ParX

The Identification of Analytical Device Models

Dr. ir. M.G. Middelhoek

Contents

Chapter 1

Introduction

Mathematical modelling plays a crucial role in all branches of science and engineering, and electronics is no exception. The main aim of modelling in the field of electronics is to accurately predict the behaviour of electronic circuits, allowing a circuit to be evaluated without requiring its implementation. This is an essential asset during the design phase of a circuit. The huge increase in the complexity and quality of electronic systems over the past several decades would therefore not have been possible without an equal advance in modelling techniques. As improvements in semiconductor technology are the driving force behind this development, considerable effort has been directed towards improving the modelling of integrated circuits, in particular of the electronic devices that form the elementary building blocks of these circuits: bipolar transistors and field-effect transistors. Because all circuit models are ultimately composed of device models, successful circuit modelling requires, first and foremost, adequate models for these devices.

Many different approaches to device modelling have been reported in literature; the mathematical representation of device models ranges from stored tables of observational data to partial differential equations [1]. However, it will be shown in Chapter 2 that not every device description can be conveniently used in all stages of the circuit-design process. Particularly for the design of high-quality analog electronic circuits, accurate analytical models for the available electronic devices have proved to be indispensable. The formulation of an effective design theory for these circuits requires a combination of high accuracy and relative simplicity of expression that only analytical device models are capable of offering. Furthermore, only the high abstraction level of the analytical model allows the full exploitation of the behavioural characteristics of devices. This explains the continuing scientific effort that is directed towards creating ever more sophisticated analytical models for all common device types.

Identification is a necessary step in the modelling of virtually any device. Model

identification implies that the values of the model parameters are extracted from the observed device behaviour. In theory, the parameter values of an analytical device model can be obtained by analysing the internal physical structure of the device [1]. In practice, however, such detailed information about the device is seldom available, so that it is necessary to tap the only source of information that is always accessible, namely the observable device behaviour. Nevertheless, the extracted parameter values must be physically meaningful.

The sequential method is the classical approach to the identification of analytical models [2–5]. A typical sequential identification method divides the model up into a sequence of sub-models by a process of approximation, so that each sub-model contains only a small number of unknown parameters, preferably just one. Each sub-model in the sequence is then identified using only a small portion of the observational data, which is taken from the limited part of the device's operating range where the simplified model is supposed to be accurate. The values of the thus extracted parameters can be used in the definition and identification of subsequent sub-models. The sub-models can usually be chosen in such a manner that when the observations are plotted on a suitable (often non-linear) scale, their parameters can be read directly from the graph, or can be found by using simple graphical techniques. In this way, the sequential method replaces a single large and hence complicated identification problem by a sequence of smaller identification problems for approximate models that can be solved by hand. Therefore, the main advantage of the sequential method is that it does not require a large computational effort.

Nevertheless, the sequential method also has a number of disadvantages. First of all, the method can only be applied to models that lend themselves to this type of decomposition. Many of the complex and extended transistor models in use today cannot easily be handled by the sequential method. Secondly, a sequential identification procedure is tailored to a particular model, and a considerable amount of work is required when models are changed or improved. Further, the approximations that are required to bring the model in a tractable form also result in approximate parameter values, which may not be accurate enough for circuit-design purposes.

Since the advent of the computer, more computation-intensive identification methods have risen to prominence in the field of device modelling. The basic approach of these methods is usually the same: a fitting criterion that expresses the distance between the theoretical model curve and the actual observations is minimized by simultaneously adjusting the values of all unknown model parameters [6–11]. As a group these methods are therefore referred to as data-fitting methods[1]. They result in an optimum set of model parameters. The data-fitting approach offers significant practical advantages over the sequential

[1] They are also known as curve fitting or non-linear regression.

method. For one, it is far more flexible because the identification procedure—more specifically, the fitting criterion and the minimization method—is now completely model independent. Hence, data fitting is ideally suited to fully automated identification, a highly desirable feature in this age of CAD/CAM. However, data-fitting methods have not succeeded in replacing the sequential methods in all areas of application. Although generally accepted as a useful tool, as is demonstrated by the large number of commercial software packages based on data fitting [12], the data-fitting approach has the reputation of being less reliable than its predecessor. The complaint most often heard is that the optimum parameter values are not physically meaningful.

In Chapter 3 of this thesis, a detailed analysis will show that the reliability of an identification method is intimately linked to the validity of the model. Although analytical models have a physical foundation, they can at best approximate the physical reality. As a result, the extent of the validity domain of an analytical model will always be limited. Therefore,

> model identification must imply the extraction of the model parameters *and* the model validity domain from the observed device behaviour.

Conventional data-fitting techniques, such as the least-squares method [13], take the validity of the model over the whole operating range of the device for granted. Consequently, their optimum parameters are meaningless whenever this assumption is not justified. Nevertheless, virtually all model-identification programs that are available to date use some form of the least-squares method.

The sequential method is potentially more robust than the least-squares method because it provides a way to identify the validity domain of a model. Since the sub-models only depend on a few parameters, they can generally be transformed into a linear model by applying a suitable non-linear transformation. When plotted on the associated non-linear scale, the observational data will then show a linear relationship in the region where the model is valid. A linear relationship is preferred because it can easily be recognized in the graphical representation of the data. Hence, it is possible to determine the validity domain of the sub-models by inspection. This information can then be used to select the observations that subsequently determine the parameter values.

The main objective of the research presented in this thesis has been to develop a method for the identification of analytical device models that combines the flexibility of a data-fitting method with the reliability of a sequential method. We are of the opinion that data fitting is indeed the answer to the problem of model identification, as only the data-fitting approach provides a consistent mathematical framework for analysing the relation between the observational

data and the model parameters. While there is presently a lot of activity in this field, it is mainly concerned with the fine tuning of minimization algorithms in order to increase the efficiency of the standard data-fitting methods. We will find, however, that it is necessary to critically re-examine the motives behind the choice of the fitting criterion to significantly improve the reliability of the data-fitting approach. This research enables us to propose a new data-fitting method for the identification of analytical device models—called *mode selection*, or MODES for short—that takes into account the limited validity domain of these models. Whereas other data-fitting methods maximize the accuracy of the model over the complete set of observations, MODES aims to satisfy a *pre-defined* accuracy requirement over the validity domain of the model. This accuracy requirement (or validity criterion) is determined by the application of the model.

The broad subject of algorithm design is covered in Chapter 4. The implementation of MODES described in this chapter concurs to a large extent with the one that can be found in the general-purpose parameter extraction program called ParX (PARameter eXtractor)—which is the ultimate practical result of the research presented in this thesis. However, as this thesis is not intended as a user's guide to this specific implementation, we prefer not to clutter the discussion with implementation-specific details. Instead we will focus on the underlying principles in order to provide the necessary insight into the algorithm. It was decided at the outset that the reliability of the identification method should be the dominant criterion for assessing the implementation. We feel that in traditional implementations of data-fitting methods the trade-off between reliability and efficiency has been too much in favour of the latter. We aim to redress this imbalance by reappraising some standard minimization algorithms. The effectiveness of our implementation of MODES will be demonstrated and discussed in Chapters 5 and 6.

Finally, although the discussion in this thesis is dedicated to the identification of electronic-device models, it is not in any way limited to it. The principles involved can easily be translated to other branches of science and engineering. Empirical science as a whole is mostly concerned with the search for analytical models of reality (in physics these are sometimes called laws of nature). Hence, the identification method developed in this thesis has a much wider field of application.

Chapter 2

Device Modelling

The principal function of any device model is to encode our knowledge of a device and to represent it in a useful form. Besides the level of abstraction of the mathematical representation, it will therefore be the source and the extent of this knowledge that determine the potential use of a model. This insight allows for a classification of the bewildering variety of model representations that are in practical use in the field of electronics into two main groups: the behavioural models and the structural models. These two groups differ mainly in the way in which they make use of the two available sources of knowledge:

1. the *a priori* knowledge, which for electronic devices consists of a description of their internal physical structure, and
2. the *a posteriori* knowledge in the form of the observed behaviour of a particular device.

The implications of the source of the knowledge for the representation and interpretation of the device models will be discussed first.

Next, the different model representations are examined with respect to our modelling goal: the design of analog electronic circuits. Although any single representation is necessarily a compromise, the result of this examination will be unequivocal: only analytical device models are able to satisfy all constraints posed by this application. However, for the interpretation of the results of this modelling methodology to be valid, new methods for the identification and validation of analytical device models will have to be developed.

2.1 Mathematical modelling

The relationship between a "real world" device and its model is established by an abstraction process. In this process, only those aspects of the device that are essential to the modelling goal are retained and represented by the model. This representation will be in a mathematical form, since mathematics is the natural language for the expression of abstraction. Mathematical modelling can thus be defined as reducing the relevant physical properties of a device to a mathematical formulation that can be used in a convenient way.

The definition of mathematical modelling does not suggest by itself a modelling methodology. We will therefore introduce the modelling methodology that is generally accepted as the foundation of empirical science [14], and which is fully compatible with the modelling goal pursued in this thesis. This methodology interprets the construction of a model as the formulation of a hypothesis. This hypothesis is then validated (or corroborated) by experiments, which in the case of a device implies observing the device behaviour. The fact that validation is only concerned with the observable behaviour of a device means that only those physical properties of the device that directly affect its external behaviour are accepted as relevant. This observation yields the following modelling criterion:

> *A device model is considered to be a valid representation of a device when the observed behaviour of the device and the behaviour predicted by the model are identical.*

However, in practice the term "identical" in the criterion will have to be qualified by specifying the required accuracy of the model. The required model quality varies as it depends on the modelling goal. Although compliance with this criterion is a minimum requirement for any model, it may not be a sufficient requirement for all models. However, it should be clear that any other criterion would have to rely on *a priori* knowledge about the device.

The fact that validation is an essential part of the modelling methodology determines the type of mathematical formalism that can be used for the construction of a model. In this process of linking the mathematical formalism to the device behaviour one can distinguish several related but distinct phases.

The initial phase in the construction of a model consists of the introduction of an interface through which the device interacts with its environment. The interface is defined as a set of variables that are chosen as representatives of physical device quantities that take on observable or even controllable values. The physical interpretation of these interface variables determines the behavioural domain of the model.

To describe the behaviour of a device, it is customary to differentiate between independent and dependent interface variables, as seen from the device side of the interface. However, this separation between independent and dependent variables is not always very strict and is often just a matter of convention or convenience. When the independent interface variables of a device are given a value by the environment, its dependent interface variables will take on definite values. This means that a device, as a result of its internal physical structure, imposes constraints on the values of the interface variables, or in mathematical terms, defines a functional relationship between the interface variables. Finding a suitable mathematical representation for this functional relationship is the fundamental phase in the construction of a model. It is often expedient to consider these mathematical representations as being composed of a structure and a set of parameters. The model structure describes the form of the representation, usually by specifying some class of mathematical equations, while the model parameters are then the unspecified coefficients in these equations. These model parameters are also called "structural" parameters to stress the fact that they are associated with a specific model structure.

Finally, in practice it is not possible to totally encompass such a complex physical phenomenon as a device by a mathematical description. Any device model will only be acceptable within a limited domain. This domain of validity of the model should always be determined.

The construction of a device model thus consists of four phases:

1. defining the model interface, i.e. the behavioural domain,
2. determining a suitable model structure,
3. determining the values of the model parameters, and
4. determining the domain of validity of the model.

The realization of the various stages depends on the modelling goal, which not only determines the relevant behavioural domains but also the interpretation of the resulting models.

2.2 Model representations

A model should at least fulfil the requirement of reproducing the actual behaviour of a device with sufficient accuracy over the relevant operating region. As this behaviour can be observed, it is possible to devise model representations that only use this source of knowledge about a device. These behavioural models describe a device as if it were a "black box" without any internal structure.

However, for most practical devices the black-box approach is not realistic as there exists a considerable body of knowledge concerning the mechanism of their internal workings. This knowledge, which is based on device physics, can be incorporated in the model, linking the physical structure of the device and the mathematical structure of the model representation. Both the behavioural and the structural models can be represented at several levels of abstraction.

2.2.1 Behavioural models

Observing the behaviour of a device implies the determination of the values of the dependent interface variables for discrete values of the independent interface variables. The discrete nature of the observed data has, through the number and distribution of the observations, a significant influence on the representation and specification of the models that are solely derived from this source of information.

Data-interpolation models

This model representation makes direct use of the observations by storing the discrete values of the independent variables and the accompanying values of the dependent variables in a multi-dimensional table [15, 16]. This table is then used to estimate the device behaviour for intermediate points by incorporating some appropriate interpolation scheme. Thus, for accurate modelling a large table and an equally large set of observations are needed.

The structure of the model representation is implicitly determined by the choice of the interpolation scheme. This structure could be described in the form of an interpolating function that passes through all the observations. The coefficients of this interpolation function are directly and uniquely determined by the observations. The observations can thus be interpreted as the parameters of the representation.

Data-fitting models

The data-interpolation models will exactly reproduce all the observations. However, the large number of parameters required for the representation of these models suggests the possibility of reducing the number of parameters by abandoning this constraint. This abstraction yields the data-fitting method which condenses the observed data by fitting it to a model that depends on a relatively small set of parameters.

The structure of these models is usually expressed in the form of an analytical function. However, without any *a priori* knowledge about the device it is advisable to choose a class of functions that is universal enough to approximate the behaviour of any device type, e.g. polynomial, spline or piecewise-linear functions [17, 18].

The values of the model parameters are determined by minimizing the difference between the observed behaviour and the behaviour predicted by the model. In general, this minimum difference will not be zero. Therefore, a measure of the above difference is needed, the so-called fitting criterion. This criterion determines the final distribution of the limited accuracy of the model over the domain space (i.e. the range of the independent variables). The resulting "best-fit" parameter values are now no longer uniquely determined by the observations, but also depend on the choice of the fitting criterion. As a consequence, the fitting criterion should be regarded as an inherent part of the model hypothesis.

2.2.2 Structural models

The incorporation of *a priori* knowledge, which for an electronic device comes from the domain of solid-state physics, means a departure from the behavioural black-box approach, as the internal structure of the device is now taken into consideration. The level of detail with which the internal structure of the device has to be described depends on the abstraction level of the model.

Physical models

At this level of abstraction a device is described by specifying how it is constructed. Therefore, a physical model relies completely on a *priori* knowledge about the device. The behaviour of the device can be deduced from this knowledge [19, 20].

The structure of the model consists of a set of partial differential equations (i.e. the Poisson and current-continuity equations) that model the basic physical processes of solid-state devices. The form of these basic equations is not device specific and can be used for a variety of device types. The model is made device specific by specifying the doping profile of the semiconductor device, together with appropriate boundary conditions associated with the device geometry and the external device contacts. These design parameters represent the physical quantities that form a description of the internal structure of the device. The basic equations can then be solved, usually with the aid of a suitable numerical method, not only yielding the behaviour of the device, but also providing insight into the physical processes taking place within the device.

The validity of a physical model depends on the validity of the *a priori* knowledge. This knowledge is stratified in accuracy as well as in generality, as even the most well-established laws of solid-state physics are but models with a domain of validity that is limited.

Analytical models

The physical models require a full and detailed description of the internal structure of a device. However, since only a small number of specific structures are actually used for the construction of practical devices, the internal structure of a device will be largely determined by the device type. This insight allows for a more abstract description whereby the internal structure of a device of a given type can be completely specified by a limited number of primary parameters. Under these restricting conditions it may be possible to solve the basic equations of the physical model and obtain an analytical expression for the device behaviour [2, 21]. This approach has resulted in various analytical models for most device types and behavioural domains, some more accurate than others depending on the simplifying assumptions that were made in their derivation.

The structure of an analytical model consists of a set of functional relationships between the interface variables. The mathematical form of these functional relationships is entirely determined by the device type, i.e. *a priori* knowledge. The functional relationships will depend on a number of structural parameters, which are derived from the primary device parameters. Therefore, the parameters of an analytical model represent those behavioural qualities of a device that are device specific. Since the primary parameters represent physical quantities, it is often admissible to assign a physical significance to the structural parameters as well. However, the validity of any interpretation of the parameters or the structure of the model depends on the validity of the premises of their derivation.

2.3 The modelling goal

In this section, the adequacy of the different model representations will be assessed with respect to the modelling goal. For this purpose, the modelling goal must be elaborated by taking a closer look at the circuit design process and the role of the device models in this process.

Circuit design can be defined as the process of achieving a circuit implementation that satisfies a given design specification [1]. This specification includes the signal-processing function that the circuit must perform and the quality aspects that have to be considered. An acceptable, preferably optimum, circuit implementation must be selected from all possible circuit implementations

that together form the design space. The extent of the design space is usually formidable, even when restricted by a large number of technological and physical constraints. Therefore, a design strategy is required to search the design space effectively. Such design strategies have been developed for many of the basic analog functions: amplifiers, oscillators, filters, etc [22–25]. The level of sophistication of these strategies varies from simple heuristic procedures to systematic algorithms. Although these strategies are thus different in nature, the image of a design strategy as a search strategy in the design space is instrumental in describing a set of design steps that all design strategies have in common.

2.3.1 Circuit design steps

A design strategy traverses and partitions the design space by a sequence of design steps. The objective of this sequence of design steps is a fully specified circuit—the topology as well as the devices—that satisfies the design specifications in the best possible way within the constraints of the design space. In other words, if all possible circuit implementations within the design space are assigned a quality measure that expresses their compliance with the design specifications, the goal of the design strategy is to find the point (circuit implementation) in the design space that has the best quality, i.e. the global optimum. When circuit design is interpreted as a (global) optimization process, the design steps can be seen to fall in three distinct categories: circuit analysis, circuit optimization and circuit synthesis. This subdivision is based on topological criteria: whether the design step involves a single point in the design space, a local region of the design space, or the design space as a whole.

Circuit analysis concerns the determination of the performance of a fully specified circuit, i.e. a single point in the design space, which implies the determination of the behaviour of the circuit in response to specified stimuli. The quality of the circuit can then be evaluated with respect to the design specifications

Circuit optimization starts from an initial circuit which is analysed to determine the deviation from the desired performance. This deviation is then reduced by an optimization procedure which involves an iterative cycle consisting of successive stages of circuit analysis, evaluation of the results, and circuit modifications based on the observed deviations from the desired performance. This cycle is repeated until the circuit satisfies the design specifications within given tolerances, or until no further improvement is obtained. The optimization process may intervene in the circuit design by modifying the topology of the circuit in a systematic way or by selecting a different device. These degrees of freedom in the circuit design are represented by a number of design variables which may be of a continuous or discrete nature. Since the initial circuit is only modified by changing a necessarily limited number of design variables,

optimization only traverses a small portion of the design space. Therefore, the resulting optimum circuit implementation will, in general, be only a local optimum.

Circuit synthesis refers to a systematic procedure which partitions the design space, with the objective of closing in on that section of the design space that contains the global optimum. If the final subsection of the design space only contains a single circuit implementation, this circuit is accepted as being optimum by construction, and its quality can be assessed by means of circuit simulation. If the final subsection is less restrictive, circuit optimization can be used to search the remaining design space. However, for this optimization to be successful, the final subsection should not contain any local optima besides the global optimum.

There is a marked difference between the abstraction levels of the two constructive design steps: circuit optimization and circuit synthesis. Circuit optimization always operates on a fully specified circuit, thus only using information that is specific to a single point in the design space or its local environment. This information is then used to induce the location of the optimum circuit. However, circuit synthesis is based on design theory which uses information that is necessarily global in nature as it is used to partition the whole design space without considering each circuit implementation in this space individually. This means that synthesis must operate on a higher abstraction level by its ability of excluding whole sets of circuit implementations on grounds that are set specific. Circuit synthesis thus deduces the location of the optimum circuit from a general design theory.

2.3.2 Device design

An important aspect of circuit design, which affects the choice of the model representation, is the level of intervention in the internal structure of the device that is possible, i.e. the degrees of freedom the designer has when choosing a device. In the case of discrete devices, the only degree of freedom is the choice of a particular device from a list of available devices. However, when designing an integrated circuit it may be possible to actually design the devices by specifying their design parameters [26, 27]. Yet in practice, device design is always limited by technological constraints and the level of intervention in the internal structure of a device usually goes no further than the specification of the device type and its geometry. But even this limited amount of freedom in device design results in a virtually unlimited number of potential devices, each of which must be modelled separately. This problem can be avoided by introducing a classification system, which reduces the number of distinct devices and accompanying models to manageable proportions. These device classes can be interpreted as differentiated device types. A particular device implementation

can then be selected by choosing its class and assigning a value to a limited number of design parameters that express the remaining degrees of design freedom (e.g. the scaling factors of a specified device geometry or other primary device parameters). The concept of the device model can then be extended to apply to a class of devices by incorporating the design parameters in the model representation.

2.3.3 Models for circuit design

Since electronic devices are the circuit primitives, device models play an important role in the circuit design process. To assess the adequacy of the different model representations that were discussed in Section 2.2 for this particular application, the requirements of each design step with respect to the validity, accuracy, generality and physicality of the model will be investigated.

Models for circuit analysis

The first requirement for performing circuit analysis is the definition of adequate models for the circuit primitives. The design specification determines the relevant behavioural domains and thus the interface through which a device interacts with its environment. As circuit analysis always deals with circuits that are completely specified, the environment of each device is known. Therefore, the only requirement that a model should fulfil is to reproduce the actual behaviour of the device with sufficient accuracy over the relevant operating region. Since this is a minimum requirement for any model, all model representations are, in principle, usable. Hence, the representation should be selected that is the most suitable for the specific analysis method.

The behavioural models have the advantage that their model structure can be adapted to the requirements of the analysis method. However, when the design specifications include multiple behavioural domains, a separate model is needed for each behavioural domain. These behavioural models are unrelated, although they all refer to the same physical device. The hidden physical relations between the different behavioural domains are thereby easily violated, which may jeopardize the accuracy of the final result of the circuit analysis in an unexpected way.

The representation of the physical model is in a form that is only amenable to the analysis of a single device. Its application is better reserved for those occasions where the detailed information it supplies about the internal operation of the device is essential.

The analytical model, however, is suitable for circuit analysis. Although its model structure is in general more complex than that of a behavioural model,

it has the advantage that the analytical models for the different behavioural domains are all derived from a single physical device structure. The physical relations between the different behavioural domains are thus included in the model structures.

Apart from their representations, the behavioural and structural models should also be compared with regard to their domains of validity. While the validity domain of a behavioural model can always be extended by simply enlarging the number of observations, the validity domain of a structural model is always limited by its physical nature. The domain of validity of a structural model can only be extented by refining the *a priori* knowledge about the internal structure of the device, or by refining the derivation of the model from this knowledge. For the analysis of arbitrary circuits, the behavioural models may therefore be the only choice. However, for analysis in the context of circuit design, the limited domain of validity of the structural models is usually not a problem.

Models for circuit optimization

Circuit optimization is based on the repeated analysis of the circuit. Hence, any model representation to be judged adequate for optimization should first of all satisfy the requirements of circuit analysis. If these requirements also suffice for circuit optimization depends on the degrees of freedom that are available in the design of the circuit. Only when these degrees of freedom include device design does optimization influence the choice of the device models. In this case the model representations must be extended by incorporating the design parameters.

For all behavioural model representations, this incorporation of the design parameters proceeds by treating them as if they were independent interface variables. The complexity and the extent of a behavioural model depends directly on the dimensionality of the model interface. Especially when multiple behavioural domains are involved, the number of interface variables may become so large that the complexity and the extent of the model exceeds practical limits, making the behavioural models unsuitable for circuit analysis. However, it should be noted that these limitations of the behavioural models are of a practical and not of a theoretical nature.

The structural models already contain the design parameters, the physical models even in explicit form. The analytical models comprise the design parameters in the form of their primary parameters. Hence, it suffices to introduce these design parameters in the model representation in the form of structural parameters. The complexity of a structural model is thus unaffected by the incorporation of the design parameters. However, there remains the problem of the validity domain of the structural models, the bounds of which also depend on the design parameters.

Models for circuit synthesis

Circuit synthesis depends critically on the formulation of a design theory that can be used to partition the design space. In general, a design theory is based on the analysis of several classes of circuits. Since only the general form of the circuits in a class is known, this analysis must be carried out at a necessarily high level of abstraction. For the representation of the individual devices, the implications of this higher level of abstraction are twofold, as neither the devices themselves are completely specified nor their circuit environments. The level of abstraction of the representation of these "generic" devices should match the analysis method. In other words, the structure of the model representation must be analytically tractable.

The data-fitting models and the analytical models are the only device descriptions that attain the required level of abstraction. Their representation is expressed in a closed analytic form that allows the standard mathematical operations and manipulations to be carried out in theoretical studies. In particular, it leads to the possibility of deriving explicit closed-form expressions for the performance of the circuit class. Different circuit classes can then be compared on the basis of these expressions with respect to the design specifications. At the device level, the expressions can be used to derive design criteria that express the influence of the behaviour of a generic device on the overall performance of the circuit. These design criteria, which still contain the unspecified model parameters and interface variables, can be used for selecting or designing the optimum devices and their environments. However, there are some subtle distinctions between the design criteria that are derived using a behavioural device model and those that are derived using a structural device model, which need elaboration.

A design criterion implements a specific measure for comparing devices with respect to their behavioural characteristics. For this comparison to be meaningful, two conditions must be satisfied:

1. A single model structure must be used for the formulation of the design criterion and the modelling of the devices.
2. The values of the model parameters must be uniquely determined by the device behaviour and vice versa, i.e. there must exist a one-to-one relationship between them.

The parameters of a data-fitting model are chosen to optimize some global accuracy criterion over a limited data set. Even when the fitted model reproduces these observations exactly, it still only reproduces the *observed* device behaviour. The parameters of a data-fitting model are purely empirical, and usually depend on the number, range and distribution of the observations. Hence,

it is often difficult to ascertain the reliability of a design criterion that is based on data-fitting models.

The structural parameters of an analytical model are derived from the primary parameters of the device. Since these primary parameters represent physical qualities, which are necessarily unique, the structural parameters must also be unique. However, an analytical model only reproduces the behaviour of the device within its validity domain, so the validity domain of the model also determines the validity domain of the design criterion. A violation of this domain suggests the need for selecting a different device, redesigning the circuit, or deriving a design criterion using a more complex analytical device model.

2.3.4 Selecting the model representation

Summarizing the preceding discussion, we conclude that for the design of analog electronic circuits the analytical model is the preferred device description. In practice, however, the ideal of an analytical model for every device is difficult to attain. The *a priori* knowledge about the device, from which an analytical model can be constructed, may not always be readily available. A lack of *a priori* knowledge can affect every stage of the modelling process.

A fully specified model structure is a prerequisite for the formulation of any design criterion. For most types of devices that are in general use (e.g. resistors, capacitors, diodes, bipolar transistors and field-effect transistors) enough *a priori* information is available to deduce the structure of an analytical model with a sizable domain of validity. However, if no reliable *a priori* knowledge about a class of devices can be obtained, a data-fitting model is the only viable alternative to an analytical model, as no other model representation is usable for circuit synthesis. A data-fitting model is also used when a structural model needs to be extended beyond its validity domain and no additional *a priori* knowledge is available. Traditionally, many device models belong to this category of regional-structural models. There is, however, an ongoing scientific effort to shift the boundary between the analytical region and the empirical region of these hybrid models in favour of the former. Nevertheless, the data-fitting model will never be completely excluded from the modelling exercise, and thus deserves some attention besides the analytical model.

The model structure is embedded in the design criterion. Consequently, the application of the design criterion requires the specification of proper values for the structural parameters and the validity domains of the given model for the individual devices.

The values of the structural parameters of an analytical device model are deduced from the primary device parameters, and thus depend on device-specific

a priori knowledge. The values of the primary device parameters that cannot be determined by direct observation are often unreliable or unknown. These device parameters can thus only be determined by indirect methods, i.e. by investigating their influence on the *a posteriori* knowledge: the observed device behaviour. In that way, the values of the unknown structural parameters are induced from the device behaviour.

The validity domain of an analytical device model depends on the validity of the *a priori* knowledge that was used in its construction, and is therefore device specific. An objective assessment of these basic assumptions may not be possible. However, checking the consequences of these assumptions, validates the *a priori* knowledge and consequently the model. This means that the validity domain of an analytical device model can also be induced from the device behaviour.

From this examination it can be inferred that when sufficient device-specific *a priori* knowledge cannot be obtained, as is usually the case, the utilization of the analytical models is made possible only by the fact that the observed behaviour of the device can be used instead. A method for extracting the parameters and the validity domain of an analytical model from the observed behaviour is therefore indispensable for reaching our modelling goal. The development of such an identification method will be the subject of the following chapters.

Chapter 3

Identification and Validation

In the previous chapter we encountered the problem of model identification, that is, the problem of determining the values of the model parameters on the basis of the observed device behaviour. To solve this problem we must study the connection between the device behaviour and the model specification in detail. This connection is established by the modelling criterion that was given in Section 2.1. The interpretation of the modelling criterion depends on the modelling goal and on an associated quantification of the concept of model validity, thereby introducing the question of model validation in the model-identification problem. The formulation of a validity criterion is therefore a prerequisite for the development of any identification method. It follows from the modelling criterion that this validity criterion should also be based on the observed device behaviour.

