Software Support for Metrology
Best Practice Guide No. 10

Discrete Model Validation

R M Barker and A B Forbes

March 2001

Software Support for Metrology
Best Practice Guide No. 10

Discrete Model Validation

R M Barker and A B Forbes
Centre for Mathematics and Scientific Computing

March 2001

ABSTRACT

Metrology, the science of measurement, involves the determination of physical quantities from experiment, along with estimates of their associated uncertainties. In this endeavour, a mathematical model of the measurement system is required in order to extract information from the experimental data. For the measurements to be reliable, these models must be validated (shown to be fit for purpose). This Best Practice Guide, a companion Guide to SS*f*M *Best Practice Guide No. 4 Discrete Modelling* looks at validation techniques for the main components of discrete modelling: building the functional and statistical model, model solving and parameter estimation methods, goodness of fit of model solutions and experimental design and measurement strategy. The techniques are illustrated in detailed case studies.

© Crown copyright 2001
Reproduced by permission of the Controller of HMSO

ISSN 1471–4124

Extracts from this guide may be reproduced provided the source is acknowledged and the extract is not taken out of context

Authorised by Dr Dave Rayner,
Head of the Centre for Mathematics and Scientific Computing

National Physical Laboratory,
Queens Road, Teddington, Middlesex, United Kingdom TW11 0LW

Contents

1 Introduction **1**
- 1.1 Modelling and model validation in metrology . . . 1
- 1.2 What is a valid model? . . . 3
- 1.3 Model validation as risk control . . . 5
- 1.4 Acknowledgements . . . 5

2 Validation of the model **6**
- 2.1 Validation of the functional model . . . 6
  - 2.1.1 Comprehensiveness of the functional model . . . 6
  - 2.1.2 Correctness of the functional model . . . 9
- 2.2 Validation of the statistical model . . . 10

3 Estimator validation **15**
- 3.1 Estimator validation issues . . . 15
- 3.2 Monte Carlo simulations . . . 17
  - 3.2.1 Monte Carlo simulation for a least squares estimator . . . 19
- 3.3 Null space benchmarking . . . 20
  - 3.3.1 Null space method for generating data for least squares problems. . . . 21
  - 3.3.2 Null space benchmarking for generalised distance regression 22
- 3.4 Estimator analysis . . . 24
  - 3.4.1 Analysis of least squares methods . . . 24

4 Validation of model solutions **26**
- 4.1 Goodness of fit and residual errors . . . 26
  - 4.1.1 Signal to noise ratio . . . 27
  - 4.1.2 Statistical model for the data . . . 27
  - 4.1.3 Residual errors and model selection . . . 28
  - 4.1.4 Validation of model outputs . . . 31

5 Validation of experimental design and measurement strategy **39**
- 5.1 System identifiability . . . 39
- 5.2 System effectiveness . . . 43

6 Case Studies **45**
- 6.1 Univariate linear regression . . . 45
  - 6.1.1 Statistical model for relative measurement error . . . 45

6.1.2 Statistical model for correlation measurement error 49
6.2 Fitting a quadratic to data 55
6.3 Fitting a Gaussian peak to data 61
6.3.1 Functional and statistical model 61
6.3.2 Estimators 61
6.3.3 Evaluation of LLS estimator using Monte Carlo Simulations 63
6.3.4 Null space benchmarking for the LLS estimator 65
6.3.5 Analysis of the LLS estimator 67
6.3.6 Evaluation of WLLS estimator 68
6.3.7 Evaluation of nonlinear effects using Monte Carlo simulations 69
6.3.8 Valid uncertainty estimates for the LLS estimator 70
6.4 Estimation of the effective area of a pressure balance 71
6.4.1 Models 71
6.4.2 Estimators 72
6.4.3 Analysis of estimator behaviour 73
6.5 Circle fitting and roundness assessment 74
6.6 Roundness assessment in the presence of form errors 82
6.6.1 Taking into account form error 83

7 Best-practice in discrete model validation: a summary 94

Chapter 1

Introduction

1.1 Modelling and model validation in metrology

Metrology, the science of measurement, involves the determination of quantitative estimates of physical quantities from experiment, along with estimates of their associated uncertainties. This process involves many or all of the following components:

Model building. Developing a mathematical model of the experimental system in terms of mathematical equations involving parameters that describe all the relevant aspects of the system, including the nature of the measurement error in experimental data.

Model solving. Determining estimates of the model parameters from the measured data by solving the mathematical equations constructed as part of model building. In general, this involves developing an algorithm that will determine the values for the parameters that best explain the data. These algorithms are often referred to as *estimators*.

Software implementation of estimators. Practically all calculations are performed by software.

Model validation. Determining whether the results produced are consistent with the input data, theoretical results, reference data, etc. All stages need to be examined. Does the model adequately encapsulate what is known about the system? Does the method of solution produce unbiased estimates of the parameters and valid estimates of uncertainties?

Generally, these steps are revisited as the model is refined and the experimental design is evolved, resulting in a better explanation of the observed behaviour and more dependable uncertainties in the quantities of interest.

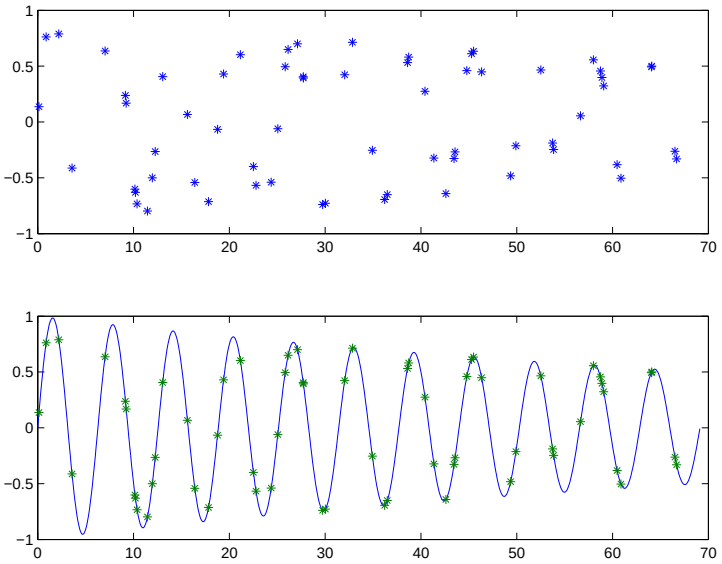

Figure 1.1: Data and fitted model for a harmonic oscillator. The data alone carries little information.

Modelling is the basis on which we extract information from measurement. This statement is illustrated by Figure 1.1. The upper graph shows a seemingly random scatter of data points while the lower graph shows the data along with the fitted model of the form $y = Ae^{-Bx}\sin(Cx+D)$. It is only from knowledge of the model that we are able to characterise the behaviour of a harmonic oscillator from the data. Techniques for discrete modelling are described in the SS*f*M *Best Practice Guide No. 4 Discrete Modelling* [4]. This document is a companion Best Practice Guide concerned with validation of the following components of discrete modelling:

Functional model consisting of:

- Problem variables representing all the quantities that are known or measured.
- Problem parameters representing the quantities that have to be determined from the measurement experiment. The problem parameters describe the possible behaviour of the system.
- The functional relationship between the variables and parameters.

Statistical model for the measurement errors consisting of:

- The error structure describing which variables are known accurately and which are subject to significant measurement error.
- The description of how the measurement errors are expected to behave.

Estimator describing a method of determining estimates of the problem parameters from the measurement data. Good estimators are unbiased, efficient and consistent.

Experimental design and measurement strategy specifying how data is to be gathered.

1.2 What is a valid model?

There are two aspects of model validity. The first is *internal consistency*. To the extent that a model represents a set of mathematical statements, its validity can be checked for mathematical correctness. Typically, a model has a set of inputs described in terms of facts and assumptions about components of the physical system. It also has a set of outputs in terms of predicted behaviour of systems to which it applies. If the model is internally consistent then the outputs are valid so long as the inputs are valid.

The second aspect of validity is *external consistency* with prior information and/or experimental results. A model is valid if information extracted from it is not contradicted by other valid information. The validation status of a model is consolidated as it is shown to be consistent with more external information. Validation of an internally consistent model focusses on the extent to which the assumptions associated with the model apply, i.e., a validation of the applicability of the model.

The main output of the modelling process is usually a measurement result, i.e., an estimate of the value of a parameter, and a statement of its uncertainty. A valid model should produce a valid measurement result. A measurement result is valid if the estimate of the parameter is close to its true value (determined by a significantly more accurate measuring system, for example) relative to the stated uncertainty. In validating a measurement result it is necessary to consider both the parameter estimate *and* its stated uncertainty. For this reason, much of this Guide is concerned not only with validation of the method by which parameter estimates are defined but also the method of uncertainty estimation covered by SSfM *Best Practice Guide No. 6 Uncertainty and Statistical Modelling* [3].

Example: a simple spring balance. Consider a simple spring balance used to estimate the weight W of an object based on the measurement of the extension of a spring. The linear extension is converted into an angular motion of an indicating needle that moves over a scale. The main inputs to the modelling process are:

- Hooke's Law for the extension $e = H(w, k)$ of a spring in terms of the applied load w and spring constant k.
- Hooke's constant k for the particular spring.
- The standard uncertainty σ_k associated with the estimate of k.

- The functional dependence of the position $x = f(e)$ of the indicating needle on the extension e of the spring.
- Measurement error with standard deviation σ_x associated with reading the position x of the needle on the scale.
- A specification of the minimum w_{min} and maximum w_{max} loads to be applied to the instrument.

The main outputs from the modelling process are:

- A function
$$w = F(x, k) \tag{1.1}$$
that converts the reading x of the indicator position into an estimate w of the artefact's weight.
- An estimate
$$\sigma_w = \sigma_x \left(\frac{\partial F}{\partial x} \right) \tag{1.2}$$
of the standard uncertainty σ_w of the measurement error associated with the weight w.

The validation of the internal consistency addresses the question: Assuming the inputs are valid, are the outputs valid? Such issues associated with this model include:

V1 The correctness of the function F derived from Hooke's law and the functional dependence $x = f(e)$.

V2 The correctness of the partial derivative of F.

V3 The expression for σ_w does not take into account the uncertainty in the value of Hooke's constant k and could therefore be overly optimistic.

V4 The expression for σ_w is based on a linear approximation to F at x. This linearisation could give an invalid estimate if the function F is highly nonlinear.

The internal consistency validation does not necessarily require any in depth knowledge of the metrological/physical system, only expertise in discrete modelling. Once the internal consistency of a model is established, the focus of the validation moves to the assumptions, explicit or implicit, associated with the model inputs. An expert in the area might wish to investigate:

V5 The applicability of Hooke's law to the type of spring.

V6 The assigned value for Hooke's constant.

V7 The functional form $x = f(e)$ relating the indicating needle to the spring extension.

V8 The dependence of the system on environmental factors such as temperature. The model as it stands depends only on the spring constant k.

In this activity, the expert is assessing the consistency of the model with prior information about the behaviour of similar systems. Finally, the validity of the model in terms of its outputs can be tested against experimental data. In this task, it is important that the statistical nature of the model outputs and the experimental data are taken into account.

1.3 Model validation as risk control

While the need for model validation seems self-evident, validation will generally require resources. It is therefore necessary to assess the *risks* associated with an invalid model and then design and implement *validation responses* that aim to control the risks to an acceptable level. In this, it is prudent to balance the risks incurred by an invalid model with the cost of validation. In the sections below we discuss the risks associated with each of the main components of the discrete modelling process and indicate suitable validation responses. The analysis of risks is discussed in SS*f*M *Best Practice Guide No. 1 Measurement System Validation* [11].

1.4 Acknowledgements

This best practice guide was produced as part of UK Department of Trade and Industry's National Measurement System *Software Support for Metrology* Programme 1998–2001.

Thanks to members of SS*f*M project teams that contributed to this project, in particular Maurice Cox, Mark Dainton, and Ian Smith. The authors are grateful to Peter Harris for his careful reading of the draft.

Chapter 2

Validation of the model

In this chapter, we look at validation of how the model is built – the specification of the functional and statistical model.

2.1 Validation of the functional model

The functional model describes the mathematical relationships between the variables and parameters of the model. Typically, the functional model will take the form

$$y = \phi(\mathbf{x}, \mathbf{a}),$$

where $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ are the independent or control variables, y is the dependent or response variable and $\mathbf{a} = (a_1, \ldots, a_n)^{\mathrm{T}}$ are the model parameters that specify the form of the function ϕ. In the case of a linear response depending on a single variable we have

$$y = a_1 + a_2 x,$$

for example. There are two main types of functional model. A *physical model* is one for which there is a theory that defines how the variables depend on each other. An *empirical model* is one in which a relationship between the variables is expected or observed but with no supporting theory. Many models have both empirical and physical components.

Validation of the functional model concerns i) its comprehensiveness, i.e., the extent to which all the main variables and parameters are covered by the model and ii) the correctness of the functional relationship between the variables and parameters.

2.1.1 Comprehensiveness of the functional model

The specification of the functional model aims to achieve an appropriate level of comprehensiveness.

Risks. The risks associated with an overly comprehensive model are mainly concerned with the unnecessary commitment of resources to:

- Develop, understand and manipulate the model.
- Monitor or control variables that may have no significant impact on system behaviour.
- Solve computational problems associated with determining the model parameters from data.

The risks associated with a minimal model are mainly concerned with the impact on how well the model performs, including:

- Scope of the model outputs.
- Validity of the model outputs.
- Flexibility of the model to cope with different hardware components or operating conditions.

From knowledge of the physical system, all the factors and/or variables that could have a potential influence on the system can be listed in a table or datasheet and a status assigned to them according to their influence:

- *Unknown parameter.* The value of the variable is unknown and has to be determined from a fit of the model to data.
- *Parameter with prior estimate.* The value of the variable is unknown and has to be determined from a fit of the model to data. A prior estimate of the variable is available along with an estimate of its uncertainty.
- *Measured subject to error.* The variable is to be measured and the uncertainty in the measurement is likely to contribute to the uncertainty in the system.
- *Accurately measured.* The variable is to be measured but the uncertainty in the measurement is not significant compared to other sources of uncertainty.
- *Constant.* The variable can be set to its nominal value and treated as exact.
- *Null.* The value of the variable has no influence on the system and can be omitted from the model.

Assigning a variable to a status high on this list is a move towards increasing the comprehensiveness. In general, it is a good idea to start with a comprehensive model and then justify and document each step in its simplification. For example, a system may be moderately sensitive to temperature so that, on a first analysis, temperature should be included in the model and assigned a status of

'measured subject to error'. However, if the system is kept in an accurately controlled temperature environment a status of 'constant' may be sufficient. This assignment can be validated by numerical simulation.

Suppose the temperature controlled environment guarantees that the temperature is kept within 1 degree of 20.0 ^{o}C and there is also available a temperature sensor that measures with a standard uncertainty 0.2 ^{o}C. If

$$y = \phi(t, \mathbf{x}, \mathbf{a})$$

represents the functional model in terms of temperature t, other variables $\mathbf{x}$ and parameters $\mathbf{a}$, then a quick estimate of the sensitivity of y to temperature can be determined by evaluating

$$\phi(t_0, \mathbf{x}, \mathbf{a}), \quad \phi(t_0 \pm \Delta t, \mathbf{x}, \mathbf{a})$$

for $t_0 = 20.0$, $\Delta t = 1.0$ and 0.2 and for typical values of the variables $\mathbf{x}$ and $\mathbf{a}$. This provides an estimate of the variation of the model values due to variation in temperature in the ranges 20.0 ± 1.0 and 20 ± 0.2. Then:

- *If* the variation in the model values for the case $\Delta t = 0.2$ is significant compared to the likely measurement error in y, then t should be assigned a status of 'measured subject to error'; *otherwise*
- *If* the variation in the model values for the case $\Delta t = 1.0$ is significant compared to the likely measurement error in y, then t should be assigned a status of 'accurately measured'; *otherwise*
- The temperature t is assigned a status of 'constant'.

The interpretation of "significant" will depend on circumstances. If the maximum variation of the model response due to a variable in its operating range is less than 1% of the estimated measurement error in the response, it may well be safe to regard the variable as a constant. However, if the system has a nonlinear response or in other ways can behave differently for different values of the variables, it is possible that a variable that has no major impact in one set of circumstances may well be significant in others. For this reason it is usually preferred to err on the side of a comprehensive model. In particular, quantities with prior calibrated values and associated uncertainties can usually be included in the model without introducing major complications at the model solving stage (see, for example, [7]).

Information about assumptions and status of the variables and parameters in a model can be recorded in the data sheet such as Table 2.1. If the design of the system changes the data sheet can be updated to reflect the changes.

Validation responses.

- Design review by an expert in the metrology field, reviewing the data sheet of model variables and checking assumptions.
- Numerical simulation comparing the behaviour of comprehensive models with simpler models.

Variable/Parameter	v_1	v_2	...	v_p
Nominal Value				
Range				
Measurement uncertainty				
Status				
Notes				

Table 2.1: Example data sheet recording assumptions about the variables and parameters in a model. Not all cells will be applicable to all variables or parameters.

2.1.2 Correctness of the functional model

Risks. The risks associated with an incorrect functional model are concerned with functionality: the model gives invalid predictions.

Often physical models are derived from models of subcomponents of the system which are aggregated, simplified or approximated. Errors can be introduced in a number of ways:

Typographical errors: errors in copying equations from source to where they are used;

Conceptualisation errors: incorrect understanding of the underlying physics;

Approximation errors: inappropriate simplifications, e.g., linearisations, significantly changing the behaviour of model;

Scoping errors: models used outside the scope of their original application.

Validation responses for physical models. Appropriate techniques for this aspect of model validation are:

- Design review by an expert in the metrology area to check the modelling of the underlying physics.
- Design review by a modelling expert to check the mathematical derivation of equations.
- Numerical simulations to check the effect of approximations, simplifications, linearisations, etc., on the model values (relative to the likely measurement error).
- Evaluation of the model at variable/parameter values for which the physical response is known accurately.
- Numerical simulations to check the qualitative behaviour of the model against expected behaviour. For example, if a response is expected to increase (or decrease) as a variable is increased, the model can be tested to verify that it exhibits this behaviour.

Empirical models are often used to represent observed or expected behaviour, for example, a calibration curve associated with a sensor's performance. Since empirical models are not derived from a physical theory their validation tends to focus on how well they represent measurement data; see Chapter 5. Although there may be no explicit physical model available, there is often concrete knowledge about the type of behaviour to be expected and it is important that the empirical model is capable of reproducing this. Empirical models such as polynomial curves have been used successfully for many years. Part of their success is due to the fact that they have proven approximation properties and can represent any smooth curve. However, they are not appropriate for modelling systems in which the response approaches a limit asymptotically (as in saturation or decay with time). For these situations, a sum of exponentials or a rational polynomial model would be more appropriate.

Validation responses for empirical models.

- Design review by a metrology and/or modelling expert to check the empirical model is appropriate for the expected type of behaviour.
- Numerical simulations to check the qualitative behaviour of the model against expected behaviour.

2.2 Validation of the statistical model

The functional model specifies the relationships between the model variables and parameters. It attempts to describe the actual behaviour of the physical system in terms of the true values of the variables and parameters. The statistical model of the system describes the relationship between the measured values of the variables and their true values. The functional model is to an extent independent of the measuring system and the same functional model can apply to different measuring systems. By constrast, the statistical model is dependent on the measuring system. For example, if we change a temperature sensor for a more accurate one, the functional model will remain unchanged but the statistical model will need to be modified.

If a typical functional model is of the form $y^* = \phi(\mathbf{x}^*, \mathbf{a})$ and describes the relationship between y^* and $\mathbf{x}^*$ in the absence of measurement error, a typical statistical model associated with measurements $X = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ is of the form

$$y_i = y_i^* + \epsilon_i, \quad \mathbf{x}_i = \mathbf{x}_i^*, \quad \epsilon_i \in N(0, \sigma^2),$$

which states that the variables $\mathbf{x}_i$ are measured accurately but that the measurements y_i are subject to independent, normally distributed measurement errors ϵ_i with standard deviation σ. Thus, a complete statement of the statistical model specifies (i) the error structure, i.e., which variables are subject to measurement error, (ii) the correlation structure of the measurement errors, i.e., which errors are statistically independent or otherwise, (iii) statistical distributions for the

Variable/Parameter	v_1	v_2	...	v_p
Nominal value				
Range				
Error distribution				
Statistical parameters				
Status				
Notes				

Table 2.2: Data sheet as in Table 2.1 expanded to record information about the functional *and* statistical model.

measurement errors and (iv) values of parameters associated with the distributions. While the derivation of the functional model can often rely on physical theory, the statistical model attempts to describe aspects of the system that are generally less well understood and, by their nature, inexact.

Risks. A poor statistical model can lead to:

- Poor choice of solution method.
- Invalid model solutions.
- Invalid uncertainties associated with the model solution.

While an incorrect functional model will often be detected when the model is fitted to measurement data, a poor statistical model can produce estimates of the uncertainty in the fitted parameters that are out by a factor of two or more without anything being obviously wrong. Since decisions about the development or purchase of an instrument are made on the basis of an uncertainty statement, the impact of a poor statistical model can be large.

Assumptions about the main components of the statistical model – error structure, correlation, distributions and their parameter values – can be recorded in the data sheet used to summarise the functional model (Table 2.2). Assumptions that need to be validated include the following.

Measurements of variables treated as exact. It is often assumed that only one variable (the response variable) is subject to measurement error. In practice, many variables may be subject to measurement error, some more significant than others. A decision to treat any one as exact has to be justified.

Measurement errors treated as having equal variance. While in many situations it is appropriate to assume that the errors associated with a set of measurements are drawn from the same distribution (and hence have the same variance), measurement errors often have a dependence on the magnitude of the variable. For example, if the measurement of y is subject to a relative error then the statistical model

$$y = y^*(1 + \epsilon), \quad \epsilon \in N(0, \sigma^2), \quad \text{or} \quad y = y^* + \epsilon, \quad \epsilon \in N(0, \sigma^2 y^2),$$

is appropriate. More generally, if estimates $y = y(\mathbf{x})$ are determined from measurements of variables $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ with associated covariance matrix $V_{\mathbf{x}}$, then the variance of y is estimated by

$$\sigma_y^2 = (\nabla_{\mathbf{x}} y)^{\mathrm{T}} V_{\mathbf{x}} (\nabla_{\mathbf{x}} y), \tag{2.1}$$

where $\nabla_{\mathbf{x}} y = (\frac{\partial y}{\partial x_1}, \ldots, \frac{\partial y}{\partial x_p})^{\mathrm{T}}$ is the vector of partial derivatives of y with respect to x_j.

Measurement errors treated as independent. The errors associated with a sequence of measurements by an individual sensor might well have a correlation structure and a time series analysis (or other methods) can be used to characterise this structure (see, e.g., [1]). Often, a more important (and more tractable) source of correlation is due to the dependence of estimates of variables on a common set of measurements. Suppose that estimates of y are determined from measurements of a variable x and constant x_0. (For example, y could be the length of a metal rod, x its temperature and x_0 the coefficient of thermal expansion for the metal.) If the measurements of x_i and x_0 are modelled as

$$x_i = x_i^* + \epsilon_i, \quad \epsilon_i \in N(0, \sigma^2), \qquad x_0 = x_0^* + \epsilon_0, \quad \epsilon_0 \in N(0, \sigma_0^2),$$

where ϵ_i and ϵ_0 represent normally distributed measurement error, then the covariance of y_i with y_j is estimated by

$$\mathrm{cov}(y_i, y_j) = \sigma_0^2 \frac{\partial y}{\partial x_0}(x_i, x_0) \frac{\partial y}{\partial x_0}(x_j, x_0).$$

This type of correlation is structural in the sense that it appears as a consequence of the error structure in the data and can be quantified and subsequently taken into account at the model solving stage.

Probability distributions assumed to be normal. The assumption of normality can usually be justified for a number of reasons including (from an empirical view point) historical information on the behaviour of a sensor and (from a theoretical view point) the Central Limit Theorem (see, e.g., [10, Chapter 5]). If the actual distribution is approximately symmetric, lightly tailed (vanishingly small probability far from the mean) and unimodal (has one peak), then an assumption of normality is usually safe. However, if it is known that the distribution is far from normal, then such an assumption could lead to an invalid model.

One area in which non-normality can be predicted is in the nonlinear transformation of a normally distributed variable. If $y = x_1 + x_2 t$ is a linear function of a variable t, and t is measured subject to error modelled as

$$t = t^* + \epsilon, \quad \epsilon \in N(0, \sigma_t^2),$$

then the statistical model for y is

$$y^* = x_1 + x_2 t^*, \quad y = y^* + \delta, \quad \delta \in N(0, x_2^2 \sigma_t^2).$$

This model is exact: if the variance of t is σ_t^2, then the variance of y is $(x_2 \sigma_t)^2$. However, if $y = y(t)$ is a *nonlinear* function of t then the variance σ_y^2 of y is

approximated by

$$\sigma_y^2 = \left(\frac{\partial y}{\partial t}\right)^2 \sigma_t^2. \tag{2.2}$$

This approximation will be good so long as the curvature of y at t is small compared with σ_t. This approximation needs to be validated as the following two examples show.

Figure 2.1 shows the curve $y = 1/t$ and 200 data points (t_i, y_i) generated according to the model

$$y_i = 1/t_i, \quad t_i = 3 + \epsilon_i, \quad \epsilon_i \in N(0, 1).$$

The estimate of the standard deviation of y_i derived from (2.2) is $\sigma_y = 1/9 \approx 0.11$ while the value obtained from the sample is 0.20. The estimate of the variance of the y_i given by (2.2) is derived using the approximating tangent line to the curve at (3,1/3). The linearised estimate of the y_i's are shown on the right hand vertical axis while the actual sample is shown on the left hand vertical axis. The linearisation fails to account for the nonlinear behaviour of the function.

Figure 2.2 presents the same type of information for the curve $y = t^2$ and 200 data points (t_i, y_i) generated according to the model

$$y_i = t_i^2, \quad t_i = 1 + \epsilon_i, \quad \epsilon_i \in N(0, 1).$$

In this case, the predictions derived from the linearisation of the model implicit in (2.2) are completely misleading.

Monte Carlo simulations (see Section 3.2 and [3]), such as those used to generate Figures 2.1 and 2.2, can be used to check the actual variation of a function $y(\mathbf{t})$ of measured quantities compared with those predicted from a formula such as (2.2) or the more general (2.1).

Validation responses for statistical models. Appropriate techniques for this aspect of model validation are:

- Design review by an expert in the metrology field, reviewing the data sheet for the statistical model for the measurement data and checking assumptions.
- Design review by modelling expert to check the statistical models for derived quantities.
- Numerical simulation to check the effect of approximations, simplifications, linearisations, etc., associated with the statistical model.
- Monte Carlo simulations to check the variation in derived quantities against the predicted variation.

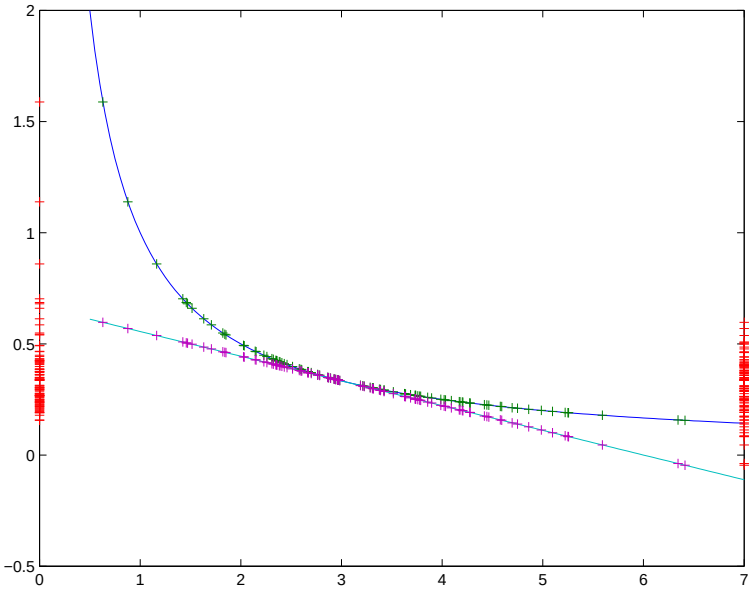

Figure 2.1: Actual (left hand vertical axis) and estimated variation (right hand vertical axis) of $y_i = 1/t_i$ where t_i are normally distributed.

