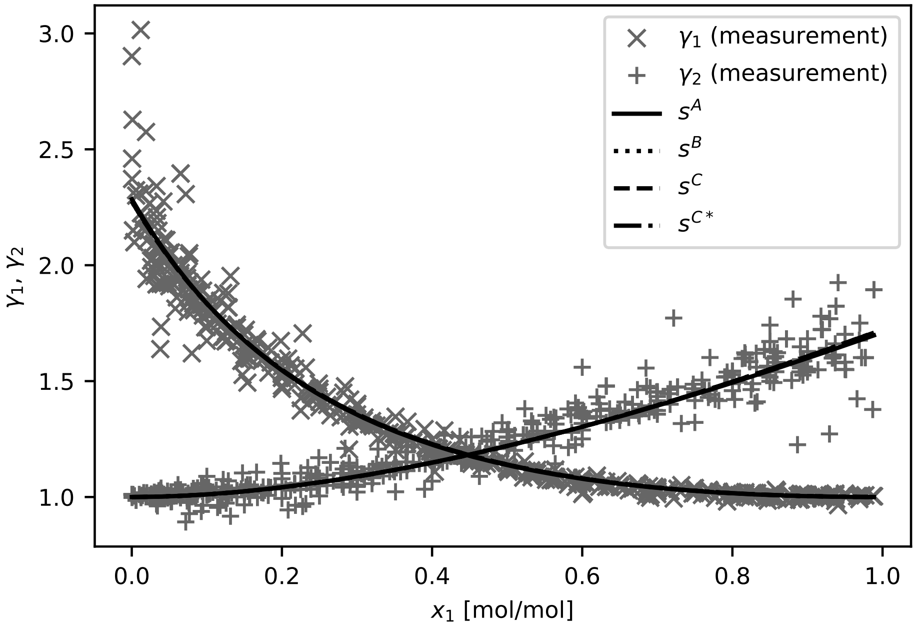


Figure 2. Comparison of activity coefficients in the system methanol (1) + water (2) for the different model parameters in Table 2, based on the investigated regression strategies.

As in the previous example, the results of all four strategies agree well with each other. While the strategies s^B, s^C, and s^{C*} use different objectives for the optimization to include the errors of input variables, the results of these three strategies can be considered identical, and even s^A was only slightly different. The effect of the input-variable errors, for which homogeneous estimate of 4% was assumed, seemed to be small in this example. In addition, the approximation used in s^C was already sufficient, such that s^{C*} did not yield an improvement. The time for all four strategies was much larger than in the first example, because the dataset consists of 540 points. This was still feasible for method s^B, but the difference in speed compared to s^C and s^{C*} was much more significant here. Therefore, the methods s^C and s^{C*} are better suited to handle cases appearing in practice that contain large datasets. Despite the large dataset, the current model was rather small, since it used only one input variable, two output variables, and two parameters. The size of the optimization problem in s^B increased with the number of input and output variables, so that much larger problems could easily occur.

In summary, all strategies to take input-variable errors into consideration agree well with each other, and differ only slightly from the ordinary least-squares strategy. However, in general, when the input errors are significant, deviations are expected. In cases where the ordinary least-squares strategy does not yield satisfactory results or it is expected that significant input variable errors are present in the data, the two other strategies are available as alternatives. The reconciliation strategy s^B is simpler than s^C and s^{C*}, but its numerical effort grows for large-size problems.

4. Multi-Criteria Parameter Estimation

The regression approaches considered thus far, in Sections 2 and 3, strongly depend on the standard deviations of the data. However, in practice, often the standard deviations of the measured data are not well known or the measuring sensors might not be reliable. Thus, it is essential to know in how far the estimated parameters depend on changes in the estimates of the variances. A solution to this problem is provided by performing parameter estimation in a multi-criteria framework. Here, one can define multiple sums of least squares as objectives to be minimized. In general, if the system is over-defined, not all objectives can be minimized simultaneously and the result is a set of Pareto optimal solutions, lying on the so-called Pareto front [26].

Exploration of the Pareto front corresponds to a sensitivity analysis of the weights of the different sums of squares. Thus, using a separate objective function for each data point, one performs a sensitivity analysis of the standard deviations. If the number of data points is large, in practice, it is more advantageous to group data points to only a few least-squares functions, e.g., according to their physical unit, order of magnitude, or the reliability of their error estimates. This kind of grouping has been proven useful in [19,20].

Furthermore, the Pareto front indicates how close the model can at all get to the data. As demonstrated via an example in Section 4.2, this reveals valuable information about systematic deviations of single measurement values from the model predictions, thus providing hints for possible gross errors. Finally, as clarified in the Introduction, a statistically rigorous estimate of the variances would, for a multivariate regression, involve solving an iterative loop between minimizing the residual sum of squares and estimating the variances. Performing this iteration is completely avoided by calculating the Pareto front, which contains all possible outcomes of the regression.

In this section, we focus on data from industrial process plants. However, the general method can be applied to arbitrary parameter estimation problems. For instance, it could be used for the examples from thermodynamics in Section 3 as well.

4.1. General Framework

In the following, we describe a general framework for multi-criteria parameter estimation using process data. Model parameters can be fitted both to measured data from one and multiple datasets (or experiments). In the context of industrial process data, multiple datasets correspond, e.g., to multiple operating points. In general, two types of model parameters appear: $\beta \in \mathbb{R}^n$ which are independent of the operating point and $x_p \in \mathbb{R}^m$, $p = 1, \ldots, N_\mathrm{P}$, which depend on the operating point, where N_P is the number of operating points. Establishing the connection to Section 2, the parameters x_p are associated with the inputs x_i in Equation (1). A single objective s_i, $i = 1, \ldots, N$ to be minimized takes the form

$$s_i = \sum_{p=1}^{N_\mathrm{P}} \sum_{j=1}^{N_{\mathrm{T}i}} \frac{1}{\sigma_{ij,p}^2} \left(f_{ij}(x_p; \beta) - y_{ij,p} \right)^2 , \tag{28}$$

where $N_{\mathrm{T}i}$ denotes the number of measured quantities (or tags) and $y_{ij,p}$ the measured value of tag j at operating point p assigned to objective i, e.g., according to the physical unit with index i. The corresponding standard deviation is denoted by $\sigma_{ij,p}$ whereas $f_{ij}(x_p; \beta)$ is the model prediction for quantity j attributed to objective i. We can define several of these objectives summarized as $s \in \mathbb{R}^N$, and the multi-criteria optimization problem to be solved reads

$$\min_{x_1, x_2, \ldots, x_{N_\mathrm{P}}, \beta} \quad s. \tag{29}$$

Pareto optimal solutions of this problem can be obtained using scalarization algorithms [27–29], which combine the objectives to a single scalar function, which is then passed to an optimization solver. In the following, we employ an adaptive method provided by the sandwiching algorithm [16,30,31].