In the present chapter, the principles of the identification and the validation of device models are introduced. For this purpose, a topological approach to model validity and model identification will be presented. This topological approach will be used to study and compare the properties of the identification methods that are discussed in the subsequent sections.

All identification methods considered in this chapter are based on data-fitting techniques. The model parameters are determined by optimizing a fitting criterion that expresses the difference between the observed device behaviour and the device behaviour predicted by the model. The choice of the fitting criterion is based (often implicitly) on assumptions about the characteristics of the model and the characteristics of the observation errors. Hence, the acceptability of an identification method depends on how realistic these assumptions are when applied to practical identification problems. It will be demonstrated that conventional identification methods, which are usually based on the standard least-squares criterion, often fail to correctly identify analytical device models. We will therefore introduce a new identification method, called *mode selection*,

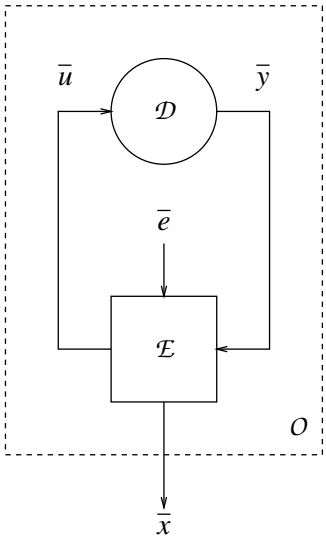

Figure 3.1: The observable system.

which uses an identification criterion that is defined in accordance with the modelling goal that was expressed in Chapter 2. This method not only determines reliable values for the model parameters, but also supplies an estimate of the model validity domain.

3.1 Basic concepts

An identification method is characterized by three elements: the observed device behaviour, the set of device models, and an identification criterion [28]. The identification problem is then to select a model in the model set that describes the observed behaviour best, in the sense of the chosen criterion. In this section, these basic elements will be described in detail by introducing the mathematical formalism and stating the assumptions regarding their properties.

3.1.1 The observed behaviour

The device behaviour is obtained from the device by performing a number of experiments. For this purpose, the device $\mathcal{D}$ is embedded in an experimentation environment $\mathcal{E}$ to form the observable system, denoted by $\mathcal{O}$, which is represented in Figure 3.1. To describe the experiments, it is customary to differentiate between the independent and dependent interface variables of the device.

The vector of independent variables $\bar{u}$ is then associated with the stimulus that is applied to the device, while the vector of dependent variables $\bar{y}$ is associated with the subsequent response of the device. The device $\mathcal{D}$ establishes a relationship between the n_u independent variables and the n_y dependent variables of the form

$$\bar{y} = h_{\mathcal{D}}(\bar{u})$$

where $h_{\mathcal{D}} \in \mathbb{R}^{n_u} \rightarrow \mathbb{R}^{n_y}$ is a deterministic function representing the behaviour of the device for that particular choice of the independent and dependent interface variables.

As the device cannot be separated from its environment, only the behaviour of the whole system $\mathcal{O}$ can actually be observed. The observations of $\mathcal{O}$ are denoted by an n_x-dimensional vector $\bar{x}$ composed of the ***observed*** values of the independent and dependent interface variables of the device:

$$\bar{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_{n_x} \end{bmatrix} = \begin{bmatrix} \bar{u} \\ \bar{y} \end{bmatrix} + \bar{e}$$

The additional stochastic vector $\bar{e}$ represents the errors that affect these observations.

The observed behaviour of the system $\mathcal{O}$ is obtained in the form of a set of N observations $\mathcal{X} = \{\bar{x}_1, \ldots, \bar{x}_N\}$. Each observation $\bar{x}_i \in \mathcal{X}$ is associated with a single experiment under the experimental conditions determined by a realization of the vector of independent variables $\bar{u}_i$. It is assumed that all the observations in $\mathcal{X}$ are independent of one another. In fact, the device interface can always be chosen in such a way that this assumption is valid.

3.1.2 The observational accuracy

The goal of device modelling is the description of the behaviour of the device $\mathcal{D}$. However, since these observations are derived from the observable system $\mathcal{O}$, each observation $\bar{x}_i$ of the device behaviour is affected by an observation error $\bar{e}_i$. Without any ***a priori*** knowledge about these observation errors, the observations are meaningless as estimates of the true values of the device interface variables. Therefore, each observation must be accompanied by a specification of its accuracy.

A full specification of the accuracy of an observation $\bar{x}_i$ would consist of the probability density function of the observation error $p(\bar{e}_i)$, or at least its first and second moments, $E(\bar{e}_i)$ and $E(\bar{e}_i \bar{e}_i^t)$. In practice, this kind of knowledge will hardly ever be available. We will therefore limit our requirements to an absolute minimum and will only presume that the accuracy of the observations

is specified in the form of the possible upper bounds of the magnitude of the observation errors, the so-called error intervals. Denoting an observation by $\bar{x}_i \pm \Delta\bar{x}_i$, the error interval $\Delta\bar{x}_i$ then determines a closed domain, centered at the observation, in which the true value is expected to lie. Since the actual realizations of the observation errors are not known, the observations $\bar{x}_i$ are accepted as the best available estimates of the true values of device interface variables.

3.1.3 The device model

In Section 2.1, a device model was introduced as a hypothesis about the observable device behaviour. As such, a device model proposes a functional relationship between the interface variables of the device. The mathematical representation of this functional relationship is composed of a structure and a set of parameters. The choice of a model structure defines the set of models, denoted by $\mathcal{M}$, from which the actual model is to be selected. The members of the set $\mathcal{M}$ are parameterized by an n_p-dimensional vector $\bar{p}$ of structural parameters. Hence, a specific member of $\mathcal{M}$ will be represented by $\mathcal{M}(\bar{p})$.

A given model $\mathcal{M}(\bar{p})$ can be interpreted as a hypothesis about the expected outcome of the experiments. A convenient mathematical formalism for expressing this expectation is as a set of n_f equality constraints on the values of the interface variables and the structural parameters

$$f_{\mathcal{M}}(\bar{p}, \bar{x}) = \bar{0} \tag{3.1}$$

The function $f_{\mathcal{M}} \in \mathbb{R}^{n_p} \times \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_f}$ specifies the structure of the class of models $\mathcal{M}$, which for analytical models as well as for data-fitting models is represented in the form of a set of analytical expressions (for more details see Appendix A).

Since a device model necessarily excludes many aspects of reality, the set of models $\mathcal{M}$ does not contain the "true" representation of the device $\mathcal{D}$, denoted by the function $h_{\mathcal{D}}$. The behaviour of $\mathcal{D}$ can only be approximated within the specified class of models. Hence, the model parameters, which have to be determined from the observed device behaviour, do not exist in any absolute sense. This qualitative vagueness of the model hypothesis will have to be compensated with a quantitative tolerance with respect to the parameters.

3.1.4 The identification criterion

According to the modelling criterion, the validity of a model hypothesis $\mathcal{M}(\bar{p})$ describing the device $\mathcal{D}$ is tested with the observed behaviour $\mathcal{X}$. In general, the

model equations (3.1) will not hold for the observations $\bar{x}_i \in \mathcal{X}$, because of the observation errors and the approximative nature of the model. This discrepancy between the model and the observations will be expressed by postulating an identification criterion, a scalar function $\mathcal{C}(\mathcal{M}(\bar{p}), \mathcal{X})$. Although the model $\mathcal{M}(\bar{p})$ and the system $\mathcal{O}$ are not in the same domain, and therefore cannot be compared directly, the criterion $\mathcal{C}$ will be regarded as a measure of the distance between them, on the basis of the available data $\mathcal{X}$. As $\mathcal{C}$ defines a distance measure, it must satisfy the condition $\mathcal{C} \geq 0$, where the equality may only hold when the model and the system are considered to be identical, i.e. when the model equations (3.1) hold for all $\bar{x}_i \in \mathcal{X}$.

For the purpose of identification, the criterion can be reformulated as a scalar function of the structural parameters. This scalar function $\mathcal{C}(\bar{p})$ is usually referred to as the objective function of the identification problem. The parameter vector that minimizes the objective function will be denoted by $\bar{p}^*$. The model $\mathcal{M}(\bar{p}^*)$ is then the best representation of the device in the model set $\mathcal{M}$ according to the given criterion $\mathcal{C}$ and the observed behaviour $\mathcal{X}$.

Since it is not possible to reduce $\mathcal{C}$ to 0 for the reasons that were stated, the form of the criterion determines the characteristics of the model that is finally adopted. The choice of the criterion $\mathcal{C}$ should therefore be guided by the *a priori* knowledge that is available about the observation errors and the accuracy of the device model.

3.2 Model validity and model accuracy

According to the modelling criterion, the validity of the model hypothesis can be ascertained by comparing the observed device behaviour with the device behaviour that is predicted by the model. The quantification of the concept of model validity therefore depends on the metric that is used for this comparison, thereby relating model validity to model accuracy.

A description of the accuracy of the model is also the key to gaining insight in the characteristics of the different identification methods. In this section we will introduce a topological interpretation of model identification, thereby establishing a relation between the identification criterion and the description of the model accuracy.

3.2.1 The model accuracy limit

The accuracy of a model in predicting the observed device behaviour is limited by the inaccuracy of the observations and by the approximative nature of the

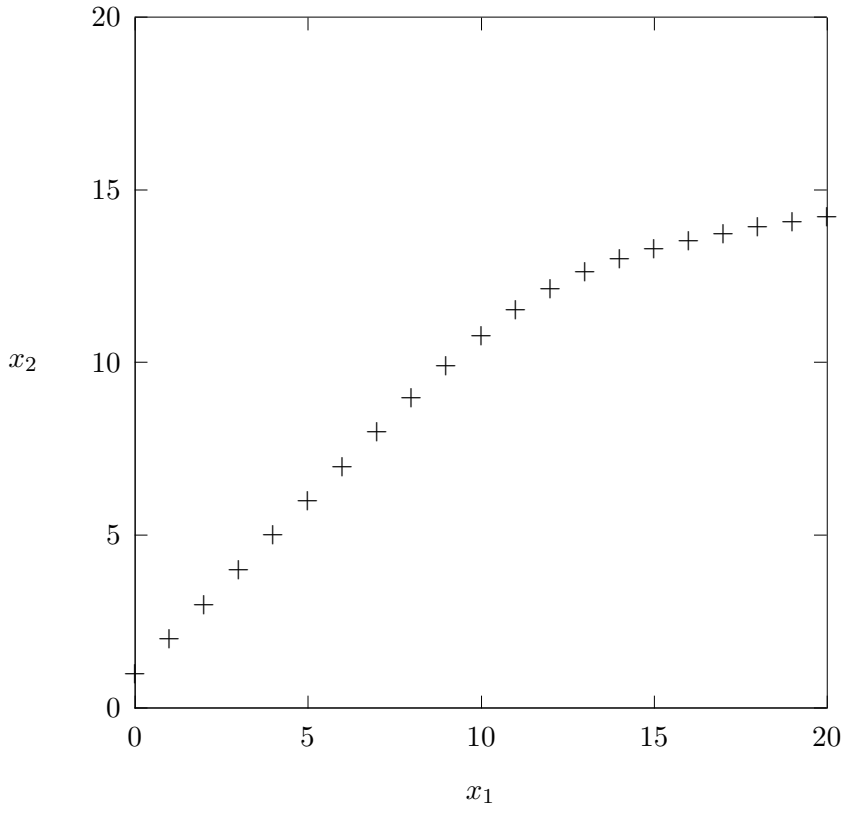

Figure 3.2: An example of an observation space ($n_x = 2$).

model itself. We will seek to determine this limit by examining the relation between the model and the observed behaviour. For this purpose, we introduce two vector spaces: the *observation space* and the *parameter space*. The observation space is the n_x-dimensional linear space in which the values of the interface variables are interpreted as the co-ordinates. In this space, the observations $\bar{x}_i$ are represented as points, the so-called data points. The parameter space is the n_p-dimensional linear space in which the values of the structural parameters are interpreted as the co-ordinates. Each model $\mathcal{M}(\bar{p})$ that is a member of the model set $\mathcal{M}$ is represented as a point in this space. The relation between the parameter space and the observation space is established by the model equations 3.1. This relation will be visualized with the help of an example.

The behaviour of a device with two interface variables, x_1 and x_2, is represented in the observation space shown in Figure 3.2, where the observations are marked $(+)$. A simple linear model $x_2 - ax_1 - b = 0$ is postulated for the description of these observations. The two parameters of this model, a and b, define the

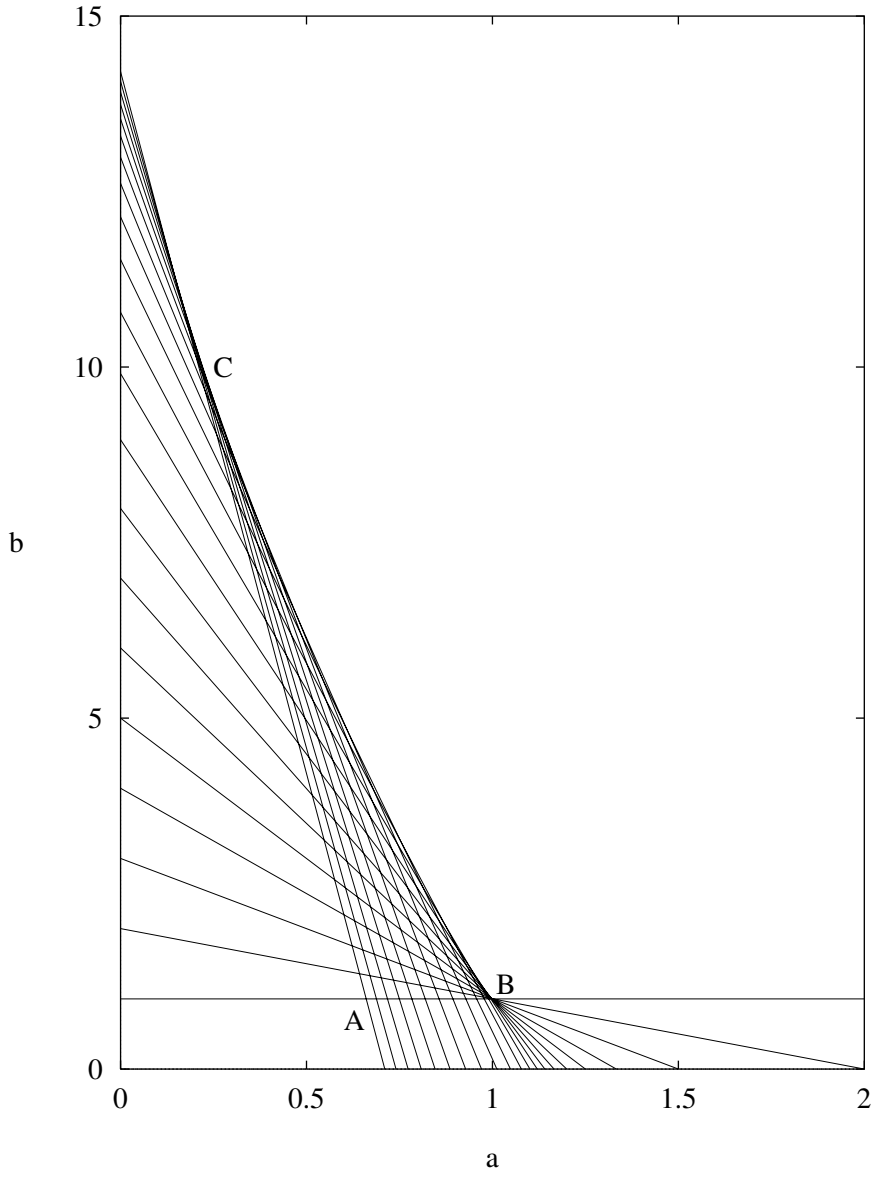

Figure 3.3: The parameter space ($n_p = 2$).

axes of the parameter space shown in Figure 3.3. For each choice of parameter values, the model is reduced to an equality constraint on the values of the interface variables. Such a constraint defines a subspace in the observation space; in this specific case the subspace can be represented by a straight line. Hence, the model associates a subspace in the observation space with each point in the parameter space. Vice versa, each point in the observation space defines an equality constraint on the values of the parameters, so that each observation is associated with a subspace in the parameter space. The subspaces associated with the observations given in Figure 3.2 are represented as straight lines in the parameter space in Figure 3.3.

From Figure 3.3 it will be clear that there is no single point (a, b) in the parameter space that satisfies all constraints. Consequently, the parameter set of the model is not uniquely determined by the observations. Instead, the observations circumscribe a compact (n_p-dimensional) subspace in the parameter space, in this example the area marked (A, B, C), the extent of which is completely determined by the observation errors and the modelling errors. We will name this subspace the ***identification space*** $\mathcal{P}_\mathcal{X}$, as all the models in this subspace can be derived from the observed behaviour. This tolerance in the parameter space corresponds to a set of model curves in the observation space, none of which passes exactly through *all* the data points.

3.2.2 The residuals

To quantify the accuracy of a model, we must first define a metric in the observation space. Since the limited accuracy of a device model will subsequently limit the precision of any design criterion or other conclusion that is based on it, the minimum model accuracy that is required for the proposed application should always be specified. The specification of the required model accuracy implies the definition of an accuracy metric. The choice of this metric, which inevitably reflects a certain weighting of features of the device behaviour, actually sets the goal of the modelling process.

The distance, with respect to the chosen metric, between a data point $\bar{x}_i$ and the subspace defined by the model $\mathcal{M}(\bar{p})$ is called the residual of the data point. This residual, which will be denoted by ϵ_i, can be regarded as a scalar function of the parameter vector $\bar{p}$; however, the term is also used for the value of ϵ_i that corresponds to a particular choice of the parameter vector $\bar{p} \in \mathcal{P}_\mathcal{X}$. Since a residual represents a distance measure, the metric axiom $\epsilon_i \geq 0$ will be satisfied. The set of residuals $\{\epsilon_1, \ldots, \epsilon_N\}$ is then a measure of the accuracy with which the model $\mathcal{M}(\bar{p})$ predicts the observed behaviour.

As the accuracy requirement imposes a definite upper bound on the distances between the model curve and each of the data points $\bar{x}_i$, the accuracy metric can

be defined in such a way that this bound is represented by the constraint $\epsilon_i \leq 1$. In this way, the required accuracy is used as a reference. The observation errors, which also contribute to the values of the residuals, determine the fundamental limit of the attainable model accuracy. Hence, the required model accuracy should not be chosen too restrictive, but allow for an observation error of $\Delta\bar{x}_i$. Under this condition, the constraint on the residuals can be interpreted as a local validity criterion: when $\epsilon \leq 1$ for some data point, the model hypothesis can be accepted as valid for that data point. However, when one of the residuals exceeds this bound, the model hypothesis is considered to be invalid for the data set $\mathcal{X}$ (or falsified [14]), and unsuitable for the proposed application.

3.2.3 The identification criterion

It is the goal of every identification method to maximize the accuracy with which the model represents the observed behaviour. As this accuracy is expressed by the set of residuals $\{\epsilon_1, \ldots, \epsilon_N\}$, this means that we should strive to minimize all the residuals ***simultaneously***. However, the model accuracy that can be attained is always limited. In Section 3.2.1 this limit was equated to the identification space $\mathcal{P}_\mathcal{X}$. That this subspace satisfies the requirement of minimizing all the residuals simultaneously is guaranteed by the following formal definition of $\mathcal{P}_\mathcal{X}$:

> A point $\bar{p}$ in the parameter space belongs to the identification space $\mathcal{P}_\mathcal{X}$ when there exists no point $\bar{p}' \neq \bar{p}$ for which the relation $\epsilon_i(\bar{p}') \leq \epsilon_i(\bar{p})$ holds for all observations $\bar{x}_i \in \mathcal{X}$.

A concise description of the identification space is therefore a pragmatic solution of the identification problem. For this purpose, the information that is represented by the set $\mathcal{P}_\mathcal{X}$ should be reduced to a small set of principal characteristics describing the location (or central tendency) and the extent of $\mathcal{P}_\mathcal{X}$ in the parameter space [13].

Now consider the following identification criterion (or objective function) over the parameter space:

$$\mathcal{C}(\bar{p}) = \sum_{i=1}^{N} l(\epsilon_i(\bar{p}))$$

where the scalar function l is a strictly increasing function of the residuals, and $l(0) = 0$. Note that each observation contributes its residual to $\mathcal{C}$ independent of all other observations, and that the weight of each contribution only depends on the size of the residual, and thus only on the definition of the local accuracy metric. Under these conditions the ***minimum point*** $\bar{p}^*$ of $\mathcal{C}(\bar{p})$ over the

parameter space will always be an element of the identification space. Hence, the parameter set $\bar{p}^*$ locates $\mathcal{P}_\mathcal{X}$ in the parameter space.

Having located $\mathcal{P}_\mathcal{X}$, we can examine the extent of $\mathcal{P}_\mathcal{X}$, i.e. the distribution of the constraints in the parameter space about $\bar{p}^*$. For this purpose, we introduce a measure of dispersion δ, a scalar quantity which is defined by

$$\delta = g(\frac{1}{N}\mathcal{C}(\bar{p}^*))$$

where g is a strictly increasing scalar function, and $g(0) = 0$. While the residuals are a measure of the local accuracy of the model, the dispersion is a measure of the accuracy of the model over the whole ensemble $\mathcal{X}$, and thus of the precision of the model in the parameter space.

For all practical purposes, the identification space is adequately described by supplying a measure of its location $\bar{p}^*$ and a measure of its dispersion δ. Which measures are eventually used depends on the choice that is made for the functions l and g.

3.3 Conventional identification methods

From our discussion thus far it follows that model identification consists of two subsequent steps:

1. the formulation of an objective function, and
2. the minimization of this function with respect to the parameters in order to determine the optimum parameter set.

In some very special cases the minimization can be done analytically. However, in most cases a numerical procedure is required. This is the basic problem which is addressed by the applied mathematical branch of non-linear programming. Although the available minimization algorithms are manifold, their applicability depends on the form of the objective function that is chosen. An objective function for which no effective minimization method exists can be of theoretical interest, but no practical identification method can be based on it. Hence, even when the form of the objective function is completely determined by *a priori* knowledge about the observation errors and the modelling errors, this form will often have to be adapted to meet the requirements of a reliable and efficient minimization technique. These computational considerations tend to influence the choice of the objective function to such an extent that a single identification method now dominates the field: the method of least squares. Most implementations of identification methods that are found in literature [6–11] or are available commercially [12] fall in this category.

3.3.1 The least-squares method

The least-squares method is usually attributed to Carl Friedrich Gauss (portrayed on the cover), who formulated its basic principle in 1809, and proposed it as a general identification method [29]. In a modelling context his principle of least squares (abbreviated LS) can be paraphrased as:

> *The most appropriate values for the unknown but desired model parameters are those for which the sum of the squares of the residuals is as small as possible.*

The parameter set that obeys this principle will therefore be the minimum point $\bar{p}^*$ of the objective function

$$\mathcal{C}_2(\bar{p}) = \sum_{i=1}^{N} \epsilon_i(\bar{p})^2 \tag{3.2}$$

Further, it is common practice to define the dispersion as

$$\delta_2 = \sqrt{\frac{1}{N}\mathcal{C}_2(\bar{p}^*)} \tag{3.3}$$

which is also known as the RMS (Root Mean Square) error of the model—strictly speaking a misnomer since the residuals are not necessarily equivalent to the modelling errors.

To gain some insight in the nature of the LS principle it is interesting to quote Gauss again [29]:

> This principle, ... , must be considered an axiom with the same propriety as the arithmetical mean of several observed values of the same quantity is adopted as the most probable value.

As a matter of fact, it is easy to show that the mean is just a special case of the application of the LS principle. Hence, we will slightly extend the concept of the mean, and state that the LS principle identifies the mean of the observational data in the parameter space. The mean is perhaps the most generally used statistical measure of location, and is in fact far older than the science of statistics to which it now belongs.

The term "most probable" in the quotation suggests an interpretation of the LS principle that has led, in the past, to confusion. The LS principle was originally derived by Gauss using probabilistic methods [13]. A LS identification method is then concerned with determining the most probable distribution

of the residuals and consequently the most probable value of the parameters. Hence, in theory, the application of a LS method requires certain probabilistic assumptions about the residuals. However, it must be emphasized here that the introduction of these assumptions, which will generally be invalid, is often only a convention for solving a practical problem.

The LS principle can be motivated independently of probability theory by interpreting the mean not as the most probable value, but merely as the most convenient measure of location. What gives the LS criterion its greatest importance in practice, is its superior mathematical tractability. The mathematical simplicity of the LS formulation appealed to Gauss and it has appealed to almost everyone concerned with the analysis of observations ever since. Powerful and elegant algorithms have been developed for the minimization of the LS objective function. The availability of these algorithms is a dominant reason for the proliferation of the LS identification method. However, since the LS method has no formal mathematical basis in and of itself, the reasons for accepting it can only be pragmatic. Therefore, as with any other systematic principle or axiom, the acceptability of the LS method depends on the acceptability of the results to which it leads.

3.3.2 The results

The outcome of an identification method is a description of the identification space, whereas most applications of device models require a single parameter set. This conflict is usually resolved by choosing the location of the identification space $\bar{p}^*$ as the identified parameter set. Figure 3.4 shows the result of the LS method in the parameter space for the example of Section 3.2.1. The mean parameter set determined by the LS method lies at the "centre of gravity" of the identification space. The associated model curve in the observation space is shown in Figure 3.5. Here, the averaging effect of the LS method has distributed the (limited) model accuracy evenly over all the observations.

The acceptability of this result depends on its accuracy. Obviously, the use of a single parameter set to represent the identification space only makes sense when the dispersion is sufficiently small. More specifically, the parameter set $\bar{p}^*$ can only be accepted when the model $\mathcal{M}(\bar{p}^*)$ is valid for the given set of observations $\mathcal{X}$. Hence, if the accuracy of the LS model curve proves to be insufficient, and the validity requirement is not met, then the mean parameter set must be rejected. The consequences of such a rejection depend on the nature of the model. Although the representations of analytical models and data-fitting models are often similar—they may even share the same model structure—the differences in their approach to the modelling problem should not be overlooked.

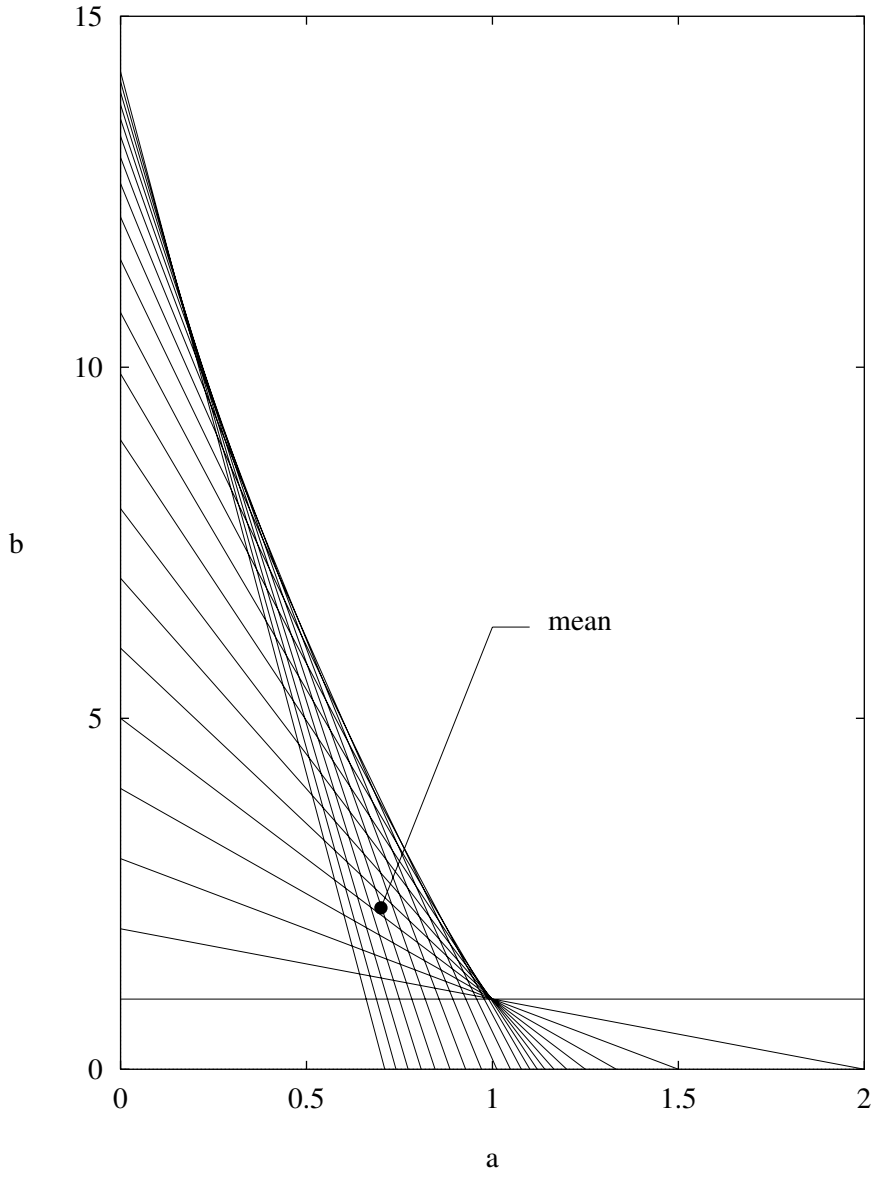

Figure 3.4: The location of the mean in $\mathcal{P}_{\mathcal{X}}$.