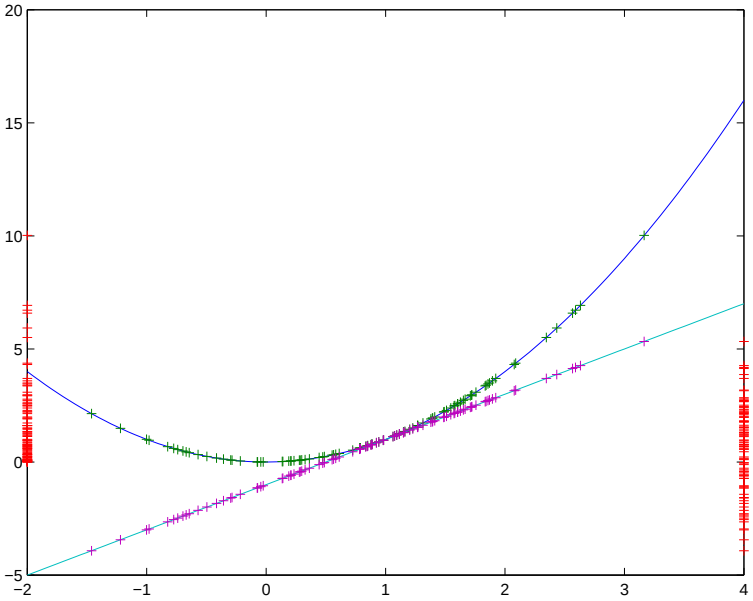

Figure 2.2: Actual (left hand vertical axis) and estimated variation (right hand vertical axis) of $y_i = t_i^2$ where t_i are normally distributed.

Chapter 3

Estimator validation

The previous chapter has been concerned with validating how the model has been built. This chapter addresses how the model is solved, i.e., the choice of estimator. It looks at the estimation of both parameters and their standard uncertainties.

Risks A poor choice of estimator can lead to

- Biased estimates of the model parameters.
- Overly optimistic or pessimistic estimates of the uncertainties in the fitted parameters.

The measurement results and their uncertainties are the outputs from the estimation process and the quality of these outputs is limited directly by the estimator. The functional and statistical models may characterise the measurement system comprehensively but this comprehensiveness will only be translated into the measurement results if the estimation process also embodies this characterisation. An (inefficient) estimator that makes poor use of the data will produce parameter estimates that have a larger uncertainty than warranted from the measurement errors in the data. A possible consequence of this is that, unnecessarily, additional measurements are undertaken or hardware upgrades are implemented in order to meet the system performance specification.

3.1 Estimator validation issues

Models of experimental systems can be complicated for a number of reasons. For example, the equations describing relationships between the variables and parameters may be quite involved or the error structure in the data may be highly correlated. For such systems, estimators which may be expected to perform well (from maximum likelihood principles) could well require the solution

of difficult optimisation problems. In these circumstances, we may wish or be forced to simplify the model and/or the optimisation problem in order to define an *approximate estimator* that is easier (or possible) to implement.

The following paragraphs outline typical situations in which estimator validation is required.

Approximations in the functional model. As considered in section 2.1, the functional model is often simplified to make it more tractable. Any such change to the functional model will have an effect on the estimates of the fitted parameters and consequently issues considered in the validation of the functional model will be directly relevant to the choice of estimator. However, two additional aspects of functional model validation are important at the model solving stage. Firstly, while at the model building stage, we may be quite comfortable with moderately complex functional models, when it comes to the practical step of implementing methods to solve the equations derived from these complex models a further degree of simplification may seem necessary over and above those already adopted.

Secondly, for well-conditioned, linear problems we can expect that a relative change in the functional model will have a similarly sized relative change in the model solution. For poorly conditioned and/or nonlinear problems the effect of an approximation of the functional model on the solution parameters is not so easy to predict and it may be that a seemingly insignificant change to the model in terms of the function values can have an unacceptably large change in the model solution. A linearised approximation of a nonlinear model can have a significantly different behaviour, qualitatively and quantitatively, from the nonlinear model. Figure 3.1 shows the linear approximation to an exponential model of the form $y = Ae^{-Bx}$. The linear model predicts that the response variable is zero at $x = 3$. For the exact model, the response is never zero.

Approximations in the statistical model. Similarly to the case of the functional model, at the model solving stage additional simplifications of the statistical model are made, either explicitly or implicitly, in order to produce a simpler computational problem to be solved. Examples include (i) measurements of variables treated as exact, (ii) input quantities treated as exact, (iii) errors in observation equations treated as independent, (iv) errors in observation equations treated as having equal variance and (v) normally distributed measurement errors.

The consequences of such approximations are in two areas. The first is that the information in the data is not used to maximum advantage and the resulting parameter estimates are not as good, in terms of their bias and variability. Secondly, for most estimators, the estimate of the uncertainty associated with the fitted model parameters is derived on the basis of a set statistical model for the data. If the statistical model does not apply, the uncertainty estimates could be completely invalid. While it is acceptable to apply a suboptimal estimator to data as long as the uncertainty calculations reflect the true statistical model and estimator behaviour, uncertainty estimates based on an approximate statistical model for any estimator can give misleading information.

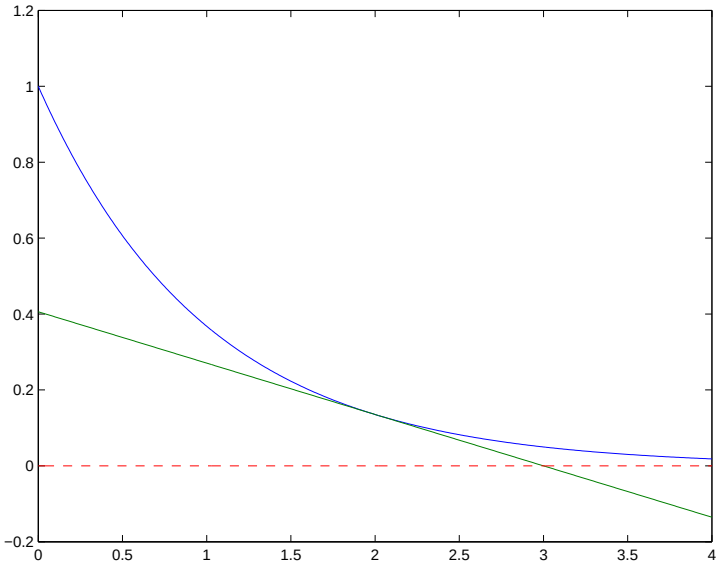

Figure 3.1: Linear approximation to an exponential model.

Approximate solution method. Approximations in the functional and statistical model are generally reflected in changes to the error function to be minimised. If an approximate solution method is employed to minimise the error function then, again, the parameter estimates can be biased and/or the uncertainty estimates invalid.

For these reasons, it is important that the choice of parameter estimation method is validated. We describe three classes of validation techniques (i) Monte Carlo simulation, (ii) null space benchmarking and (iii) estimator analysis.

3.2 Monte Carlo simulations

Monte Carlo simulation (see, e.g., [3]) is an important, general tool in modelling and analysis and can be used effectively in estimator validation. The general approach is as follows.

I Given a choice of model parameters $\mathbf{a}^* = (a_1, \ldots, a_n)^{\mathrm{T}}$, generate X^* satisfying the functional model.

II For $q = 1, 2, \ldots, N$,

II.1.q Generate perturbed data sets $X_q = X^* + \Delta_q$ where Δ_q is generated according to the statistical model using random number generators.

II.2.q Apply the estimator $\mathcal{A}$ to the data set X_q to determine an estimate

$\mathbf{a}_q = (a_{1,q}, \ldots, a_{n,q})^{\mathrm{T}}$ of the model parameters. Store $\mathbf{a}_q^{\mathrm{T}}$ as the qth row of an $N \times n$ matrix A.

III Compare statistics associated with A with $\mathbf{a}^*$.

In particular, we can form the mean estimate $\bar{\mathbf{a}}$ whose jth element $\bar{a}_j$ is the mean of the jth column of A and the sample covariance matrix

$$\bar{V}_{\mathbf{a}} = \frac{1}{N-1} \bar{A}^{\mathrm{T}} \bar{A} \tag{3.1}$$

where $\bar{A}$ is the mean-centred matrix defined by

$$\bar{A}_{ij} = A_{ij} - \bar{a}_j.$$

The difference

$$\beta_j = a_j^* - \bar{a}_j$$

between the mean estimate and the true value of the jth parameter is a measure of the bias of the estimator. The standard deviations s_j defined by

$$s_j^2 = \frac{1}{N-1} \sum_{q=1}^{N} (a_{j,q} - \bar{a}_j)^2$$

of the columns of A, that is, the square roots of the diagonal elements of $\bar{V}_{\mathbf{a}}$, are a measure of the variability of the estimator. A combined measure of estimator performance is given by the *mean squared error* (MSE)

$$\mathrm{MSE}_j = \frac{1}{N-1} \sum_{q=1}^{N} (a_{j,q} - a_j^*)^2, \tag{3.2}$$

or by the *root mean squared error*

$$\mathrm{RMSE}_j = (\mathrm{MSE}_j)^{1/2}, \tag{3.3}$$

for the jth parameter. We can also evaluate the model at $\bar{\mathbf{a}}$ or at any $\mathbf{a}_q$ and compare it with the model values at $\mathbf{a}^*$. The differences in these model values can be compared with the size of the perturbations Δ_q.

In Monte Carlo simulations, the functional model is only required to generate the exact data set X^*. This means that the behaviour of approximate estimators can be validated against data generated using only function evaluations of a comprehensive model.

As stressed in [3], the sample covariance matrix $\bar{V}_{\mathbf{a}}$ derived from the matrix A of parameter estimates $\mathbf{a}_q$ represents the covariance matrix of the *actual* distribution of the parameter estimates. No approximations or linearisations are involved in its calculation and it can be used to validate uncertainty estimates derived by other means.

The number of simulations performed is typically in the range 1000 to 100,000. However, a smaller number of simulations will often identify errors in implementations, for example. The Monte Carlo simulations can be repeated for a range of parameter values $\mathbf{a}^*$ and different measurement strategies to give a more complete picture of estimator behaviour and valuable information about the effect of measurement strategy on the parameter estimates (Chapter 5).

3.2.1 Monte Carlo simulation for a least squares estimator

See [4, Section 4.2].

To illustrate, we consider a (nonlinear) least squares approximation problem. Suppose the functional and statistical model is of the form

$$y^* = \phi(\mathbf{x}, \mathbf{a}), \quad y = y^* + \epsilon, \quad \epsilon \in N(0, \sigma^2).$$

Given a data set $X = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, least squares estimator $\mathcal{A}$ calculates (an estimate of) the solution of

$$\min_{\mathbf{a}} \sum_{i=1}^{m} (y_i - \phi(\mathbf{x}_i, \mathbf{a}))^2.$$

The estimator also estimates the covariance matrix of the fitted parameters according to

$$V_{\mathbf{a}} = \sigma^2 (J^{\mathrm{T}} J)^{-1}, \tag{3.4}$$

where J is the Jacobian matrix defined by

$$J_{ij} = \frac{\partial}{\partial a_j}(y_i - \phi(\mathbf{x}_i, \mathbf{a})), \tag{3.5}$$

evaluated at the solution.

This estimator is appropriate for the model (and in fact is a maximum likelihood estimator). However, if the function ϕ is highly nonlinear the estimator could be biased. Secondly, the estimates of the covariance matrix of the fitted parameters are based on a linearisation about the solution and could give misleading results. The aims of the Monte Carlo simulation is to estimate the bias in the parameter estimates and the validity of the uncertainty estimates.

Given a choice of $\mathbf{a}^*$ and values of the independent variables $\mathbf{x}_i$, $i = 1, \ldots, m$, the Monte Carlo simulation in this case is summarised by:

I Calculate $y_i^* = \phi(\mathbf{x}_i, \mathbf{a}^*)$ and set $X^* = \{(\mathbf{x}_i, y_i^*)\}_{i=1}^m$.

II For $q = 1, 2, \ldots, N$,

II.1.q Generate $y_{i,q} = y_i + \epsilon_{i,q}$, where $\epsilon_{i,q}$ is determined using a random number generator for the normal distribution and set $X_q = \{(\mathbf{x}_i, y_{i,q})\}_{i=1}^m$.

II.2.q Determine the least squares estimate $\mathbf{a}_q = \mathcal{A}(X_q)$ by solving

$$\min_{\mathbf{a}_q} \sum_{i=1}^{m} (y_{i,q} - \phi(\mathbf{x}_i, \mathbf{a}_q))^2,$$

and store $\mathbf{a}_q$ as the qth row of an $N \times n$ matrix A.

III Compare statistics associated with A with $\mathbf{a}^*$.

Let $\bar{\mathbf{a}}$ and $\bar{V}_{\mathbf{a}}$ be defined as above. The bias of the estimator can be measured by the difference between the mean parameter estimates and the true values

$$\boldsymbol{\beta} = \bar{\mathbf{a}} - \mathbf{a}^*,$$

and differences $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)^{\mathrm{T}}$ between the model predictions at $\bar{\mathbf{a}}$ and at $\mathbf{a}^*$ where

$$\delta_i = \phi(\mathbf{x}_i, \bar{\mathbf{a}}) - \phi(\mathbf{x}_i, \mathbf{a}^*).$$

The former can be compared directly with accuracy requirements on $\mathbf{a}$ while the latter can be compared with the size of the perturbations (simulating measurement errors) in terms of the standard deviation σ.

The sample covariance matrix $\bar{V}_{\mathbf{a}}$ can be used to validate the estimate $V_{\mathbf{a}}$ defined in (3.4).

3.3 Null space benchmarking

The Monte Carlo simulation techniques are a useful tool in validating estimator performance or in comparing one estimator with another. In this section we consider a more direct method of comparing one estimator with another, particularly in the case where we wish to compare a simpler, approximate estimator with a comprehensive, more complicated one. The main advantage of the approach is that an implementation of the comprehensive estimator is not required, but some analysis of the comprehensive estimator is required.

Suppose there are two estimators $\mathcal{A}$ and $\mathcal{B}$ associated with a measurement model with $\mathcal{A}$ representing an optimal data analysis we would ideally like to use and $\mathcal{B}$ an approximate estimator that is already implemented. We wish to know whether or not $\mathcal{B}$ is fit for purpose or if it will be necessary to implement $\mathcal{A}$.

The main steps are as follows.

I Given $\mathbf{a}$, determine a data set X (appropriate for the statistical model) for which $\mathbf{a} = \mathcal{A}(X)$, i.e., $\mathbf{a}$ is the solution supplied by estimator $\mathcal{A}$ for data set X. Often, the covariance matrix $V_{\mathbf{a}}$ for the parameters estimates $\mathbf{a}$ can also be calculated.

II Apply estimator $\mathcal{B}$ to X to determine estimates $\mathbf{b} = \mathcal{B}(X)$.

III Compare $\mathbf{b}$ with $\mathbf{a}$.

(IV) Compare the covariance matrix $V_{\mathbf{b}}$ of $\mathbf{b}$ with $V_{\mathbf{a}}$, if they are available.

The method can also be implemented in a Monte Carlo setting in which a number of data sets X_q are generated for which $\mathcal{A}(X_q) = \mathbf{a}$.

The key to the approach is step I, determining the data set X. We now describe ways to solve the problem for two important applications.

3.3.1 Null space method for generating data for least squares problems.

A common simplication in model fitting is to replace a nonlinear least squares problem with a linear least squares problem. In this section we show how to generate null space data for the nonlinear problem.

Suppose a nonlinear least squares estimation process requires the solution of

$$\min_{\mathbf{a}} \sum_{i=1}^{m} (y_i - \phi(\mathbf{x}_i, \mathbf{a}))^2. \tag{3.6}$$

The (first order) optimality conditions for $\mathbf{a}^*$ to be a local solution of (3.6) is that

$$J^{\mathrm{T}}\mathbf{f} = \mathbf{0}$$

where $J = J(\mathbf{a}^*)$ is the Jacobian matrix of partial derivatives defined by (3.5) evaluated at $\mathbf{a}^*$ and $\mathbf{f}(\mathbf{a}^*) = (f_1, \ldots, f_m)^{\mathrm{T}}$ with $f_i = y_i - \phi(\mathbf{x}_i, \mathbf{a}^*)$. This equation states that the vector of residual errors $\mathbf{f}$ lies in the *null space* of the matrix J^{T}.

The QR decomposition [4, Section 3.5], [8] can be used to generate an orthogonal basis of the null space of any $m \times n$, $m \geq n$, matrix C. Let

$$C = Q \begin{bmatrix} R \\ 0 \end{bmatrix}, \tag{3.7}$$

where Q is $m \times m$ orthogonal and R is $n \times n$ upper triangular. Let $Q_2 = \begin{bmatrix} \mathbf{q}_{n+1} & \mathbf{q}_{n+2} & \cdots & \mathbf{q}_m \end{bmatrix}$ be the submatrix of Q consisting of the last $m - n$ columns. If $\mathbf{z}$ is such that $C^{\mathrm{T}}\mathbf{z} = \mathbf{0}$ then $\mathbf{z}$ can be written uniquely as a linear combination of the columns of Q_2. That is, there exists a unique $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{m-n})^{\mathrm{T}}$ such that

$$\mathbf{z} = Q_2\boldsymbol{\nu} = \nu_1\mathbf{q}_{n+1} + \nu_2\mathbf{q}_{n+2} + \cdots + \nu_{m-n}\mathbf{q}_m, \tag{3.8}$$

and vice versa: any such linear combination $\mathbf{z}$ will be such that $C^{\mathrm{T}}\mathbf{z} = \mathbf{0}$. Since Q_2 is orthogonal the Euclidean norm of $\mathbf{z}$ in (3.8) is that of $\boldsymbol{\nu}$, i.e., $\|\mathbf{z}\| = \|\boldsymbol{\nu}\|$.

Given a choice of $\mathbf{a}^*$, σ and values of the independent variables $\mathbf{x}_i$, $i = 1, \ldots, m$, the null space method for generating reference data and results is summarised by

I Calculate $y_i^* = \phi(\mathbf{x}_i, \mathbf{a}^*)$ and set $X^* = \{(\mathbf{x}_i, y_i^*)\}_{i=1}^m$.

II Calculate the $m \times n$ Jacobian matrix J^*:

$$J_{ij}^* = \frac{\partial}{\partial a_j}(y_i^* - \phi(\mathbf{x}_i, \mathbf{a}^*)),$$

and an orthogonal basis for the null space $Q_2 = [\mathbf{q}_{n+1} \ldots \mathbf{q}_m]$ of ${J^*}^{\mathrm{T}}$. Generate $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{m-n})$ sampled from a normal distribution and normalise them so that

$$\|\boldsymbol{\nu}\|/(m-n)^{1/2} = \sigma.$$

III Set

$$\left.\begin{array}{rcl} \boldsymbol{\delta} & = & (\delta_1, \ldots, \delta_m)^{\mathrm{T}} = \sum_{k=1}^{m-n} \nu_k \mathbf{q}_{n+k}, \\ y_i & = & y_i^* + \delta_i, \\ X_{\boldsymbol{\delta}} & = & \{(\mathbf{x}_i, y_i)\}_{i=1}^m. \end{array}\right\} \tag{3.9}$$

Then if σ is small enough, $\mathbf{a}^*$ are the least squares best fit model parameters to $X_{\boldsymbol{\delta}}$ and $\boldsymbol{\delta}$ is the vector of residual errors

$$\delta_i = y_i - \phi(\mathbf{x}_i, \mathbf{a}^*)$$

and satisfies $\|\boldsymbol{\delta}\|/(m-n)^{1/2} = \sigma$. Furthermore,

$$V_{\mathbf{a}}^* = \sigma^2 (J^{*\mathrm{T}} J^*)^{-1},$$

is an estimate of the covariance matrix for the least squares estimator.

3.3.2 Null space benchmarking for generalised distance regression

The ordinary linear and nonlinear least squares estimators are appropriate if only one measured variable is subject to significant measurement error. However, in many metrological situations, more than one of the measured variables is subject to measurement error and it is important to take this into account in determining parameter estimates that are free from significant bias. A typical situation for which this is appropriate is where the response $y = \phi(x, \mathbf{a})$ is modelled as a function of the variable x and parameters $\mathbf{a}$ and both y and x are measured subject to measurement error, giving rise to a model of the form

$$y_i = \phi(x_i + \delta_i, \mathbf{a}) + \epsilon_i, \quad \delta_i \in N(0, \sigma_i^2), \quad \epsilon_i \in N(0, \rho_i^2).$$

The maximum likelihood estimate of the parameters is found by solving

$$\min_{\mathbf{a}, \{x_i^*\}} \sum_{i=1}^{m} \left\{ \alpha_i^2 (x_i - x_i^*)^2 + \beta_i^2 (y_i - \phi(x_i^*, \mathbf{a}))^2 \right\}, \tag{3.10}$$

with $\alpha_i = 1/\sigma_i$, $\beta_i = 1/\rho_i$. Note that this problem involves $\mathbf{a}$ and the *footpoint parameters* $\{x_i^*\}$ that specify the points on the curve closest to the data points in the appropriate metric. In [4], these types of problems are posed as generalisd distance regression (GDR) problems and effective methods for their solution described. However, these problems are more complex than their standard counterparts. In particular, even if ϕ is linear in the parameters $\mathbf{a}$ the optimisation problem (3.10) requires nonlinear iterative techniques for its solution. With this in view, given a problem of this type it may be worth investigating whether or not a generalised distance regression approach is required. Here we show how the null space method can be used to benchmark ordinary least squares estimators against a GDR estimator.

The GDR problem (3.10) is a nonlinear least squares problem for which the null space data generation approach described above can be applied. Let

$$\phi_i^* \quad = \quad \phi(x_i^*, \mathbf{a}^*),$$

$$\begin{aligned}\dot{\phi}_i &= \frac{\partial \phi_i}{\partial x}(x_i^*, \mathbf{a}^*),\\ s_i &= \left(\frac{\dot{\phi}_i^2}{\alpha_i^2} + \frac{1}{\beta_i^2}\right)^{1/2},\end{aligned}$$

$$d_i = \frac{1}{s_i}(-(x_i - x_i^*)\dot{\phi}_i + (y_i - \phi_i^*)), \tag{3.11}$$

and J^* the Jacobian matrix defined by

$$J_{ij}^* = -\frac{1}{s_i}\frac{\partial \phi_i}{\partial a_j}(x_i^*, \mathbf{a}^*).$$

If $\mathbf{a}^*$ and $\{x_i^*\}$ solve (3.10) for data $X = \{(x_i, y_i)\}_{i=1}^m$ then necessarily (i) there exists $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)^{\mathrm{T}}$ such that

$$x_i = x_i^* - \delta_i \frac{\dot{\phi}_i}{\alpha_i^2}, \quad y_i = \phi_i^* + \delta_i \frac{1}{\beta_i^2}, \tag{3.12}$$

and (ii)

$$J^{*\mathrm{T}}\mathbf{d} = \mathbf{0},$$

i.e.,

$$\sum_{i=1}^{m} \frac{1}{s_i}\frac{\partial \phi_i}{\partial a_j} d_i = 0, \quad j = 1, \ldots, n. \tag{3.13}$$

Substituting (3.12) in (3.11) we obtain

$$d_i = \delta_i s_i$$

so that the conditions (3.13) on $\mathbf{d}$ become

$$A^{\mathrm{T}}\boldsymbol{\delta} = \mathbf{0},$$

where $A_{ij} = \frac{\partial \phi_i}{\partial a_j}$. The perturbations δ_i can be scaled so that the estimate of the standard deviation of the residuals d_i is a pre-assigned value σ.

Taking into account this analysis, null space data can be generated according to the following scheme:

Given $\mathbf{a}^*$, σ, $\{x_i^*\}$, $\{\alpha_i\}$ and $\{\beta_i\}$,

I Set $y_i^* = \phi(x_i^*, \mathbf{a}^*)$ and $X^* = \{(x_i^*, y_i^*)\}$.

II For each i, calculate

$$\dot{\phi}_i = \frac{\partial \phi}{\partial x}(x_i^*, \mathbf{a}^*), \quad s_i = \left(\frac{\dot{\phi}_i^2}{\alpha_i^2} + \frac{1}{\beta_i^2}\right)^{1/2}.$$

III Calculate A given by

$$A_{ij} = \frac{\partial \phi_i}{\partial a_j}.$$

IV Determine $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)^{\mathrm{T}}$ such that $A^{\mathrm{T}}\boldsymbol{\delta} = 0$ (using the QR factorisation of A, for example). For each i, set

$$p_i = -\delta_i \frac{\dot{\phi}_i}{\alpha_i^2}, \quad q_i = \delta_i \frac{1}{\beta_i^2}.$$

and calculate $S = \left\{\sum_{i=1}^m (\alpha_i^2 p_i^2 + \beta_i^2 q_i^2)\right\}^{1/2}$ and $K = (m-n)^{1/2}\sigma/S$.

V For each i, set

$$x_i = x_i^* + Kp_i, \quad y_i = y_i^* + Kq_i.$$

Then $\mathbf{a}^*$ and $\{x_i^*\}$ solves (3.10) for dataset $X = \{(x_i, y_i)\}_{i=1}^m$ with

$$\frac{1}{m-n}\sum_{i=1}^m \left\{\alpha_i^2(x_i - x_i^*)^2 + \beta_i^2(y_i - \phi(x_i^*, \mathbf{a}))^2\right\} = \sigma^2.$$

Furthermore, the estimate $V_{\mathbf{a}}$ of the covariance matrix of the fitted parameters $\mathbf{a}^*$ for *the GDR estimator* is

$$V_{\mathbf{a}} = \sigma^2 (J^{\mathrm{T}} J)^{-1}, \tag{3.14}$$

where the Jacobian matrix J can be calculated from $J_{ij} = A_{ij}/s_i$. This null space approach allows us to generate data sets for which the GDR solution estimates are known along with their uncertainties without having to implement the estimator. All that is required are the evaluation ϕ, $\frac{\partial \phi}{\partial x}$ and $\frac{\partial \phi}{\partial a_j}$. We can use this information to benchmark the performance of an approximate estimator both in terms of bias and variability against the GDR estimator.

3.4 Estimator analysis

In sections 3.2 and 3.3 we have described quantitative methods for validating estimators. The main aim has been to validate the performance of an estimator given a statement of the functional and statistical model. In this section, we look at a more qualitative approach in which we start from the definition of an estimator and then consider for which class of model it is (optimally) appropriate. We can then assess if it is valid for a particular estimation problem on the basis of how well matched the model is to the specified class.

3.4.1 Analysis of least squares methods

Least squares methods are the most common estimators implemented and are appropriate for many practical model fitting problems. For linear models the following *Gauss-Markov Theorem* [9, Chapter 6] can be used to justify their use:

Gauss-Markov Theorem *For models of the form*

$$\mathbf{y} = C\mathbf{a} + \boldsymbol{\epsilon},$$

where C is an $m \times n$ full rank matrix, $m \geq n$, and for which the measurements errors $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_m)^{\mathrm{T}}$ are uncorrelated with each other and each has variance σ^2, the linear least squares estimator

$$\mathbf{a} = (C^{\mathrm{T}}C)^{-1}C^{\mathrm{T}}\mathbf{y}$$

is unbiased and has a smaller covariance matrix than that of any other linear estimator.

From this point of view, least squares estimation is optimal for these models.

Note that there is no assumption that the measurement errors are normally or even symmetrically distributed, only that they are uncorrelated and have equal variance. This generality supports the use of least squares methods. Assumptions about normality are usually only invoked when it is required to provide confidence limits associated with the fitted parameters. The theorem does not apply to nonlinear models. However, if the model is only mildly nonlinear then it can be expected that nonlinear estimation techniques will also have favourable properties. The theorem does not apply to models of the form $y = \phi(\mathbf{x}, \mathbf{a})$ if one or more of the variables x_j are subject to significant measurement error.

Whenever a least squares estimation method is proposed, it is a good idea to examine the extent to which the prior conditions of the theorem concerning linearity, correlation and variance hold. In fact, least squares methods can be adapted to deal with unequal variances, known correlation, etc., [4, Section 4] and it is important to make sure the least squares method is tuned as far as possible to the model. Often a least squares method is implemented without sufficient consideration.