4.2. Example: Distillation Column

We present the application of multi-criteria parameter estimation to a real-world industrial process. This example depicts a distillation column which is part of a downstream separation process, operated by LONZA AG at one of their production sites. The flowsheet is displayed in Figure 3.

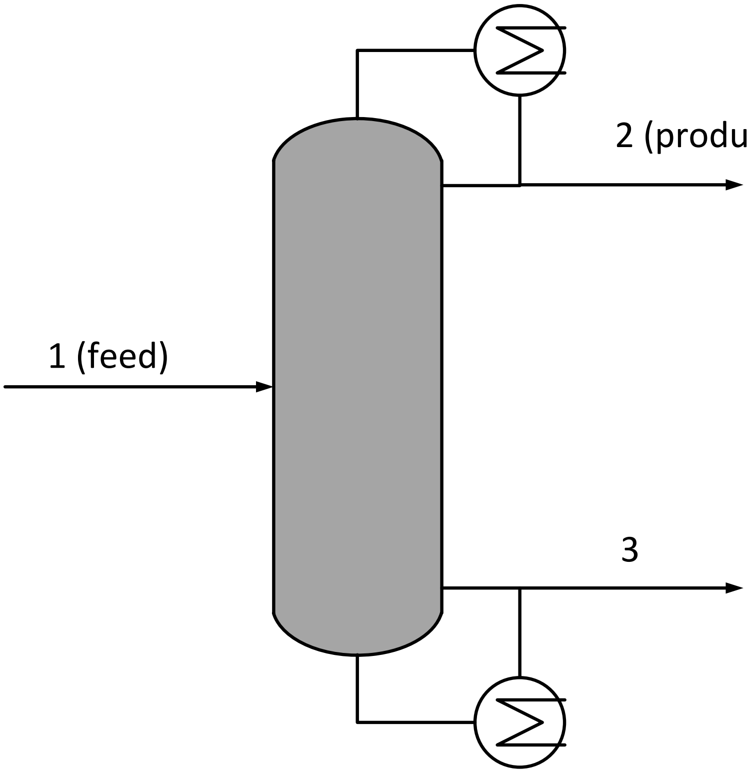


Figure 3. Flowsheet for the distillation column.

The column feed is a mixture of 15 known and some unknown components and the main product is obtained at the top drain. The process model is implemented as a steady-state simulation in the flowsheet simulator CHEMCAD. Here, the column consists of 53 stages. Available real process data are the total flow rates of the three streams 1, 2, and 3 and the temperature measurements at stages 1, 2, 19, 38, 46, 52, and 53. We adjusted the model to these data at a single operating point. The composition of the feed was measured as well and, in the simulation, was fixed to these data. It was not practical to define a separate objective for each measured quantity, since the interpretation of the results would become too involved. Therefore, we defined (only) two objectives to be minimized, one least-squares function for the flow rates and one for the temperatures:

$$\bar{f} = \sqrt{\frac{1}{3}\sum_{i=1}^{3}\frac{1}{\sigma_F^2}\left(F_i^{\text{sim}} - F_i^{\text{meas}}\right)^2}, \tag{30}$$

$$\bar{t} = \sqrt{\frac{1}{7}\sum_{i=1}^{7}\frac{1}{\sigma_T^2}\left(T_i^{\text{sim}} - T_i^{\text{meas}}\right)^2}. \tag{31}$$

Since reliable estimates for the standard deviations are not available, we set them to the default values $\sigma_F = 1$ kg/h and $\sigma_T = 1$ °C. The so-defined objectives can then be interpreted as the average deviations between simulated and measured values given in units of standard deviations. The model parameters to be adjusted are the total feed flow rate, the reflux ratio, the reboiler specification (given by the mass fraction of the main product in the bottom drain), and a global value for the tray efficiencies of all stages. Those optimization variables are summarized in Table 3.

Table 3. Model parameters to be fitted to process data.

Parameter	Initial Value	Lower Bound	Upper Bound
Total feed flow rate [kg/h]	2330	2300	2550
Reflux ratio [-]	1.4	0.5	3
Reboiler specification [-]	0.0046	0.003	0.006
Tray efficiency [-]	0.7	0.5	0.95

The multi-criteria optimization problem was solved using the sandwiching algorithm [16,30,31] together with the NLPQLP solver [32]. We obtained four Pareto points and the interpolated Pareto front is shown in Figure 4.

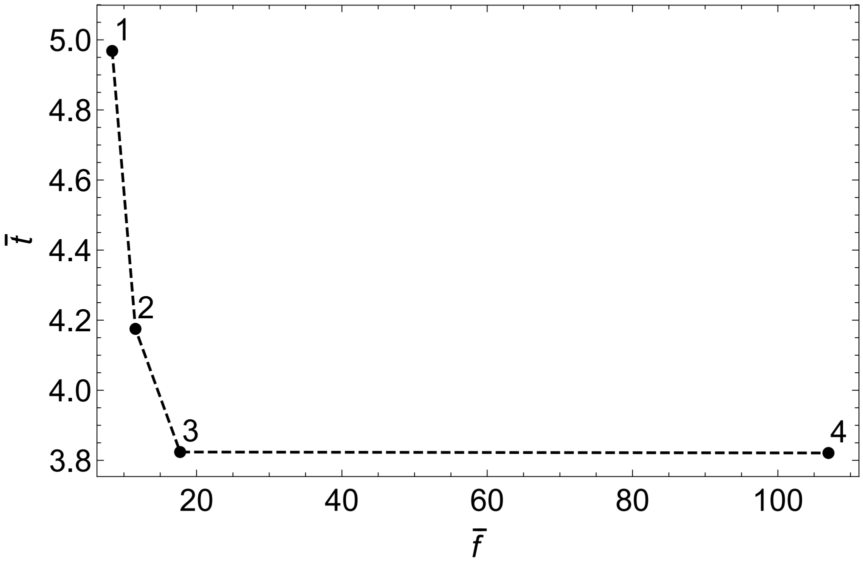


Figure 4. Calculated Pareto front for the two objectives $\bar{f}$ and $\bar{t}$. The objectives can be interpreted as the average deviations between simulated and measured values given in units of standard deviations.