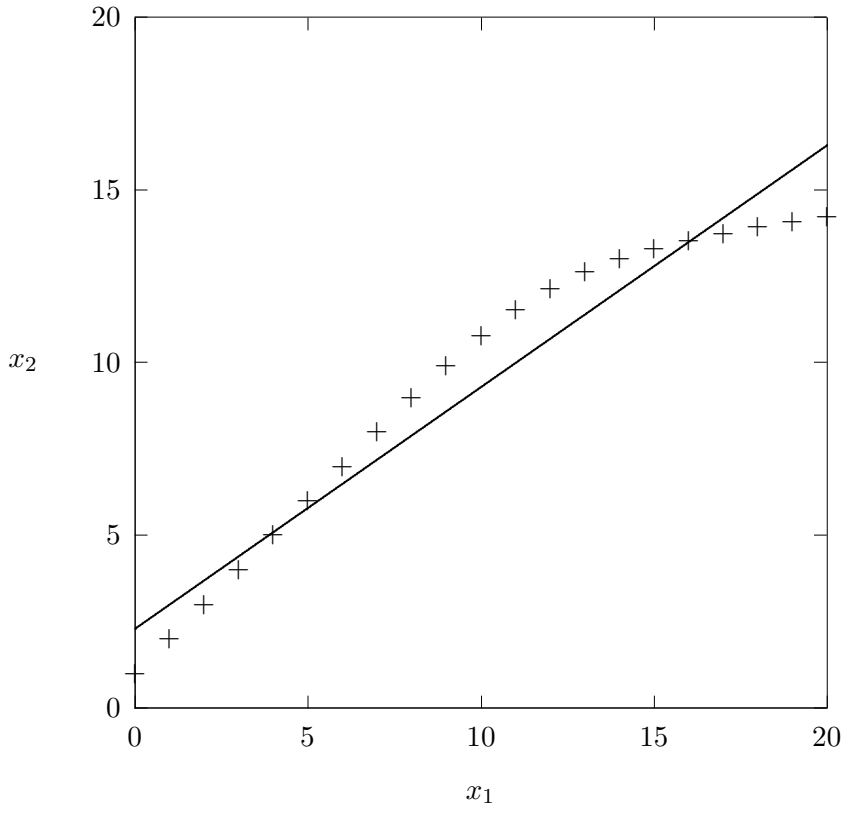

Figure 3.5: The least-squares model curve.

A data-fitting model is a compact representation of the observed behaviour, which aims to reproduce the complete set of observations $\mathcal{X}$ with maximum accuracy. By definition, maximum accuracy is achieved by minimizing a fitting criterion, which forms an integral part of the model. Here, the objective function plays the role of the fitting criterion, and the mean parameter set $\bar{p}^*$ minimizes this criterion. Hence, the rejection of the mean parameter set ensues the rejection of the data-fitting model as a whole. In order to represent the given set of observations $\mathcal{X}$ we must then make a different choice for either the model structure or the fitting criterion or both.

By contrast, in the case of an analytical model the rejection of the mean parameter set does not necessarily cast doubt on the model structure. As the validity domain of an analytical model is always finite, it may not include the whole data set $\mathcal{X}$, so we should instead reject some of the observations. By using the LS method indiscriminately, we are allowing observations that lie outside the validity domain of the model, and which contain no useful information about the

model, to bias the identified parameters in their favour. Because of this bias, the LS method will fail to determine the parameters of an analytical model with sufficient accuracy in many practical identification problems.

However, due to the convenience of the LS method, there has been a tendency to treat all analytical device models as data-fitting models with respect to identification. This approach severely limits the usefulness of these models for circuit design. Therefore, to take full advantage of the superior properties of the analytical device models, an equally convenient but more rigorous method needs to be developed for their identification.

3.4 Mode selection

When the analytical device models were introduced in Section 2.2.2 it was suggested that the parameters of these models exist in their own right as they represent physical quantities. However, since a model is an approximation of reality and the accuracy of this approximation is necessarily limited, these "true" parameters cannot exist in any absolute sense. Nevertheless, we will generally find that the true parameters of an analytical model can be defined in an asymptotical sense. It is this set of parameters that we will seek to determine.

3.4.1 The mode

The choice of a parameter set $\bar{p} \in \mathcal{P}_{\mathcal{X}}$ as an estimate of the true parameters of an analytical model requires a trade-off between the residuals of the observations. In fact, each choice for $\bar{p}$ implies a hypothesis about the distribution of the model accuracy over the observations. Not every distribution of the accuracy is equally plausible when taking into consideration the limited extent of the validity domain of an analytical model. The validity domain divides the set of observations $\mathcal{X}$ into two disjoint subsets: the set of observations that lie within the validity domain of the model $\mathcal{V}$, and the set of observations that lie outside it $\overline{\mathcal{V}}$. The residuals of all observations in $\mathcal{V}$ will satisfy the validity criterion $\epsilon \leq 1$, while the residuals of the observations in $\overline{\mathcal{V}}$ will all fail this criterion. Hence, for an analytical model the distribution of the model accuracy over the observations tends to be particularly uneven.

The uneven distribution of the model accuracy in the observation space is reflected in the parameter space, where all the observations in $\mathcal{V}$ agree on a relatively small region in the identification space, say $\phi \subset \mathcal{P}_{\mathcal{X}}$. The observations in $\overline{\mathcal{V}}$ neither agree with the parameter values in ϕ, but what is more important, nor do they have any definite relation with one another. As a result, the identification space of an analytical model does not have a homogeneous structure, but contains a cluster.

The link between the validity domain of a model and a cluster in the parameter space provides the basis for a more accurate identification method for analytical models. If, to begin with, there exists a distinct cluster in the parameter space, then it should be possible to identify the set $\mathcal{V}$ by locating this cluster. Furthermore, the location of this cluster is a good estimate of the true values of the model parameters, as it is determined only by the observations for which the model is sufficiently accurate. Now, the extent of the validity domain of an analytical model is not fixed for a device but depends on the modelling goal; a more stringent accuracy requirement will result in a smaller validity domain, and hence a smaller set $\mathcal{V}$. This will in turn result in a different estimate for the model parameters. Because this estimate is determined by observations for which the model is even more accurate, it will be an even closer approximation of the true parameters of the model, which are, of course, independent of the modelling goal. This reasoning can be taken one decisive step further by concluding that when the cluster in the identification space is located using an increasingly stricter accuracy requirement, its location will approach the true parameter set of the model. In practice, this asymptotic procedure is halted either by the magnitude of the observation errors, or by the limited number of the observations. In theory, however, the cluster could (in the limit) be reduced to a single point in the identification space. This point, which can be qualified as the *mode*[1] of the identification space, could then be defined as the true parameter set of the model. Therefore, we will state as a general principle that:

> *The mode of the observational data in the parameter space provides the most appropriate values for the unknown parameters of an analytical device model with a finite validity domain.*

Figure 3.6 shows the location of the mode of the identification space for our running example. The indicated point is clearly a focal point for the observational data in the parameter space. The associated model curve in the observation space is shown in Figure 3.7. Here, the model accuracy is unevenly distributed over the observations in comparison with the LS model (see Figure 3.5). The model accuracy is high for observations that have low values of x, while gradually deteriorating for observations that have higher values of x. When we interpret the marks $(+)$ as the validity bars of the observations (which are comparable to error bars, except that they indicate the maximum *acceptable* error), we find that the linear model hypothesis is valid for about half the observations. Moreover, the fact that these observations form an uninterrupted series suggests a validity domain for the model that can be specified as $0 \leq x \leq 10$.

[1] This usage of the term complies with the conventional use of the mode as a descriptive statistic on statistical data [13].

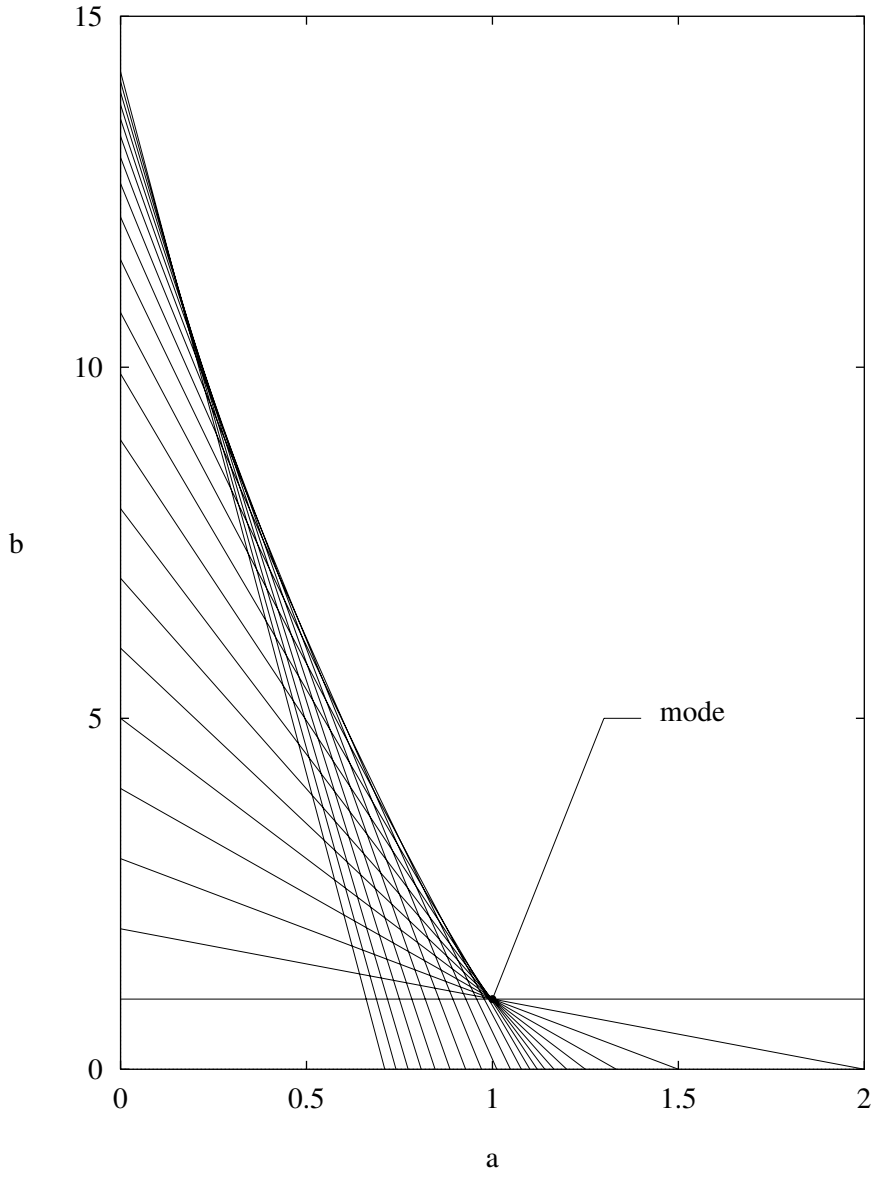

Figure 3.6: The location of the mode in $\mathcal{P}_{\mathcal{X}}$.

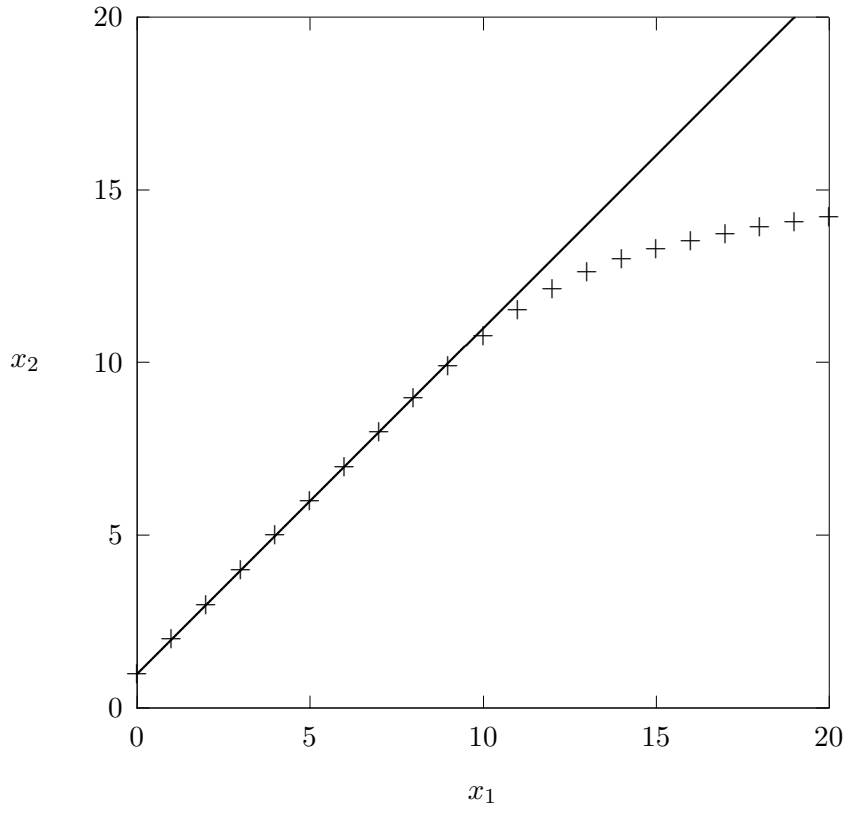

Figure 3.7: The most appropriate analytical model curve.

In the preceding discussion the mode was introduced in rather loose terms. However, a strict mathematical definition of the concept is indispensable for the development of a procedure for its identification. We therefore proceed by proposing such a definition. Consider the following function of the parameters

$$\mathcal{N}(\bar{p}) = \sum_{i=1}^{N} \begin{cases} 1 & \text{if } \epsilon_i(\bar{p}) \leq \nu \\ 0 & \text{if } \epsilon_i(\bar{p}) > \nu \end{cases} \tag{3.4}$$

This function counts the number of observations that agree on each value of $\bar{p}$, given the accuracy requirement $\epsilon \leq \nu$ (hence $0 < \nu \leq 1$). When applied to the identification space, the cluster will be characterized by large values of $\mathcal{N}$. Hence, the point (or, more precisely, the small region $\phi \subset \mathcal{P}_{\mathcal{X}}$) that maximizes $\mathcal{N}$ can serve as the location of the cluster and as an estimate of the mode. As the accuracy of this estimate depends on the value of ν, the indeterminacy in the location of the mode could be removed by reducing the value of ν to zero. Again, in practice, the choice of an acceptable lower bound for ν is a subtle

problem. Although smaller ν gives a better resolution in the parameter space and increases the chance of accurately locating a very sharp mode, smaller ν also gives poorer results in separating the true ensemble mode from chance fluctuations in the data.

Just like the mean was defined as the minimum point of the LS objective function (3.2), we can now formulate an objective function that has the mode as its minimum point. The function

$$\mathcal{C}_0(\bar{p}) = N - \mathcal{N}(\bar{p})$$

is an obvious choice (although $\mathcal{C}_0$ does not fully conform to the definition of an objective function that was given in Section 3.2.3, as the function $l(\epsilon)$ is not *strictly* increasing). However, the minimization of this objective function will be problematic. The function completely consists of flat "plateaus" and discontinuities, and usually has many local minima, or sub-modes (especially for small values of ν), none of which can be handled by the standard minimization algorithms [30]. We reiterate that the usefulness of an objective function (and ultimately of the theory behind it) stands or falls with the availability of an effective minimization method. There have been several attempts to construct objective functions that have the mode as their global minimum and can be minimized by standard minimization techniques. The resulting so-called robust estimation methods [31, 32] are, however, not completely successful. Although most of these methods have succeeded in eliminating the plateaus and the discontinuities from the objective function, unimodality of the objective function has not been achieved. This makes these methods unreliable. As a consequence, they have never acquired much popularity in the field of device model identification. Actually, by the very nature of the problem it is unlikely that a unimodal objective function for the mode even exists. The minimization of a single objective function is just not the appropriate way to solve what is basically a partitioning problem: selecting a subset of $\mathcal{X}$ for which the model is sufficiently accurate.

3.4.2 The mode selection criterion

To arrive at a practical criterion for the identification of the mode, we will in effect turn around the definition of the mode that finds expression in (3.4). There, an estimate of the mode is obtained by searching the parameter space for a parameter set that is agreed on by as many observations as possible. Here, we hope to obtain the same result by searching the set of observations $\mathcal{X}$ for the largest subset of which all the members agree on the same parameter set.

The search space is thus no longer the parameter space, but comprises all possible subsets $\mathcal{S}$ of $\mathcal{X}$. To represent these subsets we introduce the N-dimensional

selection vector $\bar{s}$. Each element s_i of the selection vector corresponds to an observation $\bar{x}_i \in \mathcal{X}$, and can only have one of two values:

$$s_i = \begin{cases} 1 & \text{if } \bar{x}_i \in \mathcal{S} \\ 0 & \text{if } \bar{x}_i \notin \mathcal{S} \end{cases}$$

In this way, each subset $\mathcal{S}$ is assigned a unique selection vector $\bar{s}$. Henceforth, we will use either notation where appropriate.

Each set $\mathcal{S} \subset \mathcal{X}$ defines an identification space $\mathcal{P}_\mathcal{S} \subset \mathcal{P}_\mathcal{X}$, the extent of which can be described using a suitable measure of dispersion such as δ_2 (3.3), which thus becomes a function of the selection vector $\delta_2(\bar{s})$. Due to the clustering in the parameter space, there will be subsets of $\mathcal{X}$, notably the set $\mathcal{V}$, that have a small value for the dispersion, not only compared with the dispersion for the complete set $\mathcal{X}$, but also compared with other subsets which contain the same number of observations. Hence, the cluster in the parameter space can be located by determining the subset in $\mathcal{X}$ that contains a large fraction of the total number of observations, while possessing a low value for the dispersion.

Since these two objectives are incompatible, the optimum subset will have to be a compromise. One way of combining both objectives in a single identification criterion for the mode is given below. The *mode selection criterion* is formulated as a constrained optimization problem over the selection vector:

$$\begin{aligned} \underset{\bar{s}}{\text{maximize}} \quad & \mathcal{N}(\bar{s}) = \sum_{i=1}^{N} s_i \\ \text{subject to} \quad & \delta_2(\bar{s}) \leq \omega \end{aligned} \tag{3.5}$$

where the bound on the dispersion ω, which in effect weighs the two objectives against each other, has some small positive value, usually $0 < \omega \leq 1$. The set of observations that solves this problem $\bar{s}^*$ (or in set notation $\mathcal{S}^*$) will be called the *mode set*.

The observations in the mode set reduce the indeterminacy of the location of the mode to the extent of their identification space $\mathcal{P}_{\mathcal{S}^*}$. When the extent of this identification space is small in comparison to the extent of $\mathcal{P}_\mathcal{X}$ (which will be the case for small values of ω), we can safely accept the location of $\mathcal{P}_{\mathcal{S}^*}$ as an estimate of the mode of $\mathcal{P}_\mathcal{X}$. However, since the extent of the $\mathcal{P}_{\mathcal{S}^*}$ is small, there is no compelling reason for choosing any particular measure of location. Hence, bearing in mind the computational convenience of the LS criterion, we will choose the mean parameters of the mode set as the estimate of the mode of $\mathcal{P}_\mathcal{X}$. The associated measure of dispersion $\delta_2(\bar{s}^*)$ can then be interpreted as a measure of the accuracy of this estimate—which is incidentally the main argument for using δ_2 as the measure of dispersion in (3.5). The role of the bound ω in (3.5) is thus comparable to that of the bound ν in (3.4).

The problem of finding the optimum of the mode selection criterion (3.5) belongs (together with most partitioning problems [33]) to the class of NP-complete problems [1]. As the number of possible subsets in $\mathcal{X}$ equals 2^N, the computational effort required for an exhaustive examination of the search space doubles for each observation that is added to $\mathcal{X}$. Therefore, an exhaustive procedure for finding the global optimum of (3.5) must be ruled out because of the prohibitive computational effort that would be required to solve most practical identification problems. Even if, for instance, the dispersion of each subset $\mathcal{S}$ could be calculated within 1 ms, an exhaustive search for a data set containing only 100 observations would already take well over 10^{19} years (which is a billion times the approximate age of the universe).

The example shows that it is imperative to employ an optimization strategy that examines only a small number of potential solutions in the search space in a systematic manner. Of course, such a systematic (and thus local) examination of the search space can only succeed when the optimization problem has some regular overall structure. However, there are good reasons to expect that for the identification problems being considered here such a structure does indeed exist. First of all, we are applying the mode selection criterion not to random data, but to observational data that has been obtained from a real "physical" device. Moreover, an analytical model is expected to approximate this data in a highly characteristic way, namely *asymptotically* (i.e. outside the validity domain, the model accuracy deteriorates only gradually). The analysis of a considerable number of practical identification problems has indeed revealed a characteristic and consistent problem structure. By exploiting this problem structure we have been able to develop a heuristic method for selecting the mode set, which has proven to be both reliable and efficient.

3.4.3 The selection space

Before presenting the optimization strategy, we will first apply an ordering to the search space. For this purpose we introduce the *selection space*, an N-dimensional binary space in which the elements of the selection vector s_i $(i = 1, \ldots, N)$ are interpreted as co-ordinates. The distance between two points in the selection space is defined as the *number* of co-ordinates at which they disagree, which is an effective measure of the similarity between the associated subsets. Hence, a step of length one in the selection space corresponds to the modification of the current set by adding or removing a single observation. A binary space in which this distance measure is defined is generally referred to as a sequence space.

The dispersion δ_2 and set size $\mathcal{N}$ are now functions over the selection space. In Figure 3.8 the selection space for $N = 4$ is represented in a form that allows the value of $\mathcal{N}$ to be displayed along the vertical axis. All points that are a distance

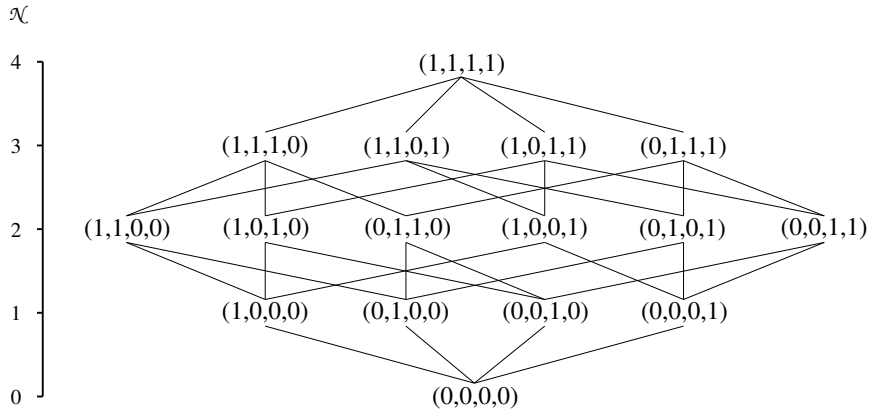

Figure 3.8: The selection space for $N = 4$.

of one apart are connected by a line. At every horizontal level $\mathcal{N} = n$ in the selection space there will be a point $\bar{s}^*(n)$ that has the lowest value for the dispersion $\delta_2^*(n)$. If we take $\omega = \delta_2^*(n)$ and then solve (3.5), we should obtain this point $\bar{s}^*(n)$ as an estimate of the mode set $\bar{s}^*$. Now the dispersion δ_2 of a set can always be reduced by removing one of its members (when $\delta_2 > 0$ such a member can always be found). Therefore, the minimum dispersion $\delta_2^*(n)$ decreases for smaller values of n. Hence, solving (3.5) for gradually decreasing values of ω would result in a whole sequence of mode-set estimates on successively lower levels in the selection space. The aim of an optimization algorithm is now to find a path (preferably the shortest) through the selection space from the initial estimate at $\bar{s} = \bar{1}$ ($\mathcal{N} = N$) to the optimum mode set $\bar{s}^*$.

3.4.4 The optimization strategy

Since the optimization problem (3.5) is expected to have a regular overall structure, we will only consider local-search algorithms [34]. These algorithms can be formulated as follows. Starting at the initial point, a sequence of iterations $\bar{s}^{(k)}$ $(k = 1, 2, \ldots)$ is generated, each consisting of a step from the current point to a point selected from the neighbourhood of the current point. Each step is selected by comparing the dispersion and set size at the current point with the dispersion and the set size at the neighbouring points. The search is terminated when a point is reached that satisfies the mode selection criterion better than any of its neighbours. As all steps are chosen on the basis of local information, this optimization strategy relies on extrapolation to find the global solution.

The computational effort that is required by the algorithm to find the solution depends on the maximum distance that it is allowed to travel per step. The

step length determines the extent of the local neighbourhoods, and hence the number of points that have to be examined per iteration. For example, if we choose the step length to be N, the local neighbourhood of each point will include the whole selection space. Although the solution can then be reached in a single step, we must perform an exhaustive search of the selection space to select the correct step, which is an $O(2^N)$ procedure. If, on the other hand, we choose a step length of *one*, the local neighbourhood of each point contains only N points. Still, the algorithm would be able to reach each point in the search space, including the solution, from any starting position in no more than N steps (of course provided that the algorithm selects the correct step every time). Hence, traversing the selection space using steps of length one can be an $O(N^2)$ procedure. However, the smaller neighbourhood also means that less information is available for the decision on the step direction. Therefore, the choice of such a small step size puts high demands on the problem structure.

A sufficient problem structure is provided by the following premise:

> The mode set $\bar{s}^*(n-1)$ is a subset of the mode set $\bar{s}^*(n)$ for all $n \in \{N, \ldots, \mathcal{N}(\bar{s}^*)\}$.

This puts the mode sets in each others neighbourhoods, so the optimum step direction can be selected by simply minimizing δ_2 over the neighbouring points. The optimization algorithm can then follow a path through the selection space consisting completely of estimates of the mode set $\bar{s}^{(k)} = \bar{s}^*(N-k)$, sequentially eliminating observations from the current set that are not in the mode set $\bar{s}^*$. In practice, we find that the premise does not hold for small n (or small ω). However, the low dispersion of small isolated subsets is likely to be caused by random fluctuations in the data. Hence, these subsets should indeed not be considered as candidates for the mode set. In other words, the premise requires the mode to be a global feature of the identification space.

By basing our optimization strategy on this premise we have in effect modified the definition of the mode. The mode for a set of observations $\mathcal{S}$ (starting with $\mathcal{S} = \mathcal{X}$) is now defined by the recursive sequence:

1. select the set $\mathcal{S} \setminus \bar{x}_i$ (for $\bar{x}_i \in \mathcal{S}$) that minimizes δ_2, and
2. determine the mode for the set $\mathcal{S} \setminus \bar{x}_i$.

By this process of stepwise refinement, henceforth referred to as the ***mode selection method*** (abbreviated MODES), we aim to locate the ensemble mode with increasing accuracy. The value of ω in the constraint on δ_2 in (3.5), which plays the role of the stopping criterion for the algorithm, then specifies the accuracy with which the location of the ensemble mode is to be approximated.

The measure of location that is defined by the MODES algorithm is often a better candidate for the mode of the observational data, at least in the sense of an intuitive interpretation of the concept, than the one given in Section 3.4.1. The standard definition of the mode (3.4) introduces the accuracy parameter ν in the identification problem. We have seen that even for the most accurate models and observations, there exist a lower bound for ν below which the grouping becomes ineffective. When the value of ν is decreased any further, the location of the maximum of $\mathcal{N}$ may suddenly shift to a different point anywhere in $\mathcal{P}_{\mathcal{X}}$. Hence, for small values of the accuracy parameter ν the point in the identification space that is designated as the mode is no longer a consistent measure of the central tendency of the data.

The MODES algorithm avoids this problem by sequentially eliminating those sections of the identification space $\mathcal{P}_{\mathcal{X}}$ that are unlikely to contain the mode. As the extent of the identification space is only reduced (and with it the indeterminacy in the location of the mode), the recursively defined mode is a highly consistent measure of location. Moreover, since the mode selection criterion is based on the least-squares criterion, it inherits its convenient mathematical structure. This mathematical structure makes it possible to develop an efficient method to calculate the location and dispersion for the selected subsets, but also to solve the selection problem itself. These properties make the MODES method particularly appropriate for the identification of analytical models. The implementation of this method will be the subject of the next chapter.

Chapter 4

Implementation

The implementation of the mode selection method (MODES) can be subdivided into three hierarchically related minimization problems:

1. the calculation of the residuals for fixed $\bar{s}$ and $\bar{p}$, minimizing a distance function with respect to $\bar{x}$,
2. the calculation of the location and dispersion for fixed $\bar{s}$, minimizing the least-squares objective function $\mathcal{C}_2$ with respect to $\bar{p}$, and finally
3. the selection of the mode set, minimizing the dispersion δ_2 with respect to $\bar{s}$.

In the subsequent sections we will develop algorithms for solving these three minimization problems that are both reliable and efficient.