Validation responses

- Monte Carlo simulations to examine the bias and variation of the solution estimates on data sets generated according to the statistical model.
- Monte Carlo simulations to compare the predicted variation of parameter estimates with the actual variation on data sets generated according to the statistical model.
- Apply the estimator to data sets for which the estimates provided by an optimal estimator are known.
- Compare the actual variation of parameter estimates on data sets generated according to the statistical model with the predicted variation for an optimal estimator.
- Compare the actual statistical model with the statistical model for which the estimator is known to perform well.

Chapter 4

Validation of model solutions

The validation of estimators considered in the previous chapter was concerned with the process of how a model solution is determined. In this chapter we are concerned with the question of validating specific model solutions to data. There are two aspects to this question, firstly, how well does the model represent the data and secondly how well does the data specify the model. This second aspect is considered further in Chapter 5.

4.1 Goodness of fit and residual errors

In solving a system of equations $C\mathbf{a} = \mathbf{y}$ where C is a square $n \times n$ matrix, a solution $\mathbf{a}$ can be validated by calculating how well it satisfies the equations. If the matrix C is full rank then (in exact arithmetic) the residual error vector $\mathbf{r} = \mathbf{y} - C\mathbf{a}$ should be identically zero. For over-determined systems in which there are more equations than parameters, not all the equations can be satisfied exactly in general and the residual error vector will be nonzero. A measure of the goodness of fit is derived from a norm of $\mathbf{r}$, e.g., the Euclidean norm

$$\|\mathbf{r}\|_2 = \left(\sum_i r_i^2 \right)^{1/2},$$

or the Chebyshev norm

$$\|\mathbf{r}\|_\infty = \max_i |r_i|.$$

The central question in determining whether a specific model solution is a valid fit to data relates to what is the expected behaviour of the residual error vector: the more sophisticated the model, the more sophisticated the validation criteria.

4.1.1 Signal to noise ratio

A minimal model associated with a set of data specifies only the likely signal to noise ratio, for example, in a statement of the form "the data is likely to be accurate to 1 part in 10^4". Suppose $X = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ represents the measurements of a response variable y_i corresponding to values $\mathbf{x}_i$ for the independent variables. A model of the form

$$y = \phi(\mathbf{x}, \mathbf{a})$$

is fitted to the data and the residual errors $r_i = y_i - \phi(\mathbf{x}_i, \mathbf{a})$ calculated. The ratio

$$R^2 = \frac{\left(\sum_i y_i^2\right) - \left(\sum_i r_i^2\right)}{\sum_i y_i^2}$$

is a measure of how much of the variation of the response y is accounted for by the model ϕ. A value of 1 indicates that all of the variation is explained by the model; value of 0 indicates that none has been accounted for. Related measures are discussed in [2].

4.1.2 Statistical model for the data

A model fit can be validated against a statistical model for the data. For example, suppose the model for data $X = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ is that the variables $\mathbf{x}$ are measured accurately but the measurements y_i of the response variable are made subject to independent measurement errors ϵ_i randomly drawn from a probability distribution with standard deviation σ:

$$y_i = y_i^* + \epsilon_i, \quad \text{var}(\epsilon_i) = \sigma^2.$$

Suppose a model of the form

$$y = \phi(\mathbf{x}, \mathbf{a})$$

specified by parameters $\mathbf{a} = (a_1, \ldots, a_n)^{\mathrm{T}}$ is fitted to the data and the residual errors $r_i = y_i - \phi(\mathbf{x}_i, \mathbf{a})$ calculated, along with the root-mean-square (RMS) error

$$\hat{\sigma} = \|\mathbf{r}\|/(m-n)^{1/2},$$

an estimate of σ derived from the residual errors. The model makes two predictions i) the expected value of $\hat{\sigma}$ is σ and ii) the residual errors should be independent of each other.

The values of $\hat{\sigma}$ and σ can be compared using the χ^2-distribution. If $\epsilon_i \in N(0,1)$ are p independent, normally distributed random variables then

$$\sum_{i=1}^{p} \epsilon_i^2 \in \chi_p^2$$

has a χ_p^2-distribution with p degrees of freedom. This distribution has an expected value of p and a variance of $2p$. Distributions for 2, 5 and 10 degrees of freedom are plotted in Figure 4.1. For small values of p, the distribution is

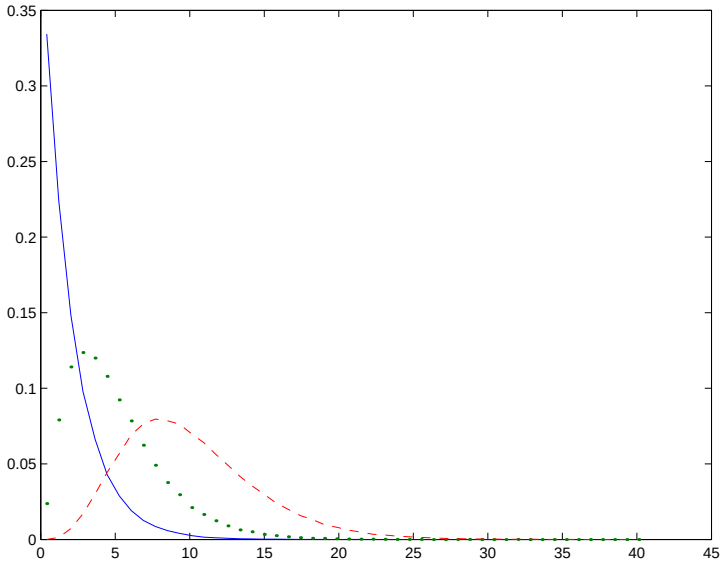

Figure 4.1: χ^2 distributions with 2 (solid), 5 (dotted) and 10 (dashed) degrees of freedom.

significantly assymmetrical. For large values, the distribution is approximated by $N(p, 2p)$, i.e., the normal distribution with mean p and variance $2p$.

If the model is valid and the distribution for the measurement errors is approximately normal, then the distribution for the sum of the squares of normalised residual errors

$$\frac{1}{\sigma^2}\sum_i r_i^2$$

is approximated by a χ^2-distribution with $m - n$ degrees of freedom. For large numbers of degrees of freedom $p = m-n$, say $p \geq 50$, the difference $(\hat{\sigma}^2-\sigma^2)/\sigma^2$ is expected to be smaller in magnitude than $2(2/p)^{1/2}$ at approximately the 95% confidence level. For small numbers of degrees of freedom it is better to compare this ratio with the 2.5% and 97.5% points of the appropriate χ^2-distribution.

The randomness of the residual errors can be tested most simply by visual inspection. For small numbers of residuals, it is usually quite difficult to show that there is a nonrandom behaviour. For larger data sets, systematic behaviour is often quite obvious.

4.1.3 Residual errors and model selection

The RMS residual error can be used to select a model from a range of model solutions, for example, in choosing the degree of a polynomial fit. To illustrate

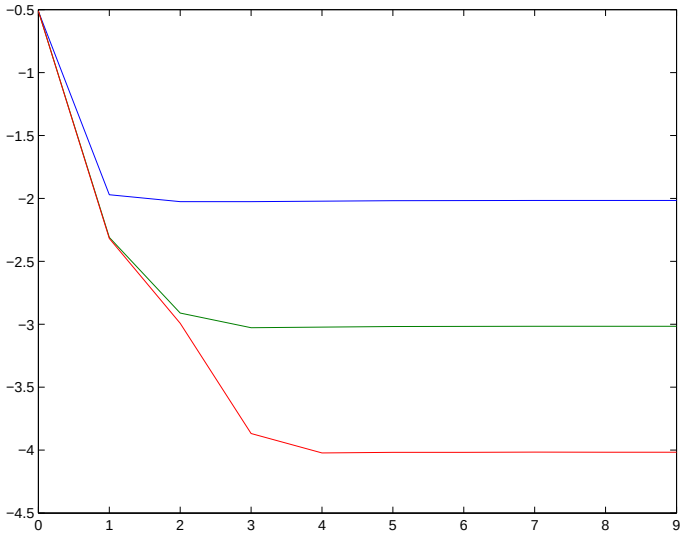

Figure 4.2: Graph of $\log \hat{\sigma}_n$ of the RMS residual error against order of polynomial fit for data generated according to the model (4.1) with $\sigma = 0.01$, 0.001 and 0.0001.

this, data (x_i, y_i) has been generated according to the model

$$y^* = x + 0.01x^2 + 0.01x^3 + 0.02x^4, \quad y = y^* + \epsilon, \quad \epsilon \in N(0, \sigma^2), \tag{4.1}$$

for x in the interval $[0,\ 1]$ and three values of $\sigma = 0.01,\ 0.001$ and 0.0001. For each value of σ, polynomials of degree 0 to 9 were fitted to the data and $\hat{\sigma}_n$ recorded for each degree n.

Figure 4.2 is the log plot of $\hat{\sigma}_n$ against degree n for each of the three values of σ. For $\sigma = 0.01$ (upper curve), the RMS plot shows that increasing the degree from 2 to 3 reduces the RMS by a small amount but increasing the degree further gives no improvement. Similarly for $\sigma = 0.001$, there is an improvement from degree 3 to 4 but none beyond. On this basis we would chose degrees 2, 3 and 4 for $\sigma = 0.01$, 0.001 and 0.0001, respectively. In each case the saturation level is $\log_{10} \sigma$. This agrees with the plots of the residual errors associated with the degree 2 (upper graph) and 3 (lower graph) fits shown in Figure 4.3. The residual errors associated with the degree 2 fit show systematic cubic behaviour while those for degree 3 appear random. A similar picture is shown in Figure 4.4 for the case $\sigma = 0.0001$ in which the residuals associated with the degree 3 fit show a quartic (W-shaped) systematic behaviour while the degree 4 residuals appear random.

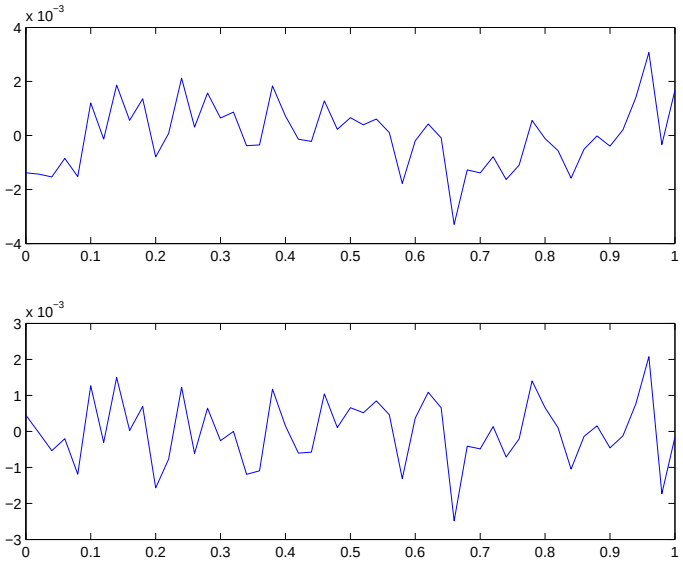

Figure 4.3: Graphs of residual error for degree 2 (upper graph) and degree 3 (lower graph) polynomial fits to data generated with $\sigma = 0.001$.

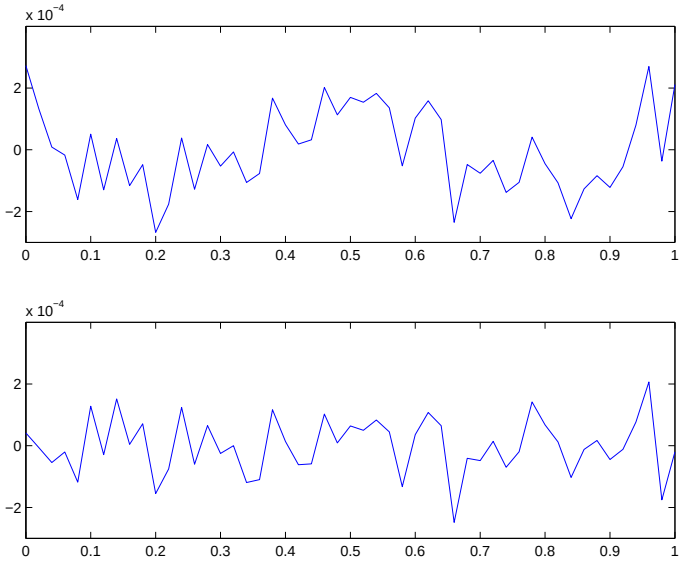

Figure 4.4: Graphs of residual error for degree 3 (upper graph) and degree 4 (lower graph) polynomial fits to data generated with $\sigma = 0.0001$.

4.1.4 Validation of model outputs

So far in this chapter we have only considered the fit of a model solution to data and its validation in terms of the goodness of fit against the statistical model for the data; we have as yet paid no attention to where the model came from or to any other characteristics of the model. However, an estimation process will generally produce (and is usual required to produce) estimates of the uncertainty in the fitted parameters and related statistical information. If this extra information has been derived in a valid way, i.e., the model and the solution process are consistent with each other, then it provides additional model outputs that can be checked against actual behaviour.

We will examine these issues in more detail for the important case of the least squares fit of linear models of the form

$$y^* = \phi(\mathbf{x}, \mathbf{a}) = \sum_{j=1}^{n} a_j \phi_j(\mathbf{x}), \quad y = y^* + \epsilon, \quad \text{var}(\epsilon) = \sigma^2.$$

Given data $X = \{(\mathbf{x}_i, y_i)\}_1^m$, $m > n$, let C be the observation matrix of basis functions ϕ_j evaluated at $\mathbf{x}_i$, i.e.,

$$C_{ij} = \phi_j(\mathbf{x}_i).$$

It is assumed that C has full column rank. The least squares model solution is determined by solving the overdetermined system of equations $C\mathbf{a} = \mathbf{y}$ in the least squares sense. The model outputs are

- Estimates of the solution parameters $\mathbf{a} = (C^{\mathrm{T}}C)^{-1}C^{\mathrm{T}}\mathbf{y}$.
- The model predictions $\hat{\mathbf{y}} = C\mathbf{a} = C(C^{\mathrm{T}}C)^{-1}C^{\mathrm{T}}\mathbf{y}$, i.e., the predicted responses $\hat{y}_i$ at values $\mathbf{x}_i$ of the independent variables.
- The residual error vector

  $$\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - C\mathbf{a} = (I - C(C^{\mathrm{T}}C)^{-1}C^{\mathrm{T}})\mathbf{y}$$

  where I is the $m \times m$ identity matrix.
- The estimate of the standard deviation of the residual errors

  $$\hat{\sigma} = \|\mathbf{r}\|/(m-n)^{1/2}.$$

- The covariance matrix of the fitted parameters. If a prior estimate of σ is available

  $$V_{\mathbf{a}} = \sigma^2(C^{\mathrm{T}}C)^{-1},$$

  otherwise, V can be estimated from

  $$V_{\mathbf{a}} = \hat{\sigma^2}(C^{\mathrm{T}}C)^{-1}.$$

- The standard uncertainties of the fitted parameters $u(a_j) = (V_{\mathbf{a}}(j,j))^{1/2}$, i.e., the square roots of the diagonal elements of the covariance matrix $V_{\mathbf{a}}$.

- The correlation matrix associated with the fitted parameters defined by

$$C_R(i,j) = \frac{V_{\mathbf{a}}(i,j)}{(V_{\mathbf{a}}(i,i)V_{\mathbf{a}}(j,j))^{1/2}}.$$

  Note that C_R is independent of the value of σ used to define the covariance matrix.

- The covariance matrix $V_{\hat{\mathbf{y}}}$ for the model predictions $\hat{\mathbf{y}}$

$$V_{\hat{\mathbf{y}}} = CV_{\mathbf{a}}C^{\mathrm{T}} = \sigma^2 C(C^{\mathrm{T}}C)^{-1}C^{\mathrm{T}}.$$

- The standard uncertainties of the model predictions $u(\hat{y}_i) = (V_{\hat{\mathbf{y}}}(i,i))^{1/2}$.

- The covariance matrix $V_{\mathbf{r}}$ of the residual errors

$$V_{\mathbf{r}} = \sigma^2(I - C(C^{\mathrm{T}}C)^{-1}C^{\mathrm{T}}).$$

- The standard uncertainties of the residual errors $u(r_i) = (V_{\mathbf{r}}(i,i))^{1/2}$.

- If $(\mathbf{z}, w)$ represents a new data point (generated from the same model but not used in defining the model fit) then the predicted model value at $\mathbf{z}$ is

$$\hat{w} = \phi(\mathbf{z}, \mathbf{a}) = \mathbf{d}^{\mathrm{T}}\mathbf{a},$$

  where $\mathbf{d} = (d_1, \ldots, d_n)^T = (\phi_1(\mathbf{z}, \mathbf{a}), \ldots, \phi_n(\mathbf{z}, \mathbf{a}))^{\mathrm{T}}$, the standard uncertainty of $\hat{w}$ is

$$u(\hat{w}) = (\mathbf{d}^{\mathrm{T}}V_{\mathbf{a}}\mathbf{d})^{1/2},$$

  the predicted residual error is $t = w - \hat{w} = w - \mathbf{d}^{\mathrm{T}}\mathbf{a}$ and its variance is

$$\mathrm{var}(t) = \sigma^2 + \mathbf{d}^{\mathrm{T}}V_{\mathbf{a}}\mathbf{d}.$$

  More generally, if $Z = \{\mathbf{z}_q\}_{q=1}^{m_Z}$ is a range of values for the independent variables and D is the corresponding matrix of basis functions evaluated at $\mathbf{z}_q$, i.e.,

$$D_{q,j} = \phi_j(\mathbf{z}_q),$$

  then the covariance matrix $V_{\mathbf{w}}$ for the model values $\mathbf{w} = (w_1, \ldots, w_{m_Z})^{\mathrm{T}}$, $w_q = \phi(\mathbf{z}_q, \mathbf{a})$, is estimated by

$$V_{\mathbf{w}} = DV_{\mathbf{a}}D^{\mathrm{T}},$$

  and the standard uncertainties $u(w_q)$ are estimated by

$$u(w_q) = \left(V_{\mathbf{w}}(q,q)\right)^{1/2}.$$

We note that if the observation matrix has QR factorisation $C = QR$ where Q is an $m \times n$ orthogonal matrix and R is an $n \times n$ upper triangular matrix and singular value decomposition (SVD) $C = USV^{\mathrm{T}}$ where U is an $m \times n$ othogonal matrix, S is $n \times n$ diagonal matrix and V is $n \times n$ orthogonal matrix, then

$$\begin{aligned}
C^TC &= R^{\mathrm{T}}R = VS^2V^{\mathrm{T}}, \\
(C^TC)^{-1} &= R^{-1}R^{-\mathrm{T}} = VS^{-2}V^{\mathrm{T}}, \\
(C^{\mathrm{T}}C)^{-1}C^{\mathrm{T}} &= R^{-1}Q^{\mathrm{T}} = VS^{-1}U^{\mathrm{T}}, \text{ and} \\
C(C^{\mathrm{T}}C)^{-1}C^{\mathrm{T}} &= QQ^{\mathrm{T}} = UU^{\mathrm{T}}.
\end{aligned}$$

These relations show that all the model outputs listed above can be calculated from C's QR factorisation or SVD. All the statistical information can be derived from $V_{\mathbf{a}}$.

If the functional and statistical model is valid then all the model predictions are valid. Conversely, if any one of these outputs is inconsistent with experimental data then there must be some aspect of the model that is invalid, putting into question the validity of all other model predictions. Each of the model predictions can be used to examine aspects of the model validity.

The residual error vector and the estimate $\hat{\sigma}$ of the standard deviation of the residual errors. Their role in model validation has already been discussed above. If there is only a small number of degrees of freedom associated with the system of equations then the expected variation in $\hat{\sigma}$ is relatively large and any statistics depending on its value are also subject to relatively large variation. This is because there is only limited information with which to validate the model. If only a small amount of data is available then more emphasis should be placed in validating the model inputs (assumptions) and its internal consistency.

If an experiment is repeated a number of times providing systems of equations $C_k\mathbf{a}_k = \mathbf{y}_k$, where C_k is $m_k \times n$, $m_k > n$, residual error vectors $\mathbf{r}_k = \mathbf{y}_k - C_k\mathbf{a}_k$ and estimates $\hat{\sigma}_k$, $k = 1, \ldots, K$, then a combined estimate of the standard deviation $\hat{\sigma}$ is given by

$$\hat{\sigma}^2 = \frac{1}{m - Kn} \sum_k \|\mathbf{r}_k\|^2 = \frac{1}{m - Kn} \sum_k (m_k - n)\hat{\sigma}_k^2.$$

where $m = \sum_k m_k$ is the total number of equations. This value can be subsequently used in the uncertainty estimation associated with the individual model fits.

The value of $\hat{\sigma}$ is also a useful measure of the quality of the data and a value much larger than expected can indicate invalid input data. For this purpose, the maximum absolute residual error $r_{max} = \max_i |r_i|$ should also be computed.

Covariance matrix and standard uncertainties of the fitted parameters $\mathbf{a}$. The statistical information associated with the fitted parameters is a key component of the model predictions. We know that, due to the stochastic nature of experimentation, if we repeat an experiment we will obtain different values for the fitted parameters. The covariance matrix indicates the likely spread of parameter estimates. So, for example, if an experiment is performed twice producing estimates $\mathbf{a}_k = (a_{k,1}, \ldots, a_{k,n})^{\mathrm{T}}$, covariance matrices $V_{\mathbf{a}_k}$ and standard uncertainties $u_k(a_j)$, $k = 1, 2$, then, assuming a valid model, the covariance matrix of the difference $\mathbf{a}_1 - \mathbf{a}_2$ is $V_{\mathbf{a}_1} + V_{\mathbf{a}_2}$ and the standard uncertainties of individual parameter differences are

$$u(a_{1,j} - a_{2,j}) = \left(u_1^2(a_j) + u_2^2(a_j)\right)^{1/2}.$$

The actual differences can be compared with these standard uncertainties (taking into account the number of degrees of freedom). For a large number of degrees of freedom, a difference in parameter estimates more than, say, three times the standard uncertainty will indicate some aspect of the model could be invalid.

The covariance matrix also provides direct information about the effectiveness of the experimental design (see Chapter 5). At a very minimum, the computation of and check on the standard uncertainties of the fitted parameters is required to verify (or otherwise) that the information in the data is sufficient to determine reliable estimates of the model parameters.

Covariance matrix and standard uncertainties of the model predictions $\hat{\mathbf{y}}$. The standard uncertainties $u(\hat{y}_i)$ of the model values $\hat{y}_i = \phi(\mathbf{x}_i, \mathbf{a})$ indicate how well the model is defined at $\mathbf{x}_i$. In many ways, the uncertainty in the model predictions is the key output of the modelling process. We wish to understand a physical system. We build a model and gather a set of measurements $X = \{(\mathbf{x}_i, y_i)\}$ to characterise the model. From this information, we hope to predict the behaviour of the system at any set of values of the independent variables $\mathbf{x}$. The standard uncertainties of the model predictions indicate how well the model has been characterised and provide a measure of the quality of the predictions.

For example, suppose we wish to characterise the response of a system $y = \phi(x, \mathbf{a})$ for x in the range $[0,\ 1]$. Measurements are made at values of the independent variable at $\mathbf{x} = (0.1, 0.2, \ldots, 0.9)^{\mathrm{T}}$. Measurements $\mathbf{y} = (y_1, \ldots, y_9)^{\mathrm{T}}$ are gathered subject to measurement error with variance σ^2. The standard uncertainty of the model predictions $u(w_q)$ are then calculated at values $\mathbf{z} = (0.00, 0.01, 0.02, \ldots, 1.00)^{\mathrm{T}}$. If these standard uncertainties are sufficiently small relative to σ, then the system is adequately characterised.

Figure 4.5 (on page 36) graphs the residual errors $r_i = y_i - \phi(x_i, \mathbf{a})$ associated with the fit of a quartic polynomial to data generated according to model (4.1) along with values $\pm 2u(w_q)$ and the band $\pm 2\sigma$. The solid curves indicate that in the range $0.1 \leq z \leq 0.9$ the uncertainty in the model fit $u(w)$ is less than the uncertainty in the measurements but outside this range the uncertainty associated with the model fit increases (quartically). The figure also graphs the model uncertainty band for a quadratic fit to the data and shows that these uncertainties are significantly larger than the measurement uncertainty indicating that a quadratic model is an inadequate representation of the system.

The uncertainty in the model predictions can be used to differentiate between two competing models. Figure 4.6 shows data generated according to the model

$$y^* = Ae^{-Bx}\sin(Cx + D), \quad y = y^* + \epsilon, \quad \epsilon \in N(0, \sigma^2)$$

with $\sigma = 0.001$. The data is fitted with two models, the physical, harmonic oscillator model used to generate the data and an empirical, polynomial model.

Figure 4.7 graphs the log of the RMS residual error $\hat{\sigma}_n$ against order n (= degree + 1) of the polynomial. On the basis of this a polynomial of degree 24 was chosen to represent the data.

Figure 4.8 shows the residual errors, the model prediction uncertainty band, i.e., $\pm 2u(\hat{y})$ and the $\pm 2\sigma$ band for the physical model fit. The figure shows that the model is a good representation of the data and is well characterised over the range, including the interval in the middle where there is no data.

Figure 4.9 presents the same information for the polynomial fit. While providing a good representation of the data, the polynomial model is poorly characterised in the middle interval where the data is missing.

Figure 4.10 illustrates why this is so. It shows two degree 24 polynomials which both fit the data well (visually) but differ significantly in the middle interval.

Validation responses.

- Goodness of fit in terms of the size of the residual errors.
- Size of the residual errors with statistical model for the measurement errors.
- Plot of the residual errors to check for random/systematic behaviour.
- Plot of the root-mean-square residual error for a number of model fits to select an appropriate model (e.g., the polynomial of appropriate degree).
- Calculate and check the covariance matrix for the fitted parameters against requirements and/or expected behaviour.
- Calculate and check the uncertainty associated with the model predictions againsts requirements and/or expected behaviour.

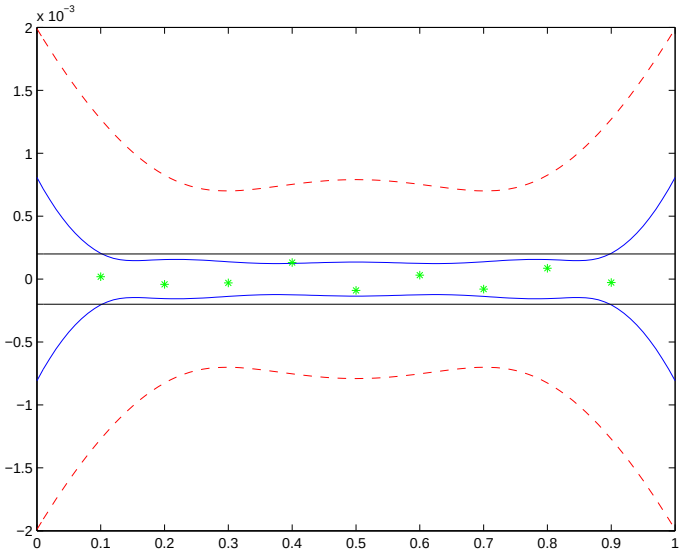

Figure 4.5: Uncertainty band for the model predictions associated with a polynomial of degree 4 (solid curves) and degree 2 (dashed curves) to data generated according to the model (4.1) compared with the uncertainty in the measured data. The residual errors '*' associated with the degree 4 fit are also plotted.