One can clearly see that it is not possible to minimize both objectives simultaneously. Navigation on the Pareto front shows the dependence of the resulting parameters on variations of the variances and allows then to choose a set of model parameters which fit either the flow rates or the temperatures better. This decision is beyond the scope of this work, since it has to be made relying on context knowledge and is therefore up to the engineer who knows which measured data are more reliable and should be described better. To study the contributions of the different measured points to the objectives, the residuals at the Pareto points are displayed in Figures 5 and 6.

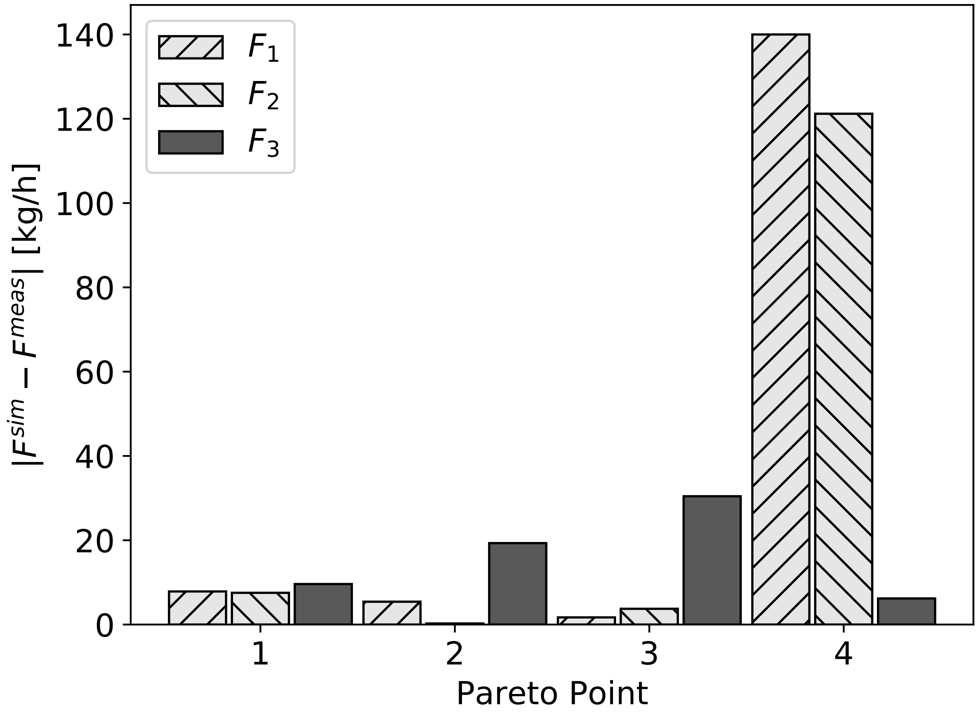


Figure 5. Residuals for the flow rates at the calculated Pareto points.

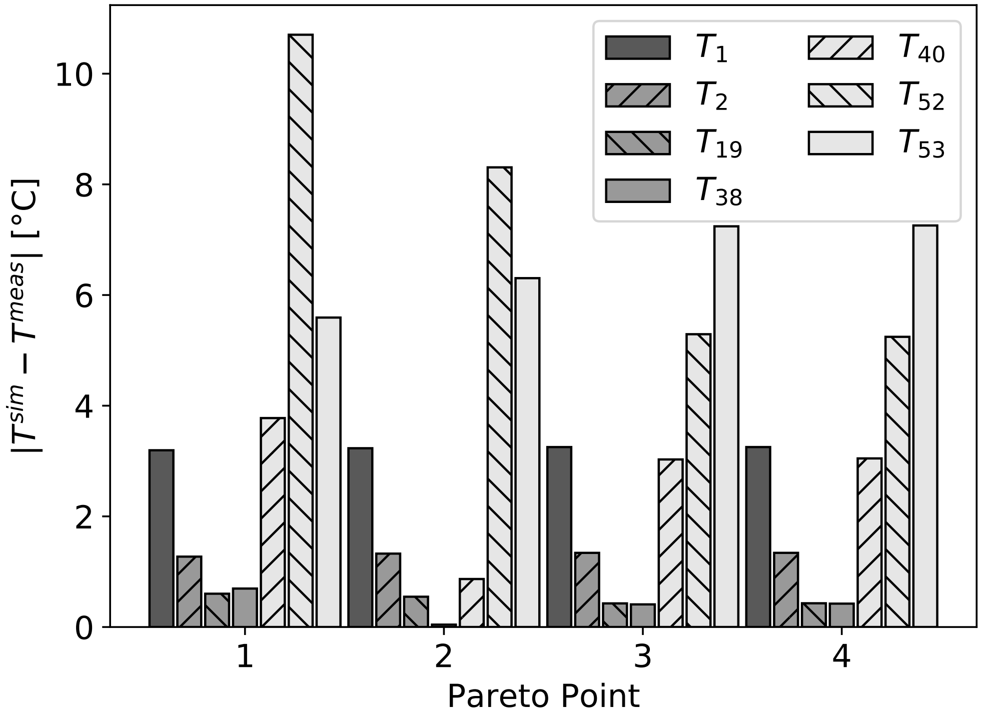


Figure 6. Residuals for the temperatures at the calculated Pareto points.

For all four Pareto points, the residuals of the temperatures showed systematic deviations for T_{52} and T_{53} at the bottom of the column. They were in agreement with the observation that the corresponding temperature sensors at the real column did not work well and might have provided erroneous results. Thus, our framework is well suited for the detection of such systematic deviations.