4.1 The residuals

The first step in evaluating the mode selection criterion (3.5) is the calculation of the residuals of the observations. In Section 3.2.2 the residual of an observation was introduced as the distance in the observation space between the observation and the model subspace. The geometric concept of distance rests on the definition of a metric. The formal definition of this metric will complete the specification of the identification criterion.

4.1.1 The accuracy metric

The weighted Euclidean distance [35] between an observation, denoted by $\bar{x}_0$, and an arbitrary point $\bar{x}$ in the observation space is defined by

$$d(\bar{x}_0, \bar{x}) \triangleq \sqrt{(\bar{x} - \bar{x}_0)^t V (\bar{x} - \bar{x}_0)}$$

where $d \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \rightarrow \mathbb{R}$ is the distance function, and the $n_x \times n_x$ matrix V represents the metric (or the local scale by which the distance is measured). To satisfy the metric axiom $d(\bar{x}_0, \bar{x}) \geq 0$, with equality if and only if $\bar{x} = \bar{x}_0$, the matrix V must be positive definite.

According to Section 3.2.2, the required model accuracy defines a reference distance and hence an accuracy metric. More specifically, all points $\bar{x}$ in the observation space that approximate the observation with sufficient accuracy must satisfy the inequality

$$d(\bar{x}_0, \bar{x}) \leq 1$$

This is the equation of a region bounded by an n_x-dimensional ellipsoid centered at the observation. Because the required model accuracy is associated with the local validity criterion this region will be called the *validity region* of the observation.

The principal axes of the ellipsoid are parallel to the eigenvectors of V, while their lengths are inversely proportional to the square roots of the corresponding eigenvalues of V. When specifying the accuracy metric of an observation it is common practice to choose the principal axes of the ellipsoid parallel to the coordinate axes of the observation space. With this simplification the accuracy metric V can be represented by a diagonal matrix

$$V = \begin{pmatrix} 1/{v_1}^2 & & & \\ & 1/{v_2}^2 & & \\ & & \ddots & \\ & & & 1/{v_{n_x}}^2 \end{pmatrix}$$

where the semi-axes of the ellipsoid have lengths v_i $(i = 1, \ldots, n_x)$ relative to the interface variables x_i $(i = 1, \ldots, n_x)$. This is illustrated in Figure 4.1 for a two-dimensional observation space.

Since the required model accuracy may depend on the observation, and since the observations are all independent, we assign a specific accuracy metric V_i $(i = 1, \ldots, N)$ to each observation.

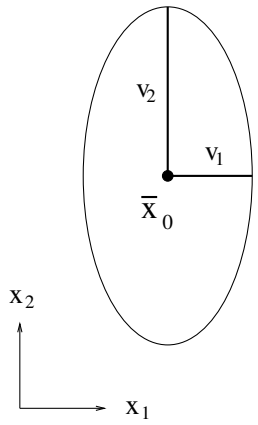

Figure 4.1: The validity region of an observation.

4.1.2 The model subspace

A device model (for a particular value of the parameter vector $\bar{p}$) defines a functional relationship between the interface variables:

$$f(\bar{x}) = \bar{0} \tag{4.1}$$

where $f \in \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_f}$. The model therefore determines an $(n_x - n_f)$-dimensional subspace in the observation space, where it is assumed that $n_x \geq n_f$, as is the case for all practical models.

In Euclidean geometry the distance between a point and a subspace is defined as the distance between this point and the nearest point in the subspace. Hence, to determine the residual of an observation we must, among all the points $\bar{x}$ that satisfy (4.1), find the one that minimizes the distance $d(\bar{x}_0, \bar{x})$. This implies solving the following equality constrained minimization problem

$$\begin{array}{ll} \underset{\bar{x}}{\text{minimize}} & d(\bar{x}_0, \bar{x}) \\ \text{subject to} & f(\bar{x}) = \bar{0} \end{array} \tag{4.2}$$

The vector $\bar{x}^*$ that solves this problem corresponds to the point in the model subspace that is closest to the observation x_0 with respect to the given metric V. Hence, the residual can be written as

$$\epsilon = d(\bar{x}_0, \bar{x}^*) = \sqrt{(\bar{x}^* - \bar{x}_0)^t V (\bar{x}^* - \bar{x}_0)} \tag{4.3}$$

If the model subspace intersects the validity region of the observation, the point $\bar{x}^*$ will be a member of the validity region. Then the residual will be less than one, satisfying the local validity criterion.

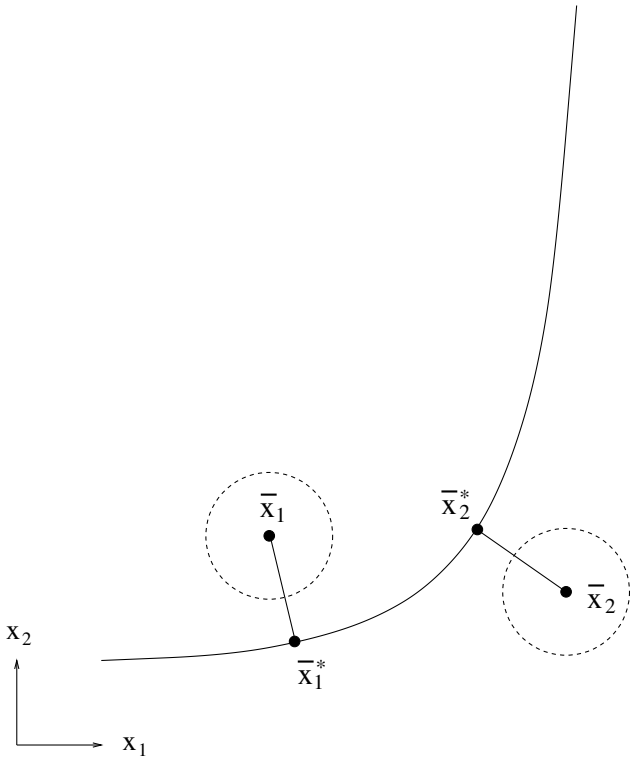

Figure 4.2: The nearest points on the model curve.

A geometrical interpretation of the residuals (for $n_x = 2$ and $n_f = 1$) is shown in Figure 4.2. Here the scale of the representation has been chosen such that the validity regions of both observations are bounded by circles. The points $\bar{x}^*$ are then the feet of the perpendiculars from the observations onto the model curve. The residuals are equal to the lengths of these perpendiculars (with respect to the chosen metric).

Calculating the residuals means solving the constrained minimization problem (4.2) for each observation. This is generally considered to be an arduous task, especially if the constraints are non-linear. Hence, there is a tendency to modify the definition of the residuals to facilitate their calculation.

4.1.3 Device simulation

If a reduction of the dimension (degrees of freedom) of the validity regions of the observations is considered acceptable, the problem of calculating the residuals can be simplified to a significant extent. To this end, the interface variables are divided into two groups: the independent variables $\bar{u} \in \mathbb{R}^{(n_x - n_f)}$, and the

dependent variables $\bar{y} \in \mathbb{R}^{n_f}$. This division may be, and often is, the same as the one that was used for the experiment. According to the implicit function theorem [36], the n_f model constraints define (under certain conditions) a function from the independent to the dependent variables of the form

$$\bar{y} = h(\bar{u})$$

where $h \in \mathbb{R}^{(n_x - n_f)} \rightarrow \mathbb{R}^{n_f}$. When now the values of the independent variables are taken to be fixed and equal to the observed values, the expression for the residual (4.3) can be reduced to

$$\epsilon = \sqrt{(h(\bar{u}_0) - \bar{y}_0)^t V_y (h(\bar{u}_0) - \bar{y}_0)} \tag{4.4}$$

where $\bar{u}_0$ and $\bar{y}_0$ stand for the observed values of the interface variables. The $n_f \times n_f$ matrix V_y, which defines the accuracy metric for the dependent variables, is a reduced form of the matrix V; all rows and columns of V associated with the independent variables have been removed, effectively setting the v_i of the independent variables to zero. The validity domains of the observations are therefore bounded by n_f-dimensional ellipsoids in the n_x-dimensional observation space.

The effect that the introduction of independent interface variables has on the residuals is illustrated in Figure 4.3. It shows the same model curve as Figure 4.2, but here the variables x_1 and x_2 have been chosen as the independent variable and the dependent variable respectively.

Calculating the value of $\bar{y}$ for a given value of $\bar{u}$ is usually referred to as device simulation, being the modelling equivalent of performing an experiment with the real device. Since the values of the independent variables are considered to be known ($\bar{u} = \bar{u}_0$), the number of variables in (4.1) is reduced to n_f. The process of calculating the value of $\bar{y}$ is then equivalent to finding the root of the set of n_f non-linear equations in n_f variables. The availability of efficient methods for solving this type of problem (usually based on the Newton-Raphson algorithm [32]), is a strong incentive for choosing this approach for calculating the residuals. This argument is reinforced by the fact that these methods are easily accessible as they are implemented in the form of general-purpose simulation programs [1]. As a consequence the "device-simulation" residual (4.4) is used in most model identification programs [6–10].

However, the *ad hoc* division of the interface variables into a dependent set and an independent set leads to theoretical and computational problems. In particular, the distinction between the variables suggests a unidirectionality which is rarely justifiable for most devices. In addition, the introduction of independent variables results in a "weighting" of the observations. The weights, which are applied to the residuals, are proportional to the gradient of the model function $\nabla h(\bar{u_0})$. Hence, when the model is strongly non-linear the weights will be

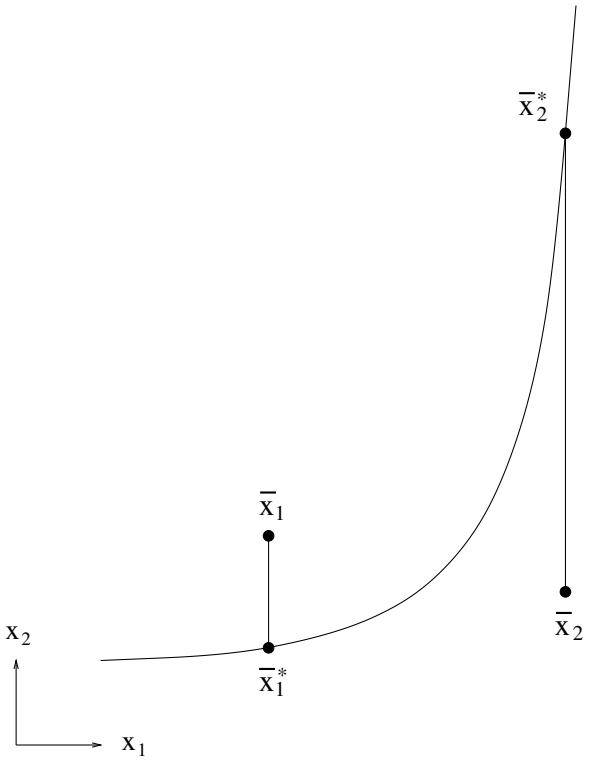

Figure 4.3: The nearest points for fixed independent variables.

spread over the observations in an highly uneven way. Consequently, the observations that are situated in the region where the model function is steep tend to dominate the objective function, effectively eliminating the other observations from the objective function. This effect is shown in Figure 4.3, where the residual of the second observation now outweighs the residual of the first observation, even though the distances from the observations to the model curve are approximately equal. These objections preclude the general use of the device-simulation residual (4.4).

4.1.4 Constrained minimization

Before introducing a method for the calculation of the residuals by directly solving the constrained minimization problem (4.2), some relevant optimization theory will be presented. We will consider the general problem of equality constrained minimization:

$$\begin{aligned} \underset{\bar{x}}{\text{minimize}} \quad & F(\bar{x}) \\ \text{subject to} \quad & f(\bar{x}) = \bar{0} \end{aligned} \tag{4.5}$$

We suppose that the functions $F \in \mathbb{R}^{n_x} \to \mathbb{R}$ and $f \in \mathbb{R}^{n_x} \to \mathbb{R}^{n_f}$ are differentiable and that first derivatives can be calculated.

The classical approach to equality constrained minimization is due to Lagrange [35]. His contribution to the theory of constrained minimization was the discovery that the solution of the constrained minimization problem (4.5) is an unconstrained critical point of the scalar function

$$L(\bar{x}, \bar{\lambda}) = F(\bar{x}) - \sum_{i=1}^{n_f} \lambda_i f_i(\bar{x})$$

The function $L \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_f} \to \mathbb{R}$ is called the Lagrangian function, and the n_f-dimensional vector $\bar{\lambda}$ is called the Lagrange multiplier vector.

The critical points of the unconstrained function L are those vectors $(\bar{x}, \bar{\lambda})$ for which the gradient of L is zero. The gradient is zero where the partial derivatives of L with respect to all variables are zero. Hence, a critical point $(\bar{x}^*, \bar{\lambda}^*)$ of L satisfies

$$f(\bar{x}) = \bar{0} \tag{4.6}$$

$$\nabla F(\bar{x}) = \sum_{i=1}^{n_f} \lambda_i \nabla f_i(\bar{x}) \tag{4.7}$$

These equations, which are referred to as the primal (4.6) and the adjoint (4.7) equation, are the first-order conditions for the constrained minimum of F.

For this result to be valid it is necessary that at the constrained minimum $\bar{x}^*$ the constraints act independently, i.e. the set $\{\nabla f_i(\bar{x}), i = 1, \ldots, n_f\}$ is linearly independent. If this regularity condition is satisfied, and we will henceforth assume that this is the case for all $\bar{x}$, the Lagrange multiplier vector $\bar{\lambda}^*$ is uniquely determined by equation (4.7) at $\bar{x}^*$. However, if the model constraints are not linearly independent in $\bar{x}^*$, the Lagrange multipliers will not be unique or may not even be finite.

The design of algorithms for solving (4.5) is usually governed by the first-order conditions. Since $(\bar{x}^*, \bar{\lambda}^*)$ is a root of (4.6) and (4.7), the obvious approach would be to solve this system of $(n_x + n_f)$ equations in $(n_x + n_f)$ variables directly. We could then use the algorithms that are available for device simulation. There are, however, two major objections to this approach. In the first place, the first-order conditions are necessary but not sufficient conditions for a solution of the constrained minimization problem (4.5). Not every critical point of L necessarily corresponds to a constrained minimum of F. A constrained maximum or a constrained saddle point of F will also contribute a critical point to L. The (non-linear) system of equations (4.6) and (4.7) will then have multiple roots. It is possible to formulate second-order conditions

for a constrained minimum of F [37], to differentiate between the different types of critical points once they have been identified. However, this ability to reject an inappropriate root still leaves the problem of finding the desired root unresolved.

The second objection to this approach is of a more practical nature. Efficient methods for solving a set of equations (such as the Newton-Raphson method) require the calculation of the first derivatives of these equations. Since the adjoint equation (4.7) already contains first derivatives of the constraints, calculation of the second derivatives of the constraints $\{\nabla^2 f_i(\bar{x}), i = 1, \ldots, n_f\}$ is required. When it is inconvenient to supply the second derivatives of the model constraints, this method of calculating the residuals is at a substantial disadvantage in comparison with the device simulation method, which does not require the second derivatives.

We will therefore propose an algorithm that only requires the calculation of the first derivatives of the model constraints. It avoids the explicit use of the first-order conditions, instead, the constrained minimum is computed by working directly with the problem functions F and f. In order to be efficient, the algorithm is designed to take full advantage of the special characteristics of problem (4.2).

4.1.5 The iteration equations

The special structure of problem (4.2) enables the direct calculation of the solution vector $\bar{x}^*$ when the model constraints f are linear in $\bar{x}$. Not only is this the case when the device model is in fact linear, but also when the distance of $\bar{x}_0$ to the model subspace gets small enough for $f(\bar{x})$ to be approximated by its linearization in $\bar{x}_0$. This will often be the case in the final stages of the identification process, as the minimization of these distances is our main goal. We will therefore linearize the constraints of problem (4.2) in an attempt to generate an algorithm in which a sequence of linearly constrained sub-problems is solved. Our aim is to construct, in the limit of an iterative process, a linearly constrained sub-problem which has $\bar{x}^*$ as a local minimum. The notation $\bar{x}^{(k)}$ (for $k = 0, 1, \ldots$) will be used for the sequence of points calculated by the iterative method, where it is usual to choose $\bar{x}^{(0)} = \bar{x}_0$.

For mathematical convenience we first reformulate the problem in a true quadratic form:

$$\begin{array}{ll} \underset{\bar{x}}{\text{minimize}} & F(\bar{x}) = \frac{1}{2} d(\bar{x}_0, \bar{x})^2 \\ \text{subject to} & f(\bar{x}) = \bar{0} \end{array} \tag{4.8}$$

This problem has the same solution $\bar{x}^*$ as problem (4.2). To produce a linearly

constrained sub-problem we linearize the constraints f in the point $\bar{x}^{(k)}$ to get

$$\begin{array}{ll} \underset{\bar{x}}{\text{minimize}} & F(\bar{x}) = \frac{1}{2}(\bar{x} - \bar{x}_0)^t V(\bar{x} - \bar{x}_0) \\ \text{subject to} & f(\bar{x}^{(k)}) + J_x\,(\bar{x} - \bar{x}^{(k)}) = \bar{0} \end{array} \tag{4.9}$$

where J_x represents the Jacobian matrix, the $n_f \times n_x$ matrix of partial derivatives of f with respect to $\bar{x}$, evaluated at $\bar{x}^{(k)}$.

This problem suggests an iterative method in which $\bar{x}^{(k+1)}$ is chosen as the solution of (4.9) and $\bar{\lambda}^{(k)}$ as the corresponding multiplier vector of the linear constraints. Hence, these vectors must satisfy the associated primal and adjoint equations:

$$f(\bar{x}^{(k)}) + J_x(\bar{x}^{(k+1)} - \bar{x}^{(k)}) = \bar{0} \tag{4.10}$$

$$V(\bar{x}^{(k+1)} - \bar{x}_0) = J_x^t \bar{\lambda}^{(k)} \tag{4.11}$$

This system of linear equations has only one root (still assuming that J_x has full rank). We can solve these equations for the correction step $\bar{\xi}^{(k)} = (\bar{x}^{(k+1)} - \bar{x}^{(k)})$ and the next estimate of Lagrange multipliers $\bar{\lambda}^{(k)}$:

$$\begin{aligned} \bar{\xi}^{(k)} &= -V^{-1}J_x^t \left(J_x V^{-1} J_x^t\right)^{-1} f(\bar{x}^{(k)}) \\ &\quad - \left\{I - V^{-1}J_x^t \left(J_x V^{-1} J_x^t\right)^{-1} J_x\right\} (\bar{x}^{(k)} - \bar{x}_0) \end{aligned} \tag{4.12}$$

$$\bar{\lambda}^{(k)} = \left(J_x V^{-1} J_x^t\right)^{-1} \left\{J_x(\bar{x}^{(k)} - \bar{x}_0) - f(\bar{x}^{(k)})\right\} \tag{4.13}$$

Equation (4.12) shows that the correction step $\bar{\xi}^{(k)}$ can be divided into two parts. This is illustrated in Figure 4.4. The first part $\bar{\xi}_n^{(k)}$ solves the linearized primal equation, calculating the nearest point to $\bar{x}^{(k)}$ (with respect to the metric V) on the constraint hyperplane. The second part $\bar{\xi}_t^{(k)}$, the projection of the vector $(\bar{x}^{(k)} - \bar{x}_0)$ on the constraint hyperplane, then minimizes $d(\bar{x}_0, \bar{x}^{(k+1)})$ while keeping the linearized constraints satisfied. The two vectors are orthogonal with respect to the given metric V, i.e. $\bar{\xi}_n^t V \bar{\xi}_t = 0$.

A particular feature of this algorithm is that it is controlled by the estimates $\bar{x}^{(k)}$ of $\bar{x}^*$. The next estimate $\bar{x}^{(k+1)}$ as well as the estimates of the Lagrange multipliers solely depend on $\bar{x}^{(k)}$. This means that if $\bar{x}^{(k)} = \bar{x}^*$, so that $\bar{\xi}^{(k)} = \bar{0}$, then $\bar{\lambda}^{(k)} = \bar{\lambda}^*$. That is to say, if $\bar{x}$ is correct, then any errors in $\bar{\lambda}$ are annihilated. Thus when the iterates $\bar{x}^{(k)}$ converge to the solution $\bar{x}^*$, the estimates $\bar{\lambda}^{(k)}$ will converge to the vector of Lagrange multipliers $\bar{\lambda}^*$. However, in the next subsection it will be shown that, instead of being a superfluous calculation, the estimates of the Lagrange multipliers $\bar{\lambda}^{(k)}$ are indispensable to the progress and convergence of the algorithm.

An interesting result is obtained when the algorithm is applied to a problem for which $n_x = n_f$, as is the case for the device simulation approach to calculating

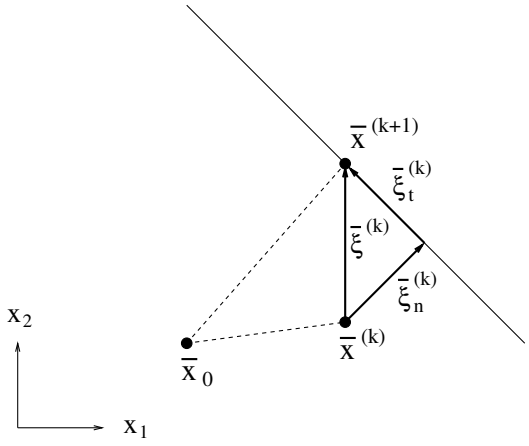

Figure 4.4: The correction step $\bar{\xi}^{(k)}$.

the residuals (see Section 4.1.3). As the Jacobian matrix J_x is then a square matrix, the formula for the correction step (4.12) reduces to

$$\bar{\xi}^{(k)} = -J_x^{-1} f(\bar{x}^{(k)}) \tag{4.14}$$

which is the formula for the correction step of the Newton-Raphson algorithm for determining the root of the model constraints $f(\bar{x})$. Even more interesting from a practical point of view is that the same result (4.14) can be obtained by setting those elements of V^{-1} in equation (4.12) to zero that correspond to the independent interface variables. Hence, the method proposed here includes the device simulation method as a special case.

4.1.6 Global convergence

To force global convergence on the minimization process some addition must be made to the basic algorithm of (4.12). An obvious approach is to impose a limit on the size of the correction step. This means that the correction step $\bar{\xi}^{(k)}$ as calculated in (4.12) is used as the step direction in the observation space

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} + \alpha\, \bar{\xi}^{(k)}$$

where α is a positive multiplier determining the step length. By controlling the value of α at every iteration, global convergence of the method can be achieved. However, it is less obvious how the value of α must be manipulated to ensure this. For instance, it is not possible to look for a sufficiently large reduction in F as a criterion for choosing α, because the amount by which the constraints f are violated must also be taken into account. These two competing aims must therefore be combined in the form of a single "cost" function.

Such a cost function together with a strategy for choosing α has been suggested by Powell [38]:

$$\Psi(\bar{x}, \bar{\mu}) = F(\bar{x}) + \sum_{i=1}^{n_f} \mu_i |f_i(\bar{x})|$$

The positive weights μ_i determine the tradeoff between the two aims. The requirement is that the correction step for every iteration satisfies the following condition

$$\Psi(\bar{x}^{(k+1)}, \bar{\mu}) < \Psi(\bar{x}^{(k)}, \bar{\mu}) \tag{4.15}$$

This condition can be obtained if the function

$$\Phi(\alpha) = \Psi(\bar{x}^{(k)} + \alpha\, \bar{\xi}^{(k)}, \bar{\mu})$$

decreases initially when α is made positive. Detailed analysis shows that this will be the case if the inequality

$$\mu_i \geq \lambda_i^{(k)} \tag{4.16}$$

is satisfied. Moreover, global convergence is assured when $\bar{\mu}$ is chosen large enough.

Some indication of the required scale of $\bar{\mu}$ can be derived from the estimates of the Lagrange multipliers. This is acknowledged by Powell, who refers to a theorem which states that condition (4.16) must hold for all iterations [38]. However, this choice of $\bar{\mu}$ is difficult to ensure, because the sequence $\bar{\lambda}^{(k)}$ $(k = 1, 2, \ldots)$ cannot be determined *a priori*. A large constant vector $\bar{\mu}$ is also inefficient, because on most iterations $\bar{\mu}$ will be much larger than necessary to obtain convergence. If too much weight is given to satisfying the constraints, the path of the iterates $\bar{x}^{(k)}$ is forced to follow the curved model subspace too tightly, which needlessly reduces the step size. Therefore, Powell recommends the following scheme for choosing the value of $\bar{\mu}$. On the first iteration we let $\mu_i = |\lambda_i^{(1)}|$. On all other iterations we apply the formula

$$\mu_i^{(k+1)} = \max\left(|\lambda_i^{(k)}|, \tfrac{1}{2}(\mu_i^{(k)} + |\lambda_i^{(k)}|)\right)$$

The main purpose of the step size reduction is to prevent divergence, hence we may accept any value of α that satisfies condition (4.15). We have developed the following procedure for obtaining a suitable value for α. The full step $\alpha = 1$ is tried first. Only when this value does not suffice, we use a one-dimensional minimization technique to determine an acceptable value for α, which is now bracketed between $0 < \alpha < 1$. Since $\Phi(\alpha)$ is not a differentiable function we have opted for a golden-section search [32]. The minimum of $\Phi(\alpha)$ needs not to be determined very accurately since the extra computations required are seldom justified by a significant further reduction in the number of iterations.

Experiments have shown that this is a very effective and above all an efficient method to prevent divergence of the algorithm. For all constraint functions that have been tried—which includes models for devices such as diodes, bipolar transistors, and field-effect transistors—the algorithm converged satisfactorily, even from remote initial points. However, because the value of μ changes on each iteration, global convergence of the given method cannot be guaranteed. An implementation of this method must therefore include a test to detect divergence (or excessively slow convergence), for example by specifying an upper bound on the number of iterations of the algorithm.

4.1.7 Multiple solutions

It follows from (4.12) and (4.13) that all points $\bar{x}$ that satisfy the primal and adjoint equation of problem (4.8)

$$\begin{aligned} f(\bar{x}) &= \bar{0} \\ V(\bar{x} - \bar{x}_0) &= J_x^t \bar{\lambda} \end{aligned} \tag{4.17}$$

are a fixed point of the algorithm, i.e. lead to $\bar{\xi} = \bar{0}$. Therefore, it seems that the proposed method offers no improvement in this respect when compared to the method of solving these equations directly. Fortunately however, not all roots of (4.17) are points of convergence of the algorithm. Since it is a descent method, the step direction tends to veer away from constrained maxima and saddle points, so that only a constrained minimum is a possible point of convergence of the algorithm.

However, there still exists the possibility that multiple local minima are present. This is often the case when the curvature of the model function is large compared to the distance between $\bar{x}_0$ and the model curve, as may be the case in the initial stages of the identification process. This is illustrated in Figure 4.5, where the points $\bar{x}_a^*$ and $\bar{x}_b^*$ are both constrained minima of $d(\bar{x}_0, \bar{x})$. There is no local test to determine if the found minimum is the global minimum. Hence, no minimization technique can ever guarantee to find the global minimum of a general problem. However, the possibility of multiple values for the same residual does not lead to theoretical problems, as long as a residual is obtained that is finite and well defined. Nevertheless, we will have to take this possibility into consideration when using the calculated residuals in the objective function, as it has some implications for the mathematical properties of the objective function. These will be accounted for in the next section.

4.2 Location and dispersion

The next step in the evaluation of the mode selection criterion (3.5) is the minimization of the least-squares (LS) objective function $\mathcal{C}_2$ (3.2) with respect to the

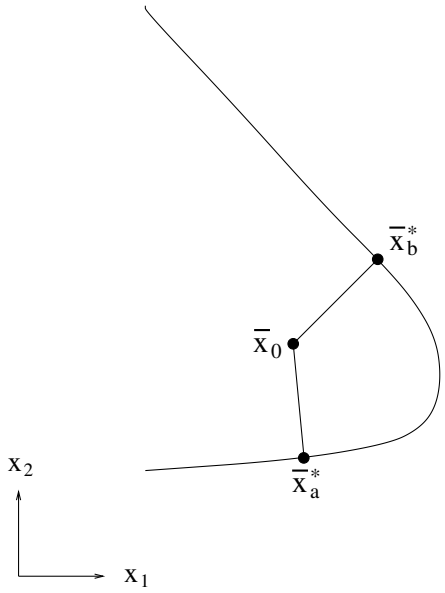

Figure 4.5: Multiple solutions.

parameters $\bar{p}$ for a particular choice of the selection vector $\bar{s}$. This is equivalent to calculating the location $\bar{p}^*(\bar{s})$ of the identification space $\mathcal{P}_\mathcal{S}$ spanned by the selected subset $\mathcal{S}$ of the set of observations $\mathcal{X}$. The calculation of the location then enables us to evaluate the dispersion $\delta_2(\bar{s})$ (3.3), which plays such a crucial role in the MODES algorithm.

The minimization of the objective function $\mathcal{C}_2$ can be formulated as a standard non-linear LS problem. Minimization algorithms that are based on the Gauss-Newton method (which is also referred to as the generalized least-squares method) are generally recognized to be the most appropriate for this type of problem [30]. However, the Gauss-Newton method in its original form does not guarantee convergence, which is clearly not acceptable. In this section, we will derive the Gauss-Newton method from the more general Newton method. This allows us to analyse the algorithm in detail to identify the cause of its instability. Based on the results of this analysis, several modifications to the Gauss-Newton algorithm will be introduced that ensure its stability.

