Chapter 1
Mathematical Modeling

In this chapter we present some of the aspects of the interface between mathematics and its applications, the so-called *mathematical modeling*. This is neither mathematics nor applied mathematics and is usually performed by scientists and engineers. We do this in the context of linear algebra, which allows easy comparison between direct and inverse problems. We also show some of the modeling stages and encourage the use of a least-squares method. An application to the study of flow in a porous medium is presented. Some key issues pertaining to the use of models in practical situations are discussed.

1.1 Models

We want to think about how we may come to understand a phenomenon, process or physical system. To simulate this endeavor, we make a *thought experiment*[1], under the assumption of conducting an investigation into the behavior of a hypothetical system, schematically represented by a black box.

Consider a *black box* whose inner working mechanisms are unknown and that, given a *stimulus* or *input*, answers with a *reaction* or *output*. See Fig. 1.1.

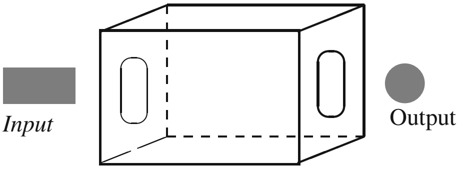

Fig. 1.1 Black box: the mental prototype

We aim to foretell the box behaviour in several distinct situations. In particular, we would like to tackle the following problems:

P_1: Given an arbitrary stimulus, tell what the corresponding reaction will be;

P_2: Given the reaction, tell what stimulus produced it.

However, we are not interested in generic predictions, but only in those based in scientific descriptions of the behaviour of the black box. To that end, we will

[1] Thought experiments have been used by several scientists, as for instance G. Galilei (1564-1642), R. Descartes (1596-1650), I. Newton (1643-1727), and A. Einstein (1879-1955), and it turns out to be a good artifact of investigation to keep in mind.

associate to the real situation a physical model and a corresponding mathematical model. Predictions will be made from the latter.

We shall call *physical model* any description of the phenomena involved using such concepts as, for example, number of items, mass, volume, energy, momentum, charge and their conservation, exchange, movement, or transference, etc.

By *mathematical model* we mean any kind of mathematical structure, such as, for example, a function, an equation, a set with an equivalence relationship, a vector space, a group, a graph, a probability distribution, a Markov chain, a neural network, a system of non-linear partial differential equations, an elliptic operator defined on a fiber space, etc.

Our guiding methodology when choosing a model, which we shall call *modeling*, is part of the *scientific method*, that we present in a somewhat *modern* language.

In practical situations we try to choose or develop a mathematical model that best describes the physical model, avoiding contradictions with reality and deviations from the phenomena of interest. We only keep the model as long as it is not refuted by experimental data[2].

1.2 Observing Reality

The sleuth that is to solve a crime, or, in other words, who is to foretell the behaviour of a person or group of persons in a given situation, must create a careful profile of those persons and know, as best he can, the episode and the circumstances. Thus, he must bear in mind the various elements of the deed, the deeds before it and its consequences. To that end he must search for information where it is available, which, logically, includes the crime scene. He must observe and carry on the investigation, questioning persons, in any way, related to the crime, and analyzing their answers[3].

Our goal is to be able to guess the behaviour of the black box, Fig. 1.1. Analogously to the observation of a crime's trail, we begin the *observational* and *experimental phase* of our project. By applying different stimuli to the black box we observe and take note of the corresponding reactions.

[2] The mathematical model will be accepted as long as its predictions do not contradict experimental facts. That is a very strong requirement of the *scientific discourse*. If the model fails, one does not have any impediment to throw it away, to reject it, no matter for how long and for how many it has been used. In fact, one is obliged to reject it. This very strong stance, singles out the scientific method. This is precisely what Johannes Kepler (1571-1630) did. At a certain point, he believed that the orbits of the planets, around the sun, were circles. However, the data collected by Tycho Brahe (1546-1601) indicated that the orbit of Mars could not be fitted by a circle. This difficulty led him to discard the hypothesis that the orbits were circles, and embrace the more accurate model of elliptical orbits.

[3] Obviously, he can also consider the forensic analysis that makes it possible to track the trajectory of a projectile, a body, or a car, and any kind of materials involved. This constitutes, in foresight, a set of inverse problems, which, naturally, uses knowledge of physics and mathematics.