5. Error Estimates

Having discussed how model parameters can be estimated from data, we now present strategies to quantify the errors of the obtained parameter values and corresponding model predictions [23]. Well-known error estimates are the parameter confidence regions, the parameter confidence intervals, and the prediction bands. They are derived from a quadratic approximation of the least-squares function $s^{A,B,C}$ around the optimal parameter values $\hat{\beta}$. At this point, the linear contribution to the approximation vanishes. From the quadratic contribution, the parameter covariance matrix C is obtained, which is a local measure of the change in the model under variations of the parameters. The parameter covariance matrix is the inverse of the Fisher information matrix (FIM), i.e., $C = (\mathrm{FIM})^{-1}$, which is calculated by

$$\mathrm{FIM} = \frac{N - N_\beta}{s} \sum_{i=1}^{N} \left(\frac{\partial s_i}{\partial \beta} \right) \left(\frac{\partial s_i}{\partial \beta} \right)^T, \tag{32}$$

where N and N_β are the number of data points and model parameters, respectively. Here, s can be any of $s^{A,B,C}$ defined in Equations (3) and (5), or Equation (17), and the s_i are the individual terms such that $s = \sum_i s_i^2$. A more general form of the FIM for multiple outputs and differently structured data points is given in Section 5.2. Before that, a short review of error estimates is given.

5.1. Confidence-Region Estimates

The parameter confidence region is a set in parameter space which contains the true parameter values with a certain probability, the so-called confidence level α, e.g., 95%. It can be obtained by statistical theory under the assumption that the data used for the regression are obtained from the model with the true parameter values and an error which is independent and normally distributed (see Section 2). The parameter confidence region is then given by [23]

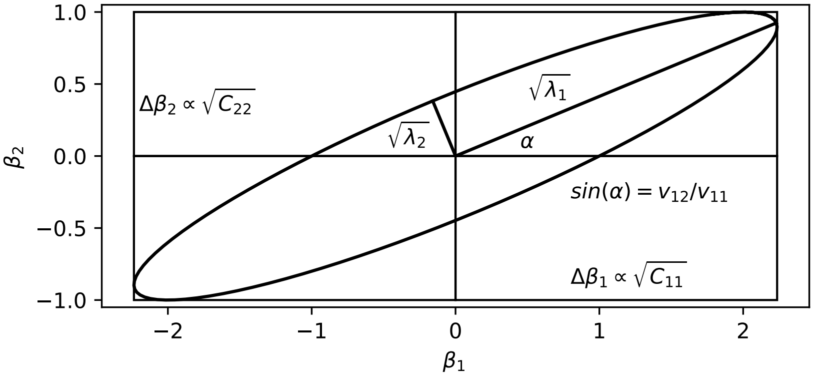
$$(\beta - \hat{\beta})^T C^{-1} (\beta - \hat{\beta}) \leq \frac{N_\beta}{N - N_\beta} F(N_\beta, N - N_\beta; \alpha), \tag{33}$$

where $F(...;\alpha)$ is the corresponding quantile of the FisherF distribution. This inequality describes an ellipsoid centered at $\hat{\beta}$, with main axes along the eigenvectors of C and proportional to the square root of the corresponding eigenvalue. The orientation of the ellipsoid in space depends on the off-diagonal elements of C.

The confidence intervals [23] provide the same information as the confidence region but for the individual parameters. They are the projections of this ellipsoid onto the coordinate axes, multiplied by the quantile of the StudentT distribution $T(...;\alpha)$ instead of the FischerF distribution

$$\Delta\beta_i = \sqrt{C_{ii}} \frac{1}{N - N_\beta} T(N - N_\beta; \alpha/2). \tag{34}$$

The confidence region and confidence intervals are illustrated in Figure 7 for a two dimensional example.



1.0
0.5
0.0
−0.5
−1.0
β_2
−2
−1
0
1
2
β_1
$\Delta\beta_2 \propto \sqrt{C_{22}}$
$\sqrt{\lambda_2}$
$\sqrt{\lambda_1}$
α
$sin(\alpha) = v_{12}/v_{11}$
$\Delta\beta_1 \propto \sqrt{C_{11}}$

Figure 7. Illustration of confidence region and confidence intervals for a 2D example. λ_1, λ_2 are the eigenvalues of the covariance matrix C. The confidence region is an ellipse, and the confidence intervals are the projections of that ellipse onto the individual coordinate axes.

In the next step, based on the parameter errors, we can determine the error of the model evaluation. The estimate for the error of the model prediction obtained from the confidence region is the prediction band [23] (also called confidence band). It provides error estimates for the evaluation of the model at arbitrary points

$$\Delta f(x;\beta) = \frac{N_\beta}{N - N_\beta} F(N_\beta, N - N_\beta; \alpha) \sqrt{\frac{\partial f}{\partial \beta}^T C \frac{\partial f}{\partial \beta}}. \tag{35}$$

It is also possible to calculate error estimates based on the confidence intervals of the individual parameters. If the covariance matrix C is (close to) diagonal, i.e., the parameters are uncorrelated, Equation (35) simplifies to

$$\Delta f(x;\beta_i) = \sqrt{\sum_{i=1}^{N_\beta} \left(\Delta\beta_i \frac{\partial f}{\partial \beta_i}\right)^2}. \tag{36}$$

On the other hand, if the covariance matrix cannot be assumed to be diagonal, an upper bound for the error is given by

$$\Delta f(x;\beta_i) = \sum_{i=1}^{N_\beta} |\Delta\beta_i \frac{\partial f}{\partial \beta_i}|, \tag{37}$$

because the off-diagonal elements of the covariance matrix are always bounded by $C_{ij} \leq \sqrt{C_{ii}C_{jj}}$. Equations (36) and (37) are the typical formulas for error propagation. These estimates, however, are very conservative in the case of large parameter correlations. The confidence intervals are practically the projections of the confidence region onto the coordinate axes, since the quantiles in Equations (33) and (34) differ only slightly numerically. Therefore, this approach corresponds to the approximation of an ellipsoid by its bounding box (aligned to the coordinate system). For large parameter correlations, the confidence region becomes more elongated and tilted away from the coordinate axes. In Figure 7, the area of the bounding box is much larger than the area of the ellipse itself. The behavior gets more pronounced in higher dimensions. This effect is demonstrated in Figure 8 for the DIPPR model described in Section 3.1 with three parameters. The prediction error band (Equation (35)) is magnified ten times to be even visible, whereas the error propagation result (Equations (34) and (36)) is shrunk by the same factor to lie inside the plot. The error estimate based on the confidence intervals is more than 100 times larger than the prediction band. As a result, for error propagation of correlated models, the prediction band method [23] should always be preferred over the simple propagation of confidence intervals.