4.2.1 Formulation of the objective function

For a particular choice of $\bar{s}$, i.e. the set $\mathcal{S}$, the objective function is given by

$$\mathcal{C}_2(\bar{p}) = \sum_{i=1}^{N} s_i \epsilon_i(\bar{p})^2 \tag{4.18}$$

This function can be written in a form that is better suited to minimization. First of all, to avoid discontinuous derivatives, the square root in the definition of ϵ (4.3) must be canceled by the square in (4.18)

$$\mathcal{C}_2(\bar{p}) = \sum_{\bar{x}_i \in \mathcal{S}} (\bar{x}_i^* - \bar{x}_i)^t V_i (\bar{x}_i^* - \bar{x}_i) \tag{4.19}$$

where the vectors $\bar{x}_i^*$ depend on the parameters. The shorthand notation $\sum_{\bar{x}_i \in \mathcal{S}}$ stands for summing only over the observations in $\mathcal{X}$ for which $s_i = 1$. This objective function has the same minimum point $\bar{p}^*$ as (4.18). Assuming for the moment that each V_i is a diagonal matrix, the objective function can again be written as the sum of squares

$$\mathcal{C}_2(\bar{p}) = \sum_{\bar{x}_i \in \mathcal{S}} \sum_{j=1}^{n_x} \left(\frac{x_{ij}^* - x_{ij}}{v_{ij}} \right)^2$$

The range of the second sum shows that each observation contributes n_x terms to $\mathcal{C}_2$, while the dimension of the range of the model constraints $f_{\mathcal{M}}$ is only n_f. This means, that the terms of $\mathcal{C}_2$ are dependent, and can be replaced by a smaller number of independent terms. To derive an expression for these independent terms, we reconsider the adjoint equation of (4.17) at the solution of problem (4.2) for the ith observation $\bar{x}_i$

$$V_i(\bar{x}_i^* - \bar{x}_i) = J_x^t \bar{\lambda}_i^*$$

where the Jacobian matrix of the model constraints J_x is evaluated at $\bar{x}_i^*$. Since J_x is assumed to be of full rank, the vector of Lagrange multipliers $\bar{\lambda}_i^*$ is unique. The corresponding term in (4.19) can then be rewritten as

$$(\bar{x}_i^* - \bar{x}_i)^t V_i (\bar{x}_i^* - \bar{x}_i) = \bar{\lambda}_i^{*t} (J_x V_i^{-1} J_x^t) \bar{\lambda}_i^* \tag{4.20}$$

This expression must now be decomposed in a sum of squares. As the square matrix $(J_x V^{-1} J_x^t)$ is positive definite, there exists a similarity transformation

$$(J_x V^{-1} J_x^t) = R D R^t$$

where the columns of the orthogonal matrix R are the normalized eigenvectors of $(J_x V^{-1} J_x^t)$, and the entries of the diagonal matrix D are the corresponding positive real eigenvalues. Equation (4.20) thus becomes

$$\bar{\lambda}_i^{*t} (R D R^t)_i \bar{\lambda}_i^* = (D^{1/2} R^t \bar{\lambda}^*)_i^t (D^{1/2} R^t \bar{\lambda}^*)_i \triangleq \bar{\rho}_i^t \bar{\rho}_i \tag{4.21}$$

substituting the n_f-dimensional vector $\bar{\rho}_i$ for the vector $(D^{1/2} R^t \bar{\lambda}^*)_i$. Hence, we have diagonalized the quadratic form (4.20) under an orthogonal change of variables.

According to (3.5) the number of observations in $\mathcal{S}$ is given by $\mathcal{N}(\bar{s})$. The vectors $\bar{\rho}_i$ can thus be combined in an $\mathcal{N}n_f$-element vector

$$\bar{r} = \begin{bmatrix} \bar{\rho}_1 \\ \vdots \\ \bar{\rho}_{\mathcal{N}} \end{bmatrix}$$

Using this last result in (4.19) yields the final expression for the objective function

$$\mathcal{C}_2(\bar{p}) = \bar{r}^t\bar{r} = \sum_{i=1}^{\mathcal{N}} \bar{\rho}_i^t\bar{\rho}_i = \sum_{i=1}^{\mathcal{N}}\sum_{j=1}^{n_f} {\rho_{ij}}^2 \tag{4.22}$$

where each observation contributes exactly n_f terms to the sum. The minimization of (4.22) is a standard non-linear LS problem.

4.2.2 Newton methods for unconstrained minimization

In order to provide a framework for discussing the characteristics of the Gauss-Newton method, we will digress slightly at this stage and first consider Newton minimization methods in general. In this section, these methods will be derived for a general unconstrained objective function of a set of parameters $F(\bar{p})$.

Consider an iterative minimization process which generates a sequence of points $\{\bar{p}^{(k)}\}$ such that the $(k+1)$st point is related to the kth point by the equation

$$\bar{p}^{(k+1)} = \bar{p}^{(k)} + \Delta\bar{p} \tag{4.23}$$

where $\Delta\bar{p}$ is referred to as the correction step. Assuming that the objective function is at least twice differentiable in the point $\bar{p}^{(k)}$ then, in the neighbourhood of the point $\bar{p}^{(k)}$, the objective function can be approximated by the truncated Taylor-series expansion

$$F(\bar{p}^{(k)} + \Delta\bar{p}) \approx F(\bar{p}^{(k)}) + \Delta\bar{p}^t\bar{g} + \tfrac{1}{2}\Delta\bar{p}^t H\Delta\bar{p} \tag{4.24}$$

where the gradient $\bar{g} = \nabla F(\bar{p}^{(k)})$ and the Hessian matrix $H = \nabla^2 F(\bar{p}^{(k)})$ are evaluated in the current point. If this quadratic function in $\Delta\bar{p}$ possesses a bounded minimum, it will satisfy the following system of equations

$$H\Delta\bar{p} = -\bar{g} \tag{4.25}$$

Equations (4.23) and (4.25) define the standard Newton method.

As the iterates $\bar{p}^{(k)}$ approach the minimum point $\bar{p}^*$ of F (where the gradient $\bar{g} = \bar{0}$), the quadratic approximation (4.24) becomes progressively more accurate. Consequently, from points in the neighbourhood of the minimum, the

Newton step (4.25) should provide a progressively increasing reduction in the error $\|\bar{p}^{(k)} - \bar{p}^*\|$. In fact, it can be shown that the Newton method should converge quadratically in the neighbourhood of a (single) solution. However, as the Newton method is based on a local approximation to F, global convergence of the Newton method is not guaranteed. To ensure global convergence some modifications will have to be made to the basic Newton algorithm.

If the objective function F is continuous and unimodal and the correction step satisfies the relation $F(\bar{p}^{(k+1)}) < F(\bar{p}^{(k)})$ for all iterations, then convergence of the sequence $\{\bar{p}^{(k)}\}$ to the minimum $\bar{p}^*$ is guaranteed. Hence, each correction step $\Delta\bar{p}$ should satisfy the condition of actually bringing about a decrease in the value of F. However, the Newton correction step given by (4.25) may fail this condition. Therefore, instead of using the full correction step, the result of (4.25) will be interpreted as a direction of search, replacing the iteration formula (4.23) by

$$\bar{p}^{(k+1)} = \bar{p}^{(k)} + \alpha\,\Delta\bar{p}$$

where the multiplier α is referred to as the step length. An iteration will now produce a reduction in the value of F if the step length α is sufficiently small and the direction of $\Delta\bar{p}$ is a direction of descent.

A search direction is a descent direction for the objective function F if its inner product with the local gradient $\bar{g} \cdot \Delta\bar{p}$ is negative [35]. For the direction of the Newton step this means that the inequality $\bar{g}^t H^{-1}\bar{g} > 0$ must hold, which can only be guaranteed if the Hessian matrix H is positive definite in $\bar{p}^{(k)}$. Although the Hessian matrix will be positive definite in a neighbourhood of the (strong) minimum $\bar{p}^*$, there is no way to ensure that H will be positive definite for all iterates $\bar{p}^{(k)}$. Hence, in order to obtain a reliable minimization method, the Hessian matrix H in equation (4.25) must be replaced by a related but guaranteed positive definite matrix G

$$G\Delta\bar{p} = -\bar{g}$$

For the convergence rate of the original Newton method to be retained, the matrix G should approach the true Hessian matrix H when $\bar{p}^{(k)}$ approaches the minimum $\bar{p}^*$. Yet, this constraint on the choice of G still leaves room for a multitude of variations on the Newton method, each with a distinct approach to approximating the Hessian of the objective function.

4.2.3 The Gauss-Newton method

The Gauss-Newton method is a Newton method that is dedicated to solving LS problems. This minimization method is designed to exploit the special structure of the Hessian matrix of a LS objective function to give improved computational efficiency and reliability. In this section, we will discuss the most

important aspects of this approach by deriving the iteration equations for our objective function $\mathcal{C}_2(\bar{p})$ (4.22).

To derive an expression for the gradient $\bar{g}$ of the function $\mathcal{C}_2$, we will first consider its elements, for $a = 1, 2, \ldots, n_p$

$$g_a = \frac{\partial \mathcal{C}_2}{\partial p_a} = 2 \sum_{i=1}^{\mathcal{N}} \bar{\rho}_i^t \frac{\partial \bar{\rho}_i}{\partial p_a} \tag{4.26}$$

As is shown in Appendix B, the partial derivatives $\partial \bar{\rho}_i / \partial p_a$ are given by

$$\frac{\partial \bar{\rho}_i}{\partial p_a} = - \left(D^{-1/2} R^t \right)_i \left. \frac{\partial f}{\partial p_a} \right|_{\bar{x}=\bar{x}_i^*} + (S_a)_i \, \bar{\rho}_i \tag{4.27}$$

where S_a is a skew-symmetric matrix comprising the second derivatives of the model constraints with respect to p_a and $\bar{x}$. Substituting this expression in (4.26) gives for the elements of the gradient

$$\frac{\partial \mathcal{C}_2}{\partial p_a} = -2 \sum_{i=1}^{\mathcal{N}} \bar{\rho}_i^t \left(D^{-1/2} R^t \frac{\partial f}{\partial p_a} \right)_i + \bar{\rho}_i^t \, (S_a)_i \, \bar{\rho}_i$$

As the matrix S_a is skew symmetric, all terms of the form $\bar{\rho}^t S_a \bar{\rho}$ are zero. Thus, we have an exact expression for the gradient of $\mathcal{C}_2$ using only the first derivatives of the model constraints with respect to $\bar{x}$ and $\bar{p}$, the Jacobian matrices J_x and J_p respectively

$$\bar{g} = -2 \sum_{i=1}^{\mathcal{N}} \bar{\rho}_i^t \left(D^{-1/2} R^t J_p \right)_i = 2 A^t \bar{r} \tag{4.28}$$

where the $\mathcal{N} n_f \times n_p$ matrix A stands for the component of the Jacobian matrix of $\bar{r}$ with respect to $\bar{p}$ that contributes to the gradient of the objective function.

To evaluate the elements of the Hessian matrix of $\mathcal{C}_2$, we take an additional partial derivative of expression (4.26), for $b = 1, 2, \ldots, n_p$

$$H_{ab} = \frac{\partial^2 \mathcal{C}_2}{\partial p_a \partial p_b} = 2 \sum_{i=1}^{\mathcal{N}} \frac{\partial \bar{\rho}_i{}^t}{\partial p_a} \frac{\partial \bar{\rho}_i}{\partial p_b} + 2 \sum_{i=1}^{\mathcal{N}} \bar{\rho}_i^t \frac{\partial^2 \bar{\rho}_i}{\partial p_a \partial p_b} \tag{4.29}$$

The elements H_{ab} of the Hessian matrix not only depend on the first derivatives but also on the second derivatives of the model constraints: the second sum on the right-hand side of (4.29), and those components of the first sum that contain the matrixes $(S_a)_i$ or $(S_b)_i$. However, all second-derivative terms in (4.29) are multiplied by the residual vector $\bar{\rho}_i$.

The Gauss-Newton method is based on the assumption that all terms of (4.29) that are multiplied by the residuals may be neglected, hence eliminating the

need for evaluating the second derivatives of the model constraints. For a successful model this approximation of the Hessian becomes increasingly more accurate when the solution is approached, and $\|\bar{r}\|$ becomes small. Applying the Gauss-Newton approximation to the Hessian of $\mathcal{C}_2$ gives for the elements of H

$$H_{ab} \approx G_{ab} = 2\sum_{i=1}^{\mathcal{N}} \left[\left(D^{-1/2} R^t \frac{\partial f}{\partial p_a} \right)^t \left(D^{-1/2} R^t \frac{\partial f}{\partial p_b} \right) \right]_i$$

Using the more concise notation introduced in (4.28), the Gauss-Newton approximation G of the Hessian matrix of $\mathcal{C}_2$ can be expressed as

$$G = 2A^t A$$

which shows an additional advantage of this approximation: if the matrix A has full rank, then the matrix G will be positive definite.

The correction step $\Delta\bar{p}$ for improving the approximate solution $\bar{p}^{(k)}$ can now be formulated as the solution of the following set of equations

$$(A^t A)\,\Delta\bar{p} = -A^t \bar{r} \tag{4.30}$$

where A and $\bar{r}$ are evaluated at $\bar{p}^{(k)}$. These equations are generally known as the ***normal equations*** of the LS problem, while the corresponding correction step is referred to as the Gauss-Newton step.

The Gauss-Newton method calculates the optimum correction step on the basis of a linearization of the model constraints with respect to $\bar{x}$ and $\bar{p}$ in $\bar{x}_i^*$ $(i = 1, \ldots, n_x)$ and $\bar{p}^{(k)}$. From this we may conclude that when only first derivatives of the model constraints are available, the Gauss-Newton method is definitely the best minimization method available. Moreover, when the model constraints are in fact linear in $\bar{x}$ and $\bar{p}$, the Gauss-Newton method will minimize the objective function in a single step. In any other case, the convergence rate of the Gauss-Newton method depends on the final accuracy of the Hessian approximation G. Usually, the approximation is accurate enough for this method to exhibit super-linear convergence in the neighbourhood of the solution.

4.2.4 Ill-conditioned problems

Solving the normal equations (4.30) directly does not provide a reliable method for obtaining the Gauss-Newton step. If the matrix $(A^t A)$ is singular or nearly singular, the computational problem of solving $\Delta\bar{p}$ from (4.30) becomes ill-conditioned or even impossible to solve. To quantify the condition of the computational problem we introduce the singular-value decomposition of A

$$A = QDB^t$$

The columns $\bar{q}_i$ and $\bar{b}_i$ ($i = 1, \ldots, n_p$) of the orthogonal matrixes Q and P are the left and right singular vectors of A, and the diagonal elements d_i of the diagonal matrix D are its singular values. As A represents the sensitivity of the objective function with respect to the parameters, the decomposition allows us to determine directions in the parameter space that do not contribute much to reducing the objective function. These "degenerate" directions are the directions of the right singular vectors that are associated with singular values that are relatively close to zero.

The Gauss-Newton step can be written as a vector sum of the right singular vectors

$$\Delta\bar{p} = -BD^{-1}Q^t\bar{r} = -\sum_{i=1}^{n_p} \frac{\bar{q}_i \cdot \bar{r}}{d_i}\bar{b}_i \tag{4.31}$$

which shows that the size of each contribution is inversely proportional to the singular value associated with its direction. When the matrix A is ill-conditioned, the resulting Gauss-Newton step will be dominated by large components in the degenerate directions. As the rounding errors that arise during the calculation of $\Delta\bar{p}$ will especially affect the precision of these components, the Gauss-Newton step tends to be ill-determined when A is ill-conditioned. As a result, the direction of the calculated Gauss-Newton step may not be a descent direction, and convergence of the minimization process is no longer guaranteed. Moreover, (4.31) also shows that the Gauss-Newton step is not even defined when the matrix A is exactly singular, as then one or more singular values d_i of A will be zero.

As ill-conditioned problems are the rule rather than an exception, we must devise a method for calculating the correction step that handles ill-conditioning explicitly. This can be achieved by basing our method on equation (4.31). Since by computing the singular-value decomposition of A we can actually identify the degenerate directions, we can directly eliminate these directions from the search direction. This yields

$$\Delta\bar{p} = -BD_r^{-1}Q^t\bar{r} = -\sum_{i=1}^{n_p^r} \frac{\bar{q}_i \cdot \bar{r}}{d_i}\bar{b}_i \tag{4.32}$$

where n_p^r ($\leq n_p$) is the number of non-degenerate directions. The matrix D_r^{-1} denotes the reduced inverse of D, which is obtained by zeroing the $(n_p - n_p^r)$ diagonal elements of D^{-1} that are the reciprocals of small valued elements in D. If enough elements of D^{-1} are zeroed in this way, the direction of the vector $\Delta\bar{p}$ is guaranteed to be a well-defined descent direction. We must, however, exercise some discretion in deciding at what threshold to zero the inverse singular values.

An objective measure for the condition of the problem is given by the condition number c of the matrix A, which is defined as the quotient of its largest and

smallest singular values

$$c = \frac{d_{\max}}{d_{\min}}$$

The condition number will be large when the matrix is ill-conditioned, or even infinite when the matrix is singular. Zeroing elements of D^{-1} means adjusting the effective condition number of the matrix A to an acceptable limit.

The inner product of the search direction with the true gradient of the objective function must always be negative. However, when the matrix A is ill-conditioned and the search direction is almost orthogonal to the calculated gradient, the inner product of the search direction and the true gradient may actually be positive. Further analysis has shown that a useful upper bound for the condition number of A is then given by the reciprocal value of the relative accuracy with which the elements of A and $\bar{r}$ are calculated (usually the square root of the machine's floating point precision). If, for example, their values are known to about six significant figures, then the maximum condition number that can be tolerated is in the order of 10^6. Hence, the reciprocals of all singular values d_i for which the inequality $d_i < 10^{-6}\, d_{\max}$ holds should then be zeroed.

Elimination of the degenerate directions from the direction of search effectively reduces the dimension of the parameter space. The correction step given by the iteration equation (4.32) can be interpreted as the projection of the Gauss-Newton step onto this reduced space. A minimization method that uses (4.32) will accordingly be referred to as a *reduced Gauss-Newton* (RGN) method.

4.2.5 Scaling of the parameters

For most device models the units of the parameters differ widely, often many orders of magnitude. It may therefore be necessary to scale the parameters, transforming them from their original representation—which may reflect the physical nature of the problem—to parameters that have certain desirable properties in terms of the identification process.

The original Gauss-Newton method can be shown to be invariant under linear transformations of the parameters [35]. However, this theoretical result does not hold for practical implementations, such as the RGN method. The RGN method modifies the Gauss-Newton direction on the basis of the condition number of the matrix A. As A represents the sensitivity of the residuals to a *unit change* in the parameter values, scaling of the parameters will influence the condition number of A by implicitly redefining these units. As a consequence, an unbalanced scaling of the parameters may cause some of them to be needlessly eliminated from the identification process. Hence, the reliability of the

RGN method (or any other minimization method for that matter) depends on the implementation of an adequate scaling technique.

As it is the purpose of identification to reduce the initial uncertainty in the values of the model parameters, it seems natural to express the values of the parameters using this initial uncertainty as the reference. A straightforward implementation of this idea requires the specification of this initial uncertainty in the form of a lower and an upper bound on the value of each parameter

$$p_i^l \leq p_i \leq p_i^u, \quad i = 1, \ldots, n_p$$

These bounds on the parameters should always enclose the identification space $\mathcal{P}_\mathcal{X}$ which represents the lower bound on the uncertainty in the parameter space based on the observational data. The scaled parameters p_i' are given by

$$p_i' = \frac{p_i}{|p_i^u - p_i^l|} \tag{4.33}$$

where a fixed scaling factor maps the range of each parameter to a range of length one.

An alternative scaling technique is obtained when the current value of the parameter $p_i^{(k)}$ is used as the reference. Here, the aim is to scale the parameters in such a way that at the beginning of every iteration the scaled parameters all have a value of one, which means that the correction step can be interpreted as a relative change in the parameter values. This dynamic approach to scaling is best implemented as

$$p_i' = \begin{cases} p_i/p_i^{(k)} & \text{if } p_i^l < p_i^{(k)} < p_i^u \\ p_i/p_i^l & \text{if } p_i^{(k)} \leq p_i^l \\ p_i/p_i^u & \text{if } p_i^{(k)} \geq p_i^u \end{cases} \tag{4.34}$$

where the bounds on the parameters are now applied to the scaling factors.

Obviously, when the range of a parameter includes zero, the dynamic scaling of (4.34) cannot be used. For the rest, the choice between these two scaling techniques depends on the range of the parameter. Static scaling (4.33) is particularly effective when the range of a parameter is relatively small. When the value of a parameter is allowed to change several orders of magnitude in the cause of the minimization process, a fixed scaling factor may not be adequate, and the dynamic scaling of (4.34) should be used instead.

4.2.6 The directional search

Although the direction of the correction step is guaranteed to be a descent direction, we do not know how far the found descent direction extends. Initially,

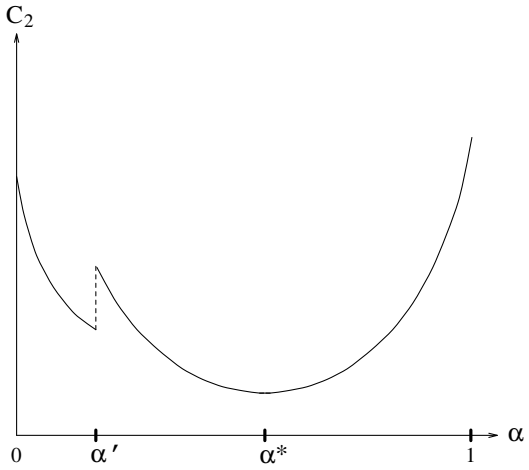

Figure 4.6: A typical discontinuity in the objective function.

it is reasonable to expect that the assumptions that were used to determine the correction step in the first place are valid, i.e. that the current point $\bar{p}^{(k)}$ is in the neighbourhood of the solution $\bar{p}^*$ where the Gauss-Newton approximation is sufficiently accurate. The given correction step (4.32) will then yield a decrease in the value of the objective function. Hence, the step length $\alpha = 1$ should be tried first, and if the objective function is sufficiently reduced, this correction step can be safely accepted. The super-linear rate of convergence of the Gauss-Newton method in the neighbourhood of the solution is thus retained.

If the trial of $\alpha = 1$ is not successful, we must assume that the current point is far from the solution, and a directional search must be carried out to reduce the step length. The aim of this directional search is to follow the search direction $\Delta\bar{p}$ as far as possible in order to provide the maximum amount of decrease in the value of the objective function. One obvious approach is to choose the step length by minimizing

$$\mathcal{C}_2(\bar{p}^{(k)} + \alpha\,\Delta\bar{p}), \quad 0 < \alpha < 1 \tag{4.35}$$

with respect to α. The minimum of (4.35) is guaranteed to exist when the function is continuous. However, when the objective function can be expected to have occasional discontinuities this approach may fail.

Discontinuities in $\mathcal{C}_2$ occur when, for a particular value of the parameters, the calculation of one of the residuals switches between two possible solutions, as was shown in Figure 4.5. The disrupting effect that such a "residual switch" can have on the directional search is illustrated in Figure 4.6, where the discontinuity at α' introduces a local minimum in the function (4.35). Minimization algorithms that only use evaluations of the function value of (4.35) may get stuck

on this type of discontinuity, whereas these discontinuities are not stationary points of the objective function (i.e. $\nabla \mathcal{C}_2(\bar{p}^{(k)} + \alpha' \Delta \bar{p}) \neq \bar{0}$). Consequently, even though the discontinuities in $\mathcal{C}_2$ are usually small in size and number, they can significantly affect the reliability of the identification process.

If the function $\mathcal{C}_2$ is not continuous, a correction step that temporarily increases the function value may be required to bring us nearer to the solution. Hence, in the presence of discontinuities, the function value alone does not provide a suitable criterion. In order to distinguish between a discontinuity of the type shown in Figure 4.6 at α', and the intended minimum at α^*, we must evaluate the derivative of (4.35) with respect to α

$$\frac{\mathrm{d}\mathcal{C}_2}{\mathrm{d}\alpha} = \nabla \mathcal{C}_2(\bar{p}^{(k)} + \alpha \Delta\, \bar{p}) \cdot \Delta \bar{p} \tag{4.36}$$

Only at the stationary point α^* does the sign of (4.36) change. Therefore, instead of minimizing (4.35) directly, the directional search should determine the root of (4.36).

To determine the initial bracket on the root of (4.36) we will, for the moment, assume that $\mathcal{C}_2$ has only a single stationary point at $\bar{p}^*$, so equation (4.36) has at the most one root in the interval $0 < \alpha \leq 1$. As $\Delta \bar{p}$ is a descent direction, the lower bound on the root is always $\alpha = 0$, but the suitability of $\alpha = 1$ as the upper bound depends on the sign of (4.36). If the sign is positive for $\alpha = 1$ we conclude that the interval contains the root, and the directional search can commence. However, if the sign is negative we conclude that the interval does not contain a root. In this case, the step length should not be reduced, accepting a temporary increase in the value of the objective function in order to skip the discontinuity.

The directional search does not need to be very exact to get good convergence properties for the overall method. However, the root of (4.36) should be isolated using a method that keeps the root strictly bracketed, such as the bisection method or the more sophisticated Van Wijngaarden-Dekker-Brent method [32]. The latter method combines the reliability of the bisection method with the efficiency of a second-order method when appropriate. After convergence of the directional search, the value of the smallest bracket is chosen for α. This value of α (which must obviously be larger than zero) ensures a decrease in the value of $\mathcal{C}_2$ if the objective function is continuous after all.

4.2.7 Bracketing the location

The reliability of the directional search rests on the assumption that the objective function has only a single stationary point. However, it is not uncommon for $\mathcal{C}_2$ to have multiple stationary points besides the one that we are searching

for. As a result, the iterations may get trapped in the "collecting region" of another strong minimum of $\mathcal{C}_2$. Hence, we should, whenever possible, confine the directional search to the neighbourhood of the required minimum of $\mathcal{C}_2$, that is, the location of the identification space $\mathcal{P}_\mathcal{S}$.

To determine a more stringent upper bound on the smallest root of (4.36), we require some indication of the scale of the minimization problem. The bounds on the parameters were introduced exactly for this purpose. These bounds bracket the identification space $\mathcal{P}_\mathcal{X}$, a region in the parameter space which contains the required minimum. Moreover, in most cases, any additional strong minima of $\mathcal{C}_2$ lie well outside $\mathcal{P}_\mathcal{X}$.

The directional search algorithm is easily modified to use the available scaling information. If the correction step $\Delta\bar{p}$ crosses any bounds, then the nearest bound determines an upper bound on the step length $\beta \leq 1$. The bounded correction step $\beta\,\Delta\bar{p}$ then replaces the initial correction step $\Delta\bar{p}$ in the directional search. However, we will have to make an exception for bounds that are exactly satisfied for the current point to ensure that $\beta > 0$. This means that when the previous iteration landed the current point on a bound, this bound should be crossed if the search direction, evaluated on this bound, points outside the bracketed region.

4.2.8 The convergence criterion

Any iteration method requires a criterion to determine when the sequence of intermediate solutions has converged. Of course, the minimum is characterized by the gradient $\bar{g}(\bar{p}^*)$ being zero, but as this point will never be reached exactly, we need to formulate a less strict criterion. Hence, the sequence should be considered to have converged when the norm of the gradient is less than some specific value. It is, however, difficult to specify an objective upper bound for the norm of the gradient. The same objection holds for a convergence criterion based on the correction step $\Delta\bar{p}$. Fortunately, we can obtain an objective reference for the change in the objective function $\Delta\mathcal{C}_2$.

In the neighbourhood of the solution the quadratic approximation of the objective function (4.24) will be accurate. For the predicted reduction in $\mathcal{C}_2$ we may then write

$$\Delta\mathcal{C}_2 = \mathcal{C}_2(\bar{p}) - \mathcal{C}_2(\bar{p} + \Delta\bar{p}) = -\Delta\bar{p}^t\bar{g} - \tfrac{1}{2}\Delta\bar{p}^t H\Delta\bar{p}$$

For $\mathcal{C}_2(\bar{p}) = \bar{r}^t\bar{r}$ and using the Gauss-Newton approximation, this expression reduces to

$$\Delta\mathcal{C}_2 = \bar{r}^t A(A^tA)^{-1}A^t\bar{r} \overset{\text{RGN}}{=} \bar{r}^t Q I_r Q^t\bar{r} \tag{4.37}$$

where I_r is the reduced identity matrix $(D_r D_r^{-1})$. When the gradient $\bar{g} = 2A^t\bar{r}$ approaches zero, $\Delta\mathcal{C}_2$ will approach zero as well. Hence, any bound on $\Delta\mathcal{C}_2$ also

specifies a bound on the gradient $\bar{g}$. We can, however, determine a reasonable convergence criterion for $\Delta\mathcal{C}_2$.