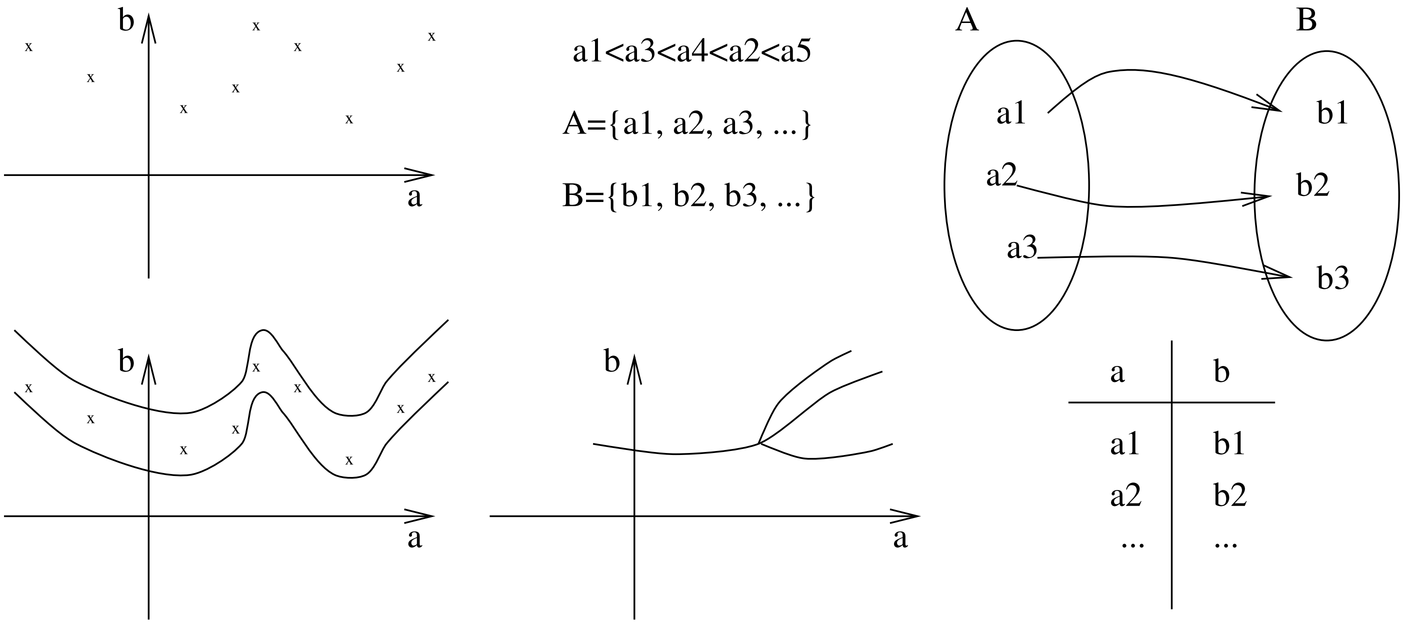

Fig. 1.2 Organization of experimental data (from left to right, and from top to bottom): (a) Scatter-plot; (b) Set *A* represents a group of stimuli and set *B* a group of reactions; (c) Representing a function by a Venn diagram; (d) Functions that approximate and bound the experimental data; (e) Forking experimental data: it does not conform to a function but conforms to a level set of a function; (f) Table

The next step is to organize the experimental data in a way suitable for analysis. We thus create a *database of real* or *experimental data*. This is critical, since the way data is organized will emphasize certain aspects at the expense of others. Information can be presented, for example, as tables, diagrams, graphs and images. Some possible data structures are sketched in Fig. 1.2. Catalogues of possible stimuli and conceivable reactions can be created (Fig. 1.2b, first row, second column). Sometimes, these catalogues correspond to the *domain* and the *codomain* of appropriate functions.

Assume we want to answer problems formulated in Section 1.1, page 7, and some other questions that may arise out of curiosity or by chance. However, we would want to avoid resorting, everytime, to experimentation. With this is mind, but still wanting to do it scientifically, we build a mathematical model of the situation.

As a matter of fact, the experimental database allows one to answer some problems. Many people would be perfectly happy to solve their problems that way. Then, why should we build a mathematical model of the situation in the first place? Among the many reasons to do so, we mention two:

- Once a mathematical model has been devised, a natural environment is available in which several settings and hypothesis can be generated and tested, i.e. a number of scenarious may be constructed. In the example, it is possible, by standard deductive reasoning (logic implication), to propose candidates for reactions due to a wider range of stimuli values;
- A remarkable advantage of a mathematical model is the *compression* of information. Once the model is obtained, additional structures within the database become apparent. Those perceived structures are enough to render almost useless large parts of the database. This results in effective compression[4] which, in turn, leads to comprehension.

1.3 The Art of Idealization: Model Characterization

A class of models is chosen according to the given situation, technological and/or pratical measuring capabilities, purpose, and so on. Choosing a class of models is known as model *characterization*. We remark that this endeavour is not an application of mathematics in itself (in the sense that it is not solely the carrying out of an algorithm); indeed, it is an *intelligent* activity, a task generally highly complex, an art that demands an educated sensibility to execute it at its best. As a matter of fact,

[4] It is relevant to realize that *mathematical models* can effectively *compress data*. To illustrate this point, assume that we are dealing with data on the price of photocopies. If one has a database, then we need to have a table relating the number of copies and its price. We would have, for example, that one copy costs 10 cents, two copies cost 20 cents, three copies cost 30 cents and so on up to, say, a hundred copies that cost 10 monetary units (m.u.), a very long table. And it could be longer! By means of a mathematical model, using the notion of function, we say that the price of n copies is $n \times 0.10$ m.u., with $n \in \mathbb{N}$.

this is so much true that, at times, in novel situations, the stage of characterization may require the development of new mathematical structures[5].

To characterize the model, for example, we can perform an *exploratory data analysis*, in which we *contemplate* the data and afterwards may point out simple relationships between them.

Let's make it concrete by returning to the black box example. Assume that the stimulus and its corresponding reaction can be quantified/described each of them by three measurings that we collect in vectors[6], the

$$\textit{input signal}, \quad \mathbf{x} = (x_1, x_2, x_3)^T \in \mathbb{R}^3 \,,$$

and the

$$\textit{output signal}, \quad \mathbf{y} = (y_1, y_2, y_3)^T \in \mathbb{R}^3 \,.$$

The description of the input/output signals as elements of $\mathbb{R}^3$ is evidently part of the characterization stage[7,8]. The summing up of the data in a table constitutes one of the simplest quantitative models, the *DB (database)* model.

An exploratory analysis of the available data may suggest as reasonable the following additional hypotheses on the black box workings:

H_1: **Reproducibility** — The repetition of the same input signal produces the same output signal[9];

[5] An illustration of this is the invention of complex numbers by Tartaglia (1500-1557) and Cardano (1501-1576) in 1545 to solve equations of the third degree. These equations were associated with practical problems, which indicated that they had three real roots. However, the available methods did not allow their solution because they required the square root of a negative number, which, at that time, had not yet been defined.

[6] Here, vectors are *vertical*, or column vectors, that is, $n \times 1$ matrices. Also, they are represented using *bold* letters. The entries of a $m \times n$ matrix, A, are denoted by A_{ij}, $i = 1, \ldots, m$, and $j = 1, \ldots, n$. For A, in particular for vertical vectors, A^T represents the *transpose* of A, that is, the matrix whose entries are given by $(A^T)_{ij} = A_{ji}$, for all $i = 1, \ldots, n$, $j = 1, \ldots, m$.

[7] We could have chosen $\mathbb{R}^4$, as the set of possible inputs if measures were related to three-dimensional space, and we also had to characterize time. Or, maybe, even higher order dimensional spaces could be necessary as is the case when, for example, we want to keep other information such as temperature, pressure, etc in the output signal.

[8] Note the non-sequential nature of modeling. Even before acquiring the experimental data of which we spoke in the previous section, we must start characterizing the model. See Section 2.8, page 44, and the Afterword, starting on page 177 for further considerations on modeling.