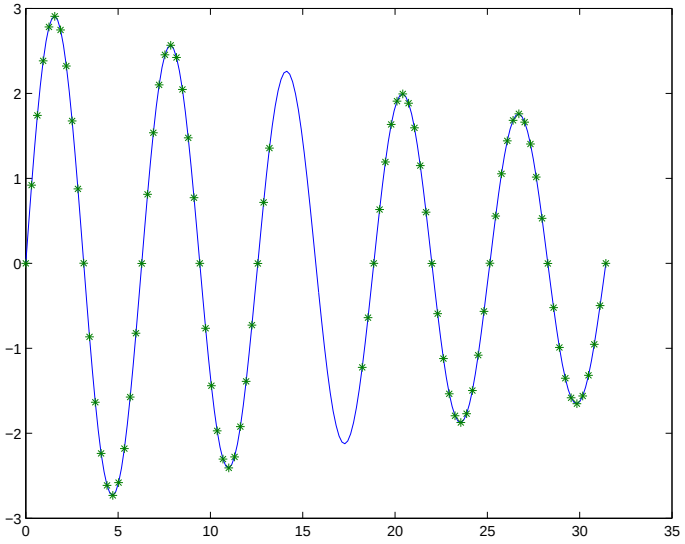

Figure 4.6: Data generated according to a harmonic oscillation model.

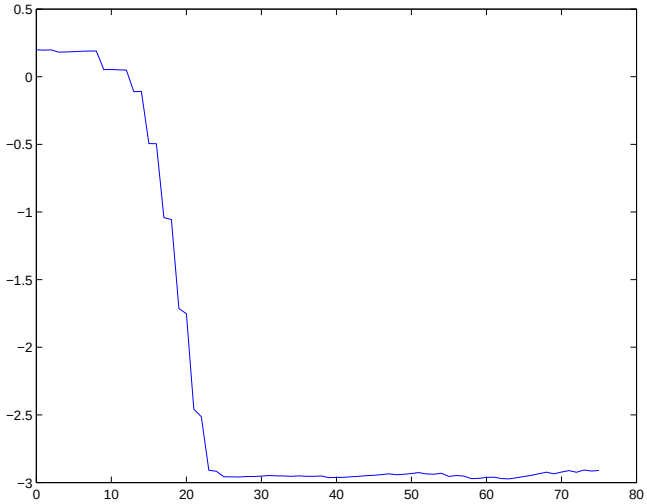

Figure 4.7: Graph of the log of the RMS residual error $\hat{\sigma}_n$ associated with a polynomial fit against order n (= degree + 1) to data generated according to a harmonic oscillator model (Figure 4.6).

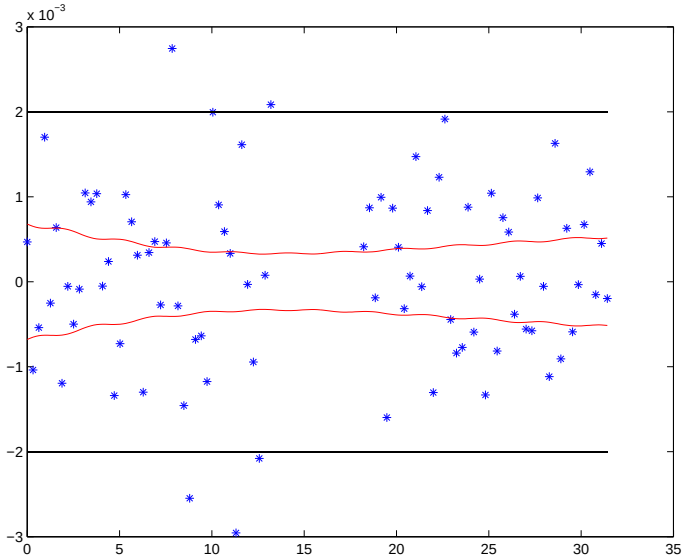

Figure 4.8: Residuals and uncertainty band associated with the fit of a physical model to data generated according to harmonic oscillator model (Figure 4.6).

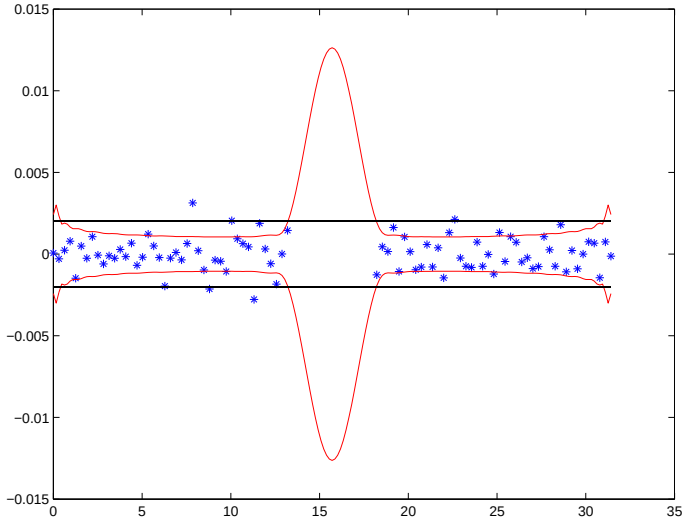

Figure 4.9: Residuals and uncertainty band associated with the fit of a polynomial model of degree 24 to data generated according to harmonic oscillator model (Figure 4.6).

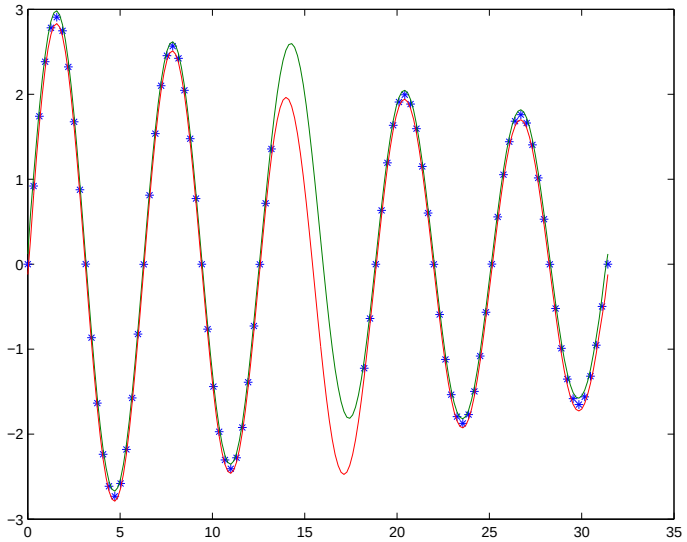

Figure 4.10: Two degree 24 polynomial fits to data generated according to harmonic oscillator model (Figure 4.6).

Chapter 5

Validation of experimental design and measurement strategy

Once the functional and statistical models and estimation method have been validated, it is possible to apply the estimator to examine issues concerning experimental design and measurement strategy. In the sections below we consider (i) system identifiability – the extent to which all the parameters can be identified from the data and (ii) system effectiveness – the extent to which parameters of the system can be determined to within uncertainty requirements.

5.1 System identifiability

A system is identifiable using a given measurement strategy if all the system parameters $\mathbf{a}$ can be determined from the resulting measurement data. Usually, the identifiability of a system corresponds to a matrix associated with the model solution being full rank.

The rank of a matrix can be examined from its singular value decomposition [8]. Any $m \times n$ matrix C can be factored as the product

$$C = USV^{\mathrm{T}}, \tag{5.1}$$

where U is an $m \times n$ orthogonal matrix, S an $n \times n$ diagonal matrix with diagonal entries $s_1 \geq s_2 \geq \ldots \geq s_n$, and V an $n \times n$ orthogonal matrix. This factorisation is known as the singular value decomposition (SVD). The columns of U (V) are the *left (right) singular vectors* and the s_j are known as the *singular values*.

If C is of column rank $p < n$ then its singular values $s_{p+1} = \ldots = s_n = 0$ and for each $\mathbf{v}_k$, $k = p+1, \ldots, n$, $C\mathbf{v}_k = \mathbf{0}$. Conversely, if $C\mathbf{b} = \mathbf{0}$, $\mathbf{b}$ can be written

uniquely as a linear combination

$$\mathbf{b} = \alpha_{p+1}\mathbf{v}_{p+1} + \ldots + \alpha_n \mathbf{v}_n.$$

Thus, $\mathbf{v}_k$, $k = p+1, \ldots, n$ provide an orthogonal set of basis vectors for the linear space of vectors $\{\mathbf{b} : C\mathbf{b} = \mathbf{0}\}$. These vectors can be used to analyse in what way a system fails to be identifiable.

Suppose C is the observation matrix in a linear least squares system of equations and $\mathbf{y}_0 = C\mathbf{a}_0$. Since $C\mathbf{v}_k = \mathbf{0}$, $k = p+1, \ldots, n$ then $C(\mathbf{a}_0 + \mathbf{v}_k) = \mathbf{y}_0$ also. In other words, the vectors $\mathbf{v}_k$ characterise the set of linear combinations of the parameters that are not determined from the data.

For more general least squares estimation problems of the form

$$\min_{\mathbf{a}} \sum_i f^2(\mathbf{x}_i, \mathbf{a}),$$

system identifiability can be examined using the following approach. Given software to calculate $\mathbf{f}$ and the associated Jacobian matrix $J_{ij} = \frac{\partial f_i}{\partial a_j}$ and nominal values $\mathbf{a}_0$ of the parameters:

1 Generate a dense sample of data points $\{\mathbf{x}_i\}$ satisfying the model equations exactly so that $f(\mathbf{x}_i, \mathbf{a}_0) = 0$.

2 Calculate the associated Jacobian matrix J.

3 Determine the SVD of $J = USV^{\mathrm{T}}$ and singular values $s_1, \ldots, s_n$.

4 If J is full rank ($s_n > 0$) then the system is identifiable.

5 If J is not full rank, so that $s_{p+1} = \ldots = s_n = 0$, examine the corresponding columns $\mathbf{v}_{p+1}, \ldots, \mathbf{v}_n$ of V to determine which combinations of parameters are not determined from the data.

Example: incorrect parametrization of a straightline. A line in two dimensions is specified uniquely by two distinct points (x_0, y_0) and (u_0, v_0) lying on the line. Assuming that $x_0 \neq u_0$, let $\mathbf{a} = (x_0, y_0, u_0, v_0)^{\mathrm{T}}$ and

$$\phi_0(x, \mathbf{a}) = y_0 + \frac{v_0 - y_0}{u_0 - x_0}(x - x_0),$$

then the set of points on the line is described by $\{(x, y) : y = \phi_0(x, \mathbf{a})\}$. However the mapping

$$L : \mathbf{a} \mapsto \{(x, y) : y = \phi_0(x, \mathbf{a})\}$$

is not a proper parameterization (we know that only two parameters are required). For a data set $X = \{(x_i, y_i)\}_{i=1}^m$ define

$$f_i(x_i, y_i, \mathbf{a}) = y_i - \phi_0(x_i, \mathbf{a})$$

and J the $m \times 4$ Jacobian matrix defined by $J_{ij} = \frac{\partial f_i}{\partial a_j}$. Below are example calculations for the case $\mathbf{a} = (1, 1, 5, 5)$ specifying the line $y = x$ and data $X = \{(1,1), (2,2), (3,3), (4,4), (5,5)\}$. In this case the Jacobian matrix J is

```
J =
    1.0000   -1.0000         0         0
    0.7500   -0.7500    0.2500   -0.2500
    0.5000   -0.5000    0.5000   -0.5000
    0.2500   -0.2500    0.7500   -0.7500
         0         0    1.0000   -1.0000
```

and has singular value decomposition $J = USV^{\mathrm{T}}$ with

```
U =
   -0.4472    0.6325    0.5410    0.2405
   -0.4472    0.3162   -0.2608   -0.6388
   -0.4472    0.0000   -0.6107    0.6441
   -0.4472   -0.3162   -0.1601   -0.3339
   -0.4472   -0.6325    0.4906    0.0881


S =
    2.2361         0         0         0
         0    1.5811         0         0
         0         0    0.0000         0
         0         0         0    0.0000


V =
   -0.5000    0.5000    0.7071         0
    0.5000   -0.5000    0.7071    0.0000
   -0.5000   -0.5000   -0.0000   -0.7071
    0.5000    0.5000    0.0000   -0.7071
```

The singular values are the diagonal elements of the matrix S and it is seen that the last two are zero (in floating point arithmetic they are of the order of 10^{-16} or smaller). The rightmost two singular vectors $\mathbf{v}_3$ and $\mathbf{v}_4$ of V indicate the two modes of behaviour of the parameters $\mathbf{a}$ that are not specified by the data. The vector $\mathbf{v}_3$ corresponds to (x_0, y_0) moving along the line $y = x$ and $\mathbf{v}_4$ corresponds to moving (u_0, v_0). These two degrees of freedom can be removed by fixing x_0 and u_0 at distinct values leaving a well defined parametrization of the line in terms of the values of the y-coordinate at x_0 and u_0.

Example: fitting a sphere to data on a circle. The least squares best fit sphere defined by centre co-ordinates $(a_1, a_2, a_3)^{\mathrm{T}}$ and radius a_4 to data $X = \{\mathbf{x}_i = (x_i, y_i, z_i)^{\mathrm{T}}\}_1^m$ is found by solving

$$\min_{\mathbf{a}} \sum_{i=1}^{m} d^2(\mathbf{x}_i, \mathbf{a})$$

where

$$d(\mathbf{x}_i, \mathbf{a}) = \{(x_i - a_1)^2 + (y_i - a_2)^2 + (z_i - a_3)^2\}^{1/2} - a_4,$$

see [4, Section 5.5]. Below are the example calculations for the case $\mathbf{a} = (0, 0, 100, 100\sqrt{2})^{\mathrm{T}}$ and data set

```
X =
  100.0000         0         0
   70.7107   70.7107         0
    0.0000  100.0000         0
  -70.7107   70.7107         0
 -100.0000    0.0000         0
  -70.7107  -70.7107         0
   -0.0000 -100.0000         0
   70.7107  -70.7107         0
```

representing eight points uniformly spaced around the circle $x^2 + y^2 = 100^2$ in the $z = 0$ plane. For this data the Jacobian matrix is

```
J =
   -0.5774         0    0.8165   -1.0000
   -0.4082   -0.4082    0.8165   -1.0000
   -0.0000   -0.5774    0.8165   -1.0000
    0.4082   -0.4082    0.8165   -1.0000
    0.5774   -0.0000    0.8165   -1.0000
    0.4082    0.4082    0.8165   -1.0000
    0.0000    0.5774    0.8165   -1.0000
   -0.4082    0.4082    0.8165   -1.0000
```

with singular values s_j and fourth right singular vector $\mathbf{v}_4$

```
    s            v4
3.6515            0
1.1547       0.0000
1.1547      -0.7746
0.0000      -0.6325
```

The vector $\mathbf{v}_4$ shows that by decreasing the z-coordinate of the sphere by 0.7746ϵ then we can ensure the sphere surface passes through the data points if we simultaneously decrease the radius by 0.6325ϵ. The system becomes fully identifiable if, for example, we add information about the radius from measurements of sphere diameter. Physically, this corresponds to the well-defined solution of a ball of radius $100\sqrt{2}$ sitting on a cylindrical hole of radius 100.

If we perform the same analysis but with $\mathbf{a} = (0, 0, 0, 100)^{\mathrm{T}}$, then we obtain singular values and fourth right singular vector

```
    s            v4
2.8284            0
2.0000            0
2.0000           -1
     0            0
```

The situation for this case is qualitatively different from the previous one, since adding information about the radius does not lead to system identifiability. Even if we know the sphere radius exactly, the z-component of the sphere centre is not determined from the data (from first order information).

5.2 System effectiveness

Given an identifiable system, i.e., one for which the Jacobian matrix is full rank, it is possible to use the Jacobian evaluation software in numerical simulations of an experiment to determine the likely uncertainties that will be obtained for a given measurement strategy. In this situation, exact measurement data $\mathbf{x}_i^*$ are generated according to a specified measurement strategy and from this, the Jacobian matrix J. Given an estimate of the standard deviation σ of the measurement errors, the covariance matrix for the model parameters is given by

$$V_{\mathbf{a}} = \sigma^2 (J^{\mathrm{T}} J)^{-1}.$$

The variances of the fitted parameters are estimated by the diagonal elements of the covariance matrix. Often we are not so much concerned with the variances of the fitted parameters but with variances and covariances of functions of the fitted parameters. If $h = \mathbf{h}^{\mathrm{T}}\mathbf{a}$ is a linear function of the fitted parameters then the variance of h is estimated by

$$\mathrm{var}(h) = \mathbf{h}^{\mathrm{T}} V_{\mathbf{a}} \mathbf{h}. \tag{5.2}$$

If $h(\mathbf{a})$ is a nonlinear function of $\mathbf{a}$ and $\mathbf{h}$ is the vector $\mathbf{h} = (\frac{\partial h}{\partial a_1}, \ldots, \frac{\partial h}{\partial a_n}^{\mathrm{T}})$ of partial derivatives of h with respect to the parameters $\mathbf{a}$, then (5.2) is an estimate of the variance of h which will be valid if $\mathbf{h}$ is not identically zero and h is not highly nonlinear at $\mathbf{a}$.

Note that if the QR factorisation of $J = QR$ or its singular value decomposition $J = USV^{\mathrm{T}}$ are available, then $V_{\mathbf{a}}$ can be calculated from

$$V_{\mathbf{a}} = \sigma^2 (R^{\mathrm{T}} R)^{-1} = \sigma^2 R^{-1} R^{-\mathrm{T}} \quad \text{or} \quad \sigma^2 V S^{-2} V^{\mathrm{T}}.$$

In either case, advantage can be taken of the form of the matrix to be inverted.

By changing the measurement strategy and monitoring the effect on the variances of the parameters of interest it is often possible to improve the experimental efficiency. Importantly, this can be achieved by using the same data analysis modules required to determine estimates of the parameters and their uncertainties from actual measurement data. In other words, with very little additional effort the model solving tools can be used to improve experimental strategy.

Example: sphere fit to data on an equatorial band. Following on from the example above, suppose we wish to fit a sphere with nominal parameters $\mathbf{a} = (0, 0, 0, 100)^{\mathrm{T}}$ to data points lying in an equatorial band at height between zero and 20 above the equator. We are limited to choosing a total 30 data points uniformly distributed on circles at heights $h_1 = 0$, $h_2 = 10$ and $h_3 = 20$. What is the best distribution of points to determine a) the radius a_4 or b) the z-coordinate a_3 of the centre? Four measurement strategies M_k, $k = 1, 2, 3, 4$, are proposed specified by the number of points on each circle:

hk	M1	M2	M3	M4
0	10	15	28	2
10	10	0	0	0
20	10	15	2	28

Assuming that the standard deviations of the residuals errors is $\sigma = 0.01$, the standard uncertainties $u(a_j) = (\mathrm{var}(a_j))^{1/2}$ for each of the four sphere parameters a_j for the four measurement strategies M_k are

	M1	M2	M3	M4	
a1	0.0026	0.0026	0.0025	0.0025	
a2	0.0026	0.0026	0.0027	0.0027	
a3	0.0224	0.0183	0.0366	0.0366	z-coordinate
a4	0.0029	0.0026	0.0019	0.0071	radius

There results show that measurement strategy M_2, placing equal numbers of data points on the $z = 0$ and $z = 20$ circles and none on the $z = 10$ circle, is best to determine the z-coordinate a_3 of the centre while strategy M_3, placing nearly all the data points on the $z = 0$ circle (i.e., the equator) is best to determine the radius. While strategy M_1 seems appropriate, M_2 outperforms it in terms of determining both radius and z-coordinate. For either requirement, M_4 is a poor strategy.

Chapter 6

Case Studies

6.1 Univariate linear regression

This case study involves the univariate linear model in which a response y is modelled as a linear function of a single variable x:

$$y = a_1 + a_2 x.$$

It is a continuation of the case study in [4, Section 6.1] which considered the behaviour of different fitting methods for different statistical models.

6.1.1 Statistical model for relative measurement error

In this section, we are concerned with a model of the form

$$y = y^* + \epsilon + \delta, \quad \epsilon \in N(0, \sigma^2), \quad \delta \in N(0, y^2\rho^2), \tag{6.1}$$

that is, the measurement error in y has two random components, one drawn from a distribution with fixed standard deviation σ, the second drawn from a distribution whose standard deviation is proportional to the measurand, representing a relative measurement error.

We consider two estimators, the linear least squares estimator (LLS) and the related weighted linear least squares estimator (WLLS). The LLS estimator determines best fit parameters $\mathbf{a} = (a_1, a_2)^{\mathrm{T}}$ to data $X = \{(x_i, y_i)\}_{i=1}^m$ by solving

$$\min_{\mathbf{a}} \sum_{i=1}^{m} (y_i - a_1 - a_2 x_i)^2.$$

Let C be the $m \times 2$ observation matrix whose ith row is $(1, x_i)$. Then, mathematically,

$$\mathbf{a} = C^+ \mathbf{y}, \quad C^+ = (C^{\mathrm{T}} C)^{-1} C^{\mathrm{T}}, \tag{6.2}$$

showing that the solution estimates $\mathbf{a}$ are a linear function of the response values $\mathbf{y} = (y_1, \ldots, y_m)^{\mathrm{T}}$. If

$$\mathbf{r} = \mathbf{y} - C\mathbf{a},$$

is the vector of residual errors and

$$\hat{\sigma} = \|\mathbf{r}\|/(m-2)^{1/2} = \left(\frac{1}{m-2} \sum_i r_i^2 \right)^{1/2} \tag{6.3}$$

is an estimate of the standard deviation of the residual errors, then the usual estimate of the covariance matrix of the fitted parameters is given by

$$V_{\mathbf{a}} = \hat{\sigma}^2 (C^{\mathrm{T}} C)^{-1}.$$

However, these uncertainty estimates are based on the assumption that the measurement errors in y are independent and have equal variance. This will be true only if the range of ρy_i is small relative to σ in (6.1). We therefore ask the questions:

V1 Does the LLS estimator produce adequate estimates of the solution parameters?

V2 Are the estimates of the uncertainties for the LLS estimator valid?

These questions can be answered by estimator analysis and/or numerical simulation. From the model (6.1), we can estimate the variance σ_i^2 associated with the ith measurement by

$$\sigma_i^2 = \sigma^2 + \rho^2 y_i^2. \tag{6.4}$$

Setting weights $w_i = 1/\sigma_i$, the statistical model for the weighted measurement data is

$$w_i y_i = w_i y_i^* + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i \in N(0,1),$$

i.e., the weighted measurement errors have equal variance. The conditions of the Gauss-Markov theorem apply (Section 3.4) so that we can say that the WLLS estimator which determines solution parameters $\tilde{\mathbf{a}}$ by solving

$$\min_{\mathbf{a}} \sum_{i=1}^{m} w_i^2 (y_i - a_1 - a_2 x_i)^2,$$

is optimal. This means we can benchmark the LLS estimator against the optimal estimator. Let $\tilde{C}$ be the $m \times 2$ observation matrix whose ith row is $(w_i, w_i x_i)$. Then

$$\tilde{\mathbf{a}} = \tilde{C}^{+} \mathbf{y}, \quad \tilde{C}^{+} = (\tilde{C}^{\mathrm{T}} \tilde{C})^{-1} \tilde{C}^{\mathrm{T}}.$$

Furthermore, the estimate of the covariance matrix $\tilde{V}_{\mathbf{a}}$ of the fitted parameters for WLLS is

$$\tilde{V}_{\mathbf{a}} = (\tilde{C}^{\mathrm{T}} \tilde{C})^{-1}. \tag{6.5}$$

This estimate is valid as it is based on the correct statistical model.

The first question (V1) is concerned with the bias and efficiency of the LLS estimator. To consider bias first, we can argue that the LLS is a weighted

version of the unbiased estimator WLLS and is therefore unbiased. The second consideration is efficiency: if the experiment is repeated many times, how will the variation in the LLS estimates compare with those of the WLLS estimates. The matrix $\tilde{V}_{\mathbf{a}}$ describes the variation for the WLLS. We also have enough information to calculate a valid estimate of the variation for the LLS estimates. The covariance matrix $V_{\mathbf{y}}$ for the measured data points is the diagonal matrix with σ_i^2 defined in (6.4) in the ith diagonal element. The solution parameters $\mathbf{a}$ are a linear combination of the response measurements as shown in (6.2). The covariance matrix $V_{\mathbf{a}}^*$ associated with these estimates is therefore

$$V_{\mathbf{a}}^* = C^+ V_{\mathbf{y}} (C^+)^{\mathrm{T}}. \tag{6.6}$$

The efficiency of the LLS estimator can be compared relative to the optimal WLLS estimator by comparing $V_{\mathbf{a}}^*$ with $\tilde{V}_{\mathbf{a}}$.

The second question (V2) concerns the validity of the estimate $V_{\mathbf{a}}$ of the covariance matrix of the fitted parameters $\mathbf{a}$ determined by the LLS estimator. This question can be answered using Monte Carlo simulations. Given values of σ, ρ, $\{x_i\}_{i=1}^m$ and the model parameters $\mathbf{a}$, we generate data sets $X_q = \{(x_i, y_{i,q})\}_{i=1}^m$, $q = 1, \ldots, N$, according to the model:

$$y_i^* = a_1 + a_2 x_i, \quad y_{i,q} = y_i^* + \epsilon_{i,q} + \delta_{i,q}, \quad \epsilon_{i,q} \in N(0, \sigma^2), \quad \delta_{i,q} \in N(0, (y_i^* \rho)^2).$$

For each data set X_q the LLS estimates $\mathbf{a}_q$ are recorded along with the estimate $\hat{\sigma}_q$ of the standard deviation of residuals calculated according to (6.3) for the qth set of residual errors $\mathbf{r}_q = \mathbf{y}_q - C\mathbf{a}_q$. For small data sets (say, number of data points $m < 50$), the estimates $\hat{\sigma}_q$ will have relatively large variation. Consequently, the estimates of the covariance matrix of the fitted parameters will also vary. An estimate of the "average" covariance matrix is given by

$$V_{\mathbf{a}} = \bar{\sigma}^2 (C^{\mathrm{T}} C)^{-1}, \quad \bar{\sigma} = \left(\frac{1}{N} \sum_{q=1}^{N} \hat{\sigma}_q^2 \right)^{1/2}. \tag{6.7}$$

We have performed 5000 Monte Carlo simulations on data generated from the exact data set

x	y*
0	1.0000
0.1000	1.5000
0.2000	2.0000
0.3000	2.5000
0.4000	3.0000
0.5000	3.5000
0.6000	4.0000
0.7000	4.5000
0.8000	5.0000
0.9000	5.5000
1.0000	6.0000

with $\mathbf{a}^* = (1, 5)^{\mathrm{T}}$ for different values of σ and ρ. For each set of experiments, we calculate

$u(\tilde{a}_j)$ the standard uncertainty of the WLLS parameter estimates calculated according to (6.5).

	$u(\tilde{a}_j)$	$\bar{u}(\tilde{a}_j)$	$u^*(a_j)$	$\bar{u}(a_j)$	$u(a_j)$
E1	$\sigma = 0.01$, $\rho = 0.005$				
a_1	0.0077	0.0076	0.0090	0.0089	0.0122
a_2	0.0187	0.0187	0.0217	0.0215	0.0206
E1	$\sigma = 0.005$, $\rho = 0.01$				
a_1	0.0090	0.0089	0.0143	0.0142	0.0217
a_2	0.0290	0.0290	0.0393	0.0389	0.0367

Table 6.1: Estimates of the standard uncertainties of the fitted parameters for the WLLS and LLS estimators for data generated according to the model (6.1).

$\bar{u}(\tilde{a}_j)$ the sample standard deviation of the WLLS estimates $\tilde{a}_{j,q}$.

$u^*(a_j)$ the standard uncertainty of the LLS parameter estimates calculated according to (6.6).

$\bar{u}(a_j)$ the sample standard deviation of the LLS estimates $a_{j,q}$.

$\hat{u}(a_j)$ the standard uncertainty of the LLS parameter estimates calculated from (6.7).

Table 6.1 shows these uncertainty estimates for two sets (E1 and E2) of values of σ and ρ. We note:

- Good agreement between the analytical estimates $u(\tilde{a_j})$ and $u^*(a_j)$ with empirical estimates $\bar{u}(\tilde{a_j})$ and $\bar{u}(a_j)$ determined from the Monte Carlo simulations. This is because the analytical estimates are based on the correct statistical model.
- There is poor agreement between the analytical estimates $u(a_j)$ and the actual variation $\bar{u}(a_j)$.
- The WLLS uncertainties are smaller than the LLS uncertainties. This is to be expected since the WLLS estimator is optimal for this model.