The accuracy with which the objective function can be calculated is known, and depends on the accuracy of the residuals. It makes no sense to continue the iteration process when the reduction that can be made in the value of the objective function is within the tolerance on that value. Therefore, at each iteration of the algorithm the value of $\Delta\mathcal{C}_2^{(k)}$ is evaluated. The minimization process is considered to have converged when the following condition is met

$$|\Delta\mathcal{C}_2^{(k)}| \leq tol_R \; \mathcal{C}_2^{(k)} + tol_A \; \mathcal{N} \tag{4.38}$$

where tol_R and tol_A are the relative and absolute tolerances on the squares of the residuals ϵ^2. Furthermore, this convergence criterion implicitly defines a convergence criterion for the correction step, since it specifies when the proposed change in the values of the parameters is no longer significant.

4.3 Mode selection

We are now ready to consider the problem of partitioning the set of observations $\mathcal{X}$ to find the subset that satisfies the mode selection criterion (3.5). The MODES method sequentially removes one observation from the current set $\mathcal{S}$. Each step in the selection space is selected with the aim of minimizing the dispersion $\delta_2(\bar{s})$.

The obvious approach to step selection would be to derive the necessary information by actually evaluating δ_2 in the neighbouring points. It was shown in Section 3.4.4 that this results in an algorithm that has a time complexity of only $O(N^2)$. More precisely, the maximum number of evaluations of δ_2 is $\frac{1}{2}(N^2+N)$. However, as each evaluation of δ_2 involves the minimization of the $\mathcal{C}_2$ objective function, the computational effort required for this algorithm can still be considerable. This was confirmed by experiments, which showed that this step selection procedure is not fast enough to be practical in solving large problems. We must therefore devise a step selection procedure that is more efficient, but without seriously compromising the reliability of the MODES algorithm. The answer to this problem lies in the use of derivative information.

4.3.1 Sensitivity analysis

The removal of an observation $\bar{x}_i$ from the current set $\mathcal{S}$ (by setting $s_i = 0$) alters the value of the dispersion to

$$\delta_2(\bar{s})|_{s_i=0} = \sqrt{\frac{\mathcal{C}_2\left(\bar{p}^*\right) - \Delta_i\mathcal{C}_2}{\mathcal{N}(\bar{s}) - 1}} \tag{4.39}$$

where $\Delta_i \mathcal{C}_2$ denotes the reduction in the minimum value of the objective function $\mathcal{C}_2$. The step selection procedure must minimize (4.39) with respect to i. This is equivalent to maximizing $\Delta_i \mathcal{C}_2$ with respect to i.

The removal of an observation affects the minimum value of the objective function in two ways. First, the residual associated with the observation is removed from the sum of squares (4.22). Second, the minimum point $\bar{p}^*$—the location of the identification space—will shift to a different location, changing the residuals of all remaining observations. Now instead of actually calculating this change in $\bar{p}^*$ by minimizing $\mathcal{C}_2$, we will derive an approximation for its contribution to $\Delta_i \mathcal{C}_2$. This approximation will be based on the linearization of the model constraints in the current location $\bar{p}^*$, as this information is available after the minimization of $\mathcal{C}_2$ for the current set of observations.

According to Section 4.2.3, the Gauss-Newton method will minimize the objective function in a single step when the model constraints are linear. Hence, the change in location can be approximated by solving the normal equations (4.30) after eliminating the elements of the residual vector $\bar{r}$ and the derivative matrix A that are associated with the observation $\bar{x}_i$. The current $\bar{r}$ and A are thus divided in two parts:

$$\bar{r} = \begin{bmatrix} \bar{r}_0 \\ \bar{\rho}_i \end{bmatrix}, \qquad A = \begin{bmatrix} A_0 \\ A_i \end{bmatrix}$$

where the elements belonging to the observation $\bar{x}_i$ are placed in $\bar{\rho}_i$ and A_i. The change in location is then given by

$$\Delta \bar{p}^* = -\left(A_0^t A_0\right)^{-1} A_0^t \bar{r}_0$$

Here it is assumed that the matrix A_0 is well-conditioned. At a later stage the results will be modified to handle ill-conditioned problems reliably. The effect of the Gauss-Newton correction step on the value of the objective function was already derived in Section 4.2.8. By substituting $\bar{r}_0$ and A_0 in equation (4.37), the reduction in the minimum value of the objective function can be expressed as

$$\Delta_i \mathcal{C}_2 = \bar{\rho}_i^t \bar{\rho}_i + \bar{r}_0^t A_0 (A_0^t A_0)^{-1} A_0^t \bar{r}_0 \tag{4.40}$$

The first term represents the direct effect of removing the ith residual, while the second term accounts for the effect of the change in location $\Delta \bar{p}^*$ on the remaining residuals.

The evaluation of (4.40) requires the calculation of the inverse of the $n_p \times n_p$ matrix $(A_0^t A_0)$ for each of the $\mathcal{N}$ observations in $\mathcal{S}$. Especially when the identification problem is ill-conditioned, this can be a substantial computational

burden. Therefore, we will seek to express (4.40) in a form that can be evaluated more efficiently. For this purpose, we apply the equalities

$$\begin{aligned} A^t A &= A_0^t A_0 + A_i^t A_i \\ A^t \bar{r} &= A_0^t \bar{r}_0 + A_i^t \bar{\rho}_i = |_{\bar{p}=\bar{p}^*} \bar{0} \end{aligned} \tag{4.41}$$

and rewrite (4.40) as

$$\Delta_i \mathcal{C}_2 = \bar{\rho}_i^t \left[I + A_i (A^t A - A_i^t A_i)^{-1} A_i^t \right] \bar{\rho}_i \tag{4.42}$$

To obtain the inverse of the modified matrix $(A^t A - A_i^t A_i)$ we use the Sherman-Morrison-Woodburry formula [1], which gives the following relationship

$$\begin{aligned} &(A^t A - A_i^t A_i)^{-1} = \\ &\quad (A^t A)^{-1} + (A^t A)^{-1} A_i^t \left[I - A_i (A^t A)^{-1} A_i^t \right]^{-1} A_i (A^t A)^{-1} \end{aligned} \tag{4.43}$$

Since the inverse of the matrix $(A^t A)$ will already have been calculated, the evaluation of (4.43) only requires the inversion of the $n_f \times n_f$ matrix between the square brackets. The substitution of (4.43) in (4.42), and a reordering of the terms, yields

$$\Delta_i \mathcal{C}_2 = \bar{\rho}_i^t \left[I - A_i (A^t A)^{-1} A_i^t \right]^{-1} \bar{\rho}_i \tag{4.44}$$

This equation[1], which uses a Gauss-Newton correction step to account for the change in location, can easily be modified to use a RGN correction step. The inverse of the matrix $(A^t A)$ must then be determined in the reduced parameter space

$$(A^t A)^{-1} = B D_r^{-1} D_r^{-1} B^t$$

Further, the matrix Q of left eigenvectors of A is divided into two parts along the same lines as A itself

$$\begin{bmatrix} A_0 \\ A_i \end{bmatrix} = \begin{bmatrix} Q_0 \\ Q_i \end{bmatrix} D B^t \quad \Longrightarrow \quad A_i = Q_i D B^t$$

Substitution in (4.44) finally yields the desired result

$$\Delta_i \mathcal{C}_2 = \bar{\rho}_i^t \left[I - Q_i I_r Q_i^t \right]^{-1} \bar{\rho}_i \tag{4.45}$$

This approach to the problem of ill-conditioning is based on the assumption that the partial matrix A_0 has approximately the same degenerate directions as the full matrix A, which is usually the case. However, if this assumption is not valid—when the elimination of an observation from $\mathcal{S}$ adds a degenerate direction to A—then the matrix $\left[I - Q_i I_r Q_i^t \right]$ can still be ill-conditioned, and the evaluation of (4.45) may fail. When this happens, we must revert to equation (4.40), and calculate the singular-value decomposition of the matrix A_0.

[1] The reader may notice the correspondence between equation (4.44) and the iteration equations of the so-called recursive least-squares method for sequential observations, a method for estimating the properties of a dynamic system in real time [39].

4.3.2 The algorithm

The flow diagram in Figure 4.7 shows how the sensitivity analysis is incorporated in the MODES algorithm. After each minimization of $\mathcal{C}_2$ with respect to $\bar{p}$, the solution is tested for compliance with the mode selection criterion (3.5). If the solution fails this test, a sensitivity analysis is performed to select the next point in the selection space. The sensitivity analysis *predicts* the reduction in the minimum value of the objective function for each step in the selection space (of length one). The optimum step is then *selected* by maximizing the predicted reduction over the current set of observations $\mathcal{S}$. However, as the prediction was based on the linearization of the model constraints, it will usually have to be *corrected* by minimizing $\mathcal{C}_2$ before the step selection procedure can be repeated—the sensitivity analysis requires the gradient of the objective function with respect to $\bar{p}$ to be zero, as is expressed in equation (4.41). The resulting prediction–selection–correction loop can be interpreted as the method of steepest descent [35] in the selection space, with the exception that the step length is always one.

When the step direction is selected on the basis of the sensitivity analysis, then for each step we save $\mathcal{N}$ evaluations of δ_2, but at the extra cost of calculating the set of $\mathcal{N}$ sensitivities $\{\Delta_i\mathcal{C}_2\}$. However, the calculation of a sensitivity $\Delta_i\mathcal{C}_2$ requires far less computational effort than the minimization of $\mathcal{C}_2$ for a different $\bar{s}$. The result is a highly efficient identification method, which at the most requires N full minimizations of the LS objective function $\mathcal{C}_2$.

As the algorithm sequentially reduces the extent of the identification space $\mathcal{P}_\mathcal{S}$, the changes in location $\Delta\bar{p}^*$ will also become smaller when the algorithm progresses. Therefore, when the minimum point of the previous minimization is used as the initial estimate of the solution of the next minimization, the subsequent minimizations will become increasingly easier to solve, requiring less and less iterations of the RGN algorithm (see the final paragraph of Section 4.2.3). In practice, we often find that the lion's share of the computing time is spent in the first minimization (for $\bar{s} = \bar{1}$).

The size of $\Delta\bar{p}^*$ also determines the reliability of the step selection procedure. The expected reductions in the minimum value of the objective function $\Delta_i\mathcal{C}_2$ were derived for linearized model constraints; an approximation which is only accurate for small parameter deviations around the current location $\bar{p}^*$. Therefore, the changes in location per selection step must be kept small. This condition is usually fulfilled when $\mathcal{N}$ is large, as then the effect of the removal of a single observation will be small. Alternatively, when $\mathcal{N}$ is small, for example in final stages of the selection process, but the extent of the identification space $\mathcal{P}_\mathcal{S}$ has also become small, then any further changes in location will again be small. Hence, the step selection procedure will be reliable when the total number of observations N is large, and a considerable section of the set of ob-

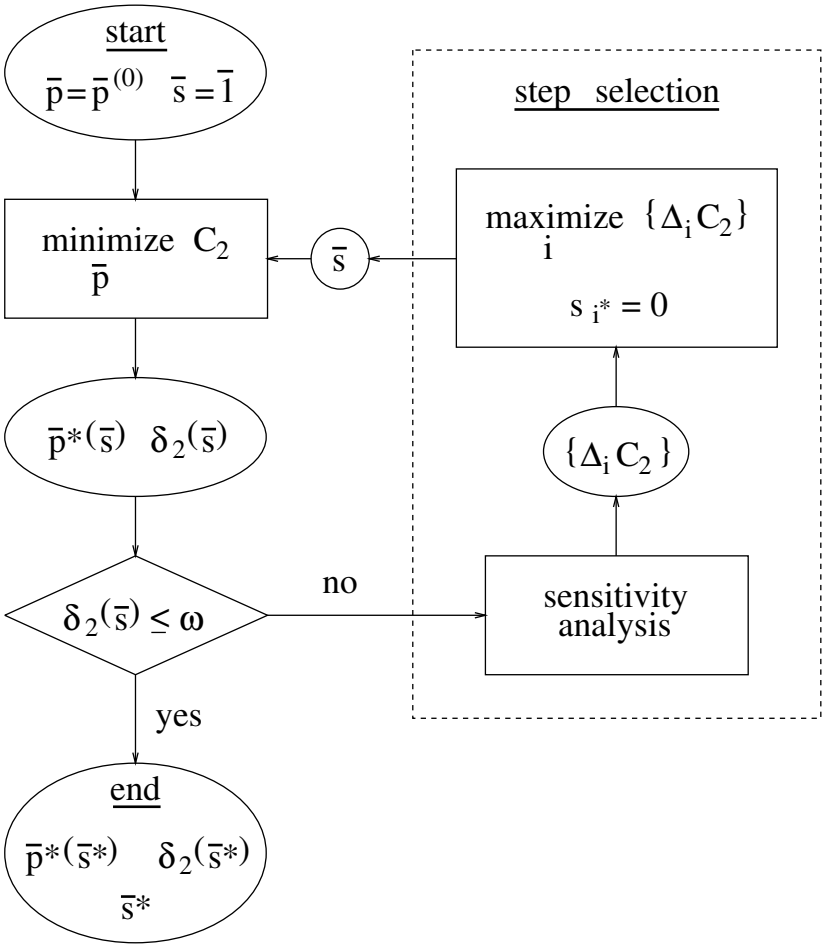

Figure 4.7: Flow diagram of the MODES algorithm.

servations lies within the validity domain of the model. Both conditions are also essential for the mode to be clearly defined.

Chapter 5

Demonstration

In order to demonstrate the effectiveness of the mode selection method for the identification of analytical device models we will present a small but representative example: the identification of the Ebers-Moll model of a bipolar transistor. For this example the results obtained by MODES will be compared with those obtained by a conventional least-squares (LS) method. The evaluation of these results can only be objective if the true values of the parameters are known *a priori*. If the observations are obtained from a real device, this would require an identification method that is more accurate than the methods under consideration. No such method is available—the sequential method, which was discussed in the introduction, can at best match the accuracy of MODES. Alternatively, the observations can be obtained by simulating a device, but using a device model for this simulation that is more accurate and has a wider validity domain than the device model that is to be identified. In this case, the true parameters are those used in the simulation, provided that these parameters play the same role in both models, i.e. have the same physical interpretation. For bipolar transistors this last condition is easily satisfied, as we have at our disposal a whole series of models of gradually increasing sophistication [2, 40]. Moreover, because the behaviour of bipolar transistors is strongly non-linear, their models are a demanding test for the robustness of our implementation of MODES.

5.1 The model

The simplest model for the DC (direct current) behaviour of a bipolar transistor that is used in circuit design is the Ebers-Moll (EM) model [2]. For the normal region of operation of the transistor the EM model is given by the following

system of non-linear equations

$$\begin{aligned} I_c &= I_S \exp(V_{be}/V_T) \\ I_b &= I_c/\beta_F \end{aligned} \tag{5.1}$$

where V_{be} is the base-emitter voltage, I_c and I_b are the collector and base currents respectively, and I_S, V_T and β_F are the model parameters.

The extent of the validity domain of the EM model is limited by several simplifying assumptions. The main assumption is that the forward current gain I_c/I_b is constant over the whole normal region, and equal to β_F. Although this assumption often holds over several current decades, especially for modern integrated transistors, the EM model will always fail for both low and high terminal currents. It is further assumed that the base-collector voltage V_{bc} does not affect the behaviour of the transistor as long as the base-collector junction remains reverse biased, which means that the base-width modulation effect [2] is ignored. Nevertheless, the model given by (5.1) is often sufficiently accurate over a large enough section of the normal operating region of a bipolar transistor for applications such as the design of linear amplifiers [22], where it serves as the DC model of the "ideal" transistor.

To some readers it may be surprising to find V_T in the list of model parameters. The thermal voltage V_T is related to the absolute temperature T (in degrees Kelvin) according to the well known formula $V_T = (kT)/q$, where q denotes the unit of electronic charge and k the Boltzmann constant. Hence, the value of V_T could be obtained by measuring the (internal) temperature of the device. However, as the behaviour of the device is very sensitive to changes in temperature, this measurement would have to be extremely accurate. Even a small measuring error in T would cause the EM model to be invalid for all observations. We have therefore chosen to extract V_T from the observed DC behaviour, together with I_S and β_F.

5.2 The data

A transistor model that is far more accurate than the EM model is the well-known Gummel-Poon (GP) model. Its model equations can be found in numerous publications [2, 21, 41], and are implemented in several circuit simulation programs. The GP model corrects many of the omissions of the EM model, which results in a significant extension of the validity domain. Most notably, the GP model improves the representation of the device in the low-current and high-current regions by incorporating the non-ideal components of the base current (adding the GP model parameters C_2 and n_{EL}), and the high-level injection effect (the parameter I_K). As a consequence, the current

gain of the transistor model is no longer constant, but varies with V_{be}. Also included are the transistor's ohmic resistances from its active region to its base and emitter terminals (the parameters r_b and r_e, respectively), which reduce the effective value of V_{be}. These resistances tend to linearize the exponential I_c and I_b versus V_{be} characteristics of the transistor in the high-current region. (As we are modelling the transistor in its normal region of operation, we decided not to include the collector resistance, which only affects the DC behaviour of the transistor in the saturation region.)

The GP model accounts for the base-width modulation effect by redefining the EM model parameters I_S and β_F (which were defined for a fixed base width [2]). The EM parameters and GP parameters of the same name[1] are only equivalent if $V_{bc} = 0$. Hence, in order to regain the specified values of the GP model parameters, the EM model must be identified using observational data that is obtained for $V_{bc} = 0$.

Parameter	Value
I_S	10.00 fA
V_T	25.50 mV
β_F	250.0
C_2	0.02
n_{EL}	1.6
I_K	0.2 A
r_b	50 Ω
r_e	0.1 Ω

Table 5.1: The values of the Gummel-Poon model parameters.

The DC behaviour of an npn transistor has been simulated with the SPICE circuit-simulation program [42], which contains the GP model. Table 5.1 lists the values of the GP model parameters that were used in this simulation. These parameter values are typical for a modern low-power transistor. The value of V_{be} was varied in steps of 0.01 V from 0 V up to 1.0 V, while V_{bc} was kept at 0 V. Figure 5.1 shows the "observed" device behaviour, which is represented in the form of a Gummel plot. We refrain from marking the individual data points because of their large number ($N = 100$) and regular distribution. The simulation accuracy, which plays here the role of the observation error, is approximately 10^{-4} times the simulated value for all interface variables.

[1] In the GP model the EM model parameter I_S is denoted by I_{SS}.

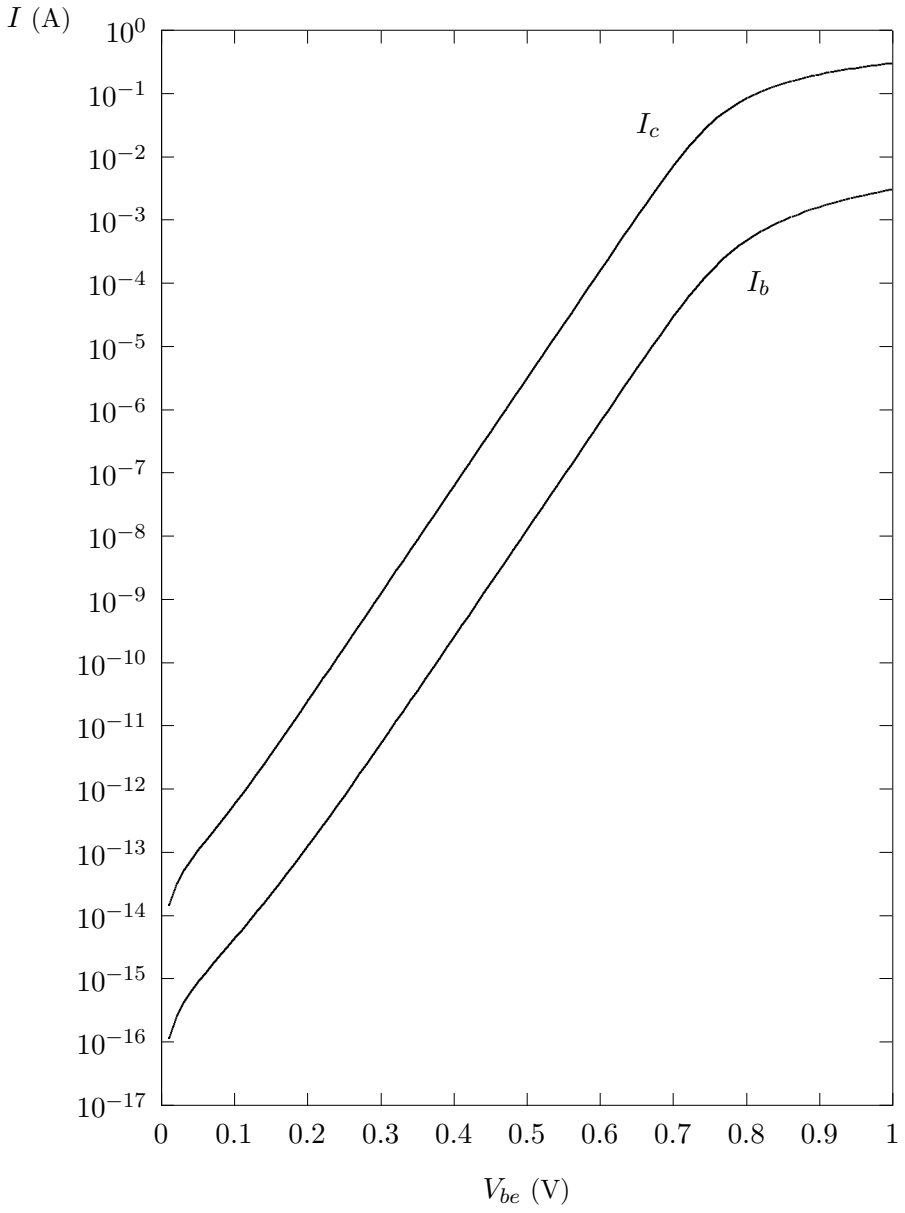

Figure 5.1: Plots of I_c and I_b versus V_{be} on a logarithmic scale (for $V_{bc} = 0$).

	LS	MODES		
		$\omega = 1.0$	$\omega = 0.1$	$\omega = 0.01$
I_S	30.13 fA	10.71 fA	10.27 fA	10.26 fA
V_T	28.43 mV	25.61 mV	25.53 mV	25.53 mV
β_F	191.2	246.9	249.3	249.6
$\mathcal{S}$	0.01–1.00 V	0.32–0.71 V	0.46–0.64 V	0.53–0.61 V
$\mathcal{N}$	100	40	19	9
δ_2	21.7	0.93	0.093	0.0092

Table 5.2: The results of the LS and MODES identification methods.

5.3 Results

The identification of the EM model has been carried out using a conventional LS method, and the new MODES method for several values of ω. The results are summarized in Table 5.2. Convergence of both the LS algorithm and the MODES algorithm was fast for most initial parameter estimates. None of the identified parameter sets contains redundant parameters; in fact, this identification problem is relatively well-conditioned as c does not exceed 500. The set $\mathcal{S}$ of selected observations is specified in the form of a domain on the V_{be} axis of the observation space. The LS parameters as well as the MODES parameters have been calculated using the least-distance residuals that were introduced in Section 4.1. The accuracy metric was set to 1% of the observed value for all interface variables (on a linear scale), so that the dispersion δ_2 is expressed in per cents RMS.

We also experimented with device-simulation residuals, taking V_{be} as the independent variable. On the whole this had an adverse effect on the efficiency and reliability of the identification process. Depending on the initial estimates of the parameters, the RGN algorithm either converged to a somewhat different solution with a significantly reduced convergence speed, or failed to converge to a meaningful solution at all. This demonstrates the superiority of the least-distance residuals for strongly non-linear device models.

Figure 5.2 illustrates the asymptotic character of the EM model that has been identified by MODES (for $\omega = 0.01$). The device behaviour predicted by the identified EM model, represented by the solid lines, is superimposed on the observed device curves, represented by the dotted lines. On the logarithmic scale of the Gummel plot the EM model always produces a pair of parallel straight lines. MODES has correctly identified the segment of the graph where the device curves are straight and parallel too. Within this region the identified EM model coincides with the device curves (i.e. the GP model), while outside this region the EM model and the device curves gradually diverge.

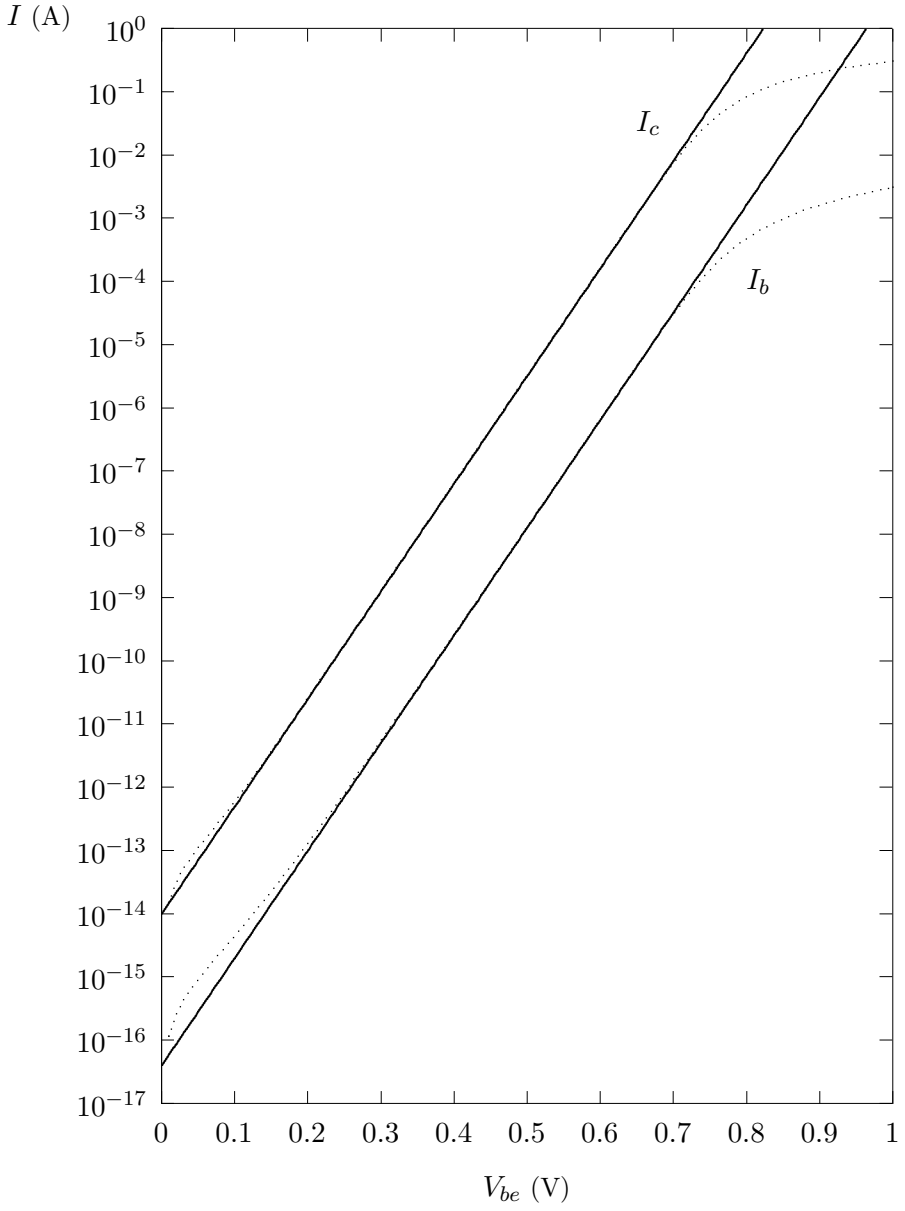

Figure 5.2: The device behaviour predicted by the identified EM model superimposed on the actual device behaviour.

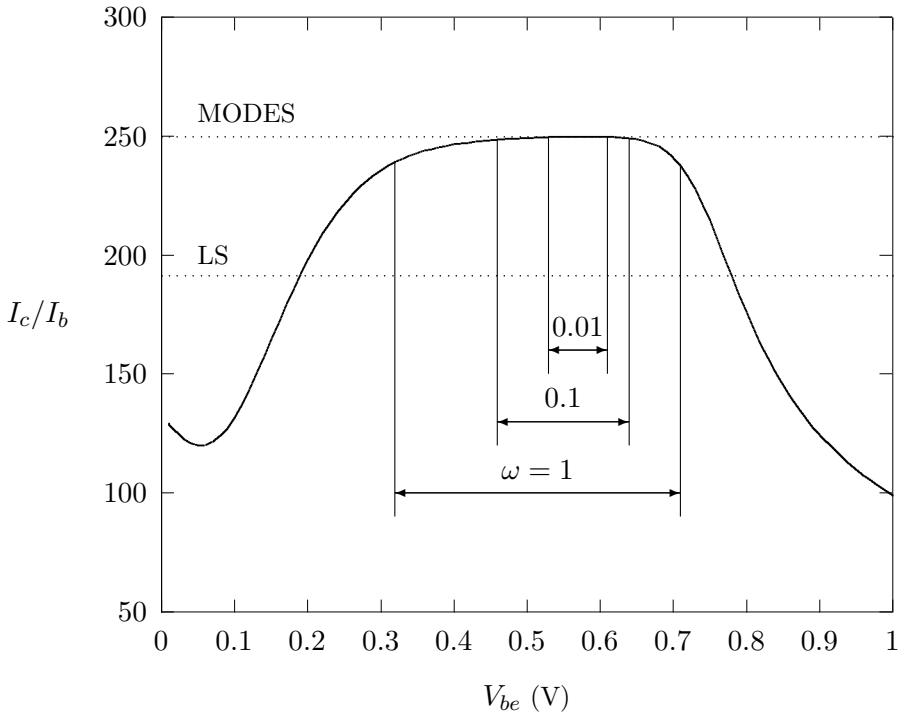

Figure 5.3: The plot of the current gain I_c/I_b versus V_{be}.