[9] It should be pointed out that *the same* here subtends *within* a certain margin of error associated with the model. The variability could be, for instance, characterized by a probabilistic random variable, or by an interval. This would require a more complete or detailed model. The characterization of a model always assumes some idealization. That is, even though the real phenomena do not strictly satisfy a certain assumption, we assume they satisfy in order to proceed with modeling. What is important is that conclusions from the model and from reality do not differ significantly. Sometimes this is a quantitative statement: the difference between numbers, coming from the model, and data, from reality, should be bounded by a certain predefined value. Modeling is an art that has to be practiced for one to have a good grasp on it.

H_2: **Proportionality** — If the input signal is amplified α times, the output signal is also amplified by that same factor;

H_3: **Superposition of effects** — If we add up two input signals, the output signal is the sum of the output signals corresponding to the individual input signals[10].

These hypotheses form an example of what we mean by a *physical model* of the real situation. Espousing these hypotheses allows us to obtain a model that is more complete than the DB model, a *descriptive model*[11]. This, in principle, contains the data set, though not necessarily in an explicit manner[12].

Mathematically, the first hypothesis means that

H_1: The attribution, *input → output*, is a function.

We shall denote that function by $\mathcal{F}$, and call it the *behaviour* function of the black box[13]. The second and third hypotheses essentially characterize $\mathcal{F}$ as a *linear function*,

H_2: $\mathcal{F}$'s evaluation commutes with scalar multiplication[14],

$$\mathcal{F}(\lambda\,\mathbf{u}) = \lambda\,\mathcal{F}(\mathbf{u}) \text{ for all } \lambda \in \mathbb{R},\ \mathbf{u} \in \mathbb{R}^3\ , \tag{1.2}$$

[10] In a sense, this means that inputs do not interact. Each one goes through the black box in its own way, ignoring the other.

[11] Descriptive models, roughly speaking, try to answer questions like: *"What happened?"* ("*What*") or *"When it happened?"* ("*When*"). Databases try to answer questions of the same nature. On the other hand, the *explanatory model* focuses in: *"How it happened?"* ("*How*"). Further discussion on the nature of models is presented in Afterword, page 177.

[12] What we mean here is that, for example, the function $\mathbb{R} \ni x \mapsto f(x) = x^2 \in \mathbb{R}$ contains the information of the following table,

x	-1	0	1	2
y=f(x)	1	0	1	4

We could say that f *encapsulates* the table.

[13] It is worth to remind that, for some functions, different inputs imply different outputs, whereas for some other functions, certain different inputs can give the same output. The former case is not the rule and deserves a special name, the function is called *injective* or a *one-to-one* function.

[14] $\mathcal{F}$ of a multiple is the multiple of $\mathcal{F}$. One may have some restrictions on this way of stating this property, and rightly so. Let us be more precise. First consider the function *multiplication by scalar* λ, given by

$$\begin{array}{rcl} \Lambda : \mathbb{R}^3 & \to & \mathbb{R}^3 \\ \mathbf{v} & \mapsto & \Lambda(\mathbf{v}) = \lambda\mathbf{v}\ . \end{array}$$

Hypothesis H_2 can be written in terms of a commutation of a composition of functions,

$$\mathcal{F} \circ \Lambda = \Lambda \circ \mathcal{F}\ , \tag{1.1}$$

which justifies the assertion at the beginning of this footnote.

H_3: The evaluation of $\mathcal{F}$ commutes with vector summation[15],

$$\mathcal{F}(\mathbf{u}+\mathbf{v}) \quad = \quad \mathcal{F}(\mathbf{u})+\mathcal{F}(\mathbf{v}) \text{ for all } \mathbf{u},\ \mathbf{v} \in \mathbb{R}^3 \,. \tag{1.4}$$

Granting the validity of the hypothesis H_1, the act of organizing the data as stimuli or reactions, as suggested in Fig. 1.2b, page 9, would correspond, respectively, to the construction of the domain and the codomain of the behaviour — characterized by a function — of the black box.

In our example, the characterization stage reaches its end when we specify the model, by saying that the black box behaviour is described by a function, which is linear.

Non-linear models are relevant in many applications and will be dealt with in other sections within this book. We only considered the linear model so far, to facilitate the understanding of the concepts presented here.

1.4 Resolution of the Idealization: Mathematics

We are assuming that the mathematical model of the black box is a linear function, from $\mathbb{R}^3$ to $\mathbb{R}^3$. This characterizes the model. However, we do not know which specific linear function it is. One needs to determine that function, or, at least, to approximate it, in order to solve the problems posed in Section 1.1. This is the stage

[15] $\mathcal{F}$ of a sum is the sum of $\mathcal{F}$'s. Similarly to the previous footnote, we can rewrite hypothesis H_3 as commutation of composition of functions. The idea of preserving the notion of commutation is a neat one. However, to do that, one sometimes has to work a bit (to adapt notions). Let $\mathcal{S}$ be the function that adds two vectors,

$$\begin{aligned} \mathcal{S} : \mathbb{R}^3 \times \mathbb{R}^3 &\quad\to\quad \mathbb{R}^3 \\ (\mathbf{u},\mathbf{v}) &\quad\mapsto\quad \mathcal{S}(\mathbf{u},\mathbf{v}) = \mathbf{u}+\mathbf{v} \,. \end{aligned}$$

Also, let $\bar{\mathcal{F}}$ be, essentially, two copies of $\mathcal{F}$, representing the function

$$\begin{aligned} \bar{\mathcal{F}} : \mathbb{R}^3 \times \mathbb{R}^3 &\quad\to\quad \mathbb{R}^3 \times \mathbb{R}^3 \\ (\mathbf{u},\mathbf{v}) &\quad\mapsto\quad \bar{\mathcal{F}}(\mathbf{u},\mathbf{v}) = (\mathcal{F}(\mathbf{u}),\mathcal{F}(\mathbf{v})) \,. \end{aligned}$$

It may seem odd to define such a function, but that is exactly what we need here. In fact, Eq. (1.4) can be rewritten as

$$\mathcal{F} \circ \mathcal{S} \quad = \quad \mathcal{S} \circ \bar{\mathcal{F}} \,. \tag{1.3}$$

Note that if we stick to the *computer science* notion of *overloading* a function, we can still denote $\bar{\mathcal{F}}$ by $\mathcal{F}$. In this case, Eq. (1.3) could be written simply as $\mathcal{F} \circ \mathcal{S} = \mathcal{S} \circ \mathcal{F}$.