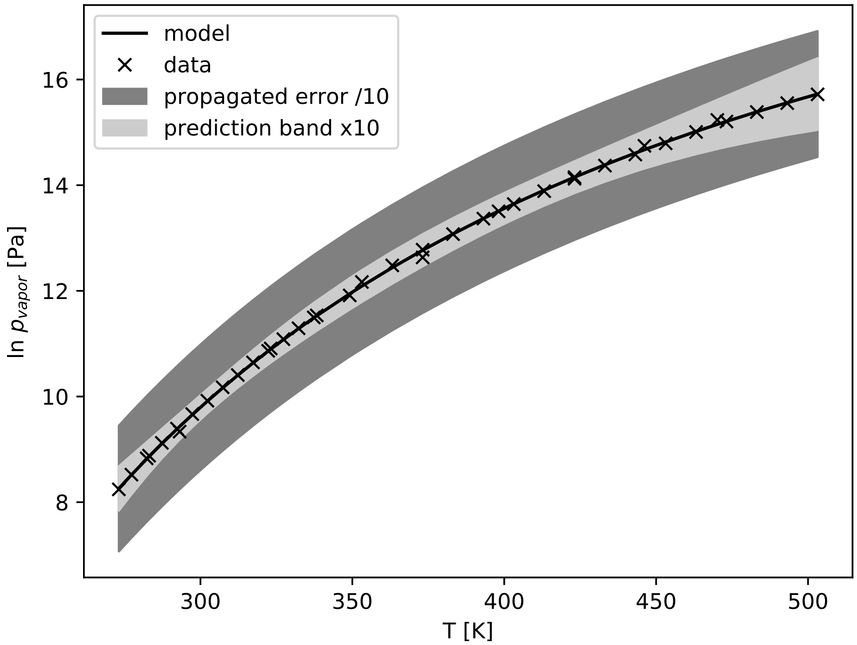


Figure 8. Errors of the model prediction based on regression results for the vapor pressure of methanol in the ordinary least-squares approach. The shaded areas depict the 95% confidence regions for the model prediction. Note that they are scaled by a factor of 10 in different directions for visibility reasons.

However, the error propagation formulas in Equations (36) and (37) are useful, when various models are employed in one calculation. Different models are usually adjusted using different datasets and by multiple independent regression procedures. Assume, that f, g, and h are models for three different properties of a system, e.g., vapor pressure, heat capacity, and activity coefficients for a binary mixture. Then, there may be a derived property that depends on all three of them $F = F(f, g, h)$, and the error of F resulting from all three parameter estimations is given by

$$\Delta F = \sqrt{\left(\Delta f \frac{\partial F}{\partial f}\right)^2 + \left(\Delta g \frac{\partial F}{\partial g}\right)^2 + \left(\Delta h \frac{\partial F}{\partial h}\right)^2}. \tag{38}$$

If the parameters are correlated, Equation (37) can be applied. Correlations are expected, e.g., if f is used in the parameter estimation for g or h, e.g., to transform some of the input data. It is also possible to derive corresponding equations for cases where only some pairs are uncorrelated, by considering a covariance matrix where some off-diagonal elements are non-zero.

5.2. General Form of the Fisher Information Matrix

Thus far, the formula for the Fisher information matrix given in Equation (32) is provided only for the single output case. However, it can be easily extended to the multiple output case, as long as every data point contains values for the output quantities. In process engineering, often multiple thermodynamic models are employed in a single calculation. We, therefore, present a strategy for treating such generalized situations in the parameter estimation process. When several inhomogeneous datasets are involved, this procedure is similar to missing-data problems [33]. However, this approach does not aim to replace the missing data, but to use as many of the available data as possible.

First, consider M models $f_i(X_i^j;\beta)$, where X_i^j is the subset of input quantities that are used by the ith model measured at the jth data point. Models with multiple outputs are treated as multiple independent single-output models which have the same set of input quantities. Then, let K_i be the subset of $\{1,...,N\}$ which contains all those data points, where all of the X_i^j and y_j^i are measured, or equivalently the set of data points that are useful for estimating the parameters of model i. Then, the total model for the parameter estimation reads

$$y_i^j = f_i(X_i^j;\beta) + \epsilon_i^j \quad \forall\, j \in K_i,\ i \in \{1,...,M\}. \tag{39}$$

The regression strategies given in Section 2 can be generalized to this model. To that end, the numbers of inputs (N_x) and outputs (N_y) become dependent on the data point, but the equations provided are independent of this fact. The generalization of the Fisher matrix reads

$$F_{kl} = \sum_{i=1}^{M} \sum_{j\in K_i} \frac{1}{\sigma_j^{(i)2}} \frac{\partial f_i(X_i^j;\beta)}{\partial \beta_k} \frac{\partial f_i(X_i^j;\beta)}{\partial \beta_l}, \tag{40}$$

where $\sigma_j^{(i)}$ denotes the corresponding standard deviation. A derivation based on probability theory is contained in Appendix C.

Equation (40) is fully additive with respect to the model functions and the data points. However, the formula contains the standard deviations of the measurements, which are rarely well-known. For the parameter estimation, approximate values might be sufficient in cases where the optimal parameters are not very sensitive to the standard deviations. For the error estimation, however, that is no longer true. Therefore, it is popular to estimate a common standard deviation σ from the residual of the objective function, $\sigma = \frac{s}{N-N_\beta}$, as done in Equation (32). If the standard deviations of the dataset depend on the data points, then only a global scaling factor for all of them is estimated. The standard deviations $\sigma_j^{(i)}$ in Equation (40), however, depend on the data point j and the model i. If just a single correction factor is estimated, the additivity of the Fisher matrix is violated, and the dependence of that factor on N and N_β is not clear. For this reason, the re-estimation of the standard deviation is performed separately for each output using the corresponding number of data points and parameters.