We can also analyse these results in terms of the estimates of the standard uncertainties $u(\hat{y})$ of the model predictions $\hat{y} = a_1 + a_2 x$ at any point x. If V is an estimate of the covariance matrix of the fitted parameters, then

$$u(\hat{y}) = \left(\begin{bmatrix} 1 & x \end{bmatrix} V \begin{bmatrix} 1 \\ x \end{bmatrix} \right)^{1/2}.$$

Figures 6.1 and 6.2 graph $\pm 2u(\hat{y})$ as a function of x determined from three estimates of the covariance matrix (i) $\tilde{V}_{\mathbf{a}}$, the covariance matrix of the fitted parameters for the WLLS estimator (Eq. (6.5), solid curve), (ii) $V^*_{\mathbf{a}}$, the correct estimate of the covariance matrix of the LLS estimates (Eq. (6.6), dashed curve) and (iii) $V_{\mathbf{a}}$, the covariance matrix for the LLS estimates assuming equal variances (Eq. (6.7), dot-dashed curve). The residual errors for an example fit are also graphed. For the case $\sigma = 0.01$, $\rho = 0.005$ (E1) shown in Figure 6.1,

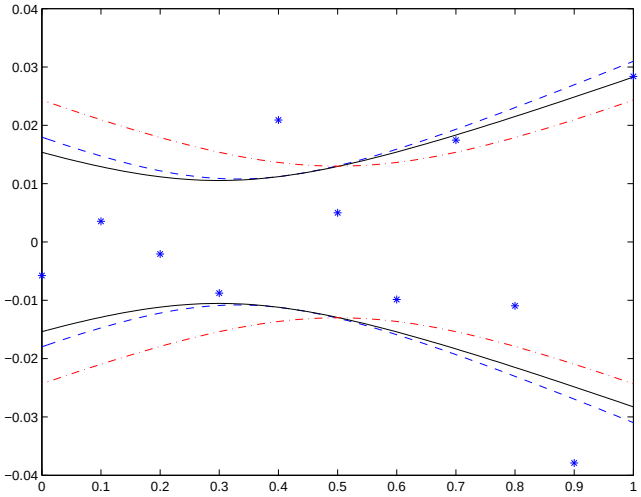

Figure 6.1: Estimates of uncertainty in the model predictions for the statistical model with $\sigma = 0.01$, $\rho = 0.005$ (E1). The three pairs of graphs present $\pm 2u(\hat{y})$ as a function of x corresponding to three estimates of the covariance matrix (i) $\tilde{V}_{\mathbf{a}}$, the covariance matrix of the fitted parameters for the WLLS estimator (Eq. (6.5), solid curve), (ii) $V^*_{\mathbf{a}}$, the correct estimate of the covariance matrix of the LLS estimates (Eq. (6.6), dashed curve) and (iii) $V_{\mathbf{a}}$, the covariance matrix for the LLS estimates assuming equal variances (Eq. (6.7), dot-dashed curve). The residual errors for an example LLS fit are also graphed ('*').

the model predictions for the WLLS estimator have smaller uncertainty than those for the LLS estimator. The estimates based on the incorrect assumption of equal variance overestimate the uncertainty near $x = 0$ but underestimate it near $x = 1$. The results in Figure 6.2 for the case $\sigma = 0.005$, $\rho = 0.01$ (E2) are qualitatively the same but the differences in the three estimates greater. This is to be expected since the statistical model is less well approximated by an equal variance model.

6.1.2 Statistical model for correlation measurement error

We now perform a similar analysis to the one above for the case of correlated measurement error. The functional model is exactly the same, but the statistical model is one in which there is structural correlation. An example application is in modelling the linear response $x = b_1 + b_2 y$ of a balance to the application of a mass y where the mass is made up from a number of component masses $\{m_k\}_{k=1}^K$ of same nominal mass along with a mass m_0 present in all measurements. We assume that the component masses m_k have been calibrated using the same equipment so that the estimate m_k is related to the true mass m_k^* through

$$m_k = m_k^* + \mu_k, \quad \mu_k \in N(0, \sigma_M^2).$$

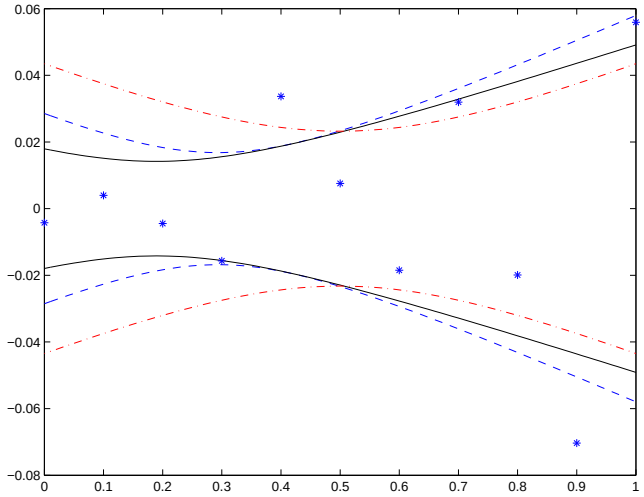

Figure 6.2: As Figure 6.1 but for the case $\sigma = 0.005$, $\rho = 0.01$ (E2).

Furthermore the "constant" mass has also been measured so that

$$m_0 = m_0^* + \mu_0, \quad \mu_0 \in N(0, \sigma_0^2).$$

If, in the ith measurement, the mass y_i is assembled from masses $\{m_{i_1}, \ldots, m_{i_q}\}$ and the response x_i is accurately measured, then an appropriate model is

$$\begin{aligned} y_i^* &= a_1 + a_2 x_i, \quad y_i^* = m_0^* + \sum_{k=1}^{q} m_{i_k}^*, \\ y_i &= y_i^* + \epsilon_i + \mu_0 + \sum_{k=1}^{q} \mu_{i_k}, \\ \epsilon_i &\in N(0, \sigma^2), \quad \mu_0 \in N(0, \sigma_0^2), \quad \mu_k \in N(0, \sigma_M^2),\ k = 1, \ldots, K, \end{aligned} \tag{6.8}$$

where μ_0 accounts for measurement error in the constant mass, μ_{i_k} that in the component masses and ϵ_i accounts for a random disturbance for that measurement (due to random air buoyancy effects, for example). The errors in y_i are correlated with each other due to their common dependence on the errors in the estimates of the component masses and that in m_0.

Since (6.8) gives the explicit dependence of y_i on the measurement of $\{m_k\}$ and m_0 the correlation can be determined from the model. For example, suppose in an experiment the masses m_k are added one at a time until all K masses are loaded and the response x accurately measured. Subsequently, the masses are removed one at a time, so that two measurements are recorded for each set of masses $m_1 + \cdots + m_q$. Let V_M be the $(K+1) \times (K+1)$ diagonal covariance

matrix with σ_M^2 in the first K diagonal elements and σ_0^2 in the $(K+1)$th:

$$V_M = \begin{bmatrix} \sigma_M^2 & & & & \\ & \sigma_M^2 & & & \\ & & \ddots & & \\ & & & \sigma_M^2 & \\ & & & & \sigma_0^2 \end{bmatrix},$$

and G the $2K \times (K+1)$ matrix with

$$G_{ij} = \frac{\partial y_i}{\partial m_j}, \quad 1 \le j \le K, \quad G_{ij} = \frac{\partial y_i}{\partial m_0}, \quad j = K+1,$$

so that G has the form

$$G = \begin{bmatrix} 1 & & & & 1 \\ 1 & 1 & & & 1 \\ \vdots & \vdots & & & \vdots \\ 1 & 1 & \cdots & 1 & 1 \\ 1 & 1 & \cdots & 1 & 1 \\ \vdots & \vdots & & & \vdots \\ 1 & 1 & & & 1 \\ 1 & & & & 1 \end{bmatrix}.$$

Then the covariance matrix for $\mathbf{y} = (y_1, \ldots, y_{2K})^T$ is

$$V_{\mathbf{y}} = G V_M G^{\mathrm{T}} + \sigma^2 I, \tag{6.9}$$

where I is the $2K \times 2K$ identity matrix. The first term is the covariance that arises from the measurements of the masses m_k and m_0 while the second is the random, uncorrelated component.

For this model, the Gauss-Markov estimator (GMLS) that determines estimates by the solving

$$\min_{\mathbf{a}} \ \mathbf{f}^{\mathrm{T}} V_{\mathbf{y}}^{-1} \mathbf{f}, \tag{6.10}$$

where $\mathbf{f} = (f_1, \ldots, f_m)^{\mathrm{T}}$ with $f_i = y_i - a_1 - a_2 x_i$, is optimal. If C is the observation matrix with $(1, x_i)$ in the ith row, the covariance matrix $\tilde{V}_{\mathbf{a}}$ of the GMLS estimates is

$$\tilde{V}_{\mathbf{a}} = (C^T V_{\mathbf{y}}^{-1} C)^{-1}. \tag{6.11}$$

However, we can also apply the simpler linear least squares estimator (LLS) that solves

$$C\mathbf{a} = \mathbf{y}$$

in the least squares sense, so that mathematically,

$$\mathbf{a} = C^{+}\mathbf{y}, \quad C^{+} = (C^{\mathrm{T}} C)^{-1} C^{\mathrm{T}}.$$

As before, and estimate of the covariance matrix of the fitted parameters is given by

$$V_{\mathbf{a}} = \hat{\sigma} (C^{\mathrm{T}} C)^{-1},$$

where $\hat{\sigma}$ is an estimate of the standard deviation of the residual errors.

We ask:

V3 Is the LLS estimator appropriate for this model?

V4 Are the estimates of uncertainties valid?

We can use exactly the same techniques of analysis and simulation as used in Section 6.1.1 to answer these questions. The true estimate of the covariance matrix for the LLS estimates for this model is

$$V_{\mathbf{a}}^* = C^+ V_{\mathbf{y}} (C^+)^{\mathrm{T}}. \tag{6.12}$$

By comparing this matrix with $\tilde{V}_{\mathbf{a}}$, we can assess the LLS estimator relative to the optimal GMLS estimator. We can also use Monte Carlo simulation to check the performance of the LLS estimator and examine the validity of the uncertainty estimates. Given values of σ_M, σ_0, σ, $\{x_i\}_{i=1}^m$ and the model parameters $\mathbf{a}$, we generate data sets $X_q = \{(x_i, y_{i,q})\}_{i=1}^m$, $q = 1, \ldots, N$, according to the model:

I Generate exact data

$$y_i^* = a_1 + a_2 x_i.$$

II$_q$ For each q,

II.1$_q$ Generate

$$\mu_{0,q} \in N(0, \sigma_0^2), \quad \mu_{k,q} \in N(0, \sigma_M^2), \quad k = 1, \ldots, K.$$

II.2$_q$ If the ith mass is assembled from masses $\{m_{i_1}, \ldots, m_{i_q}\}$, set

$$y_{i,q} = y_i^* + \epsilon_{i,q} + \mu_{0,q} + \sum_{k=1}^{q} \mu_{i_k,q}, \quad \epsilon_{i,q} \in N(0, \sigma^2).$$

For each data set X_q the LLS estimates $\mathbf{a}_q$ are recorded along with the estimate $\hat{\sigma}_q$ of the standard deviation of residuals calculated for the qth residual error $\mathbf{r}_q = \mathbf{y}_q - C\mathbf{a}_q$ according to (6.3). As before, an estimate of the average covariance matrix is given by

$$V_{\mathbf{a}} = \bar{\sigma}^2 (C^{\mathrm{T}} C)^{-1}, \quad \bar{\sigma} = \left(\frac{1}{N} \sum_{q=1}^{N} \hat{\sigma}_q^2 \right)^{1/2}. \tag{6.13}$$

We have performed 5000 Monte Carlo simulations for the data set

$$\mathbf{x}^* = (1, 2, \ldots, 8, 9, 9, 8, \ldots, 2, 1)^{\mathrm{T}}$$

and $\mathbf{a}^* = (1, 1)^{\mathrm{T}}$ and various values of σ_M, σ_0 and σ. For each set of experiments, we calculate:

$u(\tilde{a}_j)$ the standard uncertainty of the GMLS estimates calculated according to (6.11).

$u^*(a_j)$ the standard uncertainty of the LLS parameter estimates calculated according to (6.12).

	$u(\tilde{a}_j)$	$u^*(a_j)$	$\bar{u}(a_j)$	$u(a_j)$
E1	$\sigma_M = 0.001$, $\sigma_0 = 0.001$, $\sigma = 0.01$			
a_1	5.4130e-003	5.4152e-003	5.1437e-003	5.4295e-003
a_2	9.8478e-004	9.8489e-004	9.1406e-004	9.8536e-004
E2	$\sigma_M = 0.01$, $\sigma_0 = 0.001$, $\sigma = 0.001$			
a_1	1.0684e-002	1.3956e-002	4.1260e-003	1.3906e-002
a_2	3.5377e-003	3.6980e-003	7.3321e-004	3.6612e-003
E3	$\sigma_M = 0.01$, $\sigma_0 = 0.001$, $\sigma = 0.001$			
a_1	1.0081e-002	1.0109e-002	6.5555e-004	1.0135e-002
a_2	3.7093e-004	3.8079e-004	1.1650e-004	3.7730e-004

Table 6.2: Estimates of the standard uncertainties of the fitted parameters to data generated according to the statistical model (6.8).

$\bar{u}(a_j)$ the sample standard deviation of the LLS estimates $\mathbf{a}_q$.

$u(a_j)$ the standard uncertainty of the LLS parameter estimates calculated from (6.13).

Results of these calculations are presented in Table 6.2.

We can also analyse these results in terms of the estimates of the standard uncertainties $u(\hat{y})$ of the model predictions $\hat{y}$ at any point x graphed in Figures 6.3–6.5.

From the table and figures, we note

- For the case $\sigma_M = 0.001$, $\sigma_0 = 0.001$ and $\sigma = 0.01$ in which the independent random component ϵ_i dominates (E1), the behaviour of the LLS estimator is very similar to that of the GMLS estimator. The statistics based on the approximate model derived from (6.13) underestimates slightly the uncertainty in the model predictions. The pairs of residual errors for each ordinate value x_i show a random behaviour.
- For the case $\sigma_M = 0.01$, $\sigma_0 = 0.001$ and $\sigma = 0.001$ in which the error component μ_k dominates (E2), the LLS estimator performs less well than the GMLS estimator, but not drastically so. However, the uncertainties based on the approximate statistical model completely underestimate the uncertainties in the model predictions. The pairs of residual errors exhibit a strong correlation.
- For the case $\sigma_M = 0.001$, $\sigma_0 = 0.01$ and $\sigma = 0.001$ in which the error component μ_0 associated with the constant mass dominates (E3), the LLS and GMLS estimator have very similar behaviour but the uncertainties based on the approximate statistical model completely underestimate the uncertainties in the model predictions. The pairs of residual errors show a random behaviour.

This example illustrates an important point. The GMLS estimator is optimal but the LLS estimator performs quite well for the three types of experiment:

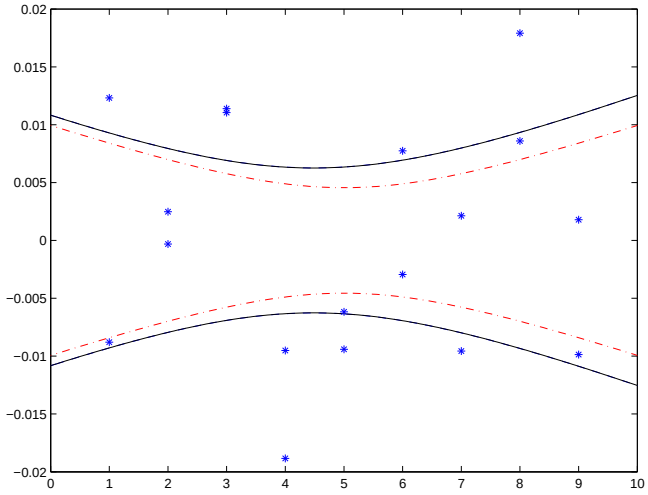

Figure 6.3: Estimates of uncertainty in the model predictions for the statistical model (6.8) with $\sigma_M = 0.001$, $\sigma_0 = 0.001$ and $\sigma = 0.01$ (E1). The three pairs of graphs present $\pm 2u(\hat{y})$ as a function of x determined from three estimates of the covariance matrix (i) $\tilde{V}_{\mathbf{a}}$, the covariance matrix of the fitted parameters for the GMLS estimator (Eq. (6.11), solid curve), (ii) $V^*_{\mathbf{a}}$, the correct estimate of the covariance matrix of the LLS estimates (Eq. (6.12), dashed curve essentially coincident with the solid curve) and (iii) $V_{\mathbf{a}}$, the covariance matrix for the LLS estimates based on approximate statistical model (Eq. (6.13), dot-dashed curve). The residual errors for an example fit are also graphed ('*').

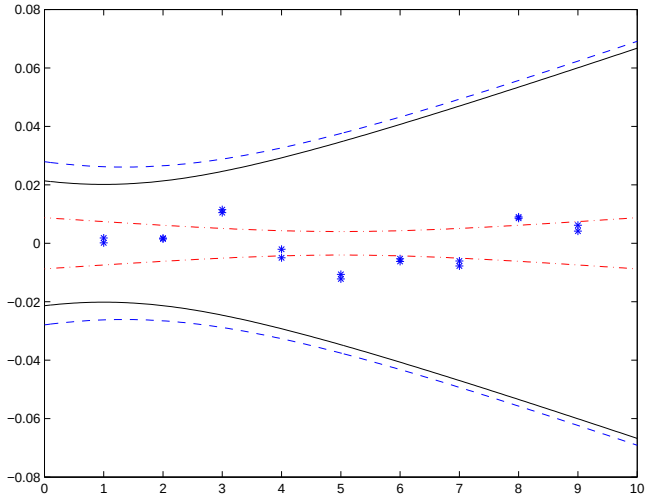

Figure 6.4: As Figure 6.3 but with $\sigma_M = 0.01$, $\sigma_0 = 0.001$ and $\sigma = 0.001$ (E2).

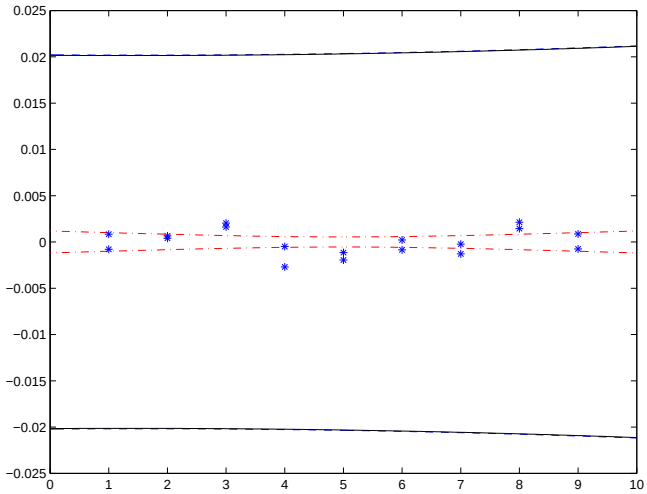

Figure 6.5: As Figure 6.3 but with $\sigma_M = 0.001$, $\sigma_0 = 0.01$ and $\sigma = 0.001$ (E3).

the variation in the LLS estimates are not drastically worse than that of the GMLS estimates. However, the estimates of uncertainties in the fitted parameters based on the approximate statistical model can be completely misleading. The uncertainties for the approximate statistical model cannot be made valid simply by introducing a fudge factor to increase the estimate $\hat{\sigma}$ of the standard uncertainty of the residual errors. The estimate of the model predictions for experiment E2 graphed in Figure 6.4 has the wrong shape compare with the actual uncertainties. Changing the value of $\hat{\sigma}$ will only change the scale, not the underlying shape. In the case of the third set of experiments (E3), the residual errors shown in Figure 6.5 are consistent with the approximate statistical model and the fit of the model to data is a good one. From an analysis of the consistency of the model with the measured data, we can be misled into believing that the model outputs, including the uncertainty estimates, are valid when in fact they are not.

6.2 Fitting a quadratic to data

In this case study, we examine the problem of fitting a quadratic model

$$y = a_1 + a_2 x + a_3 x^2$$

to data $X = \{(x_i, y_i)\}_{i=1}^m$ where both x_i and y_i are subject to measurement error. The complete model is of the form

$$\begin{aligned} y_i^* &= a_1 + a_2 x_i^* + a_3 (x_i^*)^2, \\ x_i &= x_i^* + \delta_i, \quad \delta_i \in N(0, \rho^2), \\ y_i &= y_i^* + \epsilon_i, \quad \epsilon_i \in N(0, \gamma^2). \end{aligned} \tag{6.14}$$

The preferred estimator for this type of model is the generalised distance regression estimator (GDR, [4, Section 4.3]) whose estimates are given by the solution of

$$\min_{\mathbf{a},\mathbf{x}^*} \sum_{i=1}^{m} \left\{ \alpha^2 (x_i - x_i^*)^2 + \beta^2 (y_i - \phi(x_i^*, \mathbf{a}))^2 \right\} \tag{6.15}$$

where $\phi(x, \mathbf{a}) = a_1 + a_2 x + a_3 x^2$, $\alpha = 1/\rho$ and $\beta = 1/\gamma$. However, we want to know if the linear least squares estimates (LLS) determined by the solution of the simpler problem

$$\min_{\mathbf{a}} \sum_{i=1}^{m} (y_i - \phi(x_i, \mathbf{a}))^2$$

is also suitable.

If C is the $m \times 3$ matrix with ith row given by $(1, x_i, x_i^2)$ then the LLS estimate of the parameters is defined mathematically by

$$\mathbf{a} = C^+ \mathbf{y}, \quad C^+ = (C^{\mathrm{T}} C)^{-1} C^{\mathrm{T}}. \tag{6.16}$$

As before, an estimate of the covariance matrix of the fitted parameters is given by

$$V_{\mathbf{a}} = \hat{\sigma} (C^{\mathrm{T}} C)^{-1}, \tag{6.17}$$

where $\hat{\sigma}$ is an estimate of the standard deviation of the residual errors:

$$\hat{\sigma} = \|\mathbf{r}\|/(m-3)^{1/2}, \quad \mathbf{r} = (r_1, \ldots, r_m)^{\mathrm{T}}, \quad r_i = y_i - \phi(x_i, \mathbf{a}). \tag{6.18}$$

We ask:

V1 Is the LLS estimator appropriate for the model (6.14)?

V2 Are the associated estimates of uncertainties (6.17) valid?

We can use Monte Carlo simulation and null space benchmarking described in Sections 3.2 and 3.3 to answer these questions. Given values of ρ, γ, $\{x_i^*\}_{i=1}^m$ and the model parameters $\mathbf{a}^*$, we generate data sets $X_q = \{(x_{i,q}, y_{i,q})\}_{i=1}^m$, $q = 1, \ldots, N$, according to the model:

$$\begin{aligned} y_i^* &= a_1^* + a_2^* x_i^* + a_3^* (x_i^*)^2, \\ x_{i,q} &= x_i^* + \delta_{i,q}, \quad \delta_{i,q} \in N(0, \rho^2), \\ y_{i,q} &= y_i^* + \epsilon_{i,q}, \quad \epsilon_{i,q} \in N(0, \gamma^2). \end{aligned}$$

For each data set X_q the observation matrix C_q with ith row equal to $(1, x_{i,q}, x_{i,q}^2)$ is calculated and the LLS estimates $\mathbf{a}_q$ found by solving

$$C_q \mathbf{a}_q = \mathbf{y}_q.$$

The estimate $\hat{\sigma}_q$ of the standard deviation of residuals calculated according to (6.18) for the qth set of residual errors $\mathbf{r}_q = \mathbf{y}_q - C_q \mathbf{a}_q$ is also recorded. For small data sets (number of data points $m < 50$, say), the estimates $\hat{\sigma}_q$ will have relatively large variation. Consequently, the estimates of the covariance matrix

of the fitted parameters will also vary. An estimate of the average covariance matrix is given by

$$V_{\mathbf{a}} = \bar{\sigma}^2 (C^{\mathrm{T}} C)^{-1}, \quad \bar{\sigma} = \left(\frac{1}{N} \sum_{q=1}^{N} \hat{\sigma}_q^2 \right)^{1/2}. \tag{6.19}$$

From simulations we can compare the actual variation in the LLS estimates with the variation predicted using the approximate statistical model. In this way we can answer question V2. To answer question V1, we would like to compare the variation in the LLS estimates with those for the GDR estimator. We can do this using null space benchmarking.

In section 3.3.2, we showed how to generate data according to models like (6.14) for which the GDR solution estimates are known. For the model (6.14), the data generation scheme is as follows. Given $\mathbf{a}^* = (a_1^*, a_2^*, a_3^*)^{\mathrm{T}}$, α, β, σ and $\{x_i^*\}$,

I Set $y_i^* = \phi(x_i^*, \mathbf{a}^*) = a_1^* + a_2^* x_i^* + a_3^* (x_i^*)^2$ and $X^* = \{(x_i^*, y_i^*)\}$.

II For each i, calculate

$$\dot{\phi}_i = \frac{\partial \phi}{\partial x}(x_i^*, \mathbf{a}^*) = a_2^* + 2 a_3^* x_i^*, \quad s_i = \left(\frac{\dot{\phi}_i^2}{\alpha^2} + \frac{1}{\beta^2} \right)^{1/2}.$$

III Calculate A given by

$$A_{ij} = \frac{\partial \phi_i}{\partial a_j},$$

so that the ith row of A is $(1, x_i^*, (x_i^*)^2)$.

IV Determine $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)^{\mathrm{T}}$ such that $A^{\mathrm{T}} \boldsymbol{\delta} = 0$ (using the QR factorisation of A, for example). For each i, set

$$p_i = -\delta_i \frac{\dot{\phi}_i}{\alpha^2}, \quad q_i = \delta_i \frac{1}{\beta^2}.$$

and calculate $S = \left\{ \sum_{i=1}^{m} (\alpha^2 p_i^2 + \beta^2 q_i^2) \right\}^{1/2}$ and $K = (m-n)^{1/2} \sigma / S$.

V For each i, set

$$x_i = x_i^* + K p_i, \quad y_i = y_i^* + K q_i.$$

Then $\mathbf{a}^*$ and $\{x_i^*\}$ solves (6.15) for data set $X = \{(x_i, y_i)\}_{i=1}^m$ with

$$\frac{1}{m-n} \sum_{i=1}^{m} \left\{ \alpha^2 (x_i - x_i^*)^2 + \beta^2 (y_i - \phi(x_i^*, \mathbf{a}^*))^2 \right\} = \sigma^2,$$

and the estimate $\tilde{V}_{\mathbf{a}}$ of the covariance matrix of the fitted parameters $\mathbf{a}^*$ for the GDR estimator is

$$\tilde{V}_{\mathbf{a}} = \sigma^2 (J^{\mathrm{T}} J)^{-1}, \tag{6.20}$$

where the Jacobian matrix J can be calculated from $J_{ij} = A_{ij}/s_i$.

We have performed 5000 Monte Carlo simulations for the data sets generated for the quadratic defined by $\mathbf{a}^* = (0, 0, 5)^{\mathrm{T}}$ and exact data $\{(x_i^*, y_i^*)\}$:

	$u(\tilde{a}_j)$	$\bar{u}(a_j)$	$u(a_j)$
E1	$\rho = 0.001$, $\gamma = 0.01$		
a_1	0.0048	0.0049	0.0052
a_2	0.0061	0.0063	0.0055
a_3	0.0106	0.0107	0.0098
E2	$\rho = 0.01$, $\gamma = 0.01$		
a_1	0.0084	0.0186	0.0266
a_2	0.0327	0.0404	0.0279
a_3	0.0515	0.0668	0.0499
E3	$\rho = 0.01$, $\gamma = 0.001$		
a_1	0.0010	0.0181	0.0263
a_2	0.0316	0.0401	0.0275
a_3	0.0477	0.0664	0.0492

Table 6.3: Estimates of the standard uncertainties of the fitted parameters to data generated according to the statistical model (6.14): $u(\tilde{a}_j)$ is the standard uncertainty of the GMLS parameter estimates calculated according to (6.20), $\bar{u}(a_j)$ is the actual variation in the LLS estimates and $u(a_j)$ is the predicted variation calculated assuming an approximate statistical model as in (6.19)

.

x*	y*	delta	epsilon
-1.0000	5.0000	0.0050	0.0005
-0.8000	3.2000	-0.0156	-0.0020
-0.6000	1.8000	0.0100	0.0017
-0.4000	0.8000	0.0004	0.0001
-0.2000	0.2000	-0.0018	-0.0009
0	0	0	0.0015
0.2000	0.2000	0.0010	-0.0005
0.4000	0.8000	-0.0014	0.0004
0.6000	1.8000	0.0028	-0.0005
0.8000	3.2000	0.0141	-0.0018
1.0000	5.0000	-0.0143	0.0014

for different values of γ and ρ. Also listed above are example null space perturbations δ_i and ϵ_i generated for the case $\rho = \gamma = 0.01$.