Simply put, we just want to say that, for a linear function, $\mathcal{F}$ of a sum is the sum of the $\mathcal{F}$'s, or that $\mathcal{F}$ commutes with summation. To make this statement precise, we need to do what we have just done. When you have to spell it out, before simplicity becomes simple, it may seem somewhat complex at times. Finally, we are not always blessed with commutation, so when we are, nothing more appropriate than to make an effort to recognize it.

Table 1.1 Data set: stimuli × reactions

stimuli	$\mathbf{x} \in \mathbb{R}^3$	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_3$
reactions	$\mathcal{F}(\mathbf{x}) \in \mathbb{R}^3$	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$

in which the model is *determined*, where an inverse identification problem is solved to pinpoint a specific linear function.

Example 1.1. Model identification using an ideal database. In the example we are considering, it is very easy to determine the model. We will do it first in a particular case. Let $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, be the *canonical basis* of $\mathbb{R}^3$,

$$\mathbf{e}_1 = (1,0,0)^T\,, \quad \mathbf{e}_2 = (0,1,0)^T\,, \quad \mathbf{e}_3 = (0,0,1)^T\,.$$

Assume that we know the reaction of the black box when stimuli $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ are applied. That is, let us say we have the information contained in the *ideal database*

$$\mathbf{e}_1 \overset{\mathcal{F}}{\mapsto} \mathcal{F}(\mathbf{e}_1) = \mathbf{a}_1\,, \quad \mathbf{e}_2 \overset{\mathcal{F}}{\mapsto} \mathcal{F}(\mathbf{e}_2) = \mathbf{a}_2\,, \quad \mathbf{e}_3 \overset{\mathcal{F}}{\mapsto} \mathcal{F}(\mathbf{e}_3) = \mathbf{a}_3\,, \tag{1.5}$$

where $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3 \in \mathbb{R}^3$. This dataset could also be represent by Table 1.1.

Now, construct the 3×3 matrix A whose columns are formed by the images of the vectors $\mathbf{e}_j$, $\mathcal{F}(\mathbf{e}_j)$, for $j = 1, 2, 3$, i.e.:

$$\begin{aligned} A &= \begin{pmatrix} | & | & | \\ \mathcal{F}(\mathbf{e}_1) & \mathcal{F}(\mathbf{e}_2) & \mathcal{F}(\mathbf{e}_3) \\ | & | & | \end{pmatrix} \\ &= \begin{pmatrix} | & | & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 \\ | & | & | \end{pmatrix}. \end{aligned} \tag{1.6}$$

Then, if a generic stimulus is denoted by $\mathbf{x} = (x_1, x_2, x_3)^T \in \mathbb{R}^3$, the reaction $\mathcal{F}(\mathbf{x})$ will be given by the product $A\mathbf{x} \in \mathbb{R}^3$. In fact, by $\mathcal{F}$'s linearity,

$$\begin{aligned} \mathcal{F}(\mathbf{x}) &= \mathcal{F}(x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + x_3\mathbf{e}_3) \\ &= x_1\mathcal{F}(\mathbf{e}_1) + x_2\mathcal{F}(\mathbf{e}_2) + x_3\mathcal{F}(\mathbf{e}_3) \\ &= \begin{pmatrix} | & | & | \\ \mathcal{F}(\mathbf{e}_1) & \mathcal{F}(\mathbf{e}_2) & \mathcal{F}(\mathbf{e}_3) \\ | & | & | \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \end{aligned}$$

or, simply put, the reaction $\mathcal{F}(\mathbf{x})$ is given by the product $A\mathbf{x} \in \mathbb{R}^3$,

$$\mathcal{F}(\mathbf{x}) = A\mathbf{x}\,. \tag{1.7}$$

■

In the previous example, determining the model, that is, choosing a specific linear function $\mathcal{F}$, boils down to finding A, as can be seen immediately from Eq. (1.7).

We are now in a good position to define a third general problem. Although it is not as fundamental as the two presented in Section 1.1, even so it is equally important, and it is instrumental in the resolution of the former,

P_3: From a structured data set, determine matrix A.

If, as we assume, Eq. (1.5) is known to us, i.e., we have the information contained in that equation, or equivalently in Table 1.1, we can see that, based on Eq. (1.6), it is more than automatic to obtain A.

Now, let us consider the more general situation when we know $\mathcal{F}$ in some given basis of $\mathbb{R}^3$, not necessarily the canonical one.

Example 1.2. Identification using more general ideal data. Assume the values of $\mathcal{F}$ in a basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ of $\mathbb{R}^3$ are known[16]. That is, let us say that the ideal database

$$\mathbf{u}_1 \overset{\mathcal{F}}{\mapsto} \mathcal{F}(\mathbf{u}_1) = \mathbf{v}_1\,, \quad \mathbf{u}_2 \overset{\mathcal{F}}{\mapsto} \mathcal{F}(\mathbf{u}_2) = \mathbf{v}_2\,, \quad \mathbf{u}_3 \overset{\mathcal{F}}{\mapsto} \mathcal{F}(\mathbf{u}_3) = \mathbf{v}_3\,, \tag{1.8}$$

is available to us. Clearly, this information could be organized in a table similar to Table 1.1.

This is motivated by the fact that, for practical experimental[17], economical, or even ethical reasons, it is not always possible to apply the rather special choice of stimuli $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$, but maybe it is possible to apply stimuli $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$.