Scaling the whole objective function by one (positive) scalar has no impact on the minimum value. This is no longer the case when different terms of the objective function are scaled by different factors, as in Equation (40). The correct approach would be to iteratively solve for optimal parameters and estimate the standard deviations until self-consistency is reached. Assuming that the new estimates of the individual output errors do not change their ratios by large amounts, we use the first iteration as an approximation to the self-consistent one.

Another advantage of this approach is that the simultaneous regression of disjoint models on disjoint datasets yields the same result as two independent calculations.

6. Conclusions

Different strategies to address parameter estimation problems that are relevant to applications in process engineering in general and especially in thermodynamics have been discussed. The standard least-squares approach is useful for simple situations with a univariate regression problem and measurement errors in the outputs only, not in the inputs. Measurement errors in the inputs, however, lead to constrained least-squares problems, where the number of equality constraints grows with the number of measurement points. In this situation, the Patino–Leal formulation is a valuable alternative, since it transforms the constrained least-squares problem to a non-restricted optimization problem. We applied both methods to thermodynamic regression examples.

Furthermore, we demonstrated that, for multivariate parameter estimation or reconciliation problems, a multi-criteria approach yields substantial insight into how well a model can describe multiple responses at all. To the best of our knowledge, this is the first time that the challenge of reconciling process variables was treated within a multi-criteria setting. All the described algorithms were implemented with the industrial partners.

A direction for further research is, e.g., to realize the propagation of the parameter errors from thermodynamic models all the way up to a full flowsheet simulation of a real-world example. In addition, benchmarking of the various parameter estimation strategies should be performed applying them to more complex models.

Author Contributions: Conceptualization, J.H., P.B., and M.B.; methodology, J.H., P.B., and H.-M.L.; software, J.H., P.B., P.S., and M.v.K.; resources, N.A., S.B., J.K., H.-M.L., and N.K.; writing—original draft preparation, J.H. and P.B.; writing—review and editing, J.H., P.B., M.B., J.K., and N.A.; supervision and project administration, M.B.

Funding: This research received no external funding.

Acknowledgments: This work was developed in the Fraunhofer Cluster of Excellence “Cognitive Internet Technologies”. The authors thank Jan Schwientek for helpful discussions.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Patino–Leal in Least-Squares Form

The optimization problem given in Equation (15) is not a least-squares problem, but it can be reformulated as such. To that end, each term of the sum is considered separately, so the index i can be dropped. Furthermore, we consider only the case $\xi = z$, and the general case is obtained by replacing g by $g(\xi^k, \beta) + B(z - \xi)$.

Inserting the QR decomposition of the matrix $D^T B^T = QR$ (with $DD^T = V$) into an arbitrary term of the sum in Equation (15) yields

$$g^T (BVB^T)^{-1} g = g^T (BDD^T B^T)^{-1} g = g^T (R^T Q^T QR)^{-1} g = g^T (R^T R)^{-1} g. \tag{A1}$$

The product of an inverse matrix with a vector, here $z_1 = (R^T R)^{-1} g$, is calculated by solving a linear system of equations

$$R^T R z_1 = g. \tag{A2}$$

The solution of a system of linear equations with a matrix which is the product of two triangular matrices (R and R^T) is straightforward

$$R^T z_2 = g, \tag{A3}$$

$$R z_1 = z_2. \tag{A4}$$

Equation (A2) is well formed (if R has full rank), Equation (A3) is under-determined, and Equation (A4) is over-determined. Thus, when solving Equation (A3), one has to choose the right solution to make Equation (A4) solvable. Fortunately, the corresponding solution is the one with zeros in the last N_x components, which is also the simplest one to calculate. Inserting into Equation (A1) yields

$$z_1^T g = z_1^T R^T R z_1 = z_2^T z_2. \tag{A5}$$

This is already a least-squares formulation of Equation (15), but it is advantageous to insert $Q^T Q$ in between (although the number of terms in the objective function is increased). The reason for this step is the fact that the QR decomposition of a matrix is not unique and the signs of the diagonal of R can be chosen freely. By inserting Q^TQ, the least-squares residuals become independent of that choice, which might be important for convergence or if numerical gradients are used. Equation (A5) can be used for a linearization point $\xi \neq z$ as well, and the same QR decomposition can be employed to solve the system of equations belonging to the matrix inversion in Equation (16).

Appendix B. Reformulated Exact Version

With the results of Appendix A, Equation (16) reads

$$\xi^{k+1} = z - DQR(R^TR)^{-1}\left(g + B(z - \xi^k)\right), \tag{A6}$$

where g, B (including Q, R) are evaluated at ξ^k. Thus, we need to solve

$$R^TRz_1 = g + B(z - \xi^k), \tag{A7}$$

which is done in two steps, as in Appendix A. In addition, one can insert the decomposition of B to simplify the equation. Then, the first triangular equation becomes

$$R^Tz_2 = g + R^TQ^TD^{-1}(z - \xi^k), \tag{A8}$$

and we collect the terms starting with R^T on the left-hand side

$$R^T(z_2 - Q^TD^{-1}(z - \xi^k)) = R^Tz_3 = g. \tag{A9}$$

For z_3, the last N_x entries are no longer equal to zero. This is addressed below. First, insert the identity $Q^TQ = I$:

$$R^TQ^TQz_3 = BDQz_3 = BDz_4 = Bz_5 = g, \tag{A10}$$

where

$$z_5 = Dz_4 = DQz_3 = DQz_2 - z + \xi^k. \tag{A11}$$

Thus, the solution for z_5 can be used to obtain z_2 and then solve $Rz_1 = z_2$. Note that Equation (A10) does not uniquely define z_5 and a second equation is needed, which addressed below. However, it is not necessary to solve for z_1, since Equation (A6) only requires $DQRz_1$,

$$DQRz_1 = DQz_2 = z_5 + z - \xi^k, \tag{A12}$$

and then Equation (A6) simplifies to

$$\xi^{k+1} = z - (z_5 + z - \xi^k) = -z_5 + \xi^k. \tag{A13}$$

Furthermore, the least-squares formulation is readily available, because now the definition of z_2 is equivalent to the one in Equation (A3) (except for the term on the right-hand side, which was assumed to be zero there). Thus,

$$Qz_2 = D^{-1}(z_5 + z - \xi^k) = D^{-1}(z - \xi^{k+1}), \tag{A14}$$

and the objective function in least-squares form is obtained inserting z_5

$$\begin{aligned} z_2^T z_2 &= (Q^T D^{-1}(z_5 + z - \xi))^T (Q^T D^{-1}(z_5 - z + \xi)) && \text{(A15)} \\ &= (z_5 - z + \xi)^T D^{-T} Q Q^T D^{-1} (z_5 - z + \xi) && \text{(A16)} \\ &= (z_5 - z + \xi)^T V^{-1} (z_5 - z + \xi). && \text{(A17)} \end{aligned}$$