For each set of experiments, we calculate

$u(\tilde{a}_j)$ the standard uncertainty of the GDR parameter estimates calculated according to (6.20) for data generated using the null space method.

$\bar{u}(a_j)$ the sample standard deviation of the LLS estimates a_j.

$u(a_j)$ the standard uncertainty of the LLS parameter estimates calculated from (6.19).

Results of these calculations are presented in Table 6.3. We can also analyse these results in terms of the estimates of the standard uncertainties $u(\hat{y})$ of the model predictions $\hat{y}$ at any point x graphed in Figures 6.6–6.8.

From the table and figures, we note:

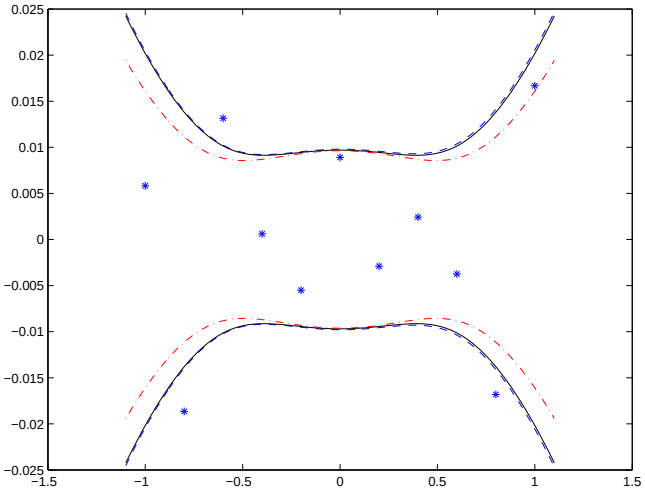

Figure 6.6: Estimates of uncertainty in the model predictions for the statistical model (6.14) with $\rho = 0.001$ and $\gamma = 0.01$ (E1). The three pairs of graphs present $\pm 2u(\hat{y})$ as a function of x determined from three estimates of the covariance matrix (i) $\tilde{V}_{\mathbf{a}}$, the covariance matrix of the fitted parameters for the GDR estimator (Eq. (6.20), solid curve), (ii) $\bar{V}_{\mathbf{a}}$, the estimate of the covariance matrix of the LLS estimates determined from the Monte Carlo simulations (dashed curve, essentially coincident with the solid curve) and (iii) $V_{\mathbf{a}}$, the covariance matrix for the LLS estimates based on an approximate statistical model (Eq. (6.19), dot-dashed curve). The residual errors for an example LLS fit are also graphed ('*').

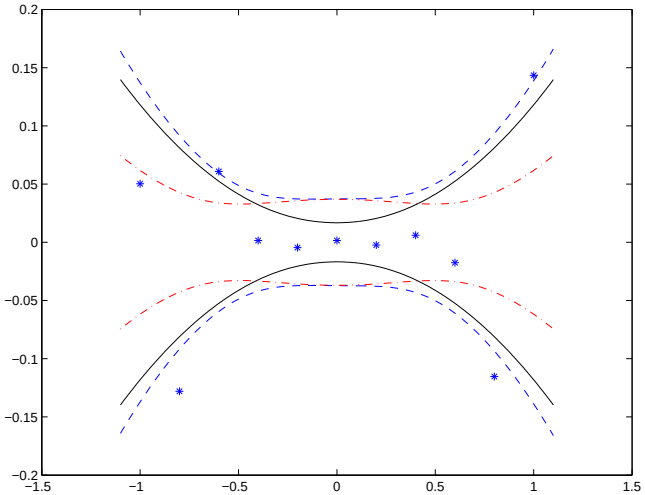

Figure 6.7: As Figure 6.6 but for $\rho = 0.01$ and $\gamma = 0.01$ (E2).

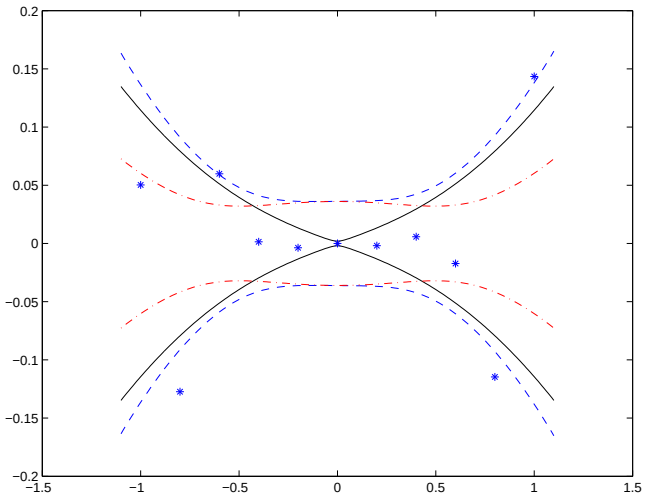

Figure 6.8: As Figure 6.6 but for $\rho = 0.01$ and $\gamma = 0.001$ (E3).

- For the case $\rho = 0.001$ and $\gamma = 0.01$ in which the measurement error in y dominates (E1), the behaviour of the LLS estimator is very similar to that of the GDR estimator. The statistics based on the approximate model derived from (6.19) underestimates slightly the uncertainty in the model predictions.
- For the case $\rho = 0.01$ and $\gamma = 0.01$ in which the standard deviation of the measurement errors in x and y are equal (E2), the LLS estimator performs less well than the GDR estimator, particularly in terms of the uncertainty of the model predictions near $x = 0$. However, the uncertainties based on the approximate statistical model significantly underestimates the uncertainty in the model predictions away from $x = 0$.
- The case $\rho = 0.01$ and $\gamma = 0.001$ in which the measurement error in x dominates (E3) is similar to that for (E2) only with larger differences in behaviour.

This analysis is similar to the analysis carried out for the case of univariate linear regression carried out in Section 6.1. In both cases, we are interested in determining how well an approximate estimator performs against an estimator that is known to be optimal. In the univariate regression case, we were able to derive correct estimates for the actual variation of the approximate estimator for the statistical model due to the fact that the estimates provided by the approximate method were a linear function of the measurement data. For the case considered in this section, the estimates defined in (6.16) are a nonlinear function of the data since the matrix C^+ also depends on the measured x variables. This means that it not so straightforward to provide valid estimates for uncertainty associated with the LLS estimator. However, Monte Carlo simulations provide a simple method of determining them.

6.3 Fitting a Gaussian peak to data

This case study concerns fitting a Gaussian peak to data.

6.3.1 Functional and statistical model

A Gaussian peak can be defined by

$$y = A \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}, \tag{6.21}$$

in terms of three parameters, the peak height A, mean μ and standard deviation σ or, more generally, in terms of quadratic coefficients $\mathbf{a} = (a_1, a_2, a_3)^{\mathrm{T}}$, as

$$y = e^{a_1 + a_2 x + a_3 x^2}. \tag{6.22}$$

(These two models are not equivalent since (6.22) can represent curves corresponding to a negative σ^2.) It is assumed that measurements of the variable x are free from measurement error but that measurements of y are subject to independent errors drawn from a normal distribution with standard deviation σ:

$$y_i^* = e^{a_1 + a_2 x_i + a_3 x_i^2}, \quad y_i = y_i^* + \epsilon_i, \quad \epsilon_i \in N(0, \sigma^2).$$

6.3.2 Estimators

We consider three methods of determining the best fit parameters to data points $X = \{(x_i, y_i)\}_1^m$.

The first is the nonlinear least squares estimator (NLLS) that estimates $\mathbf{a}_N$ by solving the nonlinear least squares problem

$$\min_{\mathbf{a}} \sum_{i=1}^{m} (y_i - e^{a_1 + a_2 x_i + a_3 x_i^2})^2.$$

From maximum likelihood principles [4, Section 3.2], we expect this estimator to perform well. This type of nonlinear optimisation problem can be solved iteratively by (variants of the) Gauss Newton algorithm [4, Section 4.2.2]. An estimate of the covariance matrix $V_{\mathbf{a}}^N$ of the fitted parameters is derived from the Jacobian matrix J evaluated at the solution. If J is the $m \times 3$ matrix

$$J_{ij} = \frac{\partial f_i}{\partial a_j}, \qquad f_i(\mathbf{a}) = y_i - e^{a_1 + a_2 x_i + a_3 x_i^2}, \quad i = 1, \ldots, m,$$

then

$$V_{\mathbf{a}}^N = \hat{\sigma}_N^2 (J^{\mathrm{T}} J)^{-1}, \tag{6.23}$$

where

$$\hat{\sigma}_N = \|\mathbf{f}\| / (m-3)^{1/2} = \left(\frac{\sum_i f_i^2}{m-3} \right)^{1/2}.$$

The second estimator arises from a transformation of (6.22). We note that if we take natural logarithms of both sides of this equation, we obtain

$$\ln y = a_1 + a_2 x + a_3 x^2, \tag{6.24}$$

an expression linear in the parameters $\mathbf{a}$. From this transformation, we define the linear least squares estimator (LLS) that determines $\mathbf{a}_L$ from the solution of

$$\min_{\mathbf{a}} \sum_{i=1}^{m} (\ln y_i - a_1 - a_2 x_i - a_3 x_i^2)^2.$$

If C is the $m \times 3$ observation matrix with $(1, x_i, x_i^2)$ in the ith row and $\mathbf{z}$ is the m-vector with $\ln y_i$ in the ith element, then $\mathbf{a}_L$ solves the matrix equation

$$C\mathbf{a}_L = \mathbf{z}$$

in the least squares sense [4, Section 4.1.3]. An estimate of the covariance of the matrix of the fitted parameters is given by

$$V_{\mathbf{a}}^L = \hat{\sigma}_L^2 (C^{\mathrm{T}} C)^{-1}, \tag{6.25}$$

where

$$\hat{\sigma}_L = \|\mathbf{r}_L\|/(m-3)^{1/2}, \quad \mathbf{r}_L = \mathbf{z} - C\mathbf{a}_L.$$

The third estimator is the weighted linear least squares estimator (WLLS) which defines $\mathbf{a}_W$ as the solution of

$$\min_{\mathbf{a}} \sum_{i=1}^{m} w_i^2 (\ln y_i - a_1 - a_2 x_i - a_3 x_i^2)^2,$$

where $w_i \geq 0$ are weights specified a priori [4, Section 4.1.5]. If $\tilde{\mathbf{z}}$ and $\tilde{C}$ are the weighted versions of $\mathbf{z}$ and C:

$$\tilde{z}_i = w_i z_i, \qquad \tilde{C}_{ij} = w_i C_{ij},$$

then $\mathbf{a}_W$ and $V_{\mathbf{a}}^W$ are estimated from

$$\tilde{C}\mathbf{a}_W = \tilde{\mathbf{z}}, \quad V_{\mathbf{a}}^W = \hat{\sigma}_W^2 (\tilde{C}^T \tilde{C})^{-1}, \tag{6.26}$$

where, in this case,

$$\hat{\sigma}_W = \|\mathbf{r}_W\|/(m-3)^{1/2}, \quad \mathbf{r}_W = \tilde{\mathbf{z}} - \tilde{C}\mathbf{a}_W.$$

While, from maximum likelihood principles, the NLLS estimator is the preferred, the LLS estimator is the easiest one to implement. We ask two questions

V1 Does the LLS estimator provide adequate solutions?

V2 Are the uncertainty estimates provided by (6.25) valid?

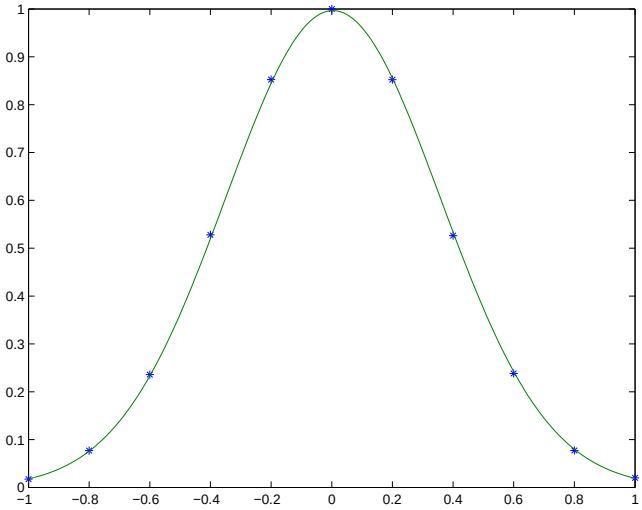

Figure 6.9: LLS fit of a Gaussian peak to simulated data.

6.3.3 Evaluation of LLS estimator using Monte Carlo Simulations

Given typical values of the parameters $\mathbf{a}^*$ and a set of values $\mathbf{x} = (x_1, \ldots, x_m)^{\mathrm{T}}$ for the independent variable, we generate data according to the model:

$$y_i^* = e^{a_1^* + a_2^* x_i + a_3^* x_i^2}, \quad y_i = y_i^* + \epsilon_i, \quad \epsilon_i \in N(0, \sigma^2),$$

and then apply the estimator to the data. Below is an example data set

x	y*	y
-1.0000	0.0183	0.0175
-0.8000	0.0773	0.0770
-0.6000	0.2369	0.2359
-0.4000	0.5273	0.5279
-0.2000	0.8521	0.8524
0	1.0000	0.9999
0.2000	0.8521	0.8522
0.4000	0.5273	0.5263
0.6000	0.2369	0.2383
0.8000	0.0773	0.0772
1.0000	0.0183	0.0200

generated with $\mathbf{a}^* = (0, 0, -4)^{\mathrm{T}}$, i.e., for the curve $y = e^{-4x^2}$, with $\sigma = 0.001$. The solution estimates provided by LLS are $\mathbf{a} = (-0.0034, 0.0324, -3.9837)^{\mathrm{T}}$. From these estimates alone, it is not easy to say whether these are adequate estimates or not. Figure 6.9 shows the fitted curve and the data points $\{(x_i, y_i)\}$. The fit looks reasonable. However, in Figure 6.10 we plot the normalised residuals errors $\mathbf{r}/\sigma$ where

$$r_i = y_i - e^{a_1 + a_2 x_i + a_3 x_i^2}. \tag{6.27}$$

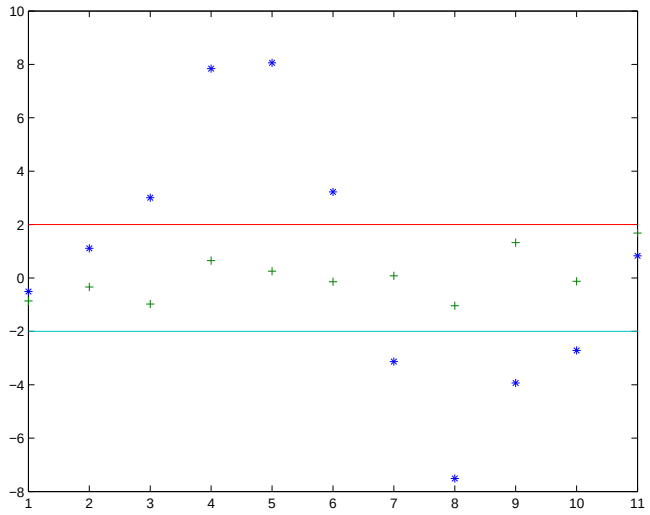

Figure 6.10: Residual errors r_i '*' for the LLS fit of a Gaussian peak to simulated data compared with the perturbations ϵ_i '+'.

From this figure, we see that the residuals are much larger than we would expect and have a non-random behaviour. Instead of most lying in the band ± 2, 8 of the 11 are outside it with three residual errors greater that 6σ in absolute value. From this view, we would say that the solution produced by the LLS estimator is poor. If we repeat the same experiment in Monte Carlo simulations, we can see if this behaviour is typical. For each of 5000 trials, we generate data sets X_q, determine the best fit parameters $\mathbf{a}_q$ and residual error vector $\mathbf{r}_q = (r_{1,q}, \ldots, r_{m,q})^{\mathrm{T}}$ calculated as in (6.27). For each $\mathbf{r}_q$ we determine an estimate

$$\hat{\sigma}_q = \|\mathbf{r}_q\|/(m-3)^{1/2}.$$

Figure 6.11 is a histogram showing the spread of the normalised estimates $\hat{\sigma}_q/\sigma$ of the standard deviation of the residual errors. This figure shows that the LLS consistently produces fits with residual errors significantly larger than those expected from the value of σ used to generate the errors ϵ_i in the data. The answer to question V1 would appear to be 'No'.

The Monte Carlo simulations can also give an answer to question V2. For the data set X, the covariance matrix $V_{\mathbf{a}}$ of the fitted parameters calculated according to (6.25) is

```
 1.2968e-004 -1.9486e-021 -1.8214e-004
-1.9486e-021  1.4207e-004 -2.2979e-021
-1.8214e-004 -2.2979e-021  4.5534e-004
```

while that $\bar{V}_{\mathbf{a}}$ derived from the estimates $\{\mathbf{a}_q\}_1^{5000}$ according to (3.1) is

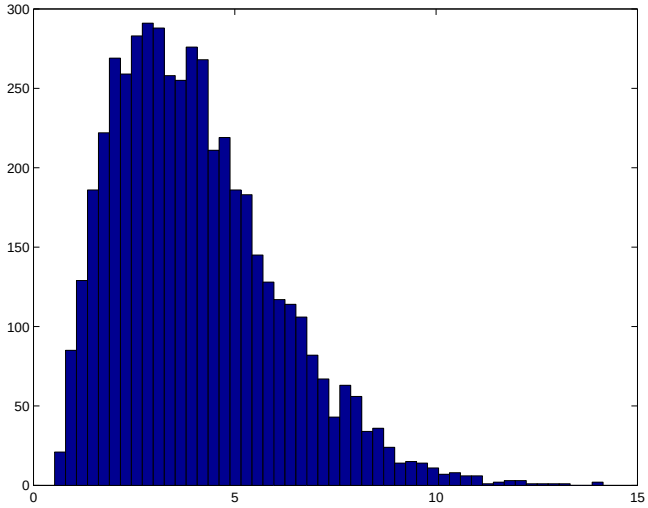

Figure 6.11: Spread of the normalised estimates $\hat{\sigma}_q/\sigma$ of the standard deviation of the residual errors for the LLS fit of a Gaussian to data over 5000 Monte Carlo simulations.

	$u(a_j)$	$\bar{u}(a_j)$
a_1	0.0114	0.0066
a_2	0.0119	0.0181
a_3	0.0213	0.0345

Table 6.4: Estimates of the standard uncertainties of the fitted parameters for the LLS estimator computed from (i) equation (6.25) (second column), (ii) 5000 Monte Carlo simulations (third column).

```
 4.4171e-005  7.5385e-007 -2.2509e-004
 7.5385e-007  3.2608e-004 -3.0955e-006
-2.2509e-004 -3.0955e-006  1.1879e-003
```

The two sets of standard uncertainties $u(a_j)$ and $\bar{u}(a_j)$) of the fitted parameters (i.e., the square roots of the diagonal elements of the corresponding covariance matrix) are given in Table 6.4.

These results show that the uncertainty estimates derived from (6.25) are not reliable.

6.3.4 Null space benchmarking for the LLS estimator

The Monte Carlo simulations have shown that the LLS estimator provides poor estimates of the parameters and their uncertainties. However, we do not know

how well this estimator compares with the preferred NLLS estimator. In this section we use null space benchmarking (Section 3.3) to answer the following questions:

V3 How much do LLS parameters estimates differ from NLLS estimates?

V4 How variable are LLS parameter estimates compared with NLLS estimates?

Given a choice of $\mathbf{a}^*$, σ and values of the independent variables x_i, $i = 1, \ldots, m$, the null space method for generating reference data for the NLLS estimator is summarised by:

I Calculate $y_i^* = e^{a_1^* + a_2^* x_i + a_3^* x_i^2}$ and set $X^* = \{(x_i, y_i^*)\}$.

II Calculate the $m \times 3$ Jacobian matrix J^*:

$$J_{ij}^* = \frac{\partial}{\partial a_j}(y_i^* - e^{a_1^* + a_2^* x_i + a_3^* x_i^2}),$$

and an orthogonal basis for the null space $Q_2 = [\mathbf{q}_4 \ldots \mathbf{q}_m]$ of ${J^*}^{\mathrm{T}}$. Compute random multipliers $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{m-3})$ satisfying

$$\|\boldsymbol{\nu}\|/(m-3)^{1/2} = \sigma.$$

III Set

$$\left.\begin{array}{rcl} \boldsymbol{\delta} & = & (\delta_1, \ldots, \delta_m)^{\mathrm{T}} = \sum_{k=1}^{m-3} \nu_k \mathbf{q}_{3+k}, \\ y_i & = & y_i^* + \delta_i, \\ X_{\boldsymbol{\delta}} & = & \{(x_i, y_i)\}_{i=1}^m. \end{array}\right\} \tag{6.28}$$

Then if σ is small enough, $\mathbf{a}^*$ are the NLLS best fit model parameters to $X_{\boldsymbol{\delta}}$ and $\boldsymbol{\delta}$ is vector of residual errors

$$\delta_i = y_i - e^{a_1^* + a_2^* x_i + a_3^* x_i^2}$$

and satisfies $\|\boldsymbol{\delta}\|/(m-3)^{1/2} = \sigma$. Furthermore,

$$V_{\mathbf{a}}^* = \sigma^2 ({J^*}^{\mathrm{T}} J^*)^{-1}, \tag{6.29}$$

is an estimate of the covariance matrix for the NLLS estimator.

In fact, the data set listed above was generated using the null space approach so that $\mathbf{a}^* = (0, 0, -4)^{\mathrm{T}}$ is the NLLS solution. The Jacobian matrix J^* for the exact data $\{(x_i, y_i^*)\}$ is

```
-0.0183    0.0183   -0.0183
-0.0773    0.0618   -0.0495
-0.2369    0.1422   -0.0853
-0.5273    0.2109   -0.0844
-0.8521    0.1704   -0.0341
-1.0000         0         0
-0.8521   -0.1704   -0.0341
-0.5273   -0.2109   -0.0844
-0.2369   -0.1422   -0.0853
-0.0773   -0.0618   -0.0495
-0.0183   -0.0183   -0.0183
```

	$u(a_j)$	$\bar{u}(a_j)$	$u^*(a_j)$	$\lvert a_j - a_j^*\rvert/u_j^*$
a_1	0.0114	0.0066	0.0007	4.8709
a_2	0.0119	0.0181	0.0023	14.3439
a_3	0.0213	0.0345	0.0064	2.5445

Table 6.5: Estimates of the standard uncertainties of the fitted parameters for the LLS estimator computed from (i) equation (6.25) (second column), (ii) 5000 Monte Carlo simulations (third column) along with estimates of the standard uncertainties for the NLSS estimator (fourth column). The fifth column shows the difference between the LLS estimate and NLLS estimate relative to the standard uncertainties $u_j^* = u^*(a_j)$.

from which we can calculate the covariance matrix $V_{\mathbf{a}}^*$

```
 4.7894e-007  5.6509e-023 -2.5569e-006
 5.6509e-023  5.1072e-006 -4.1411e-018
-2.5569e-006 -4.1411e-018  4.0917e-005
```

We can now compare the variation of the LLS estimates with those expected using the NLLS estimator. Table 6.5 shows the two sets of estimates $u(a_j)$ and $\bar{u}(a_j)$ of the standard uncertainties of the LLS fitted parameters (as in Table 6.4) along with those, $u_j^* = u^*(a_j)$ for the NLLS derived from $V_{\mathbf{a}}^*$. It is seen that actual variation $\bar{u}(a_j)$ is almost an order of magnitude greater than that predicted for the NLLS estimates. The table also shows the difference between the LLS estimate and NLLS estimate relative to the standard uncertainties $u_j^* = u^*(a_j)$. On the basis of these results, we can say that the NLLS estimator will perform much better than the LLS estimator.

6.3.5 Analysis of the LLS estimator

From the large difference between the behaviour of the LLS and NLLS estimator indicated from the null space benchmarking, it is clear that the transformation (6.24) is producing a significant change in the model. In this section, we show from an analysis of the LLS estimator why this is so. At the same time, the analysis will indicate how to define a weighted least squares estimator that can be expected to perform better.

A linear least squares estimator can be expected to perform satisfactorily if the errors in the data are uncorrelated and have equal variance (from the Gauss-Markov theorem, Section 3.4.1). We ask, therefore,

V5 To what extent are the errors associated with the model equations

$$\ln y_i = \ln(y_i^* + \epsilon_i), \qquad y_i^* = a_1 + a_2 x_i + a_3 x_i^2, \quad \epsilon_i \in N(0, \sigma^2),$$

correlated and of equal variance?

If the ϵ_i are independent of each other, then so are $\ln y_i$. However, if $\epsilon_i \in N(0, \sigma^2)$, then the variance of $z_i = \ln y_i$ is estimated by

$$\left(\frac{\partial z_i}{\partial y}\right)^2 \sigma^2 = \frac{\sigma^2}{y_i^2}, \tag{6.30}$$

using (5.2). Hence, the variances of $\ln y_i$ are approximately constant only if the $\{y_i\}$ are approximately equal.

For the data set X, the y_i's are not constant:

x	y	1/y	1/y^2
-1.0000e+000	1.7454e-002	5.7293e+001	3.2825e+003
-8.0000e-001	7.6968e-002	1.2992e+001	1.6880e+002
-6.0000e-001	2.3595e-001	4.2382e+000	1.7962e+001
-4.0000e-001	5.2794e-001	1.8941e+000	3.5878e+000
-2.0000e-001	8.5240e-001	1.1732e+000	1.3763e+000
0	9.9986e-001	1.0001e+000	1.0003e+000
2.0000e-001	8.5222e-001	1.1734e+000	1.3769e+000
4.0000e-001	5.2625e-001	1.9002e+000	3.6109e+000
6.0000e-001	2.3825e-001	4.1973e+000	1.7617e+001
8.0000e-001	7.7181e-002	1.2957e+001	1.6787e+002
1.0000e+000	1.9995e-002	5.0011e+001	2.5011e+003

Equation (6.30) indicates that with weights $w_i = y_i$, the conditions of the Gauss-Markov theorem will hold for the weighted least squares estimator

$$\min_{\mathbf{a}} \sum_{i=1}^{m} y_i^2 (\ln y_i - a_1 - a_2 x_i - a_3 x_i^2)^2, \tag{6.31}$$

and we would expect this estimator to perform better than the unweighted version.

6.3.6 Evaluation of WLLS estimator

We can repeat the validation process, this time for the WLLS estimator (6.31). Table 6.6 gives the results for 5000 Monte Carlo simulations in the same format as Table 6.5. The results show (i), there is good agreement of the WLLS estimate with the NLLS estimate (column 5), (ii) the estimates $u(a_j)$ of the standard uncertainties derived from (6.26) are in good agreement with estimates $\bar{u}(a_j)$ determined in the Monte Carlo simulations and (iii) the variation in the WLLS estimates is close to the predicted variation $u^*(a_j)$ in the NLLS estimates. From this we can conclude that WLLS estimator is fit for purpose for this type of data. If the WLLS estimator is a good approximation to the NLLS estimator, then it follows that the LLS estimator will behave similarly to the weighted nonlinear least squares estimator

$$\min_{\mathbf{a}} \sum_i \frac{1}{y_i^2} (y_i - e^{a_1 + a_2 x_i + a_3 x_i^2}).$$

This shows that the LLS estimator gives the largest weights to data points with the smallest y-values. From the values of $1/y_i$, it is seen that the first and last

	$u(a_j)$	$\bar{u}(a_j)$	$u^*(a_j)$	$\lvert a_j - a_j^*\rvert/u_j^*$
a_1	0.0007	0.0007	0.0007	0.0188
a_2	0.0023	0.0022	0.0023	-0.0093
a_3	0.0065	0.0064	0.0064	-0.0423

Table 6.6: Estimates of the standard uncertainties of the fitted parameters for the WLLS estimator computed from (i) equation (6.26) (second column), (ii) 5000 Monte Carlo simulations (third column) along with estimates of the standard uncertainties for the NLLS estimator (fourth column). The fifth column shows the difference between the WLLS estimate and NLLS estimate relative to the standard uncertainties $u_j^* = u^*(a_j)$.

data point are given a weight more that 50 times the weight attributed to middle data points. The LLS estimator is inefficient for this type of data because most of the data points are essentially ignored.