Since the columns of matrix A in Eq. (1.6) are given by the images of the canonical vectors by $\mathcal{F}$, and since we have to determine them from the information in Eq. (1.8), we must, in the first place, write the canonical vectors in terms of the elements of the basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$. We will use $\mathbf{e}_1$ as an example. Let c_1, c_2 and c_3 be the only scalars such that

$$\mathbf{e}_1 \quad = \quad c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + c_3\mathbf{u}_3\,.$$

Let U be the matrix whose columns are the vectors $\mathbf{u}_i$, $i = 1{,}2{,}3$ and let V be the matrix such that its columns are $\mathbf{v}_i$, i.e.,

$$U = \begin{pmatrix} | & | & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \\ | & | & | \end{pmatrix}, \quad \text{and} \quad V = \begin{pmatrix} | & | & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \\ | & | & | \end{pmatrix}. \tag{1.9}$$

Also, let $\mathbf{c} = (c_1, c_2, c_3)^T$. With this notation,

$$\begin{aligned} \mathbf{e}_1 &= c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + c_3\mathbf{u}_3 \\ &= \begin{pmatrix} | & | & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \\ | & | & | \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}, \end{aligned}$$

or, simply,

$$\mathbf{e}_1 = U\,\mathbf{c}\,. \tag{1.10}$$

[16] See page 189 in the Appendix A, to recall the definition of a basis of a vector space.

[17] This issue is exemplified in the last paragraph of Section 0.2, page 2.

Now, multiplying both sides of Eq. (1.10) by the inverse of matrix U, U^{-1}, and since $U^{-1}U=\mathcal{I}$ is the 3×3 identity matrix, we have

$$\mathbf{c} = U^{-1}\mathbf{e}_1 \,. \tag{1.11}$$

Notice that to obtain **c** explicitly corresponds to finding the solution of a system of linear equations (see Eq. (1.10), where the unknown is **c**). Determination of the first column of A results now from the linearity of $\mathcal{F}$, and Eqs. (1.8), (1.9), and (1.11),

$$\begin{aligned}\mathcal{F}(\mathbf{e}_1) &= \mathcal{F}(c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + c_3\mathbf{u}_3)\\ &= c_1\mathcal{F}(\mathbf{u}_1) + c_2\mathcal{F}(\mathbf{u}_2) + c_3\mathcal{F}(\mathbf{u}_3)\\ &= c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 \;=\; V\mathbf{c}\\ &= VU^{-1}\mathbf{e}_1 \,.\end{aligned}$$

If we do the same for $\mathbf{e}_2$ and $\mathbf{e}_3$, we can see that $\mathcal{F}(\mathbf{x}) = A\mathbf{x}$ with

$$A = VU^{-1} \,. \tag{1.12}$$

Thus, from Eq. (1.7) and the dataset, Eq. (1.8), we have just determined the model in this case,

$$\mathcal{F}(\mathbf{x}) = VU^{-1}\mathbf{x} \,.$$

■

It is worthwhile to pay attention to this result. Equation (1.12) justifies the following assertion[18]:

knowing a linear transformation in a basis[19] *leads to the full knowledge of it*[20].

[18] This is an interesting result. We have a function defined in $\mathbb{R}^3$. The set of all functions from $\mathbb{R}^3$ to $\mathbb{R}^3$ is a huge set. In principle, we have an enormous variety of possibilities to build a function whose domain and codomain are $\mathbb{R}^3$. Our task is simply this: we have to choose where to send through this function a very large number of points of the domain, having an equally large number of possibilities on the codomain from where to choose. Because it is a linear function, that is, because it satisfies a few simple algebraic rules, Eqs. (1.2) and (1.4), the number of possibilities to define the function is drastically reduced. In fact, knowing the value of the function in three randomly chosen points of $\mathbb{R}^3$ virtually determines the value of the function in all the other points. Of course, it is not just any three points that can be used, since, in particular, the origin cannot be chosen as one of those points. Technically, those three points must form a basis of $\mathbb{R}^3$. But, anyway, the limitation imposed by linearity is awesome: it reduces an uncountable set of information to a finite set with three data.

[19] In this example, to know a linear transformation on a basis is to have Eq. (1.8) or Eq. (1.9), (since both have the same structured information content).

[20] In the example, to know $\mathcal{F}$ is, by Eq. (1.7), to know Eq. (1.12).

1.5 Idealized Data Processing: Solution Operator

Notice that questions P_1 to P_3 are of different nature. Questions P_1 and P_2 are *natural* questions which can be posed by anyone who has devoted attention to the behaviour of the black box. However, question P_3 can only be asked by a *modeler*, that is, someone who tries to create a mathematical model to describe the behaviour of the black box.

After having characterized a model, problems P_1 to P_3 can be solved from the point of view of the mathematical model. This is very simple in this case, and that is what we will do next.

For the first one, given $\mathbf{x}$, an input signal, the solution is simply the product $A\mathbf{x}$. The second one is solved by means of the inverse of A. If $\mathbf{y}$ is the output, the input that generates it is $A^{-1}\mathbf{y}$.

Moreover, the third problem, determining A itself, can also be presented in mathematical terms. The determination of A was given in Eq. (1.12), $A = VU^{-1}$.

Schematically, we show in Table 1.2 the information required to answer each problem and the number crushing procedure that needs to be performed in these elementary linear algebra problems. It is customary to say that P_1 is a *direct problem* while P_2 and P_3 are *inverse problems*. This terminology is not in contradiction[21] with the kind of mathematical tasks involved to solve each one, see Table 1.2. Moreover, P_2 is an example of a so-called *reconstruction problem* and P_3 an *identification problem*.

It is now plain to see that, since the solutions to problems P_1 and P_2 depend on A, we need, in principle, to solve problem P_3 and only then deal with the other two.

On the other hand, as we will see in Section 1.8, knowing how to solve direct problems can be used for inverse problems resolution, if the notion of *solving* is made flexible.

In some applications, problems P_2 and P_3 can appear combined. One such example will be discussed in Chapter 4. There, one wants to recover a real image from a distorted one obtained from an experimental device. However, the distortion caused by the experimental technique is not known in advance.

1.6 Mind the Gap: Least Squares

We remark that, in the previous section, the solutions to problems P_1 to P_3 were constructed with, virtually, no contact with experiments, only in the realm of the mathematical model. In particular, it is assumed that the information in Eq. (1.8), which could come from an experiment, is ideally known, i.e., without errors.

In this section we will bring the mathematical model in contact with reality, using the physical model and the experimental data as a bridge. The criterion chosen to make such a contact possible is not a part of mathematics. However, its application is mathematical (or algorithmic).

[21] See, however, Exercise 1.3.

