mathematics

MDPI

Article

Unbiased Least-Squares Modelling

Marta Gatto and Fabio Marcuzzi *

Department of Mathematics, "Tullio Levi Civita", University of Padova, Via Trieste 63, 35131 Padova, Italy; mgatto@math.unipd.it

* Correspondence: marcuzzi@math.unipd.it

Received: 25 May 2020; Accepted: 11 June 2020; Published: 16 June 2020

Abstract: In this paper we analyze the bias in a general linear least-squares parameter estimation problem, when it is caused by deterministic variables that have not been included in the model. We propose a method to substantially reduce this bias, under the hypothesis that some a-priori information on the magnitude of the modelled and unmodelled components of the model is known. We call this method Unbiased Least-Squares (ULS) parameter estimation and present here its essential properties and some numerical results on an applied example.

Keywords: parameter estimation; physical modelling; oblique decomposition; least-squares

1. Introduction

The well known least-squares problem [1], very often used to estimate the parameters of a mathematical model, assumes an equivalence between a matrix-vector product Ax on the left, and a vector b on the right hand side: the matrix A is produced by the true model equations, evaluated at some operating conditions, the vector x contains the unknown parameters and the vector b are measurements, corrupted by white, Gaussian noise. This equivalence cannot be satisfied exactly, but the least-squares solution yields a minimum variance, maximum likelihood estimate of the parameters x, with a nice geometric interpretation: the resulting predictions Ax are at the minimum Euclidean distance from the true measurements b and the vector of residuals is orthogonal w.r.t. the subspace of all possible predictions.

Unfortunately, each violation of these assumptions produces in general a bias in the estimates. Various modifications have been introduced in the literature to cope with some of them: mainly, colored noise on b and/or A due to model error and/or colored measurement noise. The model error is often assumed as an additive stochastic term in the model, e.g., error-in-variables [2,3], with consequent solution methods like Total Least-Squares [4] and Extended Least-Squares [5], to cite a few. All these techniques let the model to be modified to describe, in some sense, the model error.

Here, instead, we assume that the model error depends from deterministic variables in a way that has not been included in the model, i.e., we suppose to use a reduced model of the real system, as it is often the case in applications. In this paper we propose a method to cope with the bias in the parameter estimates of the approximate model by exploiting the geometric properties of least-squares and using small additional a-priori information about the norm of the modelled and un-modelled components of the system response, available with some approximation in most applications. To eliminate the bias on the parameter estimates we perturb the right-hand-side without modifying the reduced model, since we assume it describes accurately one part of the true model.

2. Model Problem

In applied mathematics, physical models are often available, usually rather precise at describing quantitatively the main phenomena, but not satisfactory at the level of detail required by the application at hand. Here we refer to models described by differential equations, with ordinary and/or partial

derivatives, commonly used in engineering and applied sciences. We assume, therefore, that there are two models at hand: a true, unknown model $\mathcal{M}$ and an approximate, known model $\mathcal{M}_a$. These models are usually parametric and they must be tuned to describe a specific physical system, using a-priori knowledge about the application and experimental measurements. Model tuning, and in particular parameter estimation, is usually done with a prediction error minimization criterion that makes the model response to be a good approximation of the dynamics shown by the measured variables used in the estimation process. Assuming that the true model $\mathcal{M}$ is linear in the parameters that must be estimated, the application of this criterion brings to a linear least-squares problem:

$$\bar{x} = \underset{x' \in \mathbb{R}^n}{\operatorname{argmin}} \|Ax' - \bar{f}\|^2, \tag{1}$$

where, from here on, $\|\cdot\|$ is the Euclidean norm, $A \in \mathbb{R}^{m \times n}$ is supposed full rank $\operatorname{rank}(A) = n$, $m \geq n$, $\bar{x} \in \mathbb{R}^{n \times 1}$, Ax' are the model response values and $\bar{f}$ is the vector of experimental measurements. Usually the measured data contain noise, i.e., we measure $f = \bar{f} + \epsilon$, with ϵ a certain kind of additive noise (e.g., white Gaussian). Since we are interested here in algebraic and geometric aspects of the problem, we suppose $\epsilon = 0$ and set $f = \bar{f}$. Moreover, we assume ideally that $\bar{f} = A\bar{x}$ holds exactly. Let us consider also the estimation problem for the approximate model $\mathcal{M}_a$:

$$x^{\|} = \underset{x' \in \mathbb{R}^{n_a}}{\operatorname{argmin}} \|A_a x' - \bar{f}\|^2, \tag{2}$$

where $A_a \in \mathbb{R}^{m \times n_a}$, $x^{\|} \in \mathbb{R}^{n_a \times 1}$, with $n_a < n$. The choice of the notation for $x^{\|}$ is to remind that the least-squares solution satisfies $A_a x^{\|} = P_{A_a}(f) =: f^{\|}$, where $f^{\|}$ is the orthogonal projection of $\bar{f}$ on the subspace generated by A_a, and the residual $A_a x^{\|} - \bar{f}$ is orthogonal to this subspace. Let us suppose that A_a corresponds to the first n_a columns of A, which means that the approximate model $\mathcal{M}_a$ is exactly one part of the true model $\mathcal{M}$, i.e., $A = [A_a, A_u]$ and so the solution $\bar{x}$ of (1) can be decomposed in two parts such that

$$A\bar{x} = [A_a, A_u] \begin{bmatrix} \bar{x}_a \\ \bar{x}_u \end{bmatrix} = A_a \bar{x}_a + A_u \bar{x}_u = \bar{f}. \tag{3}$$

This means that the model error corresponds to an additive term $A_u \bar{x}_u$ in the estimation problem.

Note that the columns of A_a are linearly independent since A is supposed to be of full rank. We do not consider the case in which A_a is rank-deficient, because it would mean that the model is not well parametrized. Moreover, some noise in the data is sufficient to determine a full rank matrix.

For brevity, we will call $\mathcal{A}$ the subspace generated by the columns of A and $\mathcal{A}_a$, $\mathcal{A}_u$ the subspaces generated by the columns of A_a, A_u respectively. Note that if $\mathcal{A}_a$ and $\mathcal{A}_u$ were orthogonal, decomposition (3) would be orthogonal. However, in the following we will consider the case in which the two subspaces are not orthogonal, as it commonly happens in practice. Oblique projections, even if not as common as orthogonal ones, have a large literature, e.g., [6,7].

Now, it is well known and easy to demonstrate that, when we solve problem (2) and A_u is not orthogonal to A_a, we get a biased solution, i.e., $x^{\|} \neq \bar{x}_a$:

Lemma 1. *Given $A \in \mathbb{R}^{m \times n}$ with $n \geq 2$ and $A = [A_a, A_u]$, and given $b \in \mathbb{R}^{m \times 1} \notin \mathcal{I}_m(A_a)$, call x the least-squares solution of (2) and $\bar{x} = [\bar{x}_a, \bar{x}_u]$ the solution of (1) decomposed as in (3). Then*

(i) if $A_u \perp A_a$ then $x^{\|} = \bar{x}_a$,
(ii) if $A_u \not\perp A_a$ then $x^{\|} \neq \bar{x}_a$.

Proof. The least-squares problem $Ax = f$ boils down to finding x such that $Ax = P_{\mathcal{A}_a}(f)$. Let us consider the unique decomposition of f on $\mathcal{A}_a$ and $\mathcal{A}_a^{\perp}$ as $f = f^{\|} + f^{\perp}$ with $f^{\|} = P_{\mathcal{A}_a}(f)$ and $f^{\perp} = P_{\mathcal{A}_a^{\perp}}(f)$. Call $f = f_a + f_u$ the decomposition on $\mathcal{A}_a$ and $\mathcal{A}_u$, hence there exist two vectors $x_a \in \mathbb{R}^{n_a}, x_u \in \mathbb{R}^{n - n_a}$ such that $f_a = A_a x_a$ and $f_u = A_u x_u$. If $\mathcal{A}_u \perp \mathcal{A}_a$ then the two decompositions

are the same, hence $f^{\|} = f_a$ and so $x^{\|} = \bar{x}_a$. Otherwise, for the definition of orthogonal projection ([6], third point of Def at page 429), it must hold $x^{\|} \neq \bar{x}_a$. □

3. Analysis of the Parameter Estimation Error

The aim of this paper is to propose a method to decrease substantially the bias of the solution of the approximated problem (2), with the smallest additional information about the norms of the model error and of the modelled part responses.

In this section we will introduce sufficient conditions to remove the bias and retrieve the true solution in a unique way, as summarized in Lemma 4. Let us start with a definition.

Definition 1 (Intensity Ratio)**.** *The intensity ratio I_f between modelled and un-modelled dynamics is defined as*

$$I_f = \frac{\|A_a x_a\|}{\|A_u x_u\|}.$$

In the following we assume that a good approximation of this intensity ratio is available and that its magnitude is sufficiently big, i.e., we have an approximate model that is quite accurate. This information about the model error will be used to reduce the bias, as shown in the following sections. Moreover we will consider also the norm $N_f = \|A_a x_a\|$ (or, equivalently, the norm $\|A_u x_u\|$).

3.1. The Case of Exact Knowledge about I_f and N_f

Here we assume, initially, to know the exact values of I_f and N_f, i.e.,

$$\begin{cases} N_f = \bar{N}_f = \|A_a \bar{x}_a\|, \\ I_f = \bar{I}_f = \frac{\|A_a \bar{x}_a\|}{\|A_u \bar{x}_u\|}. \end{cases} \tag{4}$$

This ideal setting is important to figure out the problem also with more practical assumptions. First of all, let us show a nice geometric property that relates x_a and f_a under a condition like (4).