When we look at the results of MODES in Table 5.2 we find that the accuracy of the parameter estimates gradually increases for decreasing values of ω. The high accuracy that is finally attained is in stark contrast to the inaccuracy of the LS method. The obvious cause of this discrepancy is the limited extent of the validity domain of the EM model. For instance, the EM model assumes that the current gain is constant over the whole domain space, an assumption that is abandoned by the GP model. The actual current gain is shown in Figure 5.3. As expected, the current gain varies with V_{be}. The dotted lines of constant current gain represent the identified EM model for both the LS method and MODES (for $\omega = 0.01$). The LS method clearly opts for an average value for β_F, spreading the modelling error evenly over all observations. MODES focuses instead on the region where the curve is flat, and the constant current-gain hypothesis of the EM model is valid. The bounds of the selected subsets $\mathcal{S}$ for the different values of ω are indicated in Figure 5.3. For the observations in the final subset ($\omega = 0.01$) the variation in current gain is less than 0.1%.

We may conclude that for the present example MODES yields a far more accurate estimate of the parameter set of an analytical model than the LS method.

Even a sequential identification method, such as the one described in [2], could not improve on the result of MODES. However, as the GP model can never be completely reduced to the EM model, not even in the identified validity domain of the EM model, the accuracy of the parameters identified by MODES is ultimately limited. In the next chapter we will show that this limit is proportional to the final value of the dispersion δ_2, and hence to the value of ω. For the present example a further reduction of ω is not feasible. A more accurate parameter set can only be obtained by identifying a more accurate model, such as the GP model. In this case, the identification of the GP model obviously yielded the correct parameters.

Chapter 6

Discussion and Conclusions

The main objective of the research work presented in this thesis has been the design and implementation of a method for the identification of analytical device models. This method should combine the flexibility of a data-fitting method with the reliability of a sequential method. In this final chapter we will assess the mode selection method (MODES) with respect to this aim.

First and foremost, the identified model must comply with the uniqueness condition that was put forward in Section 2.3.3. It is therefore not sufficient to simply present the computed values of the model parameters. The consistency of these values must be determined as well. Another point still on the agenda is our claim that not only the parameters but also the model validity domain can be extracted from the observed device behaviour. Up to now, we have not made this process explicit. This omission will be corrected in this chapter.

We will also discuss the quality of the implementation of the MODES algorithm that was presented in Chapter 4. Many aspects of the reliability and efficiency of the algorithm were already discussed in that chapter. Here, we will compare the characteristics of our approach with those applied in other data-fitting methods.

6.1 The model parameters

The quality of the identified model is usually discussed in terms of the accuracy of the parameter values, where accuracy is then defined as conformity with true value. However, the true values of the parameters are not known, and as we are dealing with approximate models, nor are they even defined. We therefore revert to the pragmatic approach to modelling that was proposed in Chapter 2, and investigate if there is a one-to-one correspondence between the

model representation in the observation space and the model representation in the parameter space. We will investigate two aspects of this correspondence: the identifiability of the model and the consistency[1] of the identified parameters. The identifiability of the model is related to the problem of redundant parameters, and the uniqueness of the identified model in the parameter space for the given set of observations. The consistency of the identified parameters is related to their dependence on the experimental conditions, i.e. the reproducibility of the results for alternate data sets.

6.1.1 Identifiability

Each observation in the mode set $\mathcal{S}^*$ contributes some information about the possible values of the parameters. Together these observations should be able to discriminate between different members of the set of models $\mathcal{M}$. However, it happens quite often that not enough observational data is available to differentiate between the members of a subset of $\mathcal{M}$, in which case the identified model will not be unique. Instead of a single point in the parameter space we will have identified a subspace of the parameter space. There are two possible causes for this undesirable phenomenon. Firstly, there is always the danger of over-parameterization: using a model that is unnecessarily complex for describing the device behaviour in the domain of interest. For example, the model may describe a physical effect that does not take place in the actual device under the given experimental conditions. Although the associated model parameters will sometimes take on limiting values, more often this will give rise to degenerate directions in the parameter space. However, suitable values of these parameters could be determined by supplying additional observational data, for instance by extending the range of the experiments. Therefore, these parameters are only redundant in a practical sense. If, on the other hand, the model contains parameters that cannot be identified for any amount of observational data, then these parameters are mathematically redundant. Mathematical redundancy should be considered a fault in the model specification. In both cases, the redundant parameters should be separated, and subsequently eliminated from the model specification to obtain what is sometimes called a *parsimonious* model [43].

To identify the redundancies in the parameter set we will explore the objective function $\mathcal{C}_2$ in the neighbourhood of the solution $\bar{p}^*$. For small changes in the parameters around $\bar{p}^*$ the quadratic approximation (4.24) will hold. Assuming that the gradient $\bar{g}$ is zero, or at least very small, the sensitivity of the objective function to small changes in $\bar{p}$ will be governed by the Hessian matrix H (evaluated at $\bar{p}^*$). For the increase in the value of $\mathcal{C}_2$ for a small change in the value of the parameter vector $\delta\bar{p}$ we can write

$$\delta\mathcal{C}_2 = \mathcal{C}_2(\bar{p}^* + \delta\bar{p}) - \mathcal{C}_2(\bar{p}^*) = \tfrac{1}{2}\delta\bar{p}^t H \delta\bar{p}$$

[1] We use the term "consistency" without any probabilistic overtones.

Using the Gauss-Newton approximation of the Hessian matrix we obtain

$$\delta\mathcal{C}_2 = \delta\bar{p}^t(A^tA)\delta\bar{p} = (B^t\delta\bar{p})^t D^2(B^t\delta\bar{p}) = \sum_{i=1}^{n_p} {d_i}^2 \left(\bar{b}_i \cdot \delta\bar{p}\right)^2 \tag{6.1}$$

Now consider the convergence criterion of the minimization algorithm (4.38). It specifies that the true minimum point of $\mathcal{C}_2$, which will be denoted by $\bar{p}^{**}$, is surrounded by a domain in the parameter space of which any member will be accepted by the algorithm as the solution of the minimization process. This domain will be referred to as the *convergence region* of the minimization problem. If we assume that the convergence region is small, and that the Hessian matrix of the objective function does not vary significantly over this region, then the convergence region can be defined by the equation

$$\delta\mathcal{C}_2 \leq tol_C$$

where $tol_C = tol_R\,\mathcal{C}_2(\bar{p}^*) + tol_A\,\mathcal{N}$ is the tolerance on minimum value of $\mathcal{C}_2$ (see Section 4.2.8). This is the equation of an n_p-dimensional ellipsoid centered at $\bar{p}^{**}$. However, because the convergence region provides an upper bound on the difference $\bar{p}^* - \bar{p}^{**}$, its extent is also a measure for the precision of the accepted solution $\bar{p}^*$.

It follows from equation (6.1) that the principal axes of the ellipsoid are parallel to the singular vectors $\bar{b}_i$, and that the lengths of the semi-axes are equal to $\sqrt{tol_C}/d_i$. For a well-conditioned model the singular values d_i will all be of the same order of magnitude. A singular value that is relatively small may therefore imply a redundant parameter. One criterion for the acceptable size of a singular value was already introduced in Section 4.2.4

$$d_i \geq \frac{d_{\max}}{c_{\max}} \tag{6.2}$$

where $c_{\max}$ is the maximum acceptable condition number of the sensitivity matrix A. The directions of the singular vectors $\bar{b}_i$ belonging to the singular values d_i that fail (6.2) are classified as degenerate. Degenerate directions indicate which linear combinations of the parameters have no significant influence on the value of the objective function. As their singular values d_i are, strictly speaking, unknown (no significant digits) and may as well be zero, the occurrence of any degenerate directions in the solution means that the identified parameters are not unique. Instead, the solution $\bar{p}^*$ is an arbitrary point in the subspace of the parameter space that is spanned by the degenerate directions (which are orthogonal). For each degenerate direction one of the parameters involved in these directions must be declared redundant. Only when all redundant parameters are eliminated from the model, will the remaining parameters take on definite values.

The bound on the condition number of the problem provides a relative lower bound on the singular values d_i. It is also possible to specify an absolute lower bound. Note that the sensitivities (6.1) were calculated for the *scaled* parameters (see Section 4.2.5), so the absolute tolerance on the values of the elements of $\bar{p}^*$ must be smaller than unity to be significant, i.e. reduce the initial uncertainty specified by the bounds on the parameters. This means that the lengths of the semi-axes of the convergence ellipsoid should be less than one, which implies

$$d_i \geq \sqrt{tol_C} \tag{6.3}$$

The directions of the singular vectors belonging to singular values that fail this criterion should also be classified as degenerate and treated accordingly. Eliminating these additional parameters will increase the reliability of the identified location in the parameter space, at the cost of only a negligible increase in the value of the dispersion.

Model parameters that are associated with singular values d_i that fail either (6.2) or (6.3) cannot be identified. As such, the redundancy criteria define the *sensitivity threshold* of the identification procedure. This limit depends both on the problem formulation and on the implementation of the identification method. The (true) value of $d_{\max}$, as well as $\mathcal{C}_2(\bar{p}^*)$ and $\mathcal{N}$, are completely determined by the constituents of the identification problem: the model constraints, the set of observations $\mathcal{X}$, the local accuracy metric, and the scaling of the parameters. Whereas the maximum condition number $c_{\max}$ and the convergence criterion tol_C, through its components tol_R and tol_A, are implementation dependent; both account for the limited accuracy with which the elements of the sensitivity matrix A and the residual vector $\bar{r}$ can be calculated. As a consequence, the sensitivity threshold can serve as an objective measure of the quality of the implementation.

6.1.2 Consistency

It should not be forgotten that the identified model is, in the first place, a description of the available set of observations $\mathcal{X}$. The model parameters will thus depend on the experimental conditions such as the number of observations, the range of the observations, and the distribution of the observations over that range. Therefore, one must determine the ability of the identified model to predict the behaviour of the device under experiments different from those under which the model was identified. Vice versa, one should try alternate data sets for the identification of the model and check the consistency of its parameters.

Obviously, it is neither possible nor necessary to try all conceivable data sets. As we are dealing with a real device, the choice in experimental conditions will be

severely limited by practical constraints. It therefore suffices to consider alternate data sets that are similar to the given set $\mathcal{X}$. With this restriction, MODES may be expected to select out of each alternate data set a mode set that is, with respect to its dispersion and range, comparable to the identified mode set $\mathcal{S}^*$. Only the number of observations in the selected subset $\mathcal{N}$ and their distribution over the selected range will differ. The effect of variations in the experimental conditions on the identified parameters can therefore be simulated by determining the location of—systematically or randomly chosen—subsets of $\mathcal{S}^*$. As the location of each subset of $\mathcal{S}^*$ will again be a point in the identification space $\mathcal{P}_{\mathcal{S}^*}$, the result of this simulation will give an impression of the extent of the identification space.

Having established a link between the consistency of the identified parameters and the extent of identification space $\mathcal{P}_{\mathcal{S}^*}$, we proceed by formulating a practical measure of this extent. The simulation approach proposed in the preceding paragraph is not particularly practical because it requires an exhaustive search of a substantial section of the selection space, a method we rejected earlier because of its excessive use of computing time. Moreover, the maximum spread of the observational constraints in the parameter space is an overly pessimistic estimate of the consistency of the parameters, as it characterizes the ensemble by its most deviant members. Although such data sets are theoretically possible, they are not very probable, especially for larger $\mathcal{N}$.

A measure of the extent of the identification space that is readily available is the dispersion δ_2 of mode set $\mathcal{S}^*$. However, the dispersion δ_2 has been formulated so that its value is independent of the scaling of the parameters. Hence, to represent this measure in the parameter space it has to be complemented with scaling information. One possible source of scaling information that comes to mind is the matrix $(A^t A)$, which expresses the sensitivity of the objective function to changes in the parameters, though it is not immediately obvious how this scaling should be applied to δ_2.

Suppose that the dispersion is zero, so all observations in $\mathcal{S}^*$ agree perfectly on a single and unique parameter set. The consistency of the location $\bar{p}^*$ is then guaranteed. Indeed, this parameter set can, for all intents and purposes, be regarded as the true parameter set. If this true parameter set is modified by the addition of a small error vector $\delta\bar{p}$, equation (6.1) can be used to predict the ensuing value of the objective function, and thus of δ_2. Hence, with respect to the value of the dispersion, the identified model with $\delta_2 \neq 0$ is equivalent to a hypothetical true model with $\delta_2 = 0$ that has the accuracy of its parameters limited by the constraint

$$\sqrt{\frac{1}{\mathcal{N}} \sum_{i=1}^{n_p^*} {d_i}^2 (\bar{b}_i \cdot \delta\bar{p})^2} = \delta_2 \tag{6.4}$$

where n_p^* represents the number of parameters that are not redundant in the final solution of MODES. This equation provides a link between the inaccuracy of the model in the observation space and the tolerance on the identified parameter set. This is what we meant when we stated in Section 3.4.2 that the dispersion δ_2 is a measure of the accuracy of the mean as an estimate of the mode.

The ellipsoid in the parameter space described by equation (6.4) provides an upper bound (albeit a crude one) on the spread of the bulk of the observational data in $\mathcal{S}^*$. More specifically, the contour of all $\delta\bar{p}$ that satisfy equation (6.4) can be interpreted as the *standard deviation* of the observational data in the parameter space. We admit that we are stretching the term, nevertheless this interpretation of the dispersion measure δ_2 is consistent with our earlier interpretation of the minimum point of the least-squares criterion as the mean.

To find the tolerances on the individual parameters, the ellipsoid (6.4) must be projected on the co-ordinate axes of the parameter space [32]. We thus obtain for the values of the parameters

$$p_k = p_k^* \pm \delta_2 \sqrt{\mathcal{N} \sum_{i=1}^{n_p^*} \left(\frac{B_{ki}}{d_i} \right)^2} \tag{6.5}$$

This expression again demonstrates the necessity of eliminating the redundant parameters, as one small d_i will spoil the consistency of all parameters for which $B_{ki} \neq 0$. Note, however, that the validity of (6.4) and (6.5) ultimately depends on the accuracy of the approximation (6.1), which assumes that both $\delta\bar{p}$ and δ_2 are sufficiently small.

The volume of the ellipsoid described by equation (6.4) is proportional to δ_2 and to the determinant of the reduced inverse of D, which we define as

$$|D_r^{-1}| = \prod_{i=1}^{n_p^*} \frac{1}{d_i}$$

Both δ_2 and $|D_r^{-1}|$ should be small for the identified model to be consistent. Although the MODES algorithm gradually reduces δ_2, the removal of an observation from the selected subset $\mathcal{S}$ will usually give rise to an increase in the value of $|D_r^{-1}|$. However, it was found that as long as $\mathcal{P}_\mathcal{S}$ is inhomogeneous and still contains a distinct cluster, the consistency of the n_p^* dominant parameters will increase for each selection step. This is compatible with our earlier statement that the MODES algorithm gradually reduces the extent of $\mathcal{P}_\mathcal{S}$, and with it the indeterminacy in the location of the mode (see Section 3.4.4). Consequently, the final mode estimate of the MODES algorithm will be a far more consistent parameter set than the initial mode estimate: the mean parameter set.

We will end this section with a note of caution. It has already been stressed that consistency should not be mistaken for accuracy. Even when the dispersion δ_2 can be reduced to zero (while the value of $|D_r^{-1}|$ remains finite), and the consistency of the identified parameters is perfect, the identified parameter set $\bar{p}^*$ need not be equal to the true parameter set of the device $\mathcal{D}$. Remember that the identified model does not describe the behaviour of the device $\mathcal{D}$, but that of the whole observable system $\mathcal{O}$ (see Section 3.1.1). Therefore, the whole identification space might be translated due to the observation errors $\bar{e}_i$ (especially their systematic component is often overlooked). The observation errors will in the end limit the accuracy of the parameter values determined by any identification method, including MODES.

6.2 The model validity domain

The concept of model validity that has been used throughout the preceding chapters was based on the modelling criterion of Section 2.1. This criterion, however, only applies to the device behaviour that has actually been observed: the set of observations $\mathcal{X}$. The discrete nature of the observations implies that model validation is not possible in any global sense [14]. While the validity domain of a model is a subspace, a continuum, the observations can at best provide information about points in that space. Since these are totally different concepts—a line is not an infinite number of points—simply increasing the number of observations will not help. Although large $\mathcal{N}$ will generally inspire more confidence in the identified model, nothing is proven. There always is an infinite number of non-equivalent models ($\notin \mathcal{M}$) that would be valid for any finite set of observations. The correctness of the model hypothesis, at least for a limited domain, must therefore remain a presupposition. However, if we accept this fact, we can proceed by determining and describing the validity domain of the model hypothesis.

6.2.1 Interpolation

We start by reintroducing the independent and dependent interface variables $\bar{u}$ and $\bar{y}$. For convenience, we will choose the same division as was used for the experiments to obtain the observations $\mathcal{X}$ (see Section 3.1.1). The observable system $\mathcal{O}$ implements a mapping from the domain $\mathcal{U}$ defined by the observed values of the independent variables, to the range $\mathcal{Y}$ defined by the observed values of the dependent variables. As the observed values of the independent interface variables are typically controlled by the experiment, each experiment can be specified by a point $\bar{u} \in \mathcal{U}$. For any experiment, that is, for all points in the domain space, the residual can be evaluated using the identified model parameters. In theory, the validity domain of the identified model can be defined

as the closed set in $\mathcal{U}$ of all experiments that satisfy the local validity criterion $\epsilon(\bar{u}) \leq 1$. Under these experimental conditions the model is a valid representation of the observable system $\mathcal{O}$.

In practice, the value of $\epsilon(\bar{u})$ is only known for the discrete set of experiments given by $\mathcal{X}$. The validity domain divides the set $\mathcal{X}$ into the disjoint subsets $\mathcal{V}$ and $\overline{\mathcal{V}}$. The data points in $\mathcal{V}$ must now be extended into a continuous domain in $\mathcal{U}$. The obvious approach is to use interpolation. As the validity domain will, in general, not be convex or even connected [44], this interpolation must be local, i.e. the interpolation scheme should only use "neighbouring" data points. In the one-dimensional case ($n_u = 1$) an adequate interpolation scheme is easily defined because of the natural ordering of the data points. However, for $n_u \geq 2$ an interpolation scheme cannot be defined without first ordering the domain space by organizing the data points in a regular structure. Here, we will discuss the case where the domain space $\mathcal{U}$ is discretized using an n_u-dimensional rectangular mesh, and where the data points $\mathcal{X}$ are located at the vertices of this mesh. This ordering of the data points allows a compact specification of the experimental procedure and a compact representation of the resulting data set, which makes it the standard measuring strategy.

Now consider the following interpolation scheme: a mesh cell belongs to the validity domain of the model if all its 2^{n_u} vertices are elements of the valid set $\mathcal{V}$. The basic assumption is that the residual over the mesh cell is bounded from above by the largest residual at its vertices. Provided that the residual is dominated by the modelling error and the systematic component of the observation error, the mesh can always be chosen fine enough for this regularity condition on the residual to be satisfied. By applying this interpolation rule to all the cells in $\mathcal{U}$, we obtain the subspace $\mathcal{V}_C \subset \mathcal{U}$, or the complemented valid set, as an estimate of the true validity domain of the model. The validity domain of the model is thus approximated by a large number of, possibly connected, n_u-dimensional rectangular blocks. This process is illustrated for a two-dimensional example in Figure 6.1, where the true validity domain of the

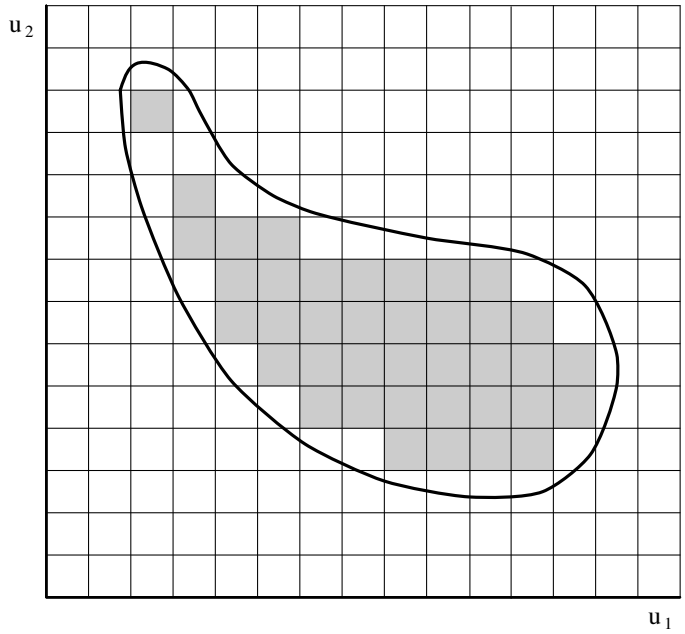

Figure 6.1: The validity domain estimate $\mathcal{V}_C$.

model is bounded by the contour, and $\mathcal{V}_C$ is indicated by the shaded area.

Interpolation tends to underestimate the extent of the validity domain by a fraction of a mesh cell at the boundaries of the domain. Hence, the accuracy of $\mathcal{V}_C$ as an approximation of the validity domain is first of all determined by the spacing of the mesh. However, the accuracy cannot be increased without limit by choosing the mesh arbitrarily fine. A fundamental limit is again posed by the accuracy of the model itself. In the observation space the model curve does not pass through the observations $\bar{x}_i \in \mathcal{V}$, but instead passes through the points $\bar{x}_i^*$, i.e. the points in the model subspace that are closest to the observations $\bar{x}_i$ with respect to the specified accuracy metrics (see Section 4.1.2). These points span a subsection of the model curve, which, when projected on the domain space, will not completely coincide with $\mathcal{V}_C$. The discrepancy between the two domains is bounded by the extent of the projected validity regions of the observations. However, as the validity regions extend symmetrically around the observations, the domain $\mathcal{V}_C$ may not only underestimate, but also overestimate the extent of the validity domain by this amount. The projected validity regions of the observations that are located at the boundary of $\mathcal{V}_C$ determine the lower bound for the accuracy of this boundary.

6.3 The implementation

The quality of our implementation of the MODES algorithm must be assessed according to the following criteria: reliability, sensitivity and efficiency—in order of priority. For those stages of the algorithm for which there are alternatives, we will do so by comparing our implementation with a more conventional choice. However, this comparison is complicated by the lack of an adequate set of benchmark problems. The reliability as well as the sensitivity of the algorithms are mainly implementation dependent, and can therefore be evaluated without referring to specific identification problems. In contrast, we find that the computing time needed to solve an identification problem is to a large extent determined by external factors, such as the formulation of the model constraints, or the accuracy of the initial estimates of the model parameters. Hence, we will mostly confine ourselves to a qualitative comparison of the different algorithms.

6.3.1 Minimizing the objective function

The MODES algorithm requires the solution of a sequence of least-squares data-fitting problems. These problems could have been solved using a standard minimization technique [30]. Instead, we have decided to develop our own algorithm: the reduced Gauss-Newton (RGN) method. This was done mainly to increase the reliability of the identification process, although efficiency and accuracy considerations also played a significant role. In this section we will compare the RGN method with its main contender: the Levenberg-Marquardt (LM) method. According to its advocates, the LM method is not only fully reliable, but also the most efficient Newton method for solving least-squares problems. As a result, the LM method has virtually become the standard minimization method for data-fitting applications, both in literature and in commercial implementations [7, 9, 10, 11, 32].

The LM method is based on the observation that when the ill-conditioning of the sensitivity matrix A causes the original Gauss-Newton method to fail, the search direction is often almost orthogonal to the direction of the negative gradient $-\bar{g}$. In order to counter this effect, the LM method modifies the normal equations (4.30) to

$$\left(A^t A + \gamma I\right) \Delta\bar{p} = -A^t \bar{r} \tag{6.6}$$

In some implementations the identity matrix in this expression is replaced by a non-singular diagonal scaling matrix. For increasing γ the search direction is gradually rotated toward the direction of the negative gradient, while the condition of the problem is gradually improved (for large γ the condition number of the left-hand side matrix approaches unity). Hence, for large enough γ the

modified normal equations (6.6) can be solved accurately, and the search direction will be a reliable descent direction. Because an increase in the value of γ also reduces the length of the correction step defined by (6.6), the Marquardt parameter γ can also be used for the purpose of achieving global convergence; there will always exist a finite value for γ so that the LM correction step will yield a reduction in the value of the objective function. In this way, the LM method avoids the directional search for the optimum step length α (see Section 4.2.6), however, only by requiring a similar algorithm for computing γ. Implementations of the LM method differ mainly in the heuristics that are used to determine an adequate value for γ.

As the LM method can solve ill-conditioned problems without actually having to determine the degenerate directions, which would require the singular-value decomposition of A, the predominant reason for the use of the LM method instead of a RGN method is its supposedly higher efficiency. One should, however, not mistake the efficiency of a single step in an algorithm, here the calculation of $\Delta\bar{p}$ from A and $\bar{r}$, for the efficiency of the whole algorithm. The number of iterations of the algorithm, or more in particular the number of evaluations of the objective function, that are needed for convergence to the solution is an equally decisive factor determining the computational efficiency of the minimization procedure. Extensive testing of both methods has shown that in this respect the RGN method will often outperform the LM method. More specifically, we have found that for problems which are ill-conditioned in the solution $\bar{p}^*$ (i.e. for models with redundant parameters) the RGN method exhibits far better convergence behaviour in the neighbourhood of the solution than the LM method. A low convergence rate for the LM method has been reported by several authors [8, 30, 37]. To illustrate this problem we will write the LM correction step as a vector sum of the right singular vectors of A:

$$\Delta\bar{p} = -\sum_{i=1}^{n_p} \frac{d_i\,(\bar{q}_i \cdot \bar{r})}{{d_i}^2 + \gamma}\bar{b}_i$$

If we compare this expression to (4.31) we find that the LM method, in order to reduce the contribution of the components of $\Delta\bar{p}$ that point in the degenerate directions, also modifies the ratio between the other components, which may spoil their accuracy. If, for instance, the model constraints are linear in $\bar{x}$ and $\bar{p}$, so the Gauss-Newton approximation is exact, the LM method, in contrast to the RGN method, will not reach the solution in a single step when $\gamma > 0$. The possible super-linear rate of convergence of the original Gauss-Newton method and the RGN method can therefore only be matched by the LM method when $\gamma = 0$ in the neighbourhood of the solution, i.e. when there are no redundant parameters.

The low convergence rate of the LM method would be extra problematic for MODES. Firstly, because MODES requires a sequence of minimization prob-

lems to be solved, each from an initial point that is already in the neighbourhood of the next solution. Therefore, slow convergence of the minimization algorithm will weigh heavily on the total efficiency. Secondly, because MODES eliminates data points and possibly whole regions of the domain space, it is very likely that some of the parameters will become redundant. Both effects were confirmed by experiments, where we found MODES to be less efficient when based on the LM algorithm than when based on the RGN algorithm, which handles ill-conditioning without degrading the convergence behaviour of the minimization process.

In respect of the sensitivity threshold of the identification method, equation (6.6) provides a poor computational approach to obtaining the LM correction step. As it is based on the normal equations, it involves the inversion of the matrix (A^tA), which has a condition number of c^2, the square of the condition number of A. Hence, if the matrix A is ill-conditioned then the matrix (A^tA) will be considerably more so. Consequently, implementations of the LM method that calculate the correction step $\Delta\bar{p}$ by solving (6.6) are inherently unsuited to solve ill-conditioned problems. An implementation of the LM method could achieve the same sensitivity threshold as the RGN method by computing the correction step using a factorization of the matrix A, for example the Householder reduction of A to a bi-diagonal form [30] (which is an intermediate stage in the singular-value decomposition of A). However, this increases the computational overhead considerably, removing the initial reason for choosing the LM method.

6.3.2 Constraints in the parameter space

Some authors have chosen to introduce linear constraints in the least-squares minimization problem [9, 10], because "the parameters tend to take on non-physical values" [9]. However, in our case, the minimum $\bar{p}^*$ is interpreted as the location of the identification space, or as the average values of the parameters. Hence, a constrained minimum of the objective function is clearly not acceptable. Moreover, the MODES algorithm which we use to obtain "physical" values for the model parameters requires an unconstrained stationary point of the objective function to function correctly. This is the motivation behind our rather casual approach to the bounds on the parameters. Even when the initial point $\bar{p}^{(0)}$ lies well within the specified bounds, in a multi-dimensional space there will always be descent paths leading to the desired minimum that temporary violate the bounds. Hence, the bounds on the parameter values are implemented as "weak" constraints. Their main purpose is to limit the length of the correction steps in order to stay within the collecting region of the desired minimum. For that purpose the bracketing technique described in Section 4.2.7 is very effective, efficient, and easy to implement.