Table 1.2 Problem's nature and its data processing features

Problem	**Problem nature**	**Problem data** Required data	**Solution-operator** Data processing task	**Algorithm**
P_1	Direct	$(\mathbf{x}, A)$	$\mathbf{y} \hookleftarrow A\mathbf{x}$	matrix by vector multiplication
P_2	Inverse Reconstruction	$(\mathbf{y}, A)$	$\mathbf{x} \hookleftarrow A^{-1}\mathbf{y}$	matrix by vector multiplication
P_3	Inverse Identification	$U = (\mathbf{u}^1, \mathbf{u}^2, \mathbf{u}^3)$, $V = (\mathbf{v}^1, \mathbf{v}^2, \mathbf{v}^3)$	$A \hookleftarrow VU^{-1}$	matrix by matrix multiplication

The introduction of reality in the model begins with the answer to question P_3, basing it not on an ideal database, Eq. (1.8), but from a real database, made up of measurements with all the ambiguity thereof, that is, with all the contradictions between the experimental data and the hypotheses H_1 to H_3, listed on page 11, that may exist by chance (compare with Fig. 1.3, page 19).

In this case, the *estimation* of *m*odel parameters or the *identification of the model* corresponds to the determination of A, from experimental data[22].

Summing up, in the characterization stage, we choose a class of models. In the stage of estimation or identification, we pick, based on real data, one of the models within that class.

Example 1.3. Identification using experimental data. Let

$$\left\{(\mathbf{x}^k, \mathbf{y}^k) \in \mathbb{R}^3 \times \mathbb{R}^3, \text{ for } k = 1, \ldots, n\right\} \tag{1.13}$$

be the experimental data set, formed by ordered pairs (couples) of input and output signals. We say that the data is *perfectly representable* (or that it *can be interpolated*) by a linear function if there exists a matrix A such that

$$A\mathbf{x}^k = \mathbf{y}^k \quad \text{for all } k = 1, \ldots, n\,. \tag{1.14}$$

Otherwise we say that the data is not perfectly representable by a linear function. The one dimensional case is illustrated in Fig.1.3.

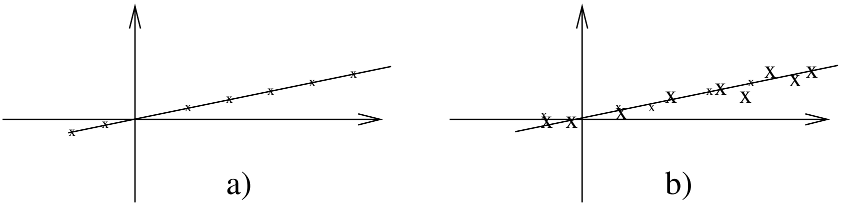

Fig. 1.3 Here are represented two data sets consisting of several pairs of real numbers, i.e., elements of $\mathbb{R}^2$. The first one may be viewed as related to an ideal data set and the second to a real, experimental, data set. (a) A set of input and output signals perfectly representable by a linear function. (b) The set of input and output signals displayed here is not perfectly representable by a linear function. However, we can choose a linear model to represent it, by means, for example, of the least squares method.

If the data is perfectly representable by a linear function we choose some input signals, $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$, that form a basis of $\mathbb{R}^3$ with the corresponding output signals, $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$, and Eqs. (1.9) and (1.12) define A.

[22] There are lots of ways to call this. Statisticians prefer, perhaps, *estimation*. You would *calibrate* the model, were you an engineer. Electrical engineers may prefer to speak of identification. Those which investigate artificial intelligence may *train* the model. Some call it *determination* of parameters, fitting the model to the data, or model selection.

If we want to keep the linear model even when the data is not perfectly representable by a linear function[23], we must relax the condition in Eq. (1.14). We consider now a way we can do it. Let

$$|\mathbf{v}| = (\sum_{j=1}^{3} v_j^2)^{\frac{1}{2}} = ((v_1)^2 + (v_2)^2 + (v_3)^2)^{\frac{1}{2}} ,$$

be the *Euclidean norm* of

$$\mathbf{v} = (v_1, v_2, v_3)^T \in \mathbb{R}^3 .$$

Also, let x_j^k be the j^{th} coordinate of $\mathbf{x}^k$, that is,

$$\mathbf{x}^k = (x_1^k, x_2^k, x_3^k)^T .$$

Denote by $M(3{,}3)$ the set of real 3×3 matrices. For $B \in M(3{,}3)$, we define half the *quadratic error function*[24],

$$\begin{aligned} E(B) &= \frac{1}{2}\sum_{k=1}^{n} |B\mathbf{x}^k - \mathbf{y}^k|^2 \\ &= \frac{1}{2}\sum_{k=1}^{n}\sum_{i=1}^{3}\left[\left(\sum_{j=1}^{3} B_{ij}x_j^k\right) - y_i^k\right]^2 . \end{aligned} \tag{1.15}$$

Note that the data is perfectly representable by A if and only if $E(A) = 0$.

We choose A, the *matrix defining the linear model*, as the one, within all $B \in M(3{,}3)$, that minimizes E,

$$A = \mathrm{argmin}_{B \in M(3,3)} E(B) ,$$

that is,

$$E(A) = \min_{B \in M(3,3)} E(B) .$$

■

This criterion is known as the *least squares method*. It bridges the gap between the mathematical model and the experimental data. We stress that the determination of A satisfying the aforementioned criterion is, strictly speaking, a mathematical problem.

[23] One reason to want to keep the linear model is its simplicity which is a value that is worth to strive for.

[24] Here, and elsewhere in this book, we use 1 / 2 for a slight simplification in the critical point equation. There is no need for it.