6.3.7 Evaluation of nonlinear effects using Monte Carlo simulations

Linearisation has so far entered into the calculations at two steps, firstly in the calculation of the covariance matrix for the NLLS estimator and secondly in the calculation of the variance of $\ln y_i$. We therefore ask

V6 Is the estimate of the variance of $\ln y_i$ determined in (6.30) valid?

V7 Are the estimates of $V_{\mathbf{a}}$ calculated in (6.23) valid?

To answer V6, we employ the following strategy.

I Given σ and a range of y_i, calculate $z_i = \ln y_i$ and an estimate $u(z_i) = \sigma/y_i$ of the standard uncertainty of z_i.

II For each i and for $q = 1, \ldots, N$ calculate

$$y_{i,q} = y_i + \epsilon_{i,q}, \quad z_{i,q} = \ln y_{i,q},$$

where $\epsilon_{i,q}$ is generated from a distribution with variance σ^2.

III For each i calculate the standard deviation $\bar{u}_i$ of $\{z_{i,q}\}_1^N$.

Table 6.7 summarises the results of $N = 10000$ Monte Carlo simulations with $\sigma = 0.001$ and perturbations drawn from rectangular (uniform) and normal distributions. For values of y far from zero relative to σ, there is good agreement between the predicted and actual variations. Differences become significant when y approaches σ in value.

The validity of the estimate $V_{\mathbf{a}}$ for the NLLS estimates can also be examined from Monte Carlo simulations. Below are the estimates $V_{\mathbf{a}}$ calculated from (6.23):

y_i	$z_i = \ln y_i$	$u(z_i)$	$\bar{u}_n(z_i)$	$\bar{u}_r(z_i)$
0.0050	-5.2983	0.2000	0.2124	0.2061
0.0100	-4.6052	0.1000	0.1016	0.1009
0.0500	-2.9957	0.0200	0.0201	0.0200
0.1000	-2.3026	0.0100	0.0100	0.0100
0.5000	-0.6931	0.0020	0.0020	0.0020
1.0000	0	0.0010	0.0010	0.0010

Table 6.7: Comparison of the predicted standard deviation $\bar{u}$ of $\ln y_i$ with actual standard deviations u_n and u_r generated by 10,000 Monte Carlo simulations with perturbations drawn from normal and rectangular distributions with standard deviation $\sigma = 0.001$.

	$u(a_j)$	$\bar{u}(a_j)$
a_1	0.0007	0.0007
a_2	0.0023	0.0022
a_3	0.0064	0.0064

Table 6.8: Estimates of the standard uncertainties of the fitted parameters for the NLLS estimator computed from (i) equation (6.23) (second column), (ii) 5000 Monte Carlo simulations (third column).

```
 4.7894e-007  5.6509e-023 -2.5569e-006
 5.6509e-023  5.1072e-006 -4.1411e-018
-2.5569e-006 -4.1411e-018  4.0917e-005
```

and $\bar{V}_{\mathbf{a}}$ calculated in 5000 Monte Carlo simulations:

```
 4.7215e-007  1.8884e-008 -2.5406e-006
 1.8884e-008  5.0213e-006 -1.2392e-007
-2.5406e-006 -1.2392e-007  4.0844e-005
```

The two sets of standard uncertainties $u(a_j)$ and $\bar{u}(a_j)$ of the fitted parameters (i.e., the square roots of the diagonal elements of the corresponding covariance matrix) are given in Table 6.8. The results show good agreement.

6.3.8 Valid uncertainty estimates for the LLS estimator

While we have found that (6.25) does not produce valid estimates of the uncertainty in the fitted parameters for the LLS estimator, we can derive an alternative estimate that is valid. The solution of the linear least squares problem

$$C\mathbf{a} = \mathbf{z}$$

is defined by

$$\mathbf{a} = C^{+}\mathbf{z}, \quad C^{+} = (C^{\mathrm{T}}C)^{-1}C^{\mathrm{T}},$$

	$u(a_j)$	$\bar{u}(a_j)$
a_1	0.0064	0.0066
a_2	0.0176	0.0181
a_3	0.0334	0.0345

Table 6.9: Estimates of the standard uncertainties of the fitted parameters for the LLS estimator computed from (i) equation (6.32) (second column), (ii) 5000 Monte Carlo simulations (third column).

and the covariance matrix of the solution parameters is estimated by

$$V_{\mathbf{a}} = (C^+)^{\mathrm{T}} V_{\mathbf{z}} C^+ \tag{6.32}$$

where $V_{\mathbf{z}}$ is the diagonal matrix with σ^2/y_i^2 in the ith diagonal element. Table 6.9 compares the standard uncertainties $u(a_j)$ calculated from (6.32) and those $\bar{u}(a_j)$) calculated from the Monte Carlo simulations. The results show good agreement (certainly compared with those in Table 6.4).

6.4 Estimation of the effective area of a pressure balance

An important step in the calibration of pressure balances is the estimation of the effective area of the piston-cylinder assembly. The effective area is a function of pressure and is usually modelled as a linear function involving two parameters. The calibration process therefore involves fitting a straight line to measurement data, at first sight, a straightforward process. However, in a typical cross-floating experiment in which a balance under test is hydrostatically compared with a reference balance, there is a large number of factors that need to be taken into account, including the uncertainties associated with (a) the calibration of the reference balance, and (b) the masses generating the applied loads. The fact that the same mass units are used in different combinations to produce different loads means that there is correlation amongst the errors associated with the mass measurements. Furthermore, there are additional loads (or equivalents) that need to be measured or estimated from the data. In this case study we analyse a number of algorithms for determining estimates of the effective area parameters of a pressure balance from measurement data recorded during a calibration experiment.

6.4.1 Models

At temperature t with applied mass m (corrected for air buoyancy), a pressure balance generates a pressure p given implicitly by

$$p = \frac{(m+c)g}{A(p,\mathbf{a})(1+\phi(t))}, \tag{6.33}$$

where ϕ is a known function of t, $|\phi(t)| \ll 1$, that accounts for a temperature correction, c is a measured constant obtained from a precise characterization of the balance, $A(p, \mathbf{a})$ describes the effective area of the balance in terms of the pressure p and calibration parameters $\mathbf{a}$, and g is gravitational acceleration. In practice, $A(p, \mathbf{a}) = a_1 + a_2 p$ with $a_1 = A_0$, $a_2 = A_0\lambda$, and $a_2 p$ small compared to a_1. Given a set of measurements p_i, m_i and t_i of pressure, applied load and temperature, and knowledge of c, calibrating a pressure balance means finding values for the parameters $\mathbf{a}$ that best-fit the model equations (6.33).

In a cross-float experiment [5], the pressure balance to be calibrated is connected to a reference balance whose effective area parameters $\mathbf{b}$ have previously been measured, and the applied loads on the two balances are adjusted so that both are in pressure and flow equilibrium. Suppose the reference balance generates a pressure given by (6.33) with m, c, $\mathbf{a}$ and $\phi(t)$ replaced by M, C, $\mathbf{b}$ and $\Phi((T))$, respectively. Then, when the balances are in equilibrium, we have

$$\frac{(m+c)g}{A(p,\mathbf{a})(1+\phi(t))} = \frac{(M+C)g}{A(p,\mathbf{b})(1+\Phi(T))}, \tag{6.34}$$

where the pressure p is estimated from the calibration of the reference balance.

6.4.2 Estimators

We consider three related estimators to determine the best fit parameters $\mathbf{a}$ to measurement data [6].

The weighted linear least squares estimator (WLLS) determines estimates of the parameters $\mathbf{a}$ by solving

$$\min_{\mathbf{a}} \sum_i w_i^2 f_i^2$$

where

$$f_i = \frac{(m_i + c)g}{1 + \phi(t_i)} - a_1 p_i - a_2 p_i^2$$

a linear function of $\mathbf{a} = (a_1, a_2)^{\mathrm{T}}$. The P-estimator (PLLS) determines parameter estimates by solving

$$\min_{\mathbf{a}} \sum_i e_i^2$$

where

$$e_i = \left[\frac{m_i + c}{M_i + C}\right]\left[\frac{1+\Phi(T_i)}{1+\phi(t_i}\right] - \frac{a_1 + a_2 p_i}{b_1 + b_2 p_i}.$$

The ΔP-method (DLLS) determines parameter estimates by solving

$$\min_{\mathbf{a}} \sum_{i>1} d_i^2$$

where

$$d_i = \frac{(m_i - m_1)g}{[p_i - p_1 + (\phi(t_i) - \phi(t_1)p_1][1 + \phi(t_i)]} - a_1 - a_2(p_i + p_1). \tag{6.35}$$

We ask the general validation question

V1 Under what circumstances are these estimators expected to perform well?

6.4.3 Analysis of estimator behaviour

All three estimators are linear least squares estimators and so we can be guided by the Gauss-Markov theorem (Section 3.4.1). Let

$$y_i = y(m_i, t_i) = \frac{(m_i + c)g}{1 + \phi(t_i)}.$$

The WLLS can be expect to perform well if (i) the pressures p_i are known accurately, (ii) the errors in y_i are independent and (iii) the weights w_i are chosen to be inversely proportional to the standard deviation of y_i. Condition (ii) implies that both c and g are known accurately and the errors in m_i and t_i are independent from those of m_j and t_j, $i \neq j$, for otherwise the errors in y_i would be correlated. If we model the measurements of m_i and t_i as

$$m_i = m_i^* + \mu_i, \quad t_i = t_i^* + \tau_i, \quad \mu_i \in N(0, \nu_i^2), \quad \tau_i \in N(0, \xi_i^2),$$

then the variance of y_i is estimated by

$$\sigma_i^2 = \left(\frac{\partial y}{\partial m}\right)^2 \nu_i^2 + \left(\frac{\partial y}{\partial t}\right)^2 \xi_i^2$$

where the partial derivatives (sensitivity coefficients) are evaluated at (m_i, t_i). If the weights w_i are set to $w_i = 1/\sigma_i$, then condition (iii) is satisfied.

If the pressures p_i are known accurately, then the $e_i \approx f_i/p_i$ and so the PLLS estimator can be expect to work well under the same conditions as WLLS with the additional condition that (iv) the standard deviation of the errors in y_i are approximately proportional to p_i. Since, p_i is approximately proportional to m_i, this last condition will hold if the error in y_i is dominated by measurement error in m_i and that the error in m_i is proportional to m_i.

The DLLS estimator is less obviously related to WLLS. The motivation for its definition is that the constant c is eliminated from the calculation and so c does not have to be estimated. We note that

$$a_1 + a_2(p_i + p_1) = \{a_1 p_i + a_2 p_i^2 - a_1 p_1 - a_2 p_1\} \times \frac{1}{p_i - p_1},$$

and the first term of the righthand side of (6.35) is an approximation to

$$(y_i - y_1) \times \frac{1}{p_i - p_1}.$$

From this, we can argue that this estimator will perform well if the measurement of y_1 is known accurately and the standard deviation of the measurement errors of y_i, $i > 1$, is proportional to $p_i - p_1$. The results of Monte Carlo simulation verifying this analysis are presented in [6]. We note that by choosing different weighting strategies the WLLS estimator can approximate the behaviour of both PLLS and DLLS.

6.5 Circle fitting and roundness assessment

This case study is a continuation of the circle fitting case study in [4, Section 6.2]. We are concerned with fitting a circle to data points $\{\mathbf{x}_i = (x_i, y_i)\}_{i=1}^m$. We assume that the measurements of x and y are each subject to uncorrelated normally distributed errors: if the true point on the circle is (x_i^*, y_i^*), then

$$(x_i, y_i) = (x_i^* + \delta_i, y_i^* + \epsilon_i), \quad \delta_i, \epsilon_i \in N(0, \sigma^2). \tag{6.36}$$

We consider two estimators. The first is the nonlinear least squares estimator (NLLS) that solves

$$\min_{\mathbf{a}} \sum_{i=1}^{m} d_i^2(\mathbf{a}) \tag{6.37}$$

where $d_i(\mathbf{a})$ is the orthogonal distance from the data point $\mathbf{x}_i = (x_i, y_i)$ to the circle specified by the parameters $\mathbf{a}$. For example, if the circle is specified by centre co-ordinates (a_1, a_2) and radius a_3, then

$$\begin{aligned} d_i &= r_i - a_3, \quad \text{where} \\ r_i &= ((x_i - a_1)^2 + (y_i - a_2)^2)^{1/2}. \end{aligned} \tag{6.38}$$

NLLS estimates can be determined using the Gauss-Newton algorithm ([4, Section 4.2]). If J is the Jacobian matrix defined by

$$J_{ij} = \frac{\partial d_i}{\partial a_j}$$

evaluated at the solution, then an estimate of the covariance matrix of the fitted parameter is given by

$$\hat{V}_{\mathbf{a}}^N = \hat{\sigma}_N^2 (J^{\mathrm{T}} J)^{-1}, \tag{6.39}$$

where

$$\hat{\sigma}_N = \|\mathbf{d}\|/(m-3)^{1/2}.$$

We also consider

$$V_{\mathbf{a}}^N = \sigma^2 (J^{\mathrm{T}} J)^{-1} \tag{6.40}$$

calculated with the prior value of the standard deviation.

The second is the linear least squares estimator (LLS)

$$\min_{\mathbf{a}} \sum_{i=1}^{m} (y_i + a_1(x_i^2 + y_i^2) + a_2 x_i + a_3)^2. \tag{6.41}$$

This estimator is derived from the implicit equation

$$y + a_1(x^2 + y^2) + a_2 x + a_3 = 0 \tag{6.42}$$

that can describe an arc of a circle approximately parallel with the x-axis. We can convert from these parameters to the circle centre co-ordinates (x_0, y_0) and radius r_0 using the equations

$$x_0 = -\frac{a_2}{2a_1}, \quad y_0 = -\frac{1}{2a_1}, \quad r_0 = \frac{(1 + a_2^2 - 4a_1a_3)^{1/2}}{2|a_1|}. \tag{6.43}$$

If C is the $m \times 3$ observation matrix with ith row equal to $(x_i^2 + y_i^2, x_i, 1)$ and $\mathbf{y} = (y_1, \ldots, y_m)^{\mathrm{T}}$, then LLS estimates are found by solving the linear least squares problem

$$C\mathbf{a} = -\mathbf{y}.$$

An estimate of the covariance matrix of fitted parameters is given by

$$V_{\mathbf{a}}^L = \hat{\sigma}_L^2 (C^{\mathrm{T}}C)^{-1}, \tag{6.44}$$

where

$$\hat{\sigma}_L = \|\mathbf{r}_L\|/(m-3)^{1,2}, \quad \mathbf{r}_L = \mathbf{y} + C\mathbf{a}.$$

We also can calculate

$$V_{\mathbf{a}}^L = \sigma^2 (C^{\mathrm{T}}C)^{-1}. \tag{6.45}$$

We are interested in how these estimators behave for two types of data, the first with data points approximately uniformly spaced around the circle as in Figure 6.12, the second with data points on a small arc of the circle (Figure 6.13). It was shown in [4, Section 6.2] that the NLLS provided appropriate estimates of the circle parameters for both types of data and the LLS provided appropriate estimates only for arc data. Here we ask the question:

V1 Are the estimates of the covariance matrices given in (6.40) and (6.45) valid?

We use Monte Carlo simulation to test these estimates.

In the first set of experiments we apply NLLS to data sets of 11 points uniformly distributed around a circle centred at the origin $(0, 0)$ with radius 1. The data sets $X_q = \{(x_{i,q}, y_{i,q})\}_{i=1}^m$ are generated from exact data (x_i^*, y_i^*) lying on the circle, and then perturbed according to the model

$$(x_{i,q}, y_{i,q}) = (x_i^* + \delta_{i,q}, y_i^* + \epsilon_{i,q}), \quad \delta_{i,q},\ \epsilon_{i,q} \in N(0, \sigma^2).$$

We consider three estimates of the standard uncertainties of the fitted parameters:

$u(a_j)$ calculated from (6.40) using the value of σ used to generate the data.

$\hat{u}(a_j)$ calculated as in (6.39) but with $\hat{\sigma}_N$ replaced by

$$\bar{\sigma}_N = \left(\frac{1}{N} \sum_{q=1}^{N} \hat{\sigma}_{N,q}^2 \right)^{1/2},$$

i.e., the root-mean-square value of the estimates of the standard deviation of the residual errors $\hat{\sigma}_{N,q}$ calculated for each data set X_q,

$\bar{u}(a_j)$ the standard deviation of the parameters estimates $\mathbf{a}_q$ determined in the Monte Carlo simulations.

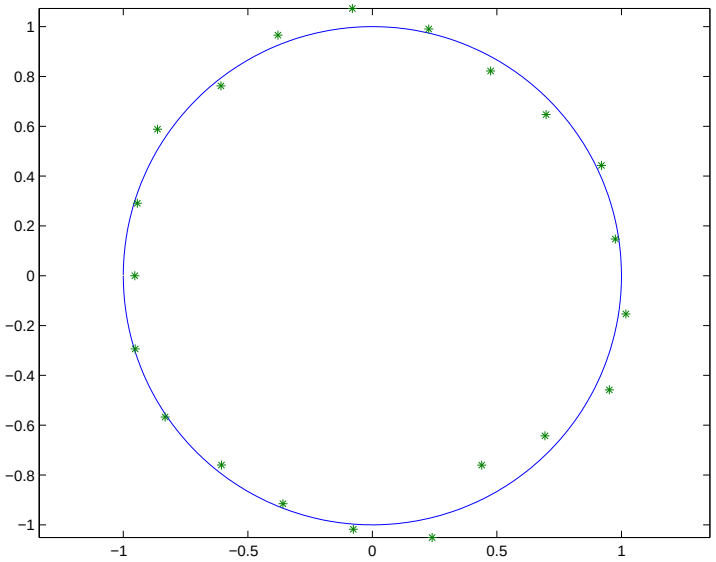

Figure 6.12: Data uniformly distributed around a circle of radius 1.

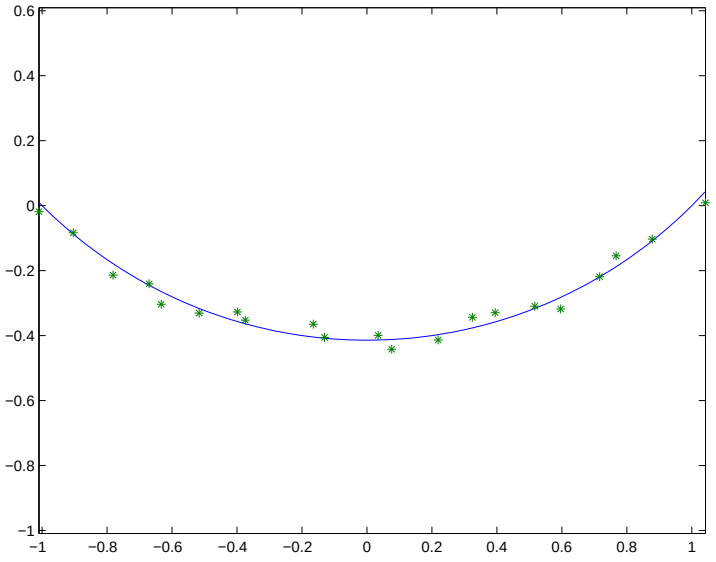

Figure 6.13: Data generated on an arc of a circle passing through $(\pm 1, 0)$.

a_j	u	$\hat{u}$	$\bar{u}$
E1	$\sigma = 0.001$		
a_1	4.2640e-004	4.2656e-004	4.3465e-004
a_2	4.2640e-004	4.2656e-004	4.2743e-004
a_3	3.0151e-004	3.0162e-004	3.0263e-004
E2	$\sigma = 0.01$		
a_1	4.2640e-003	4.2653e-003	4.3477e-003
a_2	4.2640e-003	4.2653e-003	4.2735e-003
a_3	3.0151e-003	3.0160e-003	3.0256e-003
E3	$\sigma = 0.05$		
a_1	2.1320e-002	2.1306e-002	2.1797e-002
a_2	2.1320e-002	2.1306e-002	2.1383e-002
a_3	1.5076e-002	1.5066e-002	1.5106e-002

Table 6.10: Three estimates of the standard uncertainty of the fitted parameters for the NLLS estimator for full circle data.

Table 6.10 gives these three estimates for $\sigma = 0.001$, 0.01 and 0.05. For each value, the estimates u and $\hat{u}$ are consistent with the actual variation $\bar{u}$.

In the second set of experiments we apply both the NLLS and LLS estimators to data generated from exact data $\{(x_i^*, y_i^*)\}$ lying on the arc of a circle of radius 10 between the points $(\pm 1, 0)$:

x*	y*
-1.0000	0
-0.8000	-0.0181
-0.6000	-0.0321
-0.4000	-0.0421
-0.2000	-0.0481
0	-0.0501
0.2000	-0.0481
0.4000	-0.0421
0.6000	-0.0321
0.8000	-0.0181
1.0000	0

As for the NLLS estimator, we consider three estimates of the standard uncertainty of fitted parameters for the LLS estimator:

$u_L(a_j)$ calculated from (6.45) using the value of σ used to generate the data.

$\hat{u}_L(a_j)$ calculated as in (6.44) but with $\hat{\sigma}_L$ replaced by

$$\bar{\sigma}_L = \left(\frac{1}{N} \sum_{q=1}^{N} \hat{\sigma}_{L,q}^2 \right)^{1/2},$$

i.e., the root-mean-square value of the estimates of the standard deviation of the residual errors $\hat{\sigma}_{L,q}$ calculated for each data set X_q,

$\bar{u}_L(a_j)$ the standard deviation of the parameters estimates $\mathbf{a}_q$ determined in the Monte Carlo simulations.

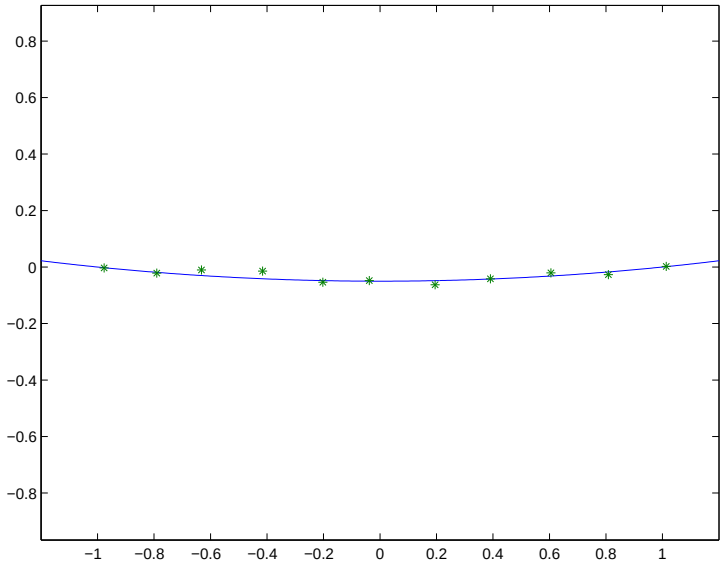

Figure 6.14: Example data generated on an arc of a circle passing through $(\pm 1, 0)$ with $\sigma = 0.015$.

Table 6.11 gives the estimates of the standard uncertainties of the circle parameters $\mathbf{a}$ (for the parametrisation of equation 6.42)) for the LLS estimator for $\sigma = 0.001$, 0.01 and 0.015. The nominal values of the parameters are $\mathbf{a}^* = (-0.0503, 0.0000, 0.0503)^{\mathrm{T}}$ (to four decimal places). A value of $\sigma = 0.015$ represents noisy data as can be seen in Figure 6.14. The table shows that the estimates u_L and $\hat{u}_L$ are consistent with $\bar{u}_L$, the standard deviation of the Monte Carlo estimates.

Table 6.12 gives the four estimates of the standard uncertainties of the circle parameters $\mathbf{a} = (x_0, y_0, r_0)^{\mathrm{T}}$, centre co-ordinates and radius for the cases $\sigma =$ 0.001, 0.01, 0.0125 and 0.015. The estimates are u, $\hat{u}$ and $\bar{u}$ defined above along with

$\bar{u}^L$ the standard deviation of the circle centre co-ordinates and radius derived from the LLS parameter estimates using (6.43).

From the results in the table we note:

- For all values of σ the estimates $\bar{u}^L$ and $\bar{u}$ are consistent with each other. This is because for this type of data the estimators produce virtually the same circle fits.
- For accurate data, $\sigma = 0.001$, there is good agreement with the predicted estimates u and $\hat{u}$ with the estimates $\bar{u}^L$ and $\bar{u}$ derived from the actual variation.

a_j	u_L	$\hat{u}_L$	$\bar{u}_L$
E3	$\sigma = 0.001$		
a_1	8.5784e-004	8.6436e-004	8.5467e-004
a_2	4.7690e-004	4.8052e-004	4.6947e-004
a_3	4.5740e-004	4.6087e-004	4.5450e-004
E3	$\sigma = 0.01$		
a_1	8.7793e-003	8.8464e-003	8.5501e-003
a_2	4.7840e-003	4.8206e-003	4.6937e-003
a_3	4.6197e-003	4.6551e-003	4.5456e-003
E3	$\sigma = 0.015$		
a_1	1.3339e-002	1.3441e-002	1.2827e-002
a_2	7.1883e-003	7.2434e-003	7.0398e-003
a_3	6.9695e-003	7.0230e-003	6.8186e-003

Table 6.11: Three estimates the standard uncertainty of the fitted parameters for the LLS estimator for arc data.

- As the value of σ becomes larger, the agreement becomes progressively worse: the estimates u and $\hat{u}$ underestimate the actual variation.

To summarise:

- We have two methods of fitting a circle to data, LLS and NLLS.
- For the type of data lying on an arc of a circle considered above, the two methods provide give essentially the same fitted circles but defined by different sets of parameters.
- The standard uncertainty estimates for the parameters determined by LLS are valid but those provided by NLLS are not.

The reason that the two methods of uncertainty estimation have different behaviour is the parameters $\mathbf{a}$ determined by the LLS estimator depend approximately linearly on the data while those $\mathbf{a} = (x_0, y_0, r_0)^{\mathrm{T}}$ determined by NLLS do not. The calculation of $V_{\mathbf{a}}^N$ in (6.39) assumes that the Jacobian matrix is approximately constant in the neighbourhood of the solution and this is not the case. In fact, we can see this from the equation for the $y-$co-ordinate of the circle centre in terms of the LLS parameters:

$$y_0 = -\frac{1}{2a_1}.$$

For simulations with $\sigma = 0.01$, we have $a_1 = -0.0503$ and, from Table 6.11, $\bar{u}(a_1) = 0.0085$. Using these values, we can estimate the standard uncertainty of y_0 from

$$u(y_0) = \left| \frac{\partial y_0}{\partial a_1} u(a_1) \right| = 1.6798.$$

This calculation assumes that the function is sufficiently linear at the value of a_0 and corresponds to the calculation of the standard uncertainty of y_0 from

a_j	u	$\bar{u}^L$	$\hat{u}$	$\bar{u}$
	$\sigma = 0.001$			
a_1	4.7673e-003	4.6706e-003	4.7794e-003	4.6706e-003
a_2	1.7027e-001	1.6938e-001	1.7070e-001	1.6938e-001
a_3	1.6993e-001	1.6904e-001	1.7036e-001	1.6904e-001
	$\sigma = 0.01$			
a_1	4.7673e-002	4.8813e-002	4.7793e-002	4.8746e-002
a_2	1.7027e+000	1.9342e+000	1.7070e+000	1.9295e+000
a_3	1.6993e+000	1.9310e+000	1.7036e+000	1.9263e+000
	$\sigma = 0.0125$			
a_1	5.9591e-002	6.3093e-002	5.9739e-002	6.2950e-002
a_2	2.1284e+000	2.6832e+000	2.1337e+000	2.6722e+000
a_3	2.1241e+000	2.6793e+000	2.1294e+000	2.6684e+000
	$\sigma = 0.015$			
a_1	7.1510e-002	8.1248e-002	7.1712e-002	8.0912e-002
a_2	2.5540e+000	4.1562e+000	2.5613e+000	4.1189e+000
a_3	2.5489e+000	4.1522e+000	2.5561e+000	4.1149e+000

Table 6.12: Four estimates the standard uncertainty of the fitted parameters (circle centre co-ordinates and radius) for circle fits to arc data. $\bar{u}^L$ is the estimate derived from variation in the parameters for the circles determined by the LLS estimator and $\bar{u}$ is that corresponding to NLLS.