Lemma 2. *The problem of finding the set of $x_a \in \mathbb{R}^n$ that give a constant, prescribed value for I_f and N_f is equivalent to that of finding the set of $f_a = A_a x_a \in \mathcal{A}_a$ of the decomposition $f = f_a + f_u$ (see the proof of Lemma 1) lying on the intersection of $\mathcal{A}_a$ and the boundaries of two n-dimensional balls in $\mathbb{R}^n$. In fact, it holds:*

$$\begin{cases} N_f = \|A_a x_a\| \\ I_f = \frac{\|A_a x_a\|}{\|A_u x_u\|} \end{cases} \iff \begin{cases} f_a \in \partial B_n(0, N_f) \\ f_a \in \partial B_n(f^{\|}, T_f) \end{cases} \quad \text{with} \quad T_f := \sqrt{\left(\frac{N_f}{I_f}\right)^2 - \|f^{\perp}\|^2}. \tag{5}$$

Proof. For every $x_a \in \mathbb{R}^{n_a}$ holds,

$$\begin{cases} N_f = \|f_a\| = \|A_a x_a\| \\ I_f = \frac{\|f_a\|}{\|f_u\|} = \frac{N_f}{\|f_u^{\perp} + f_u^{\|}\|} = \frac{N_f}{\sqrt{\|f^{\perp}\|^2 + \|f^{\|} - A_a x_a\|^2}} = \frac{N_f}{\sqrt{\|f^{\perp}\|^2 + \|f^{\|} - f_a\|^2}} \end{cases} \iff \tag{6}$$

$$\iff \begin{cases} \|f_a\| = N_f \\ \|f^{\|} - f_a\| = \sqrt{\left(\frac{N_f}{I_f}\right)^2 - \|f^{\perp}\|^2} =: T_f, \end{cases} \tag{7}$$

where we used the fact that $f_u = f_u^{\|} + f_u^{\perp}$ with $f_u^{\perp} := P_{A_a^{\perp}}(f_u) = f^{\perp}$, $f_u^{\|} := P_{A_a}(f_u) = A_a \delta x_a = f^{\|} - A_a x_a$, and $\delta x_a = (x^{\|} - x_a)$. Hence the equivalence (5) is proved. □

Given I_f and N_f, we call the feasible set of accurate model responses all the f_a that satisfy the relations (5). Now we will see that Lemma 2 allows us to reformulate problem (2) in the problem of finding a feasible f_a that, replaced to $\bar{f}$ in (2), gives as solution an unbiased estimate of $\bar{x}_a$. Indeed, it is easy to note that $A_a\bar{x}_a$ belongs to this feasible set. Moreover, since $f_a \in \mathcal{A}_a$, we can reduce the dimensionality of the problem and work on the subspace $\mathcal{A}_a$ which has dimension n_a, instead of the global space $\mathcal{A}$ of dimension n. To this aim, let us consider U_a the matrix of the SVD decomposition of A_a, $A_a = U_a S_a V_a^T$, and complete its columns to an orthonormal basis of $\mathbb{R}^n$ to obtain a matrix U. Since the vectors $f_a, f^{\parallel} \in \mathbb{R}^n$ belong to the subspace $\mathcal{A}_a$, the vectors $\tilde{f}_a, \tilde{f}^{\parallel} \in \mathbb{R}^n$ defined such that $f_a = U\tilde{f}_a$ and $f^{\parallel} = U\tilde{f}^{\parallel}$ must have zeros on the last $n - n_a$ components. Since U has orthonormal columns, it preserves the norms and so $\|f^{\parallel}\| = \|\tilde{f}^{\parallel}\|$ and $\|f_a\| = \|\tilde{f}_a\|$. If we call $\hat{f}_a, \hat{f}^{\parallel} \in \mathbb{R}^{n_a}$ the first n_a components of the vectors $\tilde{f}_a, \tilde{f}^{\parallel}$ (which have again the same norms of the full vectors in $\mathbb{R}^n$) respectively, we have

$$\begin{cases} \hat{f}_a \in \partial B_{n_a}(0, N_f), \\ \hat{f}_a \in \partial B_{n_a}(f^{\parallel}, T_f). \end{cases} \tag{8}$$

In this way the problem depends only on the dimension of the known subspace, i.e., the value of n_a, and does not depend on the dimensions $m \gg n_a$ and $n > n_a$. From (8) we can deduce the equation of the $(n_a - 2)$-dimensional boundary of an $(n_a - 1)$-ball to which the vector $f_a = A_a x_a$ must belong. In the following we discuss the various cases.

3.1.1. Case $n_a = 1$

In this case, we have one unique solution when both conditions on I_f and N_f are imposed. When only one of these two is imposed, two solutions are found, shown in Figure 1a,c. Figure 1b shows the intensity ratio I_f.

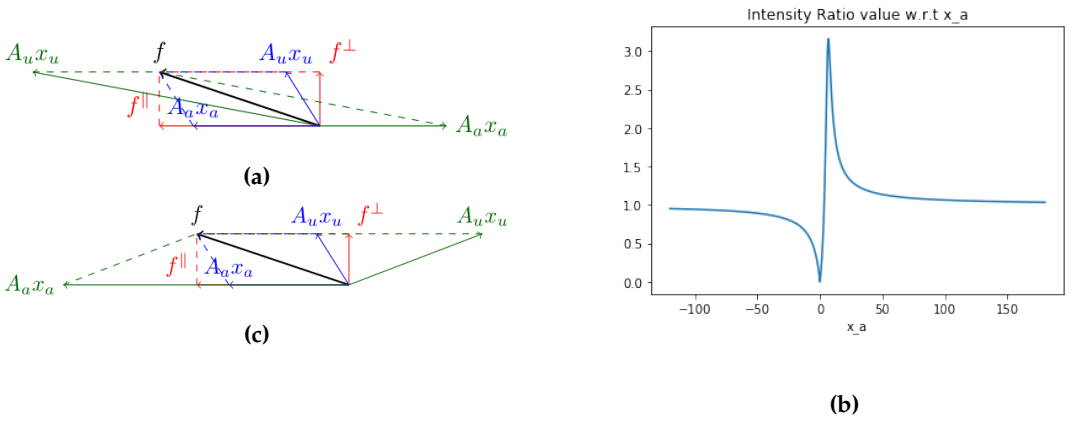


Figure 1. Case $n_a = 1$. (**a**): Case $n_a = 1$, $m = n = 2$. Solutions with the condition on N_f. In the figure: the true decomposition obtained imposing both the conditions (blue), the orthogonal decomposition (red), another possible decomposition (green) that satisfy the same norm condition N_f, but different I_f; (**b**): Case $n_a = 1$. Intensity Ratio value w.r.t the norm of the vector $A_a x_a$: given a fixed value of Intensity Ratio there can be two solution, i.e. two possible decomposition of f as sum of two vectors with the same Intensity Ratio; (**c**): Case $n_a = 1$, $m = n = 2$. Solutions with the condition on I_f. In the figure: the true decomposition obtained imposing both the conditions (blue), the orthogonal decomposition (red), another possible decomposition (green) with the same intensity ratio I_f, but different N_f.

3.1.2. Case $n_a = 2$

Consider the vectors $\hat{f}_a, \hat{f}^{\parallel} \in \mathbb{R}^{n_a=2}$ as defined previously, in particular we are looking for $\hat{f}_a = [\xi_1, \xi_2] \in \mathbb{R}^2$. Hence, conditions (8) can be written as

$$\begin{cases} \xi_1^2 + \xi_2^2 = N_f^2 \\ (\xi_1 - \hat{f}^{\parallel}_{\xi_1})^2 + (\xi_2 - \hat{f}^{\parallel}_{\xi_2})^2 = T_f^2 \end{cases} \longrightarrow \quad \mathcal{F}: \quad (\hat{f}^{\parallel}_{\xi_1})^2 - 2\hat{f}^{\parallel}_{\xi_1}\xi_1 + (\hat{f}^{\parallel}_{\xi_2})^2 - 2\hat{f}^{\parallel}_{\xi_2}\xi_2 = N_f^2 - T_f^2, \tag{9}$$

where the right equation is the $(n_a - 1) = 1$-dimensional subspace (line) $\mathcal{F}$ obtained subtracting the first equation to the second. This subspace has to be intersected with one of the beginning circumferences to obtain the feasible vectors $\hat{f}_a$, as can be seen in Figure 2a and its projection on $\mathcal{A}_a$ in Figure 2b. The intersection of the two circumferences (5) can have different solutions depending on the value of $(N_f - \|f^{\parallel}\|) - T_f$. When this value is strictly positive there are zero solutions, this means that the estimates of I_f and N_f are not correct: we are not interested in this case because we suppose the two values to be sufficiently well estimated. When the value is strictly negative there are two solutions, that coincide when the value is zero.

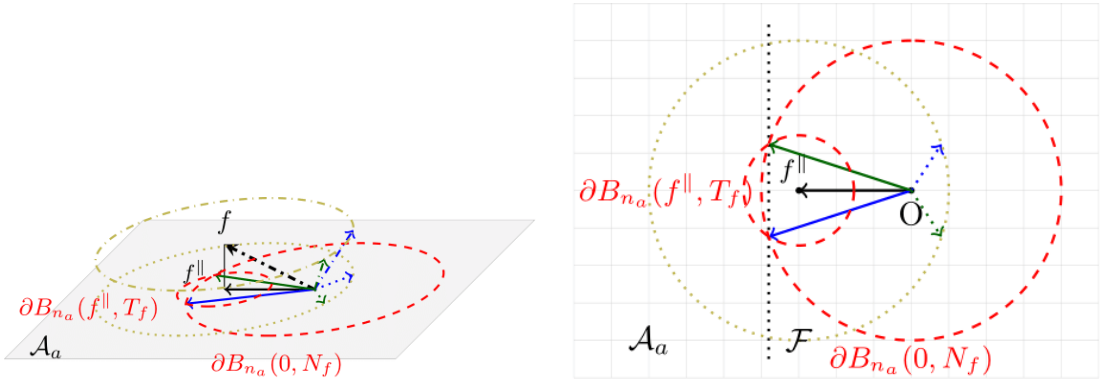


Figure 2. Case $n_a = 2$. (**a**): Case $n_a = 2$, $m = n = 3$, with $A_a x_a = [A_a(1)A_a(2)][x_a(1)x_a(2)]^T$. In the figure: the true decomposition (blue), the orthogonal decomposition (red), another possible decomposition of the infinite ones (green); (**b**): Case $n_a = 2$, $m = n = 3$. Projection of the two circumferences on the subspace $\mathcal{A}_a$, and projections of the possible decompositions of f (red, blue and green).