1.7 Optimal Solution

To obtain the optimal solution, the minimum point of E in Eq. (1.15) must be determined. We search for this minimum by means of the *critical point equation*,

$$\frac{\partial E}{\partial B_{lm}} = 0\,, \text{ for } l, m = 1, 2, 3\,.$$

We have

$$\frac{\partial B_{ij}}{\partial B_{lm}} = \delta_{il}\delta_{jm}\,,$$

where δ_{il}, the *Krönecker's delta*, is a notation[25] for the elements of the identity matrix, $\mathcal{I}$, i.e.,

$$\delta_{il} = \begin{cases} 1, & \text{if } i = l \\ 0, & \text{if } i \neq l \end{cases}\,.$$

Thus,

$$\begin{aligned}
\frac{\partial E}{\partial B_{lm}} &= \sum_{k=1}^{n}\sum_{i=1}^{3}\left[\left(\sum_{j=1}^{3} B_{ij}x_j^k\right) - y_i^k\right]\left(\sum_{j=1}^{3}\frac{\partial B_{ij}}{\partial B_{lm}}x_j^k\right) \\
&= \sum_{k=1}^{n}\sum_{i=1}^{3}\left[\left(\sum_{j=1}^{3} B_{ij}x_j^k\right) - y_i^k\right]\delta_{il}x_m^k \\
&= \sum_{k=1}^{n}\left[\left(\sum_{j=1}^{3} B_{lj}x_j^k\right) - y_l^k\right]x_m^k \\
&= \sum_{j=1}^{3} B_{lj}\left(\sum_{k=1}^{n} x_j^k x_m^k\right) - \sum_{k=1}^{n} y_l^k x_m^k \\
&= \left(\sum_{j=1}^{3} B_{lj}C_{jm}\right) - D_{lm}\,,
\end{aligned} \tag{1.16}$$

where

$$C_{jm} = \sum_{k=1}^{n} x_j^k x_m^k\,, \text{ and } D_{lm} = \sum_{k=1}^{n} y_l^k x_m^k\,. \tag{1.17}$$

The *critical point equation* for the critical point B (a matrix) is rewritten as

$$BC = D\,,$$

where C and D are known matrices, Eq.(1.17), determined from the data. The solution is then given by

$$B = DC^{-1}\,. \tag{1.18}$$

The expression for B looks like Eq. (1.12).

[25] Using this notation, the product of $\mathcal{I}$ by a matrix A, $\mathcal{I}A$ that, evidently, equals A, yields the strange looking expression: $\sum_{j=1}^{3}\delta_{ij}A_{jk} = A_{ik}$.

1.8 A Suboptimal Solution

A strategy that can be used at times when solving inverse problem P_3 involves solving the direct problem P_1 several times. One guesses or estimates values of the parameters (comprising the entries of matrix A in the example), a successive number of times, and solve P_1 for those values, selecting the value of A that best fits the data, in a previously established sense. In other words, the estimation (which corresponds to the minimization of E) can be performed by a *net search*, or *sampling*, that we describe briefly here.

Several values of the parameters, i.e., several matrices, will be chosen in succession which we shall denote by $A^1, A^2, \ldots, A^m$. They define a *grid* or *net* in $M(3,3)$. With them, the direct problem P_1 is solved successive times. That is, $A^j\,\mathbf{x}^k$ is computed, for $k=1,\ldots,n,\ j=1,\ldots,m$. We use that information to compute

$$E(A^j),\ j \quad = \quad 1,\ldots,m\,,$$

with E defined by Eq. (1.15).

The strategy to choose the next matrix in the sequence, A^{m+1}, or decide to stop the search, and to keep A^m, or any other of the previous matrices, $A^1, A^2, \ldots, A^{m-1}$, as solving the identification problem, is the defining step in several modern, non-gradient based, optimization methods. Nowadays, several methods employ different strategies in choosing the grid, as for instance in simulated annealing, [46, 50, 71], and others, so-called, metaheuristics, [69]. This is a *suboptimal* strategy, due to the fact that, since

$$\{A^1,\ldots,A^m\} \quad \subset \quad M\,,$$

we have

$$\min\nolimits_{B\in\{A^1,\ldots,A^m\}} E(B) \geq \min_{B\in M} E(B)\,. \tag{1.19}$$

Sure enough some limit theorem, when $m \to +\infty$, can be pursued. This is too far from our goal in this book.

1.9 Application to Darcy's Law

Here, we present an application of the model we have been discussing in this chapter. We trade the general black box model by the study of the flow of a fluid in a saturated *porous medium*. In such a medium, the flux, $\mathbf{u}$, and the gradient of the pressure, ∇p, satisfy *Darcy's law* ([28], [56]),

$$\mathbf{u} = -\frac{1}{\mu}K\nabla p\,, \quad \text{for } \mathbf{x} \in \Omega\,,$$

where K is the *permeability tensor* of the medium, $\mu > 0$ is the viscosity of the fluid, and $\Omega \subset \mathbb{R}^3$ is the porous region. Also, $p : \Omega \to \mathbb{R}$ and $\mathbf{u} : \Omega \to \mathbb{R}^3$. In general, the permeability $K = K(\mathbf{x})$ is a matrix-valued function,

$$K : \Omega \to M(3,3)\,,$$

where, we recall, $M(3{,}3)$ represents the set of 3×3 real matrices. In simple words, the permeability measures the *easiness* for the fluid to go through the porous medium. If K is a constant function, the porous medium is called *homogeneous*, otherwise it is a *heterogeneous porous medium*. If the medium is heterogeneous, the easiness of flow varies along the medium. If K is a scalar multiple of the identity matrix, $K = k\mathcal{I}$, then the medium is said to be *isotropic*. Otherwise, it is an *anisotropic* porous medium (in which case, the easiness of flow differs depending on the direction of the pressure gradient). Just to practice the terminology, we can have an isotropic medium, with k changing in space, in which case it is also heterogeneous.