Note that neither Q nor R is needed explicitly, since they are only auxiliary matrices to solve Equation (A10). The only point left open is to make z_5 unique. The matrix B is in general a $n \times m$ matrix, where n is equal to the number of model equations or outputs, and m equals the dimension of z (i.e., the number of inputs plus the number of outputs). Equation (A10) must be solved in a way that $Rz_1 = z_2$ is still solvable (even though the solution is not explicitly used), which is the case if and only if all components of z_2 beyond n are zero

$$\left(Q^T D^{-1}(z_5 + (z - \xi)) \right)_i = 0 \quad \forall\, i = n+1, ..., m. \tag{A18}$$

The first n entries of this vector are uniquely determined by Equation (A10), and among all possible vectors the one with zeros in the last components has the smallest norm. Furthermore, multiplication by an orthogonal matrix does not change the norm. Therefore, Equation (A10) together with

$$||D^{-1}(z_5 + (z - \xi))|| \to \min \tag{A19}$$

forms a linear least-squares problem with equality constraints. This is the alternative iteration formulation given in Equations (20)–(22).

Appendix C. Derivation of the Fisher Matrix

In probability theory, the Fisher information matrix is defined by the following expectation value

$$F_{nm} = E\left[\frac{\partial \ln p}{\partial \beta_n} \frac{\partial \ln p}{\partial \beta_m} \right], \tag{A20}$$

where the probability density p for the entire dataset (under the assumption of independent data points) reads

$$p(y|x) = \prod_{j=1}^{N} p_j(y^j|x^j). \tag{A21}$$

Here, the probability density p_j to obtain data point j given x^j (assuming that the errors follow a normal distribution) is given by

$$p_j(y^j|x^j) = \prod_{i \in \bar{K}_j} \frac{1}{\sqrt{2\pi}\sigma_j^{(i)}} e^{(y_i^j - f_i(X_i^j;\beta))^2/(2\sigma_j^{(i)2})}, \tag{A22}$$

where $\bar{K}_j \subseteq \{1, ...M\}$ is the set of all models that can contribute to the point j and is connected to K_i by the relation $\{(i,j)|\, i \in \bar{K}_j,\ j = 1, ..., N\} = \{(i,j)|\, j \in K_i,\ i = 1, ..., M\}$. The sets K_i are fixed, i.e., the selection of measured quantities is fixed and not randomly distributed. This leads to

$$\frac{\partial \ln p}{\partial \beta_k} = \sum_{i=1}^{M} \sum_{j \in K_i} \frac{y_i^j - f_i(X_i^j;\beta)}{2\sigma_j^{i2}} \frac{\partial f_i^j(X_i^j;\beta)}{\partial \beta_k}, \tag{A23}$$

and

$$\begin{aligned} F_{nm} &= E\left[\left(\sum_{i=1}^{M}\sum_{j\in K_i}\frac{y_i^j - f_i(X_i^j;\beta)}{\sigma_j^{i2}}\frac{\partial f_i(X_i^j;\beta)}{\partial \beta_n}\right)\right. \\ &\qquad \left.\times\left(\sum_{k=1}^{M}\sum_{l\in K_k}\frac{y_k^l - f_k(X_k^l;\beta)}{\sigma_l^{k2}}\frac{\partial f_k(X_k^l;\beta)}{\partial \beta_m}\right)\right]. \end{aligned} \tag{A24}$$

Due to the linearity of the expectation value

$$\begin{aligned} F_{nm} = \sum_{i=1}^{M}\sum_{k=1}^{M}\sum_{j\in K_i}\sum_{l\in K_k} E\left[\left(\frac{y_i^j - f_i(X_i^j;\beta)}{\sigma_j^i}\right)\left(\frac{y_k^l - f_k(X_k^l;\beta)}{\sigma_l^k}\right)\right] \\ \times\frac{\partial f_i(X_i^j;\beta)}{\partial \beta_n}\frac{\partial f_k(X_k^l;\beta)}{\partial \beta_m}\frac{1}{\sigma_j^i\sigma_l^k}. \end{aligned} \tag{A25}$$

The expectation value term is the correlation matrix of the measurement errors of the individual data points, and is the identity matrix in cases where the measurement errors are uncorrelated. However, if the dataset does not consist of direct measurement results but of values calculated thereof, the correlations should be included. Under the assumption of independent measurement errors, Equation (A25) simplifies to

$$F_{kl} = \sum_{i=1}^{M}\sum_{j\in K_i}\frac{1}{\sigma_j^{i2}}\frac{\partial f_i(X_i^j;\beta)}{\partial \beta_k}\frac{f_i(X_i^j;\beta)}{\partial \beta_l}. \tag{A26}$$