When there are two solutions, we have no sufficient information to determine which one of the two solutions is the true one, i.e., the one that gives $f_a = A_a \bar{x}_a$: we cannot choose the one that has minimum residual, neither the vector f_a that has the minimum angle with f, because both solutions have the same values of these two quantities. However, since we are supposing the linear system to be originated by an input/output system, where the matrix A_a is a function also of the input and f are the measurements of the output, we can take two tests with different inputs. Since all the solution sets contain the true parameter vector, we can determine the true solution from their intersection, unless the solutions of the two tests are coincident. The condition for coincidence is expressed in Lemma 3.

Let us call $A_{a,i} \in \mathbb{R}^{n \times n_a}$ the matrix of the test $i = 1, 2$, to which correspond a vector f_i. The line on which lie the two feasible vectors f_a of the same test i is $\mathcal{F}_i$ and $\mathcal{S}_i = A^{\dagger}_{a,i}\mathcal{F}_i$ is the line through the two solution points. To have two tests with non-coincident solutions, we need that these two lines $\mathcal{S}_1, \mathcal{S}_2$ do not have more than one common point, that in the case $n_a = 2$ is equivalent to $\mathcal{S}_1 \neq \mathcal{S}_2$, i.e., $A^{\dagger}_{a,1}\mathcal{F}_1 \neq A^{\dagger}_{a,2}\mathcal{F}_2$, i.e., $\mathcal{F}_1 \neq A_{a,1}A^{\dagger}_{a,2}\mathcal{F}_2 =: \mathcal{F}_{12}$. We represent the lines $\mathcal{F}_i$ by means of their orthogonal vector from the origin $f^{ort,i} = l_{ort,i}\frac{f_i^{\parallel}}{\|f_i^{\parallel}\|}$. We introduce the matrices C_a, C_f, C_{fp} such that $A_{a,2} = C_a A_{a,1}$, $f_2 = C_f f_1$, $f_2^{\parallel} = C_{fp} f_1^{\parallel}$ and k_f such that $\|f_2^{\parallel}\| = k_f \|f_1^{\parallel}\|$.

Lemma 3. *Consider two tests $i = 1, 2$ from the same system with $n_a = 2$ with the above notation. Then it holds $\mathcal{F}_1 = \mathcal{F}_{12}$ if and only if $C_a = C_{fp}$.*

Proof. From the relation $f_i^{\|} = \mathcal{P}_{\mathcal{A}_{a,i}}(f_i) = A_{a,i}(A_{a,i}^T A_{a,i})^{-1} A_{a,i}^T f_i$, we have

$$f_2^{\|} = A_{a,2}(A_{a,2}^T A_{a,2})^{-1} A_{a,2}^T f_2 = C_a A_{a,1}(A_{a,1}^T C_a^T C_a A_{a,1})^{-1} A_{a,1}^T C_a^T C_f f_1. \tag{10}$$

It holds $\mathcal{F}_1 = \mathcal{F}_{12} \iff f^{ort,1} = f^{ort,12} := A_{a,1} A_{a,2}^{\dagger} f^{ort,2}$, hence we will show this second equivalence. We note that $l_{ort,2} = k_f l_{ort,1}$ and calculate

$$f^{ort,12} = A_{a,1} A_{a,2}^{\dagger} f^{ort,2} = A_{a,1} A_{a,1}^{\dagger} C_a^{\dagger} \left(l_{ort,2} \frac{f_2^{\|}}{\|f_2^{\|}\|} \right) = A_{a,1} A_{a,1}^{\dagger} C_a^{\dagger} \left(k_f l_{ort,1} \frac{C_{fp} f_1^{\|}}{k_f \|f_1^{\|}\|} \right) = A_{a,1} A_{a,1}^{\dagger} C_a^{\dagger} C_{fp} f^{ort,1}. \tag{11}$$

Now let us call $s^{ort,1}$ the vector such that $f^{ort,1} = A_{a,1} s^{ort,1}$, then, using the fact that $C_a = C_{fp}$ we obtain

$$f^{ort,12} = A_{a,1} A_{a,1}^{\dagger} C_a^{\dagger} C_{fp} A_{a,1} s^{ort,1} = A_{a,1}(A_{a,1}^{\dagger} A_{a,1}) s^{ort,1} = (\text{since} A_{a,1}^{\dagger} A_{a,1} = I_{n_a}) = A_{a,1} s^{ort,1} \tag{12}$$

Hence we have $\mathcal{F}_{12} = \mathcal{F}_1 \iff A_{a,1} A_{a,1}^{\dagger} C_a^{\dagger} C_{fp} f^{ort,1} = f^{ort,1} \iff C_a^{\dagger} C_{fp} = I$. □

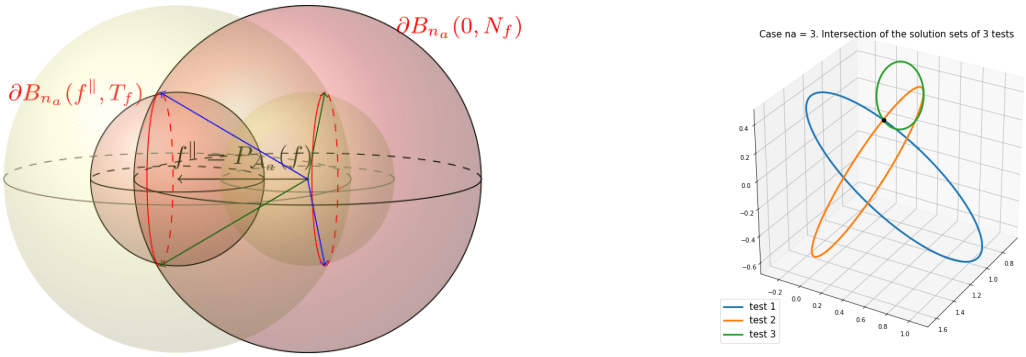
3.1.3. Case $n_a \geq 3$

More generally, for the case $n_a \geq 3$, consider the vectors $\hat{f}_a, \hat{f}^{\|} \in \mathbb{R}^{n_a}$ as defined previously, in particular we are looking for $\hat{f}_a = [\xi_1, \ldots, \xi_{n_a}] \in \mathbb{R}^{n_a}$. Conditions (8) can be written as

$$\begin{cases} \sum_{i=1}^{n_a} \xi_i^2 = N_f^2 \\ \sum_{i=1}^{n_a} (\xi_i - \hat{f}_{\xi_i}^{\|})^2 = T_f^2 \end{cases} \longrightarrow \quad \mathcal{F}: \quad \sum_{i=1}^{n_a} ((\hat{f}_{\xi_i}^{\|})^2 - 2\hat{f}_{\xi_i}^{\|} \xi_i) = N_f^2 - T_f^2, \tag{13}$$

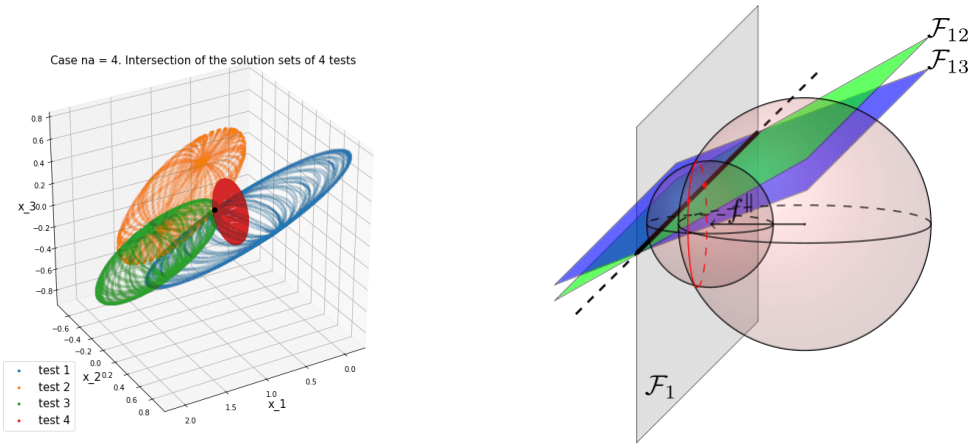
where the two equations on the left are two $(n_a - 1)$-spheres, i.e., the boundaries of two n_a-dimensional balls. Analogously to the case $n_a = 2$, the intersection of these equations can be empty, one point or the boundary of a $(n_a - 1)$-dimensional ball (with the same conditions on $(N_f - \|f^{\|}\|) - T_f$). The equation on the right of (13) is the $(n_a - 1)$-dimensional subspace $\mathcal{F}$ on which lies the boundary of the $(n_a - 1)$-dimensional ball of the feasible vectors f_a, and is obtained subtracting the first equation to the second one. In Figure 3a the graphical representation of the decomposition $f^{\|} = f_a + f_u^{\|}$ for the case $n_a = 3$ is shown, and in Figure 3b the solution ellipsoids of 3 tests whose intersection is one point. Figure 4a shows the solution hyperellipsoids of 4 tests whose intersection is one point, in the case $n_a = 4$.

We note that, to obtain one unique solution x_a we must intersect the solutions of at least two tests. Let us give a more precise idea of what happens in general. Given $i = 1, \ldots, n_a$ tests we call, as in the previous case, $f^{ort,i}$ the vector orthogonal to the $(n_a - 1)$-dimensional subspace $\mathcal{F}_i$ that contains the feasible f_a, and $\mathcal{S}_i = A_{a,i}^{\dagger} \mathcal{F}_i$. We project this subspace on $\mathcal{A}_{a,1}$ and obtain $\mathcal{F}_{1i} = A_{a,1} A_{a,i}^{\dagger} \mathcal{F}_i$ that we describe through its orthogonal vector $f^{ort,1i} = A_{a,1} A_{a,i}^{\dagger} f^{ort,i}$. If the vectors $f^{ort,1}, f^{ort,12}, \ldots f^{ort,1n_a}$ are linearly independent, it means that the $(n_a - 1)$-dimensional subspaces $\mathcal{F}_1, \mathcal{F}_{12}, \ldots \mathcal{F}_{1n_a}$ intersect themselves in one point. In Figure 4b it is shown an example in which, in the case $n_a = 3$ the vectors $f^{ort,1}, f^{ort,12}, f^{ort,13}$ are not linearly independent. The three solution sets of this example will intersect in two points, hence, for $n_a = 3$, three tests are not always sufficient to determine a unique solution.



(a) **(b)**

Figure 3. Case $n_a = 3$. (**a**): Case $n_a = 3$, $m = n = 4$, $n - n_a = 1$: in the picture $\bar{f}^{\parallel}$, i.e., the projection of f on $\mathcal{A}_a$. The decompositions that satisfies the conditions on I_f and N_f are the ones with f_a that lies on the red circumference on the left. The spheres determined by the conditions are shown in yellow for the vector f_a and in blue for the vector $f^{\parallel} - a_a$. Two feasible decompositions are shown in blue and green; (**b**): Case $n_a = 3$. Intersection of three hyperellipsoids, set of the solutions x_a of three different tests, in the space $\mathbb{R}^{n_a=3}$.



(a) **(b)**

Figure 4. Case $n_a \geq 3$. (**a**): Case $n_a = 4$. Intersection of four hyperellipsoids, set of the solutions x_a of four different tests, in the space $\mathbb{R}^{n_a=4}$; (**b**): Case $n_a = 3$. Example of three tests for which the solution has an intersection bigger than one single point. The three $(n_a - 1)$-dimensional subspaces $\mathcal{F}_1, \mathcal{F}_{12}, \mathcal{F}_{13}$ in the space generated by $A_{a,1}$ intersect in a line and their three orthogonal vectors are not linearly independent.

Lemma 4. *For all $n_a > 1$, the condition that, given $i = 1, \ldots, n_a$ tests, the n_a hyperplanes $\mathcal{S}_i = A_{a,i}^{\dagger}\mathcal{F}_i$ previously defined have linearly independent normal vectors is sufficient to determine one unique intersection, i.e., one unique solution vector $\bar{x}_a$, that satisfies the system of conditions (4) for each test.*

Proof. The intersection of n_a independent hyperplanes in $\mathbb{R}^{n_a}$ is a point. Given a test i and $\mathcal{S}_i = A_{a,i}^{\dagger}\mathcal{F}_i$ the affine subspace of that test

$$\mathcal{S}_i = v_i + W_i = \{v_i + w \in \mathbb{R}^{n_a} \ : \ w \cdot \mathbf{n}_i = 0\} = \{x \in \mathbb{R}^{n_a} \ : \ \mathbf{n}_i^T(x - v_i) = 0\},$$

where $\mathbf{n}_i$ is the normal vector of the linear subspace and v_i the translation with respect to the origin.

The conditions on $\mathcal{S}_i$ relative to n_a tests correspond to a linear system $Ax = b$, where $\mathbf{n}_i$ is the i-th row of A and each component of the vector b given by $b_i = \mathbf{n}_i^T v_i$. The matrix A has full rank because of the linear independence condition of the vectors $\mathbf{n}_i$, hence the solution of the linear system is unique.

The unique intersection is due to the hypothesis of full column rank of the matrices $A_{a,i}$: this condition implies that the matrices $A_{a,i}$ map the surfaces $\mathcal{F}_i$ to hyperplanes $\mathcal{S}_i = A_{a,i}\mathcal{F}_i$. □

For example, with $n_a = 2$ (Lemma 3) this condition is equal to considering two tests with non-coincident lines $\mathcal{S}_1, \mathcal{S}_2$, i.e., two non-coincident $\mathcal{F}_1, \mathcal{F}_{12}$.

3.2. The Case of Approximate Knowledge of I_f and N_f Values

Let us consider N tests and call $I_{f,i}, N_{f,i}$ and $T_{f,i}$ the values as defined in Lemma 2, relative to test i. Since the system of conditions

$$\begin{cases} N_{f,i} = \|A_{a,i}x_a\| \\ I_{f,i} = \frac{\|A_{a,i}x_a\|}{\|z_i - A_{a,i}x_a\|} \end{cases} \quad \text{and} \quad \begin{cases} N_{f,i} = \|A_{a,i}x_a\| \\ T_{f,i} = \|f_i^{\|} - A_{a,i}x_a\| \end{cases} \tag{14}$$

is equivalent, as shown in Lemma 2, we will take into account the system on the right for its simplicity: the equation on $T_{f,i}$ represents an hyperellipsoid, translated with respect to the origin.

In a real application, we can assume to know only an interval in which the true values of I_f is contained and, analogously, an interval for N_f values. Supposing we know the bounds on I_f and N_f, then the bounds on T_f can be easily computed. Let us call these extreme values $N_f^{max}, N_f^{min}, T_f^{max}, T_f^{min}$, we will assume it always holds

$$\begin{cases} N_f^{max} \geq max_i(N_{f,i}), \\ N_f^{min} \leq min_i(N_{f,i}), \end{cases} \quad \text{and} \quad \begin{cases} T_f^{max} \geq max_i(T_{f,i}), \\ T_f^{min} \leq min_i(T_{f,i}), \end{cases} \tag{15}$$

for each i-th test of the considered set $i = 0, \ldots, N$.

Condition (4) is now relaxed as follows: the true solution $\bar{x}_a$ satisfies

$$\begin{cases} \|A_{a,i}\bar{x}_a\| \leq N_f^{max}, \\ \|A_{a,i}\bar{x}_a\| \geq N_f^{min}, \end{cases} \quad \text{and} \quad \begin{cases} \|A_{a,i}\bar{x}_a - f_i^{\|}\| \leq T_f^{max}, \\ \|A_{a,i}\bar{x}_a - f_i^{\|}\| \geq T_f^{min}, \end{cases} \tag{16}$$

for each i-th test of the considered set $i = 0, \ldots, N$.

Assuming the extremes to be non-coincident ($N_f^{min} \neq N_f^{max}$ and $T_f^{min} \neq T_f^{max}$), these conditions do not define a single point, i.e., the unique solution $\bar{x}_a$ (as in (4) of Section 3.1), but an entire closed region of the space that may be even not connected, and contains infinite possible solutions x different from $\bar{x}_a$.

In Figure 5 two examples, with $n_a = 2$, of the conditions for a single test are shown: on the left in the case of exact knowledge of the $N_{f,i}$ and $T_{f,i}$ values, and on the right with the knowledge of two intervals containing the right values.

Given a single test, the conditions (16) on a point x can be easily characterized. Given the condition

$$\|f_a\| = \|A_a x_a\| = N_f,$$

we write $x_a = \sum \chi_i v_i$ with v_i the vectors of the orthogonal basis, given by the columns V of the SVD decomposition $A_a = USV^T$. Then

$$f_a = A_a x_a = USV^T(\sum_i \chi_i v_i) = US(\sum_i \chi_i e_i) = U(\sum_i s_i \chi_i e_i) = \sum_i s_i \chi_i \mathbf{u}_i.$$

Since the norm condition $\|f_a\|^2 = \sum_i (s_i\chi_i)^2 = N_f^2$ holds, then we obtain the equation of the hyperellipsoid for x_a as:

$$\sum_i (s_i\chi_i)^2 = \sum_i \frac{\chi_i^2}{(\frac{1}{s_i})^2} = N_f^2. \tag{17}$$

The bounded conditions hence gives the region inside the two hyperellipsoids centered in the origin:

$$N_f^{min} \leq \sum_i \frac{\chi_i^2}{(\frac{1}{s_i})^2} \leq N_f^{max}. \tag{18}$$

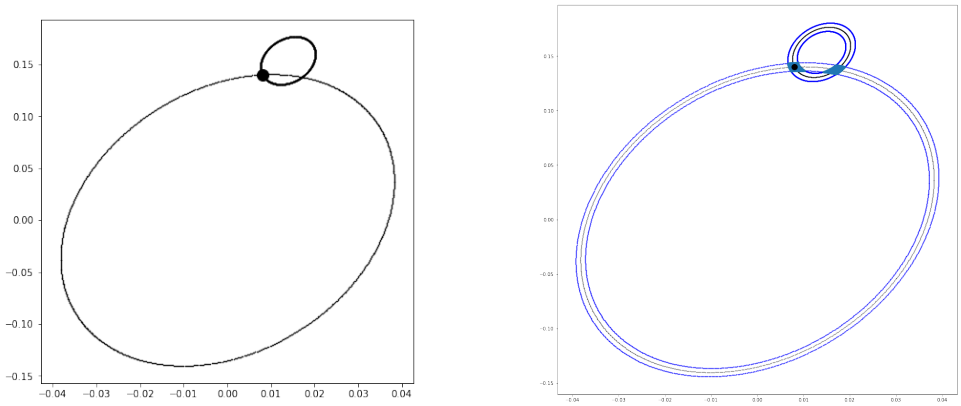
Analogously for the I_f condition, the region inside the two translated hyperellipsoids:

$$T_f^{min} \leq \sum_i \frac{\chi_i^2}{(\frac{1}{s_i})^2} - f^{\|} \leq T_f^{max}. \tag{19}$$

Given a test i, each of the conditions (18) and (19), constrain $\bar{x}_a$ to lie inside a thick hyperellipsoid, i.e., the region between the two concentric hyperellipsoids. The intersection of these two conditions for test i is a zero-residual region that we call Z_{r_i}

$$Z_{r_i} = \{x \in \mathbb{R}^{n_a} \mid (18) \text{ and } (19) \text{ hold } \}. \tag{20}$$

It is easy to verify that if $N_{f,i}$ is equal to the assumed N_f^{min} or N_f^{max}, or $T_{f,i}$ is equal to the assumed T_f^{min} or T_f^{max}, the true solution will be on a border of the region Z_{r_i}, and if it holds for both $N_{f,i}$ and $T_{f,i}$ it will lie on a vertex.