An additional reason for introducing constraints that is peculiar to the LM method is to limit and effectively fix the values of any redundant parameters. When the identification problem is ill-conditioned, the LM correction step will have components of an arbitrary but often substantial length in the degenerate directions. Over many iterations these components can accumulate and cause numerical problems, degrading the reliability of the identification process. The RGN method avoids this problem by always setting the lengths of these components to zero.

6.3.3 Calculating the residuals

The efficiency of the algorithm that is used for calculating the residuals will, to a large extent, determine the efficiency of the whole identification method. This is the reason why in Section 4.1.6 the tradeoff between the reliability and the efficiency of the algorithm was slightly in favour of the latter. A tradeoff which is acceptable because the failure of a single residual calculation (or possibly a few) does not necessarily lead to the failure of the whole identification process. The temporary removal of the offending observations from the data set $\mathcal{X}$, which is easily implemented in MODES, will often allow further progress to be made. At a later stage of the identification process, when the model subspace has been moved closer to these observations, the removed observations can often be reintroduced.

One aspect of the efficiency of this minimization method is its rate of convergence. A comprehensive analysis, which can be found in Appendix C, shows that the convergence rate of the proposed method is linear. This compares unfavourably with several closely related methods, such as Wilson's SOLVER method [37] and the "variable metric" method developed by Powell [38]. These methods accumulate information from successive iterations in order to construct an approximation of the second derivatives of the model constraints. If this approximation is successful, these methods will exhibit super-linear convergence in the final stages of the iteration process. It is possible to extend the present method in the same direction. However, here the asymptotic rate of convergence of the minimization algorithm is not always the significant criterion. Only when the evaluation of the model equations dominates the total computing time—as might be the case for complex circuit models (see Appendix A)—would the lower iteration count of these methods offer a real benefit. For the relatively small but highly non-linear device models that are preferred for circuit synthesis, and which form the main field of application of MODES, experiments have shown that the gain in convergence rate does not outweigh the extra computational complexity involved in these methods. On the whole we found that the proposed minimization method represents a good compromise between the amount of computation required per iteration and the total number of iterations required for convergence.

6.3.4 The model constraints

The evaluation of the model constraints forms the basis of the whole identification process. Therefore, some care should be taken when formulating the model constraints. An awkward formulation can lead to complex topologies in the observation space and parameter space, which are difficult to search when calculating the residuals or minimizing the objective function. Any mathematically redundant parameters are best eliminated in advance. Also recall the assumption concerning the independence of the model constraints in Section 4.1.4. Most of the efficiency and reliability problems that occur during identification can be traced back to neglecting this important initial stage of the modelling exercise.

The MODES algorithm has been designed to use the system of model constraints $f_{\mathcal{M}}$ in its linearized form:

$$f_{\mathcal{M}}(\bar{p}, \bar{x}) \approx f_{\mathcal{M}}(\bar{p}_0, \bar{x}_0) + J_x\,(\bar{x} - \bar{x}_0) + J_p\,(\bar{p} - \bar{p}_0)$$

where the function value at the current point $f(\bar{p}_0, \bar{x}_0)$, and the Jacobian matrices $J_x = \partial f_{\mathcal{M}}/\partial \bar{x}$ and $J_p = \partial f_{\mathcal{M}}/\partial \bar{p}$ at the current point, must be supplied. In developing the minimization algorithms we tacitly assumed that the model constraints and their first derivatives could be evaluated with unrestricted accuracy. In reality, the accuracy with which the model constraints are evaluated is crucial to the quality of the whole identification process. First of all, the accuracy of $f_{\mathcal{M}}(\bar{p}_0, \bar{x}_0)$ and J_x determines the accuracy with which the residuals $\bar{\rho}_i$ can be calculated. The accuracy of the residual vector $\bar{r}$ and the accuracy of the sensitivity matrix A in their turn are relied upon by the RGN method to determine the degenerate directions in the parameter space, and by the step selection procedure (see Section 4.3.1) to single out the correct observation to be removed from the current mode set.

According to Section 4.2.4 it is the accuracy of the calculated gradient $\bar{g}$ of the objective function $\mathcal{C}_2$ that determines $c_{\max}$, the maximum acceptable condition number of A. From (4.21) and (4.28) we can derive that the accuracy of $\bar{g}$ depends equally on the accuracy of $f_{\mathcal{M}}(\bar{p}_0, \bar{x}_0)$ and the accuracy of J_x and J_p. Since $f_{\mathcal{M}}$ is available in analytical form, its current value can, at least in theory, be evaluated with machine accuracy. The sensitivity of the identification method is then limited by the accuracy of the supplied derivative information. Preferably, the accuracy with which the derivatives J_x and J_p are evaluated should therefore be comparable to the accuracy with which the constraint function itself is evaluated.

The constraint derivatives J_x and J_p can either be supplied in analytical form, or approximated numerically by finite differences [37]. For easy implementation of new models it is generally considered desirable to avoid the sometimes cum-

bersome process of expressing analytically the derivatives of the model equations. However, when the accuracy of the finite difference approximation is poor, as is often the case when the model is strongly non-linear, convergence of the minimization methods that use this approximation cannot be guaranteed. Inaccurate derivative information in fact nullifies the theory underlying their iteration equations, which may give rise to unusable search directions. This was found to be a major source of convergence problems and inaccurate solutions in several conventional identification methods. Therefore, whenever possible, analytical derivatives of the model constraints should be supplied. Analytical derivatives, with their inherent accuracy, contribute significantly to the quality of the identification procedure.

6.4 Conclusions

In this thesis we have presented a unified approach to the identification of analytical device models. The improved theoretical understanding of the identification problem has resulted in the mode selection method (MODES), which can replace the sequential method as the identification method of choice for all applications where the validity of the model cannot be guaranteed *a priori*.

The reliability of MODES at least equals that of the sequential method, as was demonstrated in the preceding chapter. The reliability of both these identification methods arises from the fact that they take into consideration the limited extent of the validity domain of the analytical device model when determining the model parameters. However, MODES has a sound theoretical basis which the sequential method lacks. MODES only relies on the asymptotic character of the model, an assumption which is justified for most analytical device models. In contrast, the definition of a sequential method is usually a somewhat haphazard affair, involving additional model approximations and non-linear transformations of the data, which require assumptions that are far more precarious. Hence, a sequential method may effectively modify the model structure and identify a model that differs from the intended model. As MODES always works directly with the model itself, it may even surpass the reliability of a sequential method.

MODES has all the flexibility of a data-fitting method. It can be applied without any modification to analytical models of arbitrary complexity. The often problematic linearization of the model curve, and the inaccurate estimation of the validity domain by eye, are dispensed with. The validity domain of the model is extracted automatically from the observed device behaviour, using a well-defined validity criterion. Supplying the validity domain exposes and localizes the model's deficiencies, showing where the model needs to be extended. This feature makes MODES a particularly useful tool for the development of new models and the improvement of existing models.

MODES provides realistic consistency estimates for the identified parameters. These estimates are indispensable to all applications where devices are compared with respect to their model parameters. These applications range from circuit synthesis to (statistical) process characterization. The sequential method does not usually provide consistency estimates, while conventional data-fitting methods tend to base their estimates on unrealistic assumptions about the validity of the model or the stochastic properties of the observation errors.

The implementation of MODES presented in this thesis combines robustness with efficiency. Compared to other data-fitting algorithms, our algorithm is marked by its proficient handling of strongly non-linear models and its ability to deal effectively with over-parameterized models. As these contributions can improve the robustness of data-fitting algorithms in general, they should be of interest to a wider audience. Finally, because MODES is based on the least-squares method, existing least-squares parameter-extraction programs can be readily upgraded to MODES.

Appendix A

The Model Equations

Throughout this thesis, analytical device models are represented by a set of equality constraints

$$f_{\mathcal{M}}(\bar{p}, \bar{x}) = \bar{0}, \qquad f \in \mathbb{R}^{n_p} \times \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_f} \tag{A.1}$$

where $\bar{p}$ is the vector of structural parameters, and $\bar{x}$ is the vector of interface variables. However, many device models do not come in this form. In this appendix, therefore, we will introduce a broader class of analytical model specifications, and discuss a number of methods for deriving model specifications of the form (A.1).

Any device model (structural as well as behavioural) that deserves the epithet "analytical" can be specified in the form of a set of n_e analytical expressions

$$\begin{aligned} e_1(\bar{p}, \bar{x}, \bar{a}) &= 0 \\ e_2(\bar{p}, \bar{x}, \bar{a}) &= 0 \\ &\cdots \\ e_{n_e}(\bar{p}, \bar{x}, \bar{a}) &= 0 \end{aligned} \tag{A.2}$$

where $\bar{a}$ represents a vector of n_a additional variables, which will be referred to as the "internal" variables as opposed to the "external" interface variables $\bar{x}$, and $n_e \leq n_x + n_a$. This representation also includes the so-called circuit models, which model a device by an equivalent circuit from a small set of circuit-theoretical primitives, such as ideal linear components, diodes, and non-linear sources. The internal variables then represent electrical quantities associated with the internal nodes of the equivalent circuit.

To obtain a model representation of the form (A.1) it is necessary to eliminate the n_a internal variables together with n_a of the equations. The set of equations (A.2) defines the internal variables as an implicit function of the parameters and

the interface variables: $\bar{a}(\bar{p}, \bar{x})$. (Although this function will not be unique when the set of equations is underdetermined, i.e. $n_e < n_x + n_a$.) When this function is substituted in (A.2), the resulting set of equations will be dependent, and exactly n_a equations can be removed from the set. Hence, the set of equations (A.2) implicitly defines a (again not unique) set of model constraints $f_{\mathcal{M}}$, where $n_f = n_e - n_a$. Since the analytical tractability of the model structure is important to our modelling goal (recall Section 2.3.3), we would prefer $\bar{a}(\bar{p}, \bar{x})$ to be expressed in analytical form. This is feasible for many of the popular analytical device models. Moreover, these manipulations on the set of equations can usually be performed by one of the mathematical expert systems that are currently available [45].

In case it is not possible to obtain $\bar{a}(\bar{p}, \bar{x})$, and hence $f_{\mathcal{M}}$, in analytical form, we can revert to a numerical evaluation of $f_{\mathcal{M}}$. However, it is then necessary to ensure that effectively the same non-linear transformation is used to obtain (A.1) from (A.2) for all possible values of $\bar{p}$ and $\bar{x}$. When the number of internal variables that cannot be eliminated analytically is small, this is best achieved by defining a function $\bar{a}(\bar{p}, \bar{x})$ for the remaining variables, but now in the form of a numerical procedure. Alternatively, when the number of internal variables is large—for example in the case of a large circuit model—it will be more efficient to use device simulation (see Section 4.1.3) to construct a set of model constraints. After designating $n_x - n_f$ of the interface variables as independent variables, denoted by $\bar{u}$, the set of $n_e = n_f + n_a$ equations (A.2) has exactly $n_f + n_a$ remaining free variables (for fixed $\bar{p}$). This set of equations can then be solved numerically, in the case of a circuit model using one of the available circuit simulators, which allows us to determine the values of the dependent interface variables, denoted by $\bar{y}$, as a function of the independent interface variables: $\bar{y}(\bar{p}, \bar{u})$. The set of model constraints can be expressed as

$$f_{\mathcal{M}}(\bar{p}, \bar{x}) = \bar{y} - \bar{y}(\bar{p}, \bar{u}), \qquad \text{where } \bar{x} = \begin{bmatrix} \bar{u} \\ \bar{y} \end{bmatrix}$$

However, as the use of numerical techniques, such as device simulation, in the formulation of the model constraints severely limits the analytical tractability of the model, it is often better (in view of the modelling goal) to modify the model equations by making some additional assumptions and approximations.

Appendix B

Derivatives of the Residuals

The purpose of this appendix is to derive an expression for the partial derivatives of the residuals with respect to the parameters, thereby proving the result that was given in equation (4.27).

For $\bar{p} = \bar{p}^{(k)}$ and $\bar{x} = \bar{x}^*$ the following relations must hold

$$f(\bar{p}, \bar{x}) = \bar{0} \tag{B.1}$$
$$V(\bar{x} - \bar{x}_0) = J_x^t \bar{\lambda} \tag{B.2}$$

where J_x is the $n_f \times n_x$ Jacobian matrix of partial derivatives $\partial f / \partial \bar{x}$. It is further assumed that J_x is of full rank, in which case the vector of Lagrange multipliers $\bar{\lambda}$ is unique. Taking partial derivatives with respect to an element p of the parameter vector $\bar{p}$ of the equations (B.1) and (B.2) gives

$$\begin{aligned} \partial_p f + J_x \, \partial_p \bar{x} &= \bar{0} \\ V \, \partial_p \bar{x} &= J_x^t \, \partial_p \bar{\lambda} + (\partial_p J_x)^t \, \bar{\lambda} \end{aligned} \tag{B.3}$$

where ∂_p is a shorthand notation for the operator $\partial / \partial p$. Equation (B.3) contains second derivatives of the model constraints because the adjoint equation (B.2) already contains first derivatives. Solving this set of equations for $\partial_p \bar{\lambda}$ yields

$$\partial_p \bar{\lambda} = - \left(J_x V^{-1} J_x^t \right)^{-1} \left[\partial_p f + J_x V^{-1} (\partial_p J_x)^t \bar{\lambda} \right]$$

The residual vector $\bar{\rho}$ is defined as

$$\bar{\rho} = \left(D^{1/2} R^t \right) \bar{\lambda}$$

where the diagonal matrix D and the orthogonal matrix R (which means that $R^t = R^{-1}$) follow from the relation

$$J_x V^{-1} J_x^t = R D R^t = \left(D^{1/2} R^t \right)^t \left(D^{1/2} R^t \right) \tag{B.4}$$

The partial derivate of $\bar{\rho}$ is given by

$$\partial_p \bar{\rho} = \left(D^{1/2} R^t\right) \partial_p \bar{\lambda} + \partial_p \left(D^{1/2} R^t\right) \bar{\lambda}$$

This equation can be rewritten as

$$\begin{aligned} \partial_p \bar{\rho} &= -\left(D^{1/2} R^t\right) \left(R D^{-1} R^t\right) \left[\partial_p f + J_x V^{-1} (\partial_p J_x)^t \left(D^{-1/2} R^t\right) \bar{\rho}\right] \\ &\quad + \partial_p \left(D^{1/2} R^t\right) \left(D^{-1/2} R^t\right) \bar{\rho} \\ &= -\left(D^{-1/2} R^t\right) \partial_p f + \left(D^{-1/2} R^t\right) Z \left(D^{-1/2} R^t\right)^t \bar{\rho} \end{aligned} \tag{B.5}$$

where the matrix Z is defined as

$$Z = \left[\left(D^{1/2} R^t\right)^t \partial_p \left(D^{1/2} R^t\right) - J_x V^{-1} (\partial_p J_x)^t\right]$$

To prove that the matrix Z is skew-symmetric (which means that $Z = -Z^t$), we use the derivative of equation (B.4)

$$\begin{aligned} &J_x V^{-1} (\partial_p J_x)^t + (\partial_p J_x) V^{-1} J_x^t = \\ &\quad \left(D^{1/2} R^t\right)^t \partial_p \left(D^{1/2} R^t\right) + \partial_p \left(D^{1/2} R^t\right)^t \left(D^{1/2} R^t\right) \end{aligned}$$

Reordering of the terms results in

$$\begin{aligned} &\left[\left(D^{1/2} R^t\right)^t \partial_p \left(D^{1/2} R^t\right) - J_x V^{-1} (\partial_p J_x)^t\right] = \\ &\quad - \left[\partial_p \left(D^{1/2} R^t\right)^t \left(D^{1/2} R^t\right) - (\partial_p J_x) V^{-1} J_x^t\right] = \\ &\quad - \left[\left(D^{1/2} R^t\right)^t \partial_p \left(D^{1/2} R^t\right) - J_x V^{-1} (\partial_p J_x)^t\right]^t \end{aligned}$$

Since the matrix Z is skew-symmetric, the matrix S which is defined as

$$S = \left(D^{-1/2} R^t\right) Z \left(D^{-1/2} R^t\right)^t$$

is also skew-symmetric. It is thus possible to reduce equation (B.5) to

$$\partial_p \bar{\rho} = -\left(D^{-1/2} R^t\right) \partial_p f + S\, \bar{\rho}$$

where S is a skew-symmetric matrix, which is the desired result.

Appendix C

Convergence of the Residual Calculation

In this appendix we will determine the rate of convergence of the method for constrained minimization that was developed in Section 4.1.5.

As point of departure we take the iteration equations of the algorithm, the primal and adjoint equations (4.10) and (4.11), and write them in the form of a block-matrix equation that is better suited to the analysis that will follow. The relation between two consecutive iterations of the algorithm can be expressed as

$$\begin{bmatrix} -V & J_x^t \\ J_x & 0 \end{bmatrix} \begin{bmatrix} \bar{x}^{(k+1)} \\ \bar{\lambda}^{(k)} \end{bmatrix} = \begin{bmatrix} -V\bar{x}_0 \\ -f(\bar{x}^{(k)}) + J_x\bar{x}^{(k)} \end{bmatrix}$$

From the iteration equation of the algorithm we can derive the following expression for the accuracy—the deviation from the solution $(\bar{x}^*, \bar{\lambda}^*)$—of the $(k+1)$st iterate

$$\begin{bmatrix} -V & J_x^t \\ J_x & 0 \end{bmatrix} \begin{bmatrix} (\bar{x}^{(k+1)} - \bar{x}^*) \\ (\bar{\lambda}^{(k)} - \bar{\lambda}^*) \end{bmatrix} = \begin{bmatrix} V(\bar{x}^* - \bar{x}_0) - J_x^t\bar{\lambda}^* \\ -f(\bar{x}^{(k)}) + J_x(\bar{x}^{(k)} - \bar{x}^*) \end{bmatrix} \tag{C.1}$$

To determine the convergence rate of the algorithm, we must express the right-hand side of (C.1) in terms of the accuracy of the former iterate and derivatives of the model constraints at the solution. For this purpose, we expand the Jacobian matrix J_x in a second-order Taylor series about the solution. The rows of J_x are the gradients of the individual model constraints ∇f_i, hence for $i = 1, \ldots, n_f$

$$\nabla f_i(\bar{x}^{(k)}) = \nabla f_i(\bar{x}^*) + \nabla^2 f_i(\bar{x}^*)(\bar{x}^{(k)} - \bar{x}^*) + O(\|\bar{x}^{(k)} - \bar{x}^*\|^2)$$

Substitution in the upper part of the right-hand side of (C.1) yields

$$\begin{aligned} V(\bar{x}^* - \bar{x}_0) - J_x^t \bar{\lambda}^* &= V(\bar{x}^* - \bar{x}_0) - \sum_{i=1}^{n_f} \lambda_i^* \nabla f_i(\bar{x}^{(k)}) \\ &= \left[V(\bar{x}^* - \bar{x}_0) - \sum_{i=1}^{n_f} \lambda_i^* \nabla f_i(\bar{x}^*) \right] \\ &\quad + \sum_{i=1}^{n_f} \lambda_i^* \nabla^2 f_i(\bar{x}^*)(\bar{x}^{(k)} - \bar{x}^*) + O(\|\bar{x}^{(k)} - \bar{x}^*\|^2) \\ &= O(\|\bar{x}^{(k)} - \bar{x}^*\|) \end{aligned}$$

because the term between the square brackets is the adjoint equation in the solution, and thus equals zero.

A second-order Taylor series expansion of the model constraints f, but now about $\bar{x}^{(k)}$, yields

$$f(\bar{x}^*) = f(\bar{x}^{(k)}) + J_x(\bar{x}^* - \bar{x}^{(k)}) + O(\|\bar{x}^* - \bar{x}^{(k)}\|^2)$$

As $f(\bar{x}^*) = \bar{0}$, the terms can be reordered to obtain an expression for the lower part of the right-hand side of (C.1)

$$-f(\bar{x}^{(k)}) + J_x(\bar{x}^{(k)} - \bar{x}^*) = O(\|\bar{x}^k - \bar{x}^*\|^2)$$

Collecting the results, we finally arrive at the following expression for the error equation of the algorithm:

$$\begin{bmatrix} -V & J_x^t \\ J_x & 0 \end{bmatrix} \begin{bmatrix} (\bar{x}^{(k+1)} - \bar{x}^*) \\ (\bar{\lambda}^{(k)} - \bar{\lambda}^*) \end{bmatrix} = \begin{bmatrix} O(\|\bar{x}^{(k)} - \bar{x}^*\|) \\ O(\|\bar{x}^{(k)} - \bar{x}^*\|^2) \end{bmatrix} \tag{C.2}$$

$$= \begin{bmatrix} \sum_{i=1}^{n_f} \lambda_i^* \nabla^2 f_i(\bar{x}^*)(\bar{x}^{(k)} - \bar{x}^*) + O(\|\bar{x}^{(k)} - \bar{x}^*\|^2) \\ O(\|\bar{x}^{(k)} - \bar{x}^*\|^2) \end{bmatrix} \tag{C.3}$$

Equation (C.2) shows that if the algorithm converges, it does so only linearly. The more detailed expression (C.3) also shows that the size of the first-order term, and consequently the convergence factor of the algorithm, is determined by the product of the curvature of the model constraints $\nabla^2 f_i$, and the size of the Lagrange multipliers at the solution λ_i^*. Hence, when the model constraints are almost linear, or when the observation $\bar{x}_0$ is close to the model curve so $\bar{\lambda}^*$ is small, the convergence factor will be small and the rate of convergence of the algorithm can still be high in the neighbourhood of the solution. Both effects were confirmed in experiments.

Bibliography

[1] A.F. Schwarz: *Computer-Aided Design of Microelectronic Circuits and Systems.* Academic Press, London, 1987.

[2] I.E. Getreu: *Modelling the Bipolar Transistor.* Elsevier, Amsterdam, 1978.

[3] T.J. Krutsick, M.H. White, H. Wong and R.V.H. Booth: An improved method of MOSFET modeling and parameter extraction. *IEEE Trans. Electron Devices,* vol. ED-34, no. 8, pp. 1676–1680, 1987.

[4] A. Ibarra and J. Gracia: Stategy for DC parameter extraction in bipolar transistors. *IEE Proc. part G,* vol. 137, no. 1, pp. 5–11, 1990.

[5] K.K. Ng and J.R. Brews: Measuring the effective channel length of MOSFETs. *IEEE Circuits & Devices,* vol. 6, no. 6, pp. 33–38, 1990.

[6] R. Rohrer, S. Fan and L. Claudio: Automated bipolar junction transistor DC model parameter determination. *IEEE J. Solid-State Circuits,* vol. SC-6, no. 4, pp. 260–262, 1971.

[7] D.E. Ward and K. Doganis: Optimized extraction of MOS model parameters. *IEEE Trans. Computer-Aided Design,* vol. CAD-1, no. 4, pp. 163–168, 1982.

[8] P. Yang and P.K. Chatterjee: An optimal parameter extraction program for MOSFET models. *IEEE Trans. Electron Devices,* vol. ED-30, no. 9, pp. 1214–1219, 1983.

[9] K. Doganis and D.L. Scharfetter: General optimization and extraction of IC device model parameters. *IEEE Trans. Electron Devices,* vol. ED-30, no. 9, pp. 1219–1228, 1983.

[10] P.B. Wolbert: *Promea Report.* Technical Report, University of Twente, IC-Technology and Electronics Group (ICE), 1986.

[11] E. Khalily: Transistor electrical characterization and analysis program (TECAP). *Hewlett-Packard J.,* vol. 32, pp. 16–18, 1981.

[12] J. Domitrowich: Choosing parameter extraction software. *VLSI Systems Design,* vol. July 1987, pp. 64–68.

[13] M.G. Kendall and A. Stuart: *The Advanced Theory of Statistics,* vol. 2. Charles Griffin, London, 1967.

[14] K.R. Popper: *The Logic of Scientific Discovery.* Unwin Hyman, London, 1980.

[15] T. Shima, T. Sugawara, S. Moriyama, and H. Yamada: Three dimensional table look-up MOSFET model for precise circuit simulation. *IEEE J. Solid-State Circuits,* vol. SC-17, no. 3, pp. 449–454, 1982.

[16] M. Yanilmaz and V. Eveleigh: Table look-up MOSFET modeling for electronic circuit simulation. *Proc. 30th Midwest Symposium on Circuit and Systems,* pp. 1074–1077, 1987.

[17] M.E. Daniel: Development of mathematical models of semiconductor devices for computer aided circuit analysis. *Proc. IEEE,* vol. 55, no. 11, pp. 1913–1920, 1967.

[18] L.O. Chua and A. Deng: Canonical piecewise-linear modeling. *IEEE Trans. Circuits and Systems,* vol. CAS-33, no. 5, pp. 511–525, 1986.

[19] D.C. D'Avanzo, M. Vanzi and R.W. Dutton: *One-Dimensional Semiconductor Device Analysis (SEDAN).* Technical Report no. G-201-5, Stanford University of Technology, 1979.

[20] P.B. Wolbert: *Modeling and Simulation of Semiconductor Devices in TRENDY.* Ph.D. Thesis, University of Twente, Enschede, 1991.

[21] R. Benumof and J. Zoutendyk: Theoretical values of various parameters in the Gummel-Poon model of a bipolar junction transistor. *J. Appl. Phys.,* vol. 59, no. 2, pp. 636–644, 1986.

[22] E.H. Nordholt: *The Design of High-Performance Negative-Feedback Amplifiers.* Elsevier, Amsterdam, 1983.

[23] J. Stoffels: *Automation in High-Performance Negative Feedback Amplifier Design.* Ph.D. Thesis, Delft University of Technology, Delft, 1988.

[24] C.J.M. Verhoeven: *First Order Oscillators.* Ph.D. Thesis, Delft University of Technology, Delft, 1990.

[25] J.A. Hegt: *Contributions to switched capacitor filter synthesis.* Ph.D. Thesis, Eindhoven University of Technology, Eindhoven, 1988.

[26] H.N. Ghosh, F.H. De La Moneda and N.R. Dono: Computer-aided transistor design, characterization, and optimization. *Proc. IEEE,* vol. 55, no. 11, pp. 1897–1912, 1967.

[27] S.W. Director, W.A. Bristol and A.J. Brodersen: Fabrication-based optimization of linear integrated circuits. *IEEE Trans. Circuit Theory,* vol. CT-20, no. 6, pp. 690–697, 1973.

[28] L. Ljung: Convergence analysis of parametric identification methods. *IEEE Trans. Automatic Control,* vol. AC-23, no. 5, pp. 770–783, 1978.

[29] C.F. Gauss: *Teoria Motus Corporum Coelestium.* 1809. Reprinted translation: *Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections.* Dover, New York, 1963.

[30] R. Fletcher: *Practical Methods of Optimization,* vol. 1. Wiley, Chichester, 1980.

[31] H. Bunke and O. Bunke: *Nonlinear Regression, Functional Relations and Robust Methods.* Wiley, Chichester, 1989.

[32] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling: *Numerical Recipes in C.* Cambridge University Press, Cambridge, 1988.

[33] B.W. Kernighan and S. Lin: An effective heuristic procedure for partitioning graphs. *Bell Syst. Techn. J.,* vol. 49, pp. 291–307, 1970.

[34] D.H. Ackley: *A Connectionist Machine for Genetic Hillclimbing.* Kluwer, Dordrecht, 1987.

[35] C.N. Dorny: *A Vector Space Approach to Models and Optimization.* Wiley, New York, 1975.

[36] E. Hille: *Methods in Classical and Functional Analysis.* Addison-Wesley, Reading, 1972.

[37] P.E. Gill and W. Murray (eds.): *Numerical Methods for Constrained Optimization.* Academic Press, London, 1974.

[38] M.J.D. Powell: A fast algorithm for nonlinearly constrained optimization calculations. In *Numerical Analysis.* G.A. Watson (ed.), Springer, Berlin, pp. 144–157, 1978.

[39] V. Strejc: Least squares and regression methods. In *Trends and Progress in System Identification.* P. Eykhoff (ed.), Pergamon, Oxford, pp. 103–144, 1981.

[40] F.A. Lindholm, S.W. Director and D.L. Bowler: Assessing model adequacy and selecting model complexity in integrated-circuit simulation. *IEEE J. Solid-State Circuits,* vol. SC-6, no. 4, pp. 213–222, 1971.

[41] H.K. Gummel and H.C. Poon: An integral charge control model of bipolar transistors. *Bell Syst. Tech. J.,* vol. 49, pp. 827–852, 1970.

[42] L.W. Nagel: *SPICE2: a Computer Program to Simulate Semiconductor Circuits.* Memorandum no. ERL-M520, Electronics Res. Lab., University of California, Berkeley, 1975.

[43] K.J. Åström: Maximum likelihood and prediction error methods. In *Trends and Progress in System Identification.* P. Eykhoff (ed.), Pergamon, Oxford, pp. 145–168, 1981.

[44] B. Martos: *Nonlinear Programming, Theory and Methods.* North-Holland, Amsterdam, 1975.

[45] B.W. Char, K.O. Geddes, G.H. Gonnet, M.B. Monagan and S.M. Watt: *Maple Reference Manual.* Watcom Publications, Waterloo, 1988.