(6.40). However, we can use a simple Monte Carlo simulation to estimate the uncertainty by generating

$$a_{1,q} = a_1 + \epsilon_q, \quad \epsilon_q \in N(0, 0.0085^2)$$

and calculating the standard deviation of the results. For a Monte Carlo trial with 5000 simulations, we obtain an estimate of $\bar{u}(y_0) = 1.9108$. This shows that the estimate based on linearisation is an underestimate. The situation is illustrated by Figure 6.15 which graphs $y_0 = -1/(2a)$ and its tangent line at $(-0.0503, 9.9404)$. The figure shows how a normally distributed set of a-values (displayed on the lower horizontal axis) is mapped to a set of y-co-ordinate values (displayed on the righthand vertical axis) and a linearised set of estimates of y-co-ordinates projected from the tangent line (displayed on the lefthand vertical axis). The rapid change in curvature of the graph $y_0 = -1/(2a)$ relative to the standard deviation of the a-values accounts for the discrepancy for the two sets of estimates.

The nonlinearity associated with the estimation of the radius means that for this type of data, the standard uncertainties provided by (6.40) can be unreliable. However, even if the correct covariances are calculated (from Monte Carlo simulations for example) the estimates of standard uncertainty alone do not provide a complete picture of the distribution of the radius values due to its asymmetry. For example, Figure 6.16 shows the empirical (un-normalised) distribution of radius values derived from 5000 Monte Carlo simulations for the case $\sigma = 0.0125$.

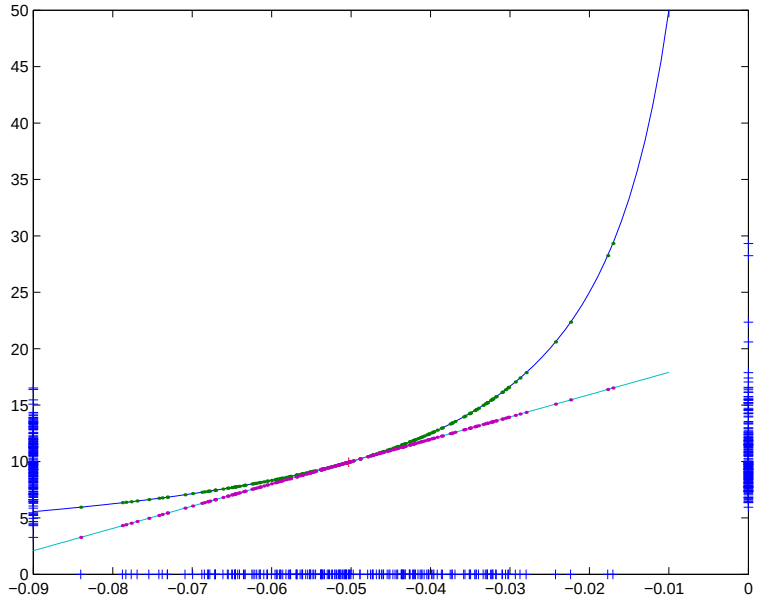

Figure 6.15: Uncertainty estimates for the y-co-ordinate of a circle fitted to data covering an arc of a circle.

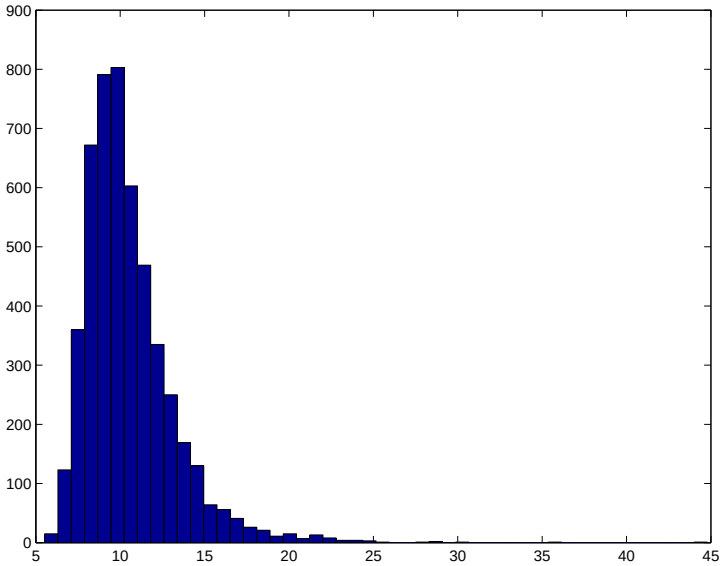

Figure 6.16: Empirical (un-normalised) distribution of radius values derived from 5000 Monte Carlo simulations for the case $\sigma = 0.0125$.

6.6 Roundness assessment in the presence of form errors

Form error is a measure of the departure of a workpiece from its nominal, ideal geometry. It is a quantity that is estimated from measurements and, in common with other measured quantities, valid ways of providing such estimates and evaluating their associated uncertainties are required. This case study is concerned with the calculation of the radius and form error and their uncertainties from measurements.

We first consider exactly the same model (6.36) and apply the NLLS estimator (6.37) to determine estimates $\mathbf{a}$ and associated estimate of the covariance matrix of the fitted parameters given by

$$\hat{V}_{\mathbf{a}} = \hat{\sigma}^2 (J^{\mathrm{T}} J)^{-1}, \tag{6.46}$$

where $\hat{\sigma}$ is an estimate of the standard deviation of the residual errors $\mathbf{d} = (d_1, \ldots, d_m)^{\mathrm{T}}$ at the solution calculated according to

$$\hat{\sigma} = \|\mathbf{d}\|/(m-3)^{1/2}.$$

We also consider the estimate

$$V_{\mathbf{a}} = \sigma^2 (J^{\mathrm{T}} J)^{-1}. \tag{6.47}$$

Suppose a circular artefact of nominal radius $a_3 = 100$ is measured by experimenter A using a co-ordinate measuring machine (CMM). From previous experiments, it is known that random measurement errors associated with the CMM are approximately normal distributed with standard deviation $\sigma = 0.001$.

Gathering data points uniformly space around the artefact, the experiment calculates the best-fit circle parameters, residual errors $\mathbf{d}$ and estimate $\hat{\sigma}$ of the standard deviation of the residual errors and standard uncertainties $u(a_j)$ using (6.47) and $\hat{u}(a_j)$ using (6.46) of the fitted parameters. In Section 6.5, it was shown that these uncertainty estimates are reliable for this type of data and model.

The expected value of $\hat{\sigma}$ is σ but the actual estimate is significantly larger. In order to obtain more information, the measurements are repeated from which mean estimates $\bar{\mathbf{a}}$ of the parameters $\mathbf{a}$ and estimates $\bar{u}(a_j)$ of their uncertainties can be determined. There are now three estimates of the uncertainty of the radius of the artefact. Which, if any, are valid? In order to resolve this question, experimenter B is asked to repeat the measurements, calculating the uncertainties using the three approaches. These uncertainty estimates for A and B are shown in Table 6.6. While for experimenter A the repeat experiments seem to confirm the estimates u derived from (6.47) based on a prior estimate of the uncertainty of the CMM measurement uncertainty σ, for experimenter B the repeat experiments indicate that both estimates $\hat{u}$ and u underestimate the actual variation.

The six sets of estimates produce a confusing picture. The explanation of why they are so different comes from two sources a) form error and b) measurement

	$\hat{u}$	u	$\bar{u}$
A	8.6474e-004	2.5000e-004	2.2082e-004
B	3.2264e-004	5.0000e-004	1.7325e-003

Table 6.13: Two sets of estimates of the uncertainty in the radius of artefact: $\hat{u}$ calculated from (6.46), u calculated from (6.47) and $\bar{u}$ calculated from 5 repeat experiments.

strategy. Figure 6.17 shows the actual shape of the circular artefact (with the form error greatly magnified) and example measurement strategies for experimenters A (data points marked '*') and B (marked '+').

The model (6.36) assumes that the departures from circularity in the data are due purely to random, independent CMM measurement error and that the artefact is perfectly circular. In fact, the artefact has a profile, expressed in polar co-ordinates, of the form

$$r(\theta, \mathbf{b}) = r_0 + s(\theta, \mathbf{b})$$

where $s(\theta, \mathbf{b})$ is a periodic function depending on parameters $\mathbf{b}$. A better model is therefore

$$\begin{aligned} x_i^* &= a_1 + r(\theta_i^*, \mathbf{b}) \cos\theta_i^*, \\ y_i^* &= a_2 + r(\theta_i^*, \mathbf{b}) \sin\theta_i^*, \\ x_i &= x_i^* + \epsilon_i, \quad y_i = y_i^* + \delta_i, \quad \epsilon_i,\ \delta_i \in N(0, \sigma^2), \end{aligned}$$

where the angles θ_i^* specify where the artefact is measured. When experimenter A first measures the artefact, the systematic errors associated with the departure from circularity lead to much larger residual errors and hence a larger estimate of $\hat{\sigma}$ and standard uncertainties than those expected from the prior estimate of σ. When A remeasures the artefact, the artefact is kept in its original position on the CMM and the same measurement strategy is executed so that the measured data is gathered from the same locations on the artefact. This means that differences between A's repeat measurements are accounted for by the random errors associated with the CMM. For this reason, the estimates $\bar{u}$ are consistent with u.

When B measures the artefact, only four measurement points are chosen on the artefact instead of the 16 chosen by A. For B's first measurement, the systematic errors associated with the form of the circle happen to have a small impact on the residual errors and the estimate $\hat{u}$ is smaller than anticipated by u. When B repeats the measurement, the artefact is replaced at random orientations so that new sets of four points are measured on the artefact, leading a much larger range of estimates for the radius and hence a larger $\bar{u}$.

6.6.1 Taking into account form error

The inconsistent information in Table 6.6 arises from differences in measurement strategy and experimental design in the presence of systematic effects that have

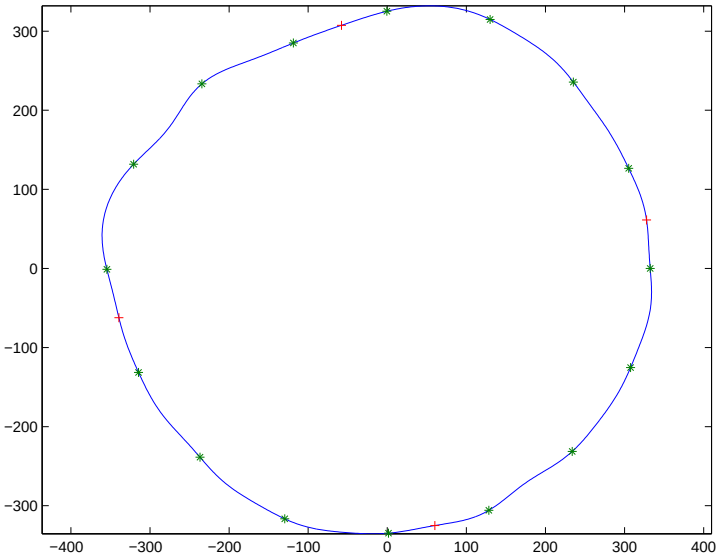

Figure 6.17: Circle form and measurement data for experimenters A ('*') and B ('+').

not been included in the model. This situation illustrates a general problem in metrology: how to account for small systematic errors in uncertainty estimation. In this section we look more closely at the problems of determining estimates of radius and form error of a nominally circular artefact.

The first issue is what do we mean by form error. A mathematical definition of roundness, i.e., form error for a circle, is as follows. Let $P = \{\mathbf{x} = (x, y)\}$ be the set of points lying on the profile of a nominally circular artefact. Given a circle $C(\mathbf{a})$ defined in terms of centre co-ordinates (a_1, a_2) and radius a_3, let $e(\mathbf{a}) = \max_{\mathbf{x}\in P} d(\mathbf{x}, \mathbf{a})$ be the maximum distance from a point in P to the circle, measured orthogonally to the circle as in (6.38). Then the form error (peak to valley) is $2e^*$ where $e^* = e(\mathbf{a}^*)$ is the value for e for the circle specified by $\mathbf{a}^*$ for which $e(\mathbf{a})$ is minimised. In other words, the form error is determined in terms of the best-fit circle that minimises the maximum deviation.

Note that in order to determine an artefact's roundness an accurate knowledge of the complete profile is required. In general, only a finite number of points gathered subject to measurement error is available. The estimate of form error derived from a finite set of accurate data is likely to be an underestimate of the form error. On the other hand, if the measurement errors are comparable to the likely form error, it is possible that a significant component of the form error determined from the data is due solely to measurement error, potentially producing an overestimate.

In many circumstances, it is expected that the uncertainties due to the probing strategy, as a consequence of form error, constitute the major contribution to

measurement uncertainty, as the example above shows. However, the relationship between probing strategy and the reliability of the resulting estimates is often complex. This leads to two fundamental problems: (a) how to assess the uncertainties associated with a given choice of probing strategy, and (b) given a measurement task, how to determine a probing strategy that is fit for purpose.

A solution to these problems can be found if we have a valid model of the likely form and measurement errors.

Example: circularity of a lobed circle. Consider the problem of determining the roundness of a lobed circle. This problem has practical importance since many manufacturing processes introduce lobing to surfaces of revolution. A circle with q lobes is described in polar coordinates by

$$r(\theta) = r_0 + A\cos(q\theta + \theta_0);$$

and the roundness of such a circle is $2A$. If we sample the circle at angles $\{\theta_i\}_{i=1}^m$, then the estimate of circularity obtained from the points is bounded above by

$$2a = \max_i A\cos(q\theta_i + \theta_0) - \min_i A\cos(q\theta_i + \theta_0).$$

We wish to choose θ_i so that this bound is as close to $2A$ as possible. If there are m points uniformly spaced around the circle, it is not difficult to show that if m and q have no common factor, then

$$\begin{aligned} A \geq \quad a \quad &\geq A\cos\frac{\pi}{m}, \quad m \text{ even}, \\ A\cos^2\frac{\pi}{2m} \geq \quad a \quad &\geq A\cos\frac{\pi}{2m}, \quad m \text{ odd}. \end{aligned}$$

We note that for m odd, $2a$ underestimates the circularity by at least a factor of $\cos\frac{\pi}{2m}$ and we can therefore take as our estimate of A

$$\hat{a} = a/\cos\frac{\pi}{2m} \geq A\cos\frac{\pi}{2m}.$$

Table 6.14 shows the value of $\cos\frac{\pi}{2m}$ for m small and prime. Five points will detect 95% of the lobing if the number of lobes q is not a multiple of 5, while eleven points will detect 99% of the lobing if q is not multiple of 11. The use of an even number of points for detecting lobing is not recommended. Figure 6.18 shows two distributions of six points on a three lobed circle; one set marked $*$ detects 100% of the lobing while the other set marked o fails to detect any lobing. By contrast the seven points shown in Figure 6.19 detects 97.5% of the lobing, slightly above the theoretical minimum given in Table 6.14.

The analysis in the example above exploits a known model for the expected form error, namely lobing. In many situations, it is possible to derive a model for the likely form error either from an assessment of the manufacturing process, past history or from the detailed analysis of a number of typical workpieces.

Figure 6.20 shows the profile of a circular artefact in terms of its departure for circularity at 2000 uniformly spaced data points. We wish to determine the roundness of artefacts with similar profile characteristics from measurements using a CMM.

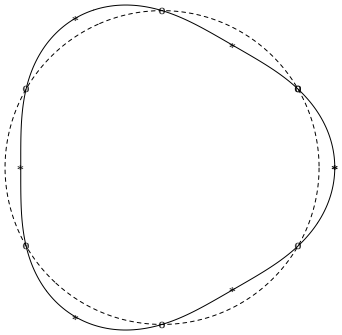

Figure 6.18: Two sets of six points uniformly spaced on a three-lobed circle. The points marked $*$ detect 100% of the lobing while the points marked $\circ$ fail to detect any form deviation.

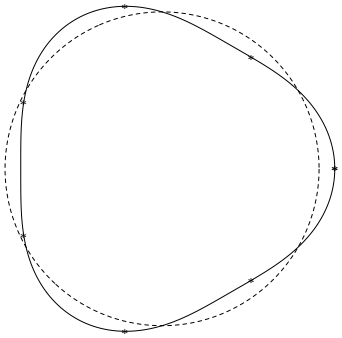

Figure 6.19: Seven points uniformly spaced on a three-lobed circle detect at least 97% of the lobing.

m	$\cos\frac{\pi}{2m}$	m	$\cos\frac{\pi}{2m}$
5	0.9511	19	0.9966
7	0.9749	23	0.9977
11	0.9898	29	0.9985
13	0.9927	31	0.9987
17	0.9957	37	0.9991

Table 6.14: Values of $\cos\frac{\pi}{2m}$ for small primes.

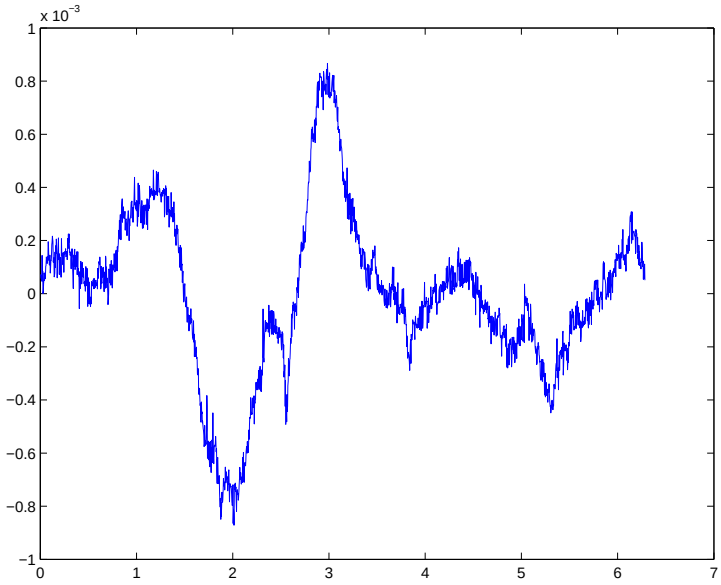

Figure 6.20: Profile of a circular artefact determined from 2000 measured points.

V1 How do we assign uncertainties to these roundness estimates?

V2 How do we determine a measurement strategy sufficient to provide accurate roundness estimates?

From the measurements, we define a virtual feature of the form

$$x^* = a_1 + r(\theta, \mathbf{b})\cos\theta, \quad y^* = a_2 + r(\theta, \mathbf{b})\sin\theta, \tag{6.48}$$

where $r(\theta, \mathbf{b}) = a_3 + s(\theta, \mathbf{b})$ is a best-fit periodic function to the profile. In this situation it is appropriate represent the profile as a Fourier series of the form

$$s_n(\theta, \mathbf{b}) = b_1 + \sum_{j=1}^{n} [b_{2j}\sin j\theta + b_{2j+1}\cos j\theta]$$

parametrized by the $2n+1$ coefficients $\mathbf{b} = (b_1, \ldots, b_{2n+1})^{\mathrm{T}}$. Fourier series are a class of empirical functions that are well suited to modelling periodic empirical functions. The Fast Fourier Transform algorithm (FFT) enables the least squares approximation of data by Fourier series to be implemented efficiently and in a numerically stable way. In order to determine a suitable fit it is necessary to select the number n of harmonics.

V3 What is a good choice of number n of harmonics for a Fourier series representation of the data represented in Figure 6.20?

This question can be answered using the strategy outlined in section 4.1:

- For $n = 1, 2, \ldots$, fit a Fourier series $s_n = s_n(\theta, \mathbf{b}_n)$ of n harmonics to the data (θ_i, r_i) and residual errors $\mathbf{f}_n = (f_1, \ldots, f_m)^{\mathrm{T}}$ where
$$f_i = r_i - s_n(\theta_i, \mathbf{b}_n).$$
- For each n, determine the estimate
$$\hat{\sigma}_n = \|\mathbf{f}_n\|/(m - 2n - 1)^{1/2}.$$
of the standard deviation of residual errors.
- Plot $\hat{\sigma}_n$ and choose n where at a point where the decrease in $\hat{\sigma}_n$ is judged to have levelled off.

Figure 6.21 shows the values of $\hat{\sigma}_n$ up to 50 harmonics. From this figure, 18 harmonics is judged suitable. Figure 6.22 shows the fits of Fourier series with 5 and 18 harmonics to the data in Figures 6.20. Figure 6.23 shows the residual errors f_i corresponding to a fit to the profile by a Fourier series with 18 harmonics. These residual errors look largely random, certainly in comparison with those associated with a circle fit in Figure 6.20.

With $r(\theta, \mathbf{b}) = a_3 + s_{18}(\theta, \mathbf{b}_{18})$ the virtual feature (6.48) is defined and can be used to generate data sets simulating measurements of the artefact according to a proposed measurement strategy. For example, suppose a measurement strategy of eight data points uniformly distribution around the circle is to be tested. Data sets can be generated according to

$$\begin{aligned} x_{i,q} &= a_1 + r(\theta_{i,q}, \mathbf{b}) \cos, \theta_{i,q} + \delta_{i,q}, \quad \delta_{i,q} \in N(0, \sigma^2) \\ y_{i,q} &= a_2 + r(\theta_{i,q}, \mathbf{b}) \sin \theta_{i,q} + \epsilon_{i,q}, \quad \epsilon_{i,q} \in N(0, \sigma^2), \end{aligned}$$

with

$$\theta_{i,q} = i\pi/4 + \phi_q, \quad 0 \le \phi_q \le \pi/4, \quad i = 1, \ldots, 8,$$

where ϕ_q is a randomly chosen phase angle specifying where on the circle the measurements are made and $\delta_{i,q}$ and $\epsilon_{i,q}$ represent random measurement errors.

Circle fitting software is then applied to these data sets to calculate the circle parameters $\mathbf{a}_q$, residual errors $\mathbf{d}_q = (d_{1,q}, \ldots, d_{8,q})^{\mathrm{T}}$ and estimate

$$e_q = \max_i d_{i,q} - \min_i d_{i,q}$$

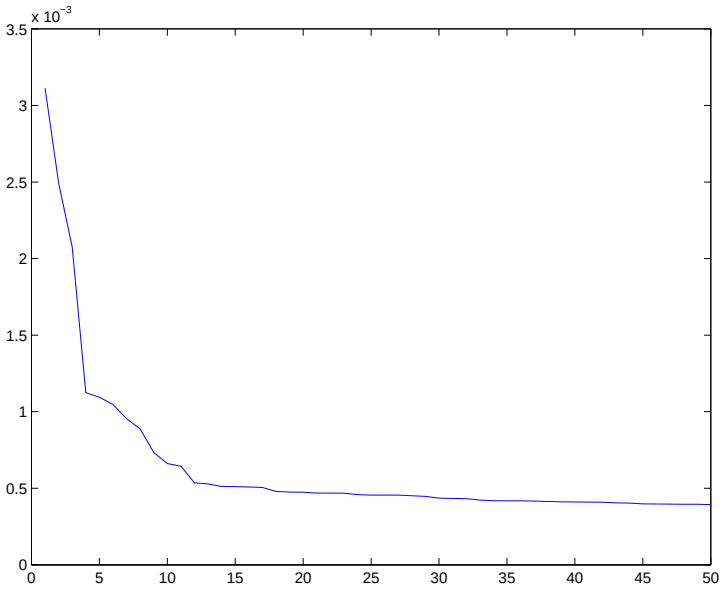

Figure 6.21: Estimates of the RMS residual errors for fits of Fourier series up to 50 harmonics to profile data in Figure 6.20.

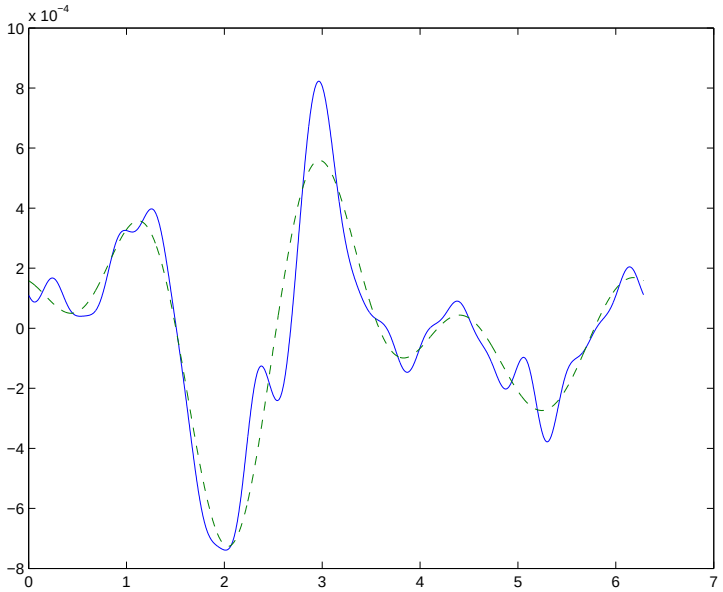

Figure 6.22: Fourier series with 5 (dashed) and 18 (solid) harmonics fitted to the profile of a circular artefact.

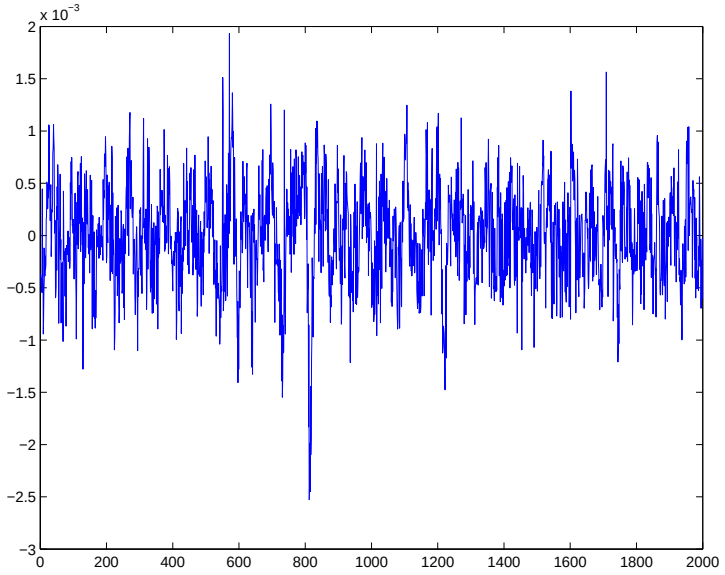

Figure 6.23: Residuals associated with the fit of a Fourier series of degree 18 to the profile of a circular artefact.

of the maximum form error for the qth data set. From a number of repeat Monte Carlo simulations we can determine the mean estimate $\bar{e}$ of the form error and its standard deviation $\bar{u}(e)$. Table 6.6.1 presents these estimates for measurement strategies with different numbers m and values $\sigma = 10^{-7}$ representing accurate measurement data and $\sigma = 0.0005$ for data generated for a circle with form error $e^* = 0.01$. The results show that for the case $\sigma = 1.0^{-7}$, the estimates of the form error e approach the true value of 0.01 and the standard deviation approaches zero as the number of points m increases. For the case $\sigma = 0.0005$, the estimate e overshoots the true value while the standard deviation levels off at approximately 0.0005 as m increases. This data in represented graphically in Figures 6.24 and 6.25.

These results were generated from just one example profile (shown in Figure 6.20). The experiments can be repeated for other profiles either based on measurements of workpieces or synthesised from empirical models. For example, we can generate profiles P_q represented by Fourier series with up to twenty harmonics with Fourier coefficients $\mathbf{b}_q$ selected at random and normalised to represent form error $e = 0.01$. The measurement strategies can then be tested on these range of profiles. Figure 6.26 presents the same information as Figure 6.24 calculated for 1000 randomly generated