(a) **(b)**

Figure 5. Examples of the exact and approximated conditions on a test with $n_a = 2$. In the left equation the two black ellipsoids are the two constraints of the right system of (14), while in the right figure the two couples of concentric ellipsoids are the borders of the thick ellipsoids defined by (16) and the blue region Z_{r_i} is the intersection of (18) and (19). The black dot in both the figures is the true solution. (**a**): Exact conditions on N_f and T_f; (**b**): Approximated conditions on N_f and T_f.

When more tests $i = 1, \ldots, N$ are put together, we have to consider the points that belong to the intersection of all these regions Z_{r_i}, i.e.,

$$I_{zr} = \bigcap_{i=0,\ldots,N} Z_{r_i}. \tag{21}$$

These points minimize, with zero residual, the following optimization problem:

$$\begin{aligned} \min_x \sum_{i=1}^{N} min(0, \|A_{a,i}x\| - N_f^{min})^2 + \sum_{i=1}^{N} max(0, \|A_{a,i}x\| - N_f^{max})^2 + \\ + \sum_{i=1}^{N} min(0, \|A_{a,i}x - f_i^{\|}\| - T_f^{min})^2 + \sum_{i=1}^{N} max(0, \|A_{a,i}x - f_i^{\|}\| - T_f^{max})^2. \end{aligned} \tag{22}$$

It is also easy to verify that, if the true solution lies on an edge/vertex of one of the regions Z_{r_i}, it will lie on an edge/vertex of their intersection.

The intersected region I_{zr} tends to monotonically shrink in a way that depends from the properties of the added tests. We are interested to study the conditions that make it reduce to a point, or at least to a small region. A sufficient condition to obtain a point is given in Theorem 1.

Let us first consider the function that, given a point in the space $\mathbb{R}^{n_a}$, returns the squared norm of its image through the matrix A_a:

$$\begin{aligned} N_f^2(x) &= \|A_a x\|_2^2 = \|U\Sigma V^T x\|_2^2 = \|\Sigma V^T x\|_2^2 = (\Sigma V^T x)^T(\Sigma V^T x) = x^T(V\Sigma^T\Sigma V^T)x = \\ &= \| \begin{bmatrix} \sigma_1 v_1^T x \\ \sigma_2 v_2^T x \\ \vdots \end{bmatrix} \|_2^2 = \sigma_1^2 (v_1^T x)^2 + \sigma_2^2 (v_2^T x)^2 + \dots, \end{aligned} \tag{23}$$

where v_i are the columns of V and $x = [x(1)\, x(2) \dots, x(n_a)]$.

The direction of maximum increase of this function is given by its gradient

$$\nabla N_f^2(x) = 2(V\Sigma^2 V^T)x = \begin{bmatrix} 2\sigma_1^2 v_1^T x v_1(1) + 2\sigma_2^2 v_2^T x v_2(1) + \dots + 2\sigma_{n_a}^2 v_{n_a}^T x v_{n_a}(1) \\ 2\sigma_1^2 v_1^T x v_1(2) + 2\sigma_2^2 v_2^T x v_2(2) + \dots + 2\sigma_{n_a}^2 v_{n_a}^T x v_{n_a}(2) \\ \vdots \end{bmatrix}. \tag{24}$$

Analogously, define the function $T_f^2(x)$ as

$$\begin{aligned} T_f^2(x) &= \|A_a x - f^{\|}\|_2^2 = \|U\Sigma V^T x - f^{\|}\|_2^2 = \|\Sigma V^T x - f^{\|}\|_2^2 = \\ &= (\Sigma V^T x - f^{\|})^T(\Sigma V^T x - f^{\|}) = (\Sigma V^T x)^T(\Sigma V^T x) - 2(\Sigma V^T x)^T f^{\|} + (f^{\|})^T(f^{\|}) \\ &= x(V\Sigma^2 V^T)x - 2(x)^T V\Sigma f^{\|} + (f^{\|})^T(f^{\|}) = \\ &= \left\| \begin{bmatrix} \sigma_1 v_1^T x \\ \sigma_2 v_2^T x \\ \vdots \end{bmatrix} - f^{\|} \right\|_2^2 \end{aligned} \tag{25}$$

with gradient

$$\begin{aligned} \nabla T_f^2(x) &= 2(V\Sigma^2 V^T)x - 2V\Sigma f^{\|} = \\ &= \begin{bmatrix} 2\sigma_1^2 v_1^T x v_1(1) + 2\sigma_2^2 v_2^T x v_2(1) + \dots + 2\sigma_{n_a}^2 v_{n_a}^T x v_{n_a}(1) \\ \vdots \\ 2\sigma_1^2 v_1^T x v_1(j) + 2\sigma_2^2 v_2^T x v_2(j) + \dots + 2\sigma_{n_a}^2 v_{n_a}^T x v_{n_a}(j) \\ \vdots \end{bmatrix} - \begin{bmatrix} -2\sigma_i^2 \sum_i f^{\|}(i) v_i(1) \\ \vdots \\ -2\sigma_i^2 \sum_i f^{\|}(i) v_i(j) \\ \vdots \end{bmatrix}. \end{aligned} \tag{26}$$

Definition 2. *(Upward/Downward Outgoing Gradients) Take a test i, and the functions $N_f^2(x)$ and $T_f^2(x)$ as in (23) and (25), with the formulas of the gradient vectors of these two functions $\nabla N_{f,i}(x)$, $\nabla T_{f,i}(x)$ as in (24) and (26). Given the two extreme values $N_f^{min/max}$ and $T_f^{min/max}$ for each test, let us define*

- *the downward outgoing gradients as the set of gradients calculated on the points on the minimum hyperellipsoid*

$$\{-\nabla N_{f,i}(x) \mid N_{f,i}(x) = N_f^{min}\} \quad \text{and} \quad \{-\nabla T_{f,i}(x) \mid T_{f,i}(x) = T_f^{min}\} \tag{27}$$

  they point inward to the region of the thick hyperellipsoid.
- *the Upward Outgoing Gradients as the set of negative gradients of points on the maximum hyperellipsoid*

$$\{\nabla N_{f,i}(x) \mid N_{f,i}(x) = N_f^{max}\} \quad \text{and} \quad \{\nabla T_{f,i}(x) \mid T_{f,i}(x) = T_f^{max}\} \tag{28}$$

they point outward the region.

Note that the upward/downward outgoing gradient of function $N_f^2(x)$ (or $T_f^2(x)$) on point x is the normal vector to the tangent plane on the hyperellipsoid on which the point lies. Moreover, these vectors point outward the region defined by Equation (18) (and (19) respectively). In Figure 6, an example of some upward/downward outgoing gradients of function $N_f^2(x)$ is shown.

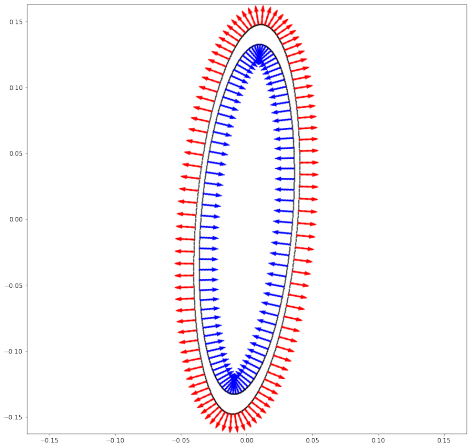


Figure 6. In the figure some upward/downward outgoing gradients are shown: the blue internal ones are downward outgoing gradients calculated on points x on the internal ellipsoid with $N_{f,i}(x) = N_f^{min}$, while the external red ones are upward outgoing gradients calculated on points x on the external ellipsoid with $N_{f,i}(x) = N_f^{max}$.

Theorem 1. *Given N tests with values $I_{f,i}$ and $N_{f,i}$ in the closed intervals $[I_f^{min}, I_f^{max}]$ and $[N_f^{min}, N_f^{max}]$, take the set of all the upward/downward outgoing gradients of functions $N_{f,i}^2(x)$ and $T_{f,i}^2(x)$ calculated in the true solution $\bar{x}_a$, i.e.,*

$$\begin{aligned}\{\nabla N_{f,i}(\bar{x}_a) \text{ for } i = 1,\dots,N \mid N_{f,i}(\bar{x}_a) = N_f^{max}\} \cup \{\nabla N_{f,i}(\bar{x}_a) \text{ for } i = 1,\dots,N \mid N_{f,i}(\bar{x}_a) = N_f^{min}\} \cup \\ \cup \{\nabla T_{f,i}(\bar{x}_a) \text{ for } i = 1,\dots,N \mid T_{f,i}(\bar{x}_a) = T_f^{max}\} \cup \{\nabla T_{f,i}(\bar{x}_a) \text{ for } i = 1,\dots,N \mid T_{f,i}(\bar{x}_a) = T_f^{min}\}.\end{aligned} \tag{29}$$

If there is at least one outgoing gradient of this set in each orthant of $\mathbb{R}^{n_a}$, then the intersection region I_{zr} of Equation (21) reduces to a point.

Proof. What we want to show is that given any perturbation δ_x of the real solution $\bar{x}_a$, there exists at least one condition among (18) and (19) that is not satisfied by the new perturbed point $\bar{x}_a + \delta_x$.

Any sufficiently small perturbation δ_x in an orthant in which lies an upward/downward outgoing gradient (from now on "Gradient"), determines an increase/decrease in the value of the hyperellipsoid function relative to that Gradient, that makes the relative condition to be unsatisfied.

Hence, if the Gradient in the orthant considered is upward, it satisfies $N_{f,i}(\bar{x}_a) = N_f^{max}$ (or analogously with $T_{f,i}$) and for each perturbation δ_x in the same orthant we obtain

$$N_{f,i}(\bar{x}_a + \delta_x) > N_{f,i}(\bar{x}_a) = N_f^{max}$$

(or analogously with $T_{f,i}$). In the same way, if the Gradient is downward we obtain

$$N_{f,i}(\bar{x}_a + \delta_x) < N_{f,i}(\bar{x}_a) = N_f^{min}$$

(or analogously with $T_{f,i}$).

When in one orthant there are more than one Gradient, it means that more than one condition will be unsatisfied by the perturbed point $\bar{x}_a + \delta_x$ for a sufficiently small δ_x in that orthant. □

4. Problem Solution

The theory previously presented allows us to build a solution algorithm that can deal with different a-priori information. We will start in Section 4.1 with the ideal case, i.e., with exact knowledge of I_f and N_f. Then, we generalize to a more practical setting, where we suppose to know an interval that contains the T_f values of all the experiments considered and an interval for the N_f values. Hence, the estimate solution will satisfy Equations (18) and (19). In this case we describe an algorithm for computing an estimate of the solution, that we will test in Section 5 against a toy model.

4.1. Exact Knowledge of I_f and N_f

When the information about I_f and N_f is exact, with the minimum amount of experiments indicated in Section 3 we can find the unbiased parameter estimate as the intersection I_{zr} of the zero-residual sets Z_{r_i} corresponding to each experiment. In principle this could be done also following the proof of Lemma 4, but the computation of the v_i vectors is quite cumbersome. Since this is an ideal case, we solve it by simply imposing the satisfaction of the various N_f and T_f conditions (Equation (14)) as an optimization problem:

$$\min_x F(x) \quad \text{with} \quad F(x) = \sum_{i=1}^{N} (\|A_{a,i}x\| - N_{f,i})^2 + \sum_{i=1}^{N} (\|A_{a,i}x - f_i^{\|}\| - T_{f,i})^2. \tag{30}$$

The solution of this problem is unique when the tests are in a sufficient number and satisfies the conditions of Lemma 4.

This nonlinear least-squares problem can be solved using a general nonlinear optimization algorithm, like Gauss–Newton method or Levenberg–Marquardt [8].

4.2. Approximate Knowledge of I_f and N_f

In practice, as already pointed out in Section 3.2, it is more realistic to know the two intervals that contain all the $N_{f,i}$ and $I_{f,i}$ values for each test i. Then, we know that within the region I_{zr} there is also the exact unbiased parameter solution $\bar{x}_a$, that we want at least to approximate. We introduce here an Unbiased Least-Squares (ULS) Algorithm 1 for the computation of this estimate.

Algorithm 1 An Unbiased Least-Squares (ULS) algorithm.

1: Given a number n_{tests} of available tests, indexed with a number between 1 and n_{tests}, and two intervals, $\left[I_f^{min}, I_f^{max}\right]$ and $\left[N_f^{min}, N_f^{max}\right]$, containing the I_f and N_f values of all tests.
2: At each iteration we will consider the tests indexed by the interval $[1, i_t]$; set initially $i_t = n_a$.
3: **while** $i_t \leq n_{tests}$ **do**
4: 1) compute a solution with zero residual of the problem (22) with a nonlinear least-squares optimization algorithm,
5: 2) estimate the size of the zero-residual region as described below in (31),
6: 3) increment by one the number i_t of tests.
7: **end while**
8: Accept the final solution if the estimated region diameter is sufficiently small.

In general, the zero-residual region Z_{r_i} of each test contains the true point of the parameters vector, while the estimated iterates with the local optimization usually start from a point outside this region and converge to a point on the boundary of the region.

The ULS estimate can converge to the true solution in two cases:

1. the true solution lies on the border of the region I_{zr} and the estimate reach the border on that point;
2. the region I_{zr} reduces to a dimension smaller than the required accuracy, or reduces to a point.

The size of the intersection set I_{zr}, of the zero-residual regions Z_{r_i}, is estimated in the following way.

Let us define an index, that we call region shrinkage estimate, as follows:

$$\hat{s}(x) = min\{n \mid \sum_{\delta \in \mathcal{P}} \Delta_{I_{zr}}(x + \mu^{-n}\delta) > 0\}, \tag{31}$$

where we used $\mu = 1.5$ in the experiments below, $\mathcal{P} = \{\delta \in \mathbb{R}^{n_a} \mid \delta(i) \in (-1, 0, 1)\ \forall i = 1, \ldots, n_a\}$ and $\Delta_{I_{zr}}$ is the Dirac function of the set I_{zr}.

5. Numerical Examples

Let us consider a classical application example, the equations of a DC motor with a mechanical load, where the electrical variables are governed by the following ordinary differential equation

$$\begin{cases} L\dot{I}(t) & = -K\omega(t) - RI(t) + V(t) - f_u(t) \\ I(t_0) & = I_0, \end{cases} \tag{32}$$

where I is the motor current, ω the motor angular speed, V the applied voltage, and $f_u(t)$ a possible unmodelled component

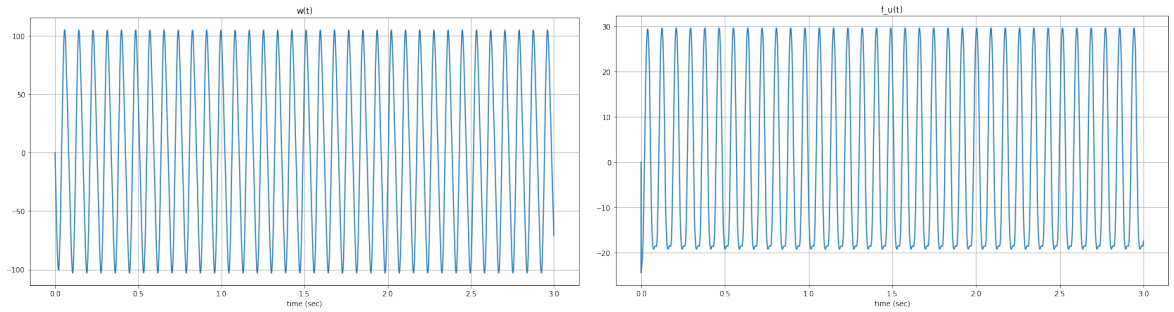
$$f_u(t) = -m_{err} cos(n_{poles}\theta(t)), \tag{33}$$

where n_{poles} is the number of poles of the motor, i.e., the number of windings or magnets [9], m_{err} the magnitude of the error model and θ the angle, given by the system

$$\begin{cases} \dot{\omega}(t) & = \theta(t) \\ \omega(t_0) & = \omega_0. \end{cases} \tag{34}$$

Note that the unknown component f_u of this example can be seen as a difference in the potential that is not described by the approximated model. We are interested in the estimation of parameters $[L, K, R]$. In our test the true values were constant values $[L = 0.0035, K = 0.14, R = 0.53]$.

We suppose to know the measurements of I and ω at equally spaced times $t_0, \ldots, t_{\bar{N}}$ with step h, such that $t_k = t_0 + kh$, and $t_{k+1} = t_k + h$. In Figure 7 we see the plots of the motor speed ω and of the unknown component f_u for this experiment.



(a) **(b)**

Figure 7. The plots of (**a**) $\omega(t)$ and (**b**) $f_u(t)$ in the experiment.

We compute the approximation of the derivative of the current signal $\hat{\dot{I}}(t_k)$ with the forward finite difference formula of order one

$$\hat{\dot{I}}(t_k) = \frac{I(t_k) - I(t_{k-1})}{h}, \quad \text{for} \quad t_k = t_1, \ldots, t_{\bar{N}}$$

with a step $h = 4 \times 10^{-4}$. The applied voltage is held constant to the value $V(t) = 30.0$.

To obtain a more accurate estimate, or to allow the possibility of using higher step size values h, finite differences of higher order can be used, for example the fourth order difference formula

$$\hat{\dot{I}}(t_k) = \frac{I(t_k - 2h) - 8I(t_k - h) + 8I(t_k + h) - I(t_k + 2h)}{12h}, \qquad \text{for} \quad t_k = t_2, \ldots, t_{\bar{N}-2}.$$

With the choice of the finite difference formula, we obtain the discretized equations

$$L\hat{\dot{I}}(t_k) = -K\omega(t_k) - RI(t_k) + V(t_k) - f_u(t_k), \qquad \text{for} \quad t_k = t_1, \ldots, t_{\bar{N}}. \tag{35}$$

We will show a possible implementation of the method explained in the previous sections, and the results we get with this toy-model example. The comparison is made against the standard least-squares. In particular, we will show that when the information about I_f and N_f is exact, we have an exact removal of the bias. In case this information is only approximate, which is common in a real application, we will show how the bias asymptotically disappears when the number of experiments increases.

We build each test taking the Equation (35) for n samples in the range $t_1, \ldots, t_{\bar{N}}$, obtaining the linear system

$$\begin{bmatrix} \hat{\dot{I}}(t_k) & \omega(t_k) & I(t_k) \\ \hat{\dot{I}}(t_{k+1}) & \omega(t_{k+1}) & I(t_{k+1}) \\ \vdots & \vdots & \vdots \\ \hat{\dot{I}}(t_{k+n}) & \omega(t_{k+n}) & I(t_{k+n}) \end{bmatrix} \begin{bmatrix} L \\ K \\ R \end{bmatrix} + \begin{bmatrix} f_u(t_k) \\ f_u(t_{k+1}) \\ \vdots \\ f_u(t_{k+n}) \end{bmatrix} = \begin{bmatrix} V(t_k) \\ V(t_{k+1}) \\ \vdots \\ V(t_{k+n}) \end{bmatrix} \tag{36}$$

so that the first matrix in the equation is $A_a \in \mathbb{R}^{n \times n_a}$ with $n_a = 3$, the number of parameters to be estimated.

To measure the estimation relative error $\hat{e}_{rel}$ we will use the following formula, where $\hat{x}_a$ is the parameter estimate:

$$\hat{e}_{rel} = \frac{1}{n_a} \sum_{i=1}^{n_a} \frac{||\hat{x}_a(i) - \bar{x}_a(i)||_2}{||\bar{x}_a(i)||_2}. \tag{37}$$

Note that the tests that we built in the numerical experiments below are simply small chunks of consecutive data, taken from one single simulation for each experiment.

The results have been obtained with a Python code developed by the authors, using NumPy for linear algebra computations and `scipy.optimize` for the nonlinear least-squares optimization.

5.1. Exact Knowledge of I_f and N_f

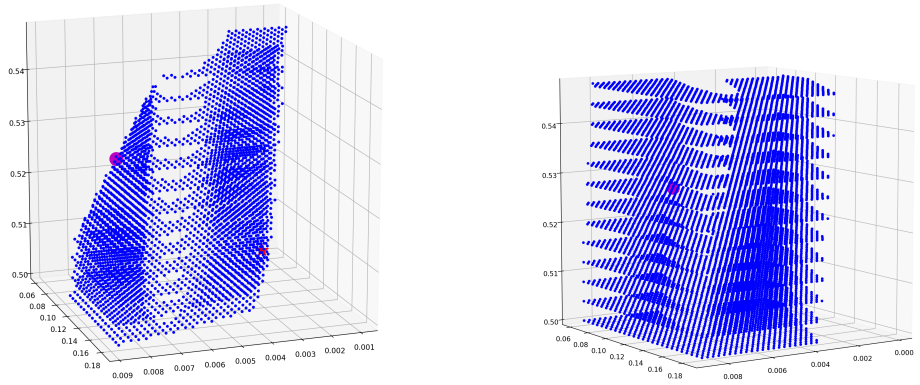
As analyzed in Section 4.1, the solution of the minimization problem (30) is computed with a local optimization algorithm.

Here the obtained results show an error $\hat{e}_{rel}$ with an order of magnitude of 10^{-7} in every test we made. Note that it is also possible to construct geometrically the solution, with exact results.

5.2. Approximate Knowledge of I_f and N_f

When I_f and N_f are known only approximately, i.e., we know only an interval that contains all the I_f values and an interval that contains all the N_f values, we lose the unique intersection of Lemma 4, that would require only n_a tests. Moreover, with a finite number of tests we cannot guarantee in general to satisfy the exact hypotheses of Theorem 1. As a consequence, various issues open up. Let's start by showing in Figure 8 that when all the four conditions of (15) hold with equality, the true solution lies on the boundary of the region I_{zr} as already mentioned in Section 3.2. If this happens, then with the conditions of Theorem 1 on the upward/downward outgoing gradients, the region I_{zr} is

a point. When all the four conditions of (15) hold with strict inequalities, the true solution lies inside the region I_{zr} (Figure 8b). From a theoretical point of view this distinction has a big importance, since it means that the zero-residual region can or cannot be reduced to a single point. From a practical point of view it becomes less important, for the moment, since we cannot guarantee that the available tests will reduce I_{zr} exactly to a single point and we will arrive most of the times to an approximate estimate. This can be more or less accurate, but this depends on the specific application, and this is out of the scope of the present work.

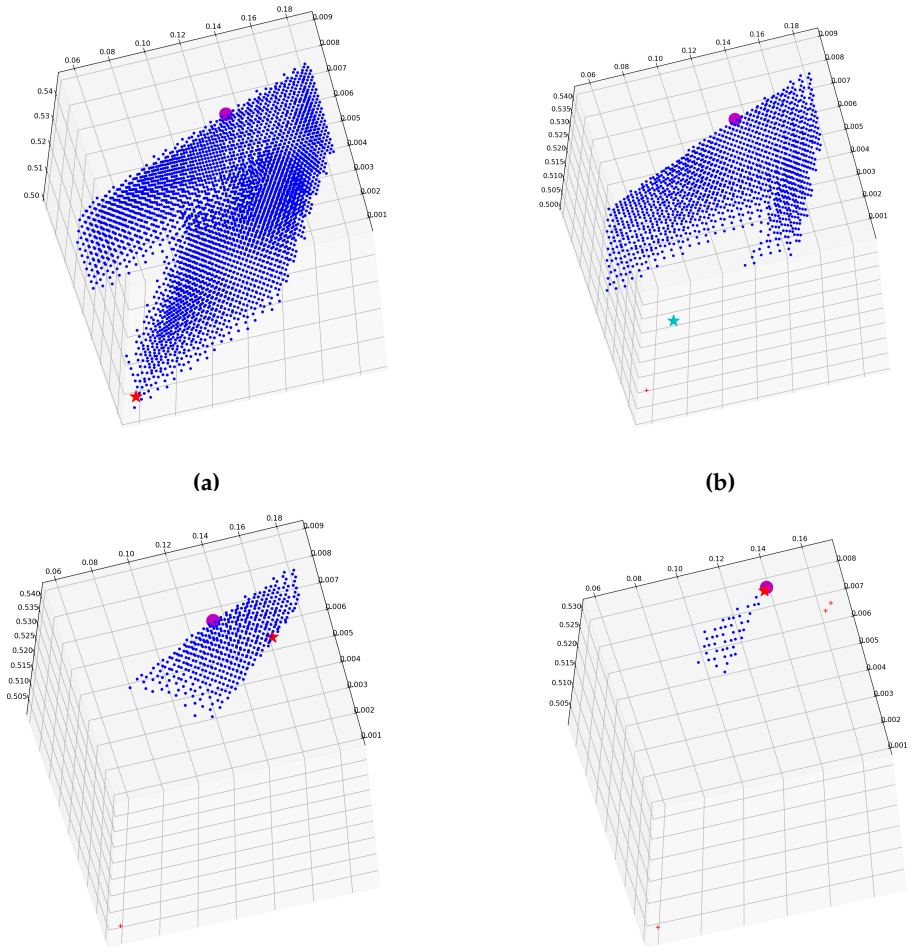


(a) **(b)**

Figure 8. Two examples of (zero-residual) intersection regions $I_{zr} \subset \mathbb{R}^3$ with different location of the true solution: inside the region or on its border. For graphical reasons the region has been discretized and the dots are the grid nodes; the bigger ball (thick point) is the true solution. (**a**): The true solution (ball) is on the border of I_{zr}; (**b**): The true solution (ball) is internal to I_{zr}.

To be more precise, when the conditions of Theorem 1 are not satisfied, there is an entire region of the parameters space which satisfies exactly problem (30), but only one point of this region is the true solution $\bar{x}_a$. As more tests are added and intersected together, the zero-residual region I_{zr} tends to reduce, simply because it must satisfy an increasing number of inequalities. In Figure 9 we can see four iterations taken from an example, precisely with 3, 5, 9 and 20 tests intersected and $m_{err} = 19$. With only three tests (Figure 9a), there is a big region I_{zr} (described by the mesh of small dots), and here we see that the true solution (thick point) and the current estimate (star) stay on opposite sides of the region, as accidentally happens. With five tests (Figure 9a) the region has shrunk considerably and the estimate is reaching the boundary (in the plot it is still half-way), and even more with nine tests (Figure 9c). The convergence arrives here before the region collapses to a single point, because accidentally the estimate has approached the region boundary at the same point where the true solution is located.

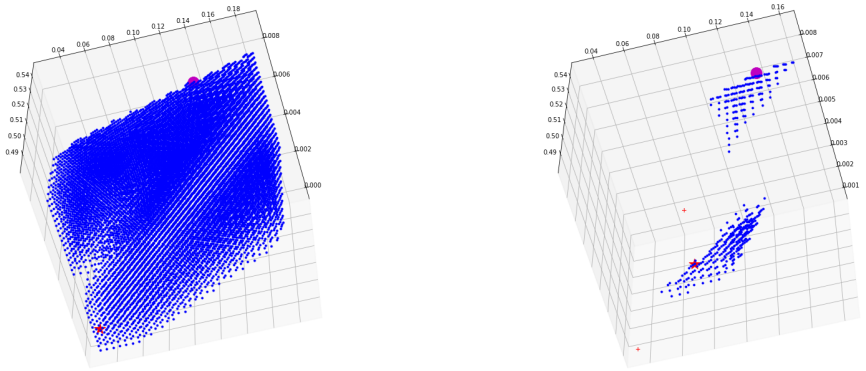
In general, the zero-residual region Z_{r_i} (20) of each test contains the true solution, while the estimate arrives from outside the region and stops when it bumps the border of the intersection region I_{zr} (21). For this reason we have convergence when the region that contains the true solution is reduced to a single point, and the current estimate $\hat{x}_a$ does not lie in a disconnected sub-region of I_{zr} different from the one in which the true solution lies. Figure 10 shows an example of an intersection region I_{zr} which is the union of two closed disconnected regions: this case creates a local minimum in problem (30).



(a) **(b)**

(c) **(d)**

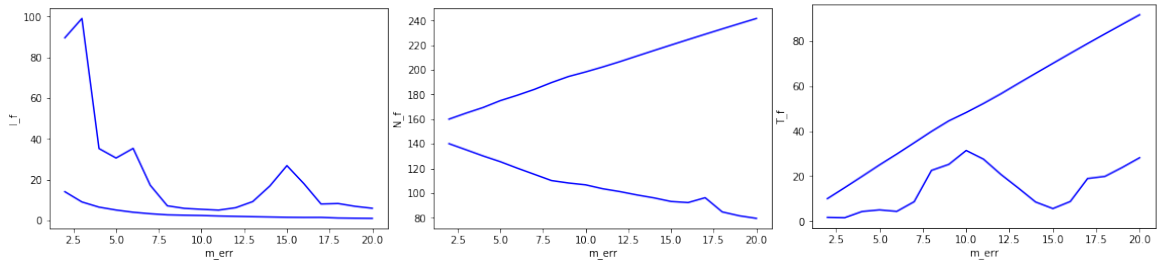
Figure 9. The intersection region $I_{zr} \subset \mathbb{R}^3$ at different number of tests involved. For graphical reasons the region has been discretized and the dots are the grid nodes; the bigger ball is the true solution and the star is the current estimate in the experiment. (**a**) 3 tests; (**b**) 5 tests; (**c**) 9 tests; (**d**) 20 tests.



(a) **(b)**

Figure 10. The intersection region $I_{zr} \subset \mathbb{R}^3$ at different number of tests involved. On the left a few tests have created a single connected region while, on the right, adding more tests have splitted it into two subregions. For graphical reasons the region has been discretized and the dots are the grid nodes; the bigger ball is the true solution and the star is the current estimate in the experiment. (**a**) A (portion of a) connected region I_{zr}; (**b**) A region I_{zr} split into two not connected sub regions.

In Figure 11 we see the differences $N_f^{max} - N_f^{min}$ and $T_f^{max} - T_f^{min}$ vs. m_{err}. The differences are bigger for higher values of the model error. It seems that this is the cause of a more frequent creation of local minima.



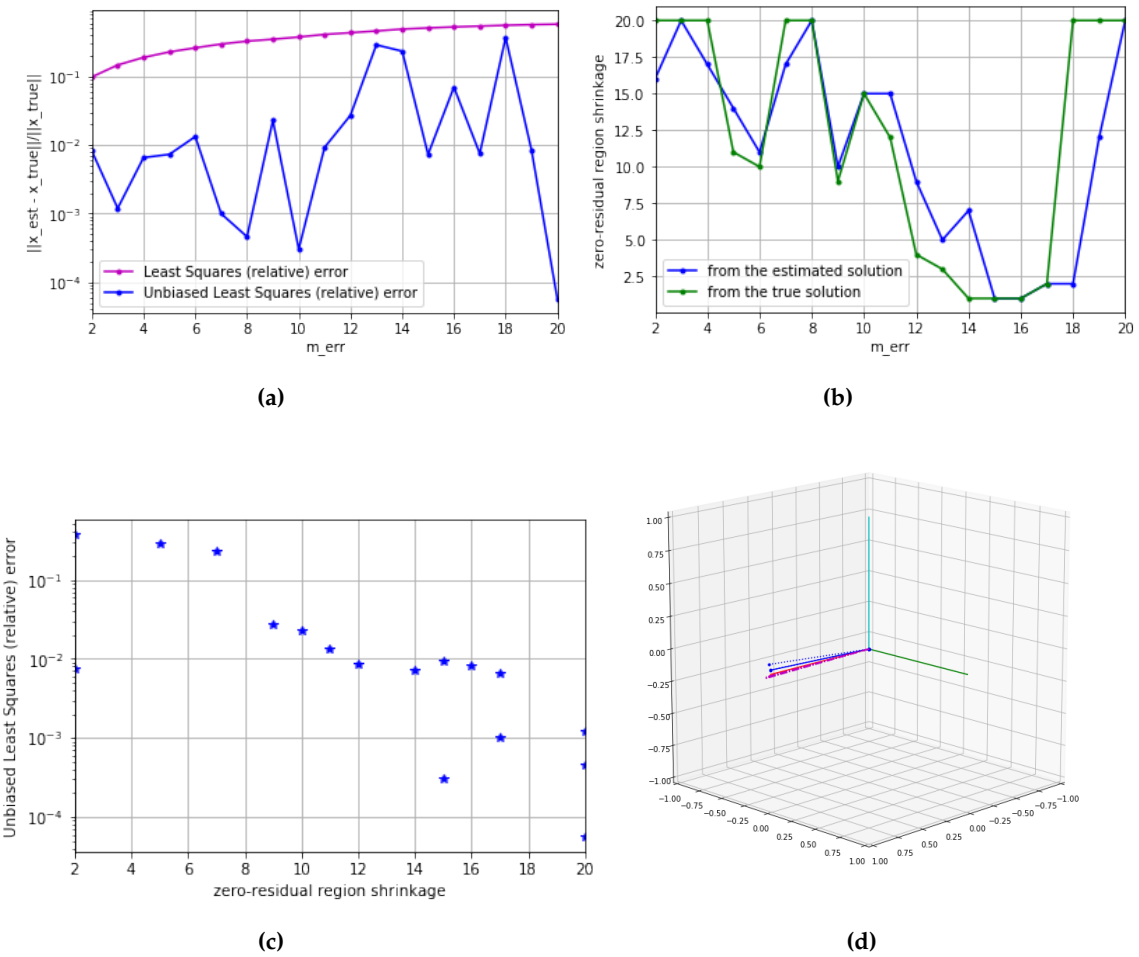
(a) **(b)** **(c)**

Figure 11. The three plots show the values assumed by the extreme values (15) as a function of m_{err}. **(a)**: $\{I_f^{min}, I_f^{max}\}$ vs. m_{err}; **(b)**: $\{N_f^{min}, N_f^{max}\}$ vs. m_{err}; **(c)** $\{T_f^{min}, T_f^{max}\}$ vs. m_{err}.

Figure 12 synthesizes the main results that we have experienced with this new approach. Globally it shows a great reduction of the bias contained in the standard least-squares estimates; indeed, we had to use the logarithmic scale to enhance the differences in the behaviour of the proposed method while varying m_{err}. In particular,

- with considerable levels of modelling error, let us say m_{err} between 2 and 12, the parameter estimation error $\hat{e}_{rel}$ is at least one order of magnitude smaller that that of least-squares; this is accompanied by high levels of shrinkage of the zero-residual region (Figure 12b);
- with higher levels of m_{err}, we see a low shrinkage of the zero-residual region and consequently an estimate whose error is highly oscillating, depending on where the optimization algorithm has brought it to get in contact with the zero-residual region;
- at $m_{err} = 18$ we see the presence of a local minimum, due to the falling to pieces of the zero-residual region as in Figure 10: the shrinkage at the true solution is estimated to be very high, while at the estimated solution it is quite low, since it is attached to a disconnected, wider sub-region.
- the shrinking of the zero-residual region is related to the distribution of the outgoing gradients, as stated by Theorem 1: in Figure 12d we see that in the experiment with $m_{err} = 18$ they occupy only three of eight orthants, while in the best results of the other experiments the gradients distribute themselves in almost all orthants (not shown).

It is evident from these results that for lower values of modelling error m_{err}, it is much easier to produce tests that reduce the zero-residual region to a quite small interval of R^{n_a}, while for high values of m_{err} it is much more difficult and the region I_{zr} can even fall to pieces, thus creating local minima. It is also evident that a simple estimate of the I_{zr} region size, like (31), can reliably assess the quality of the estimate produced by the approach here proposed, as summarized in Figure 12c.



(a) **(b)**

(c) **(d)**

Figure 12. The plots summarize the results obtained by the ULS approach to parameter estimation no the model problem explained at the beginning of this section. (**a**): The relative estimation error (37) vs. m_{err}; (**b**): The I_{zr} region shrinkage estimate (31) vs. m_{err}; (**c**): The relative estimation error (37) vs. the estimate of the I_{zr} region shrinkage, considering the experiments with $m_{err} \in [2, 20]$; (**d**): A three dimensional view of the Outgoing Gradients at the last iteration of the experiment with $m_{err} = 18$.

6. Conclusions

In this paper we have analyzed the bias commonly arising in parameter estimation problems where the model is lacking some deterministic part of the system. This result is useful in applications where an accurate estimation of parameters is important, e.g., in physical (grey-box) modelling typically arising in the model-based design of multi-physical systems, see e.g., the motivations that the authors did experience in the design of digital twins of controlled systems [10–12] for virtual prototyping, among an actually huge literature.

At this point, the method should be tested in a variety of applications, since the ULS approach here proposed is not applicable black-box as Least-Squares are. Indeed, it requires some additional a priori information. Moreover, since the computational complexity of the method here presented is relevant, efficient computational methods must be considered and will be a major issue in future investigations.

Another aspect that is even worth to deepen is also the possibility to design tests that contribute optimally to the reduction of the zero-residual region.

Author Contributions: Conceptualization, methodology, validation, formal analysis, investigation, software, resources, data curation, writing—original draft preparation, writing—review and editing, visualization: M.G. and F.M.; supervision, project administration, funding acquisition: F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the project DOR1983079/19 from the University of Padova and by the doctoral grant "Calcolo ad alte prestazioni per il Model Based Design" from Electrolux Italia s.p.a.

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

References

1. Björck, A. *Numerical Methods for Least Squares Problems*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 1996. [CrossRef]
2. Huffel, S.V.; Markovsky, I.; Vaccaro, R.J.; Söderström, T. Total least squares and errors-in-variables modeling. *Signal Process.* **2007**, *87*, 2281–2282. [CrossRef]
3. Söderström, T.; Soverini, U.; Mahata, K. Perspectives on errors-in-variables estimation for dynamic systems. *Signal Process.* **2002**, *82*, 1139–1154. [CrossRef]
4. Van Huffel, S.; Vandewalle, J. *The Total Least Squares Problem: Computational Aspects and Analysis*; Frontiers in Applied Mathematics (Book 9); Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 1991.
5. Peck, C.C.; Beal, S.L.; Sheiner, L.B.; Nichols, A.I. Extended least squares nonlinear regression: A possible solution to the "choice of weights" problem in analysis of individual pharmacokinetic data. *J. Pharmacokinet. Biopharm.* **1984**, *12*, 545–558. [CrossRef] [PubMed]
6. Meyer, C.D. *Matrix Analysis and Applied Linear Algebra*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2000.
7. Hansen, P.C. Oblique projections and standard-form transformations for discrete inverse problems. *Numer. Linear Algebra Appl.* **2013**, *20*, 250–258. [CrossRef]
8. Nocedal, J.; Wright, S. *Numerical Optimization*; Springer: Berlin, Germany, 1999.
9. Krause, P.C. *Analysis of Electric Machinery*; McGraw Hill: New York, NY, USA, 1986.
10. Beghi, A.; Marcuzzi, F.; Rampazzo, M.; Virgulin, M. Enhancing the Simulation-Centric Design of Cyber-Physical and Multi-physics Systems through Co-simulation. In Proceedings of the 2014 17th Euromicro Conference on Digital System Design, Verona, Italy, 27–29 August 2014; pp. 687–690. [CrossRef]
11. Beghi, A.; Marcuzzi, F.; Rampazzo, M. A Virtual Laboratory for the Prototyping of Cyber-Physical Systems. *IFAC-PapersOnLine* **2016**, *49*, 63–68. [CrossRef]
12. Beghi, A.; Marcuzzi, F.; Martin, P.; Tinazzi, F.; Zigliotto, M. Virtual prototyping of embedded control software in mechatronic systems: A case study. *Mechatronics* **2017**, *43*, 99–111. [CrossRef]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).