References

1. Fahrmeir, L.; Kneib, T.; Lang, S. *Regression*; Springer: Berlin/Heidelberg, Germany, 2007. [CrossRef]
2. Peter, J.; Rousseeuw, A.M.L. *Robust Regression and Outlier Detection*; John Wiley & Sons: Hoboken, NJ, USA, 2003.
3. Lätgering-Lin, O.; Schäniger, A.; Nowak, W.; Gross, J. Bayesian Model Selection Helps To Choose Objectively between Thermodynamic Models: A Demonstration of Selecting a Viscosity Model Based on Entropy Scaling. *Ind. Eng. Chem. Res.* **2016**, *55*, 10191–10207. [CrossRef]
4. Von Toussaint, U. Bayesian inference in physics. *Rev. Mod. Phys.* **2011**, *83*, 943–999. [CrossRef]
5. Kravaris, C.; Hahn, J.; Chu, Y. Advances and selected recent developments in state and parameter estimation. *Comput. Chem. Eng.* **2013**, *51*, 111–123. [CrossRef]
6. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin, Germany, 2006.
7. Renon, H.; Prausnitz, J.M. Local compositions in thermodynamic excess functions for liquid mixtures. *AIChE J.* **1968**, *14*, 135–144. [CrossRef]
8. Li, Z.; Mumford, K.A.; Shang, Y.; Smith, K.H.; Chen, J.; Wang, Y.; Stevens, G.W. Analysis of the Nonrandom Two-Liquid Model for Prediction of Liquid–liquid Equilibria. *J. Chem. Eng. Data* **2014**, *59*, 2485–2489. [CrossRef]
9. Gross, J.; Sadowski, G. Application of perturbation theory to a hard-chain reference fluid: An equation of state for square-well chains. *Fluid Phase Equilib.* **2000**, *168*, 183–199. [CrossRef]
10. Gross, J.; Sadowski, G. Perturbed-Chain SAFT: An Equation of State Based on a Perturbation Theory for Chain Molecules. *Ind. Eng. Chem. Res.* **2001**, *40*, 1244–1260. [CrossRef]
11. Kontogeorgis, G.M.; Folas, G.K. *Thermodynamic Models for Industrial Applications: From Classical and Advanced Mixing Rules to Association Theories*; John Wiley & Sons: Hoboken, NJ, USA, 2009. [CrossRef]
12. Narasimhan, S.; Jordache, C. *Data Reconciliation & Gross Error Detection*; Gulf Publishing Company: Houston, TX, USA, 2000.
13. Arora, N.; Biegler, L. Redescending estimators for data reconciliation and parameter estimation. *Comput. Chem. Eng.* **2001**, *25*, 1585–1599. [CrossRef]

14. Reilly, P.M.; Patino-Leal, H. A Bayesian Study of the Error-in-Variables Model. *Technometrics* **1981**, *23*, 221–231. [CrossRef]
15. Patino-Leal, H.; Reilly, P.M. Statistical estimation of parameters in vapor-liquid equilibrium. *AIChE J.* **1982**, *28*, 580–587. [CrossRef]
16. Bortz, M.; Burger, J.; Asprion, N.; Blagov, S.; Böttcher, R.; Nowak, U.; Scheithauer, A.; Welke, R.; Küfer, K.H.; Hasse, H. Multi-criteria optimization in chemical process design and decision support by navigation on Pareto sets. *Comput. Chem. Eng.* **2014**, *60*, 354–363. [CrossRef]
17. Burger, J.; Asprion, N.; Blagov, S.; Bortz, M. Simple Perturbation Scheme to Consider Uncertainty in Equations of State for the Use in Process Simulation. *J. Chem. Eng. Data* **2017**, *62*, 268–274. [CrossRef]
18. Forte, E.; Jirasek, F.; Bortz, M.; Burger, J.; Vrabec, J.; Hasse, H. Digitalization in Thermodynamics. *Chem. Ing. Tech.* **2019**, *91*, 201–214. [CrossRef]
19. Stöbener, K.; Klein, P.; Horsch, M.; Küfer, K.; Hasse, H. Parametrization of two-center Lennard-Jones plus point-quadrupole force field models by multicriteria optimization. *Fluid Phase Equilib.* **2016**, *411*, 33–42. [CrossRef]
20. Forte, E.; Burger, J.; Langenbach, K.; Hasse, H.; Bortz, M. Multi-criteria optimization for parameterization of SAFT-type equations of state for water. *AIChE J.* **2018**, *64*, 226–237. [CrossRef]
21. López, C.; Diana, C.; Barz, T.; Peñuela, M.; Villegas, A.; Ochoa, S.; Wozny, G. Model-based identifiable parameter determination applied to a simultaneous saccharification and fermentation process model for bio-ethanol production. *Biotechnol. Prog.* **2013**, *29*, 1064–1082. [CrossRef] [PubMed]
22. Müller, D.; Esche, E.; López, C.; Diana, C.; Wozny, G. An algorithm for the identification and estimation of relevant parameters for optimization under uncertainty. *Comput. Chem. Eng.* **2014**, *71*, 94–103. [CrossRef]
23. Bates, D.M.; Watts, D.G. *Nonlinear Regression Analysis and Its Applications*; Wiley: Weinheim, Germany, 2007.
24. Spedicato, E.; Vespucci, M.T. Numerical experiments with variations of the Gauss-Newton algorithm for nonlinear least squares. *J. Optim. Theory Appl.* **1988**, *57*, 323–339. [CrossRef]
25. Schittkowski, K. *NLPLSQ: A Fortran Implementation of An SQP-Gauss-Newton Algorithm for Least Squares Optimization*; User Documentation; Department of Computer Science, University of Bayreuth: Bayreuth, Germany, 2007.
26. Geoffrion, A.M. Proper efficiency and the theory of vector maximization. *J. Math. Anal. Appl.* **1968**, *22*, 618–630. [CrossRef]
27. Miettinen, K.M. *Nonlinear Multiobjective Optimization*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1999.
28. Hugo, A.; Ciumei, C.; Buxton, A.; Pistikopoulos, E.N. Environmental impact minimization through material substitution: A multi-objective optimization approach. *Green Chem.* **2004**, *6*, 407–417. [CrossRef]
29. Ehrgott, M. *Multicriteria Optimization*; Springer: Berlin, Germany, 2005.
30. Klamroth, K.; Tind, J.; Wiecek, M.M. Unbiased approximation in multicriteria optimization. *Math. Methods Oper. Res.* **2003**, *56*, 413–437. [CrossRef]
31. Hernandez, J.I.S. Multi-objective optimization in Mixed Integer Problems with application to the Beam Selection Optimization Problem in IMRT. Ph.D. Thesis, Technical University Kaiserslautern, Kaiserslautern, Germany, 2012.
32. Schittkowski, K. *NLPQLP: A Fortran Implementation of a Sequential Quadratic Programming Algorithm with Distributed and Non-Monotone Line Search*; User Documentation; Department of Computer Science, University of Bayreuth: Bayreuth, Germany, 2010.
33. Schmitt, P.; Mandel, J.; Guedj, M. A Comparison of Six Methods for Missing Data Imputation. *J. Biom. Biostat.* **2015**. [CrossRef]