Assume that we are investigating the permeability of a homogeneous anisotropic porous medium, and that we have a table containing several measurements of (vector) fluid flows, $\mathbf{u}^k$ for given applied pressure gradient, $\xi^k = -\nabla p$,

$$\xi^k \in \mathbb{R}^3 \,, \mathbf{u}^k \in \mathbb{R}^3 \,, \qquad k = 1, \ldots, n \ .$$

Assume also that the fluid viscosity, μ, is known. Then, following Eq. (1.15), we define half the quadratic error function,

$$\begin{aligned} E(K) &= \frac{1}{2}\sum_{k=1}^{n} |\frac{1}{\mu}K\xi^k + \mathbf{u}^k|^2 \\ &= \frac{1}{2}\sum_{k=1}^{n}\sum_{i=1}^{3}\left[\left(\sum_{j=1}^{3}\frac{1}{\mu}K_{ij}\xi_j^k\right) + u_i^k\right]^2 \ . \end{aligned} \tag{1.20}$$

This function is to be minimized to obtain the permeability tensor of the porous medium, i.e., the permeability tensor K_* satisfies

$$K_* = \text{argmin}_{K\in M(3,3)} E(K) \,.$$

Exercises

1.1. Show that: (a) Eq. (1.1) is equivalent to Eq. (1.2); (b) Eq. (1.3) is equivalent to Eq. (1.4).

1.2. Verify the validity of Eq. (1.12).

1.3. Assume that the mathematical model relating stimuli $\mathbf{x} \in \mathbb{R}^3$ and reactions $\mathbf{y} \in \mathbb{R}^3$ is given by $B\mathbf{y} = \mathbf{x}$, where B is a 3×3 matrix. Construct a table similar to Table 1.2 in this case. Read critically the paragraph where the footnote 21 is called, page 17.

1.4. (a) Verify the assertion in the sentence following Eq. (1.15). (b) Give conditions on the data, Eq. (1.13), such that there is only one solution to $E(A) = 0$.

1.5. (a) Compute $\sum_{j=1}^{3} \delta_{ij}A_{jk}$.

(b) Check the details on the derivation of Eq. (1.16).

(c) From data, Eq. (1.13), define matrices $X = (\mathbf{x}^1, \ldots, \mathbf{x}^n)$ and $Y = (\mathbf{y}^1, \ldots, \mathbf{y}^n)$. Show that $C = XX^T$ and $D = YX^T$, where C and D are defined in Eq. (1.17).

(d) If X is invertible (in particular, X must be a 3×3 matrix), show that the expression for B, Eq. (1.18), reduces to Eq. (1.12).

1.6. Determine the critical point equation for function $E = E(K)$ given by Eq. (1.20).

1.7. Simple regression model. Consider the *simple regression model*

$$y = a + bx\,, \tag{1.21}$$

where a and b are called *regression coefficients*, a is the *intercept* and b is the *slope*, x is called *regressor*, predictor or independent variable, and y is the *response* or dependent variable. Assume that you have pairs of measurements $(x_i, y_i) \in \mathbb{R}^2$, $i = 1, \ldots, n$, of regressor and response variables. The *residual* equation is

$$r_i = y_i - (a + bx_i)\,,$$

and half the *quadratic error* function is

$$E(a,b) = \frac{1}{2}\sum_{i=1}^{n} r_i^2 = \frac{1}{2}\sum_{i=1}^{n} [y_i - (a + bx_i)]^2\ .$$

(a) Obtain the *critical point* equation

$$\nabla E = \left(\frac{\partial E}{\partial a}, \frac{\partial E}{\partial b}\right) = (0, 0)\,. \tag{1.22}$$

(b) Denote the solution of Eq. (1.22) by $(\hat{a}, \hat{b})$, and show that

$$\hat{a} = \frac{\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}\,, \tag{1.23a}$$

$$\hat{b} = \frac{n\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}\,. \tag{1.23b}$$

Hint. Write the equations in matrix form and make use of the expression for the inverse of a 2×2 matrix,

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \frac{1}{AD - BC}\begin{pmatrix} D & -B \\ -C & A \end{pmatrix}.$$

(c) It is worthwhile to verify the dimensional correctness of Eq. (1.23). Check this.

Hint. Let the units of a variable x be denoted by $[x]$. Typical values of $[x]$ would be L for length, M for mass, and T for time. Then, if $[x] = X$ and $[y] = Y$, one can check from Eq. (1.23) that $[\hat{a}] = Y$ and $[\hat{b}] = Y/X$. This is compatible with Eq. (1.21) since b is multiplied by x to, partly, produce y, $[b][x] = [y]$, and then, $[b] = [y]/[x] = Y/X$.

(d) In the setting of statistics, the following notations are usual,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,,$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\,, \text{ and}$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\,.$$

Show that

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2\,, \text{ and } S_{xy} = \sum_{i=1}^{n} y_i(x_i - \bar{x})\,.$$

(e) Show that

$$\hat{a} = \bar{y} - \hat{b}\bar{x}\,,$$

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,(\bar{x})^2} = \frac{S_{xy}}{S_{xx}}\,.$$

(f) Let $\mathcal{H}(E)$ denote the *Hessian* of E, that is, the matrix of the second order derivatives of E,

$$\mathcal{H}(E) = \begin{pmatrix} \frac{\partial^2 E}{\partial a^2} & \frac{\partial^2 E}{\partial a \partial b} \\ \frac{\partial^2 E}{\partial b \partial a} & \frac{\partial^2 E}{\partial b^2} \end{pmatrix}\,.$$

Compute $\mathcal{H}(E)\mid_{(\hat{a},\hat{b})}$.

(g) Determine an expression for the *eigenvalues*[26] of $\mathcal{H}(E)$.

(h) Show that the eigenvalues of $\mathcal{H}(E)$ are positive, whenever $n \geq 2$.
Hint. Show that $\alpha \pm \sqrt{\alpha^2 - \beta} > 0$ whenever $\alpha > 0$ and $\alpha^2 \geq \beta$.

(i) Use the spectral theorem to show that $(\hat{a}, \hat{b})$ is a minimum point of E since $\mathcal{H}(E)$ is a positive-definite matrix[27].

[26] For a reminder on how to compute eigenvalues in simple cases, see Example A.1, page 192.

[27] Further information about regression models can be seen in [55]. In particular, the modeling of residues as a probabilistic distribution is discussed. Usually, the normal distribution is used, which amounts to introducing another parameter, σ, in the model present in Eq. (1.21), corresponding to the standard deviation of the normal, leading to the model $y = a + bx + \epsilon$, where $\epsilon \sim N(0,\sigma^2)$ stands for the normal distribution with zero mean and variance σ^2.

1.8. Consider Darcy's law for a homogeneous, isotropic porous medium, in one dimension. Use previous exercise with $a = 0$ in Eq. (1.21), to model the relationship between flux and pressure gradient. Determine an appropriate expression for an estimator of the scalar permeability, by mimicking the steps proposed in Exercise 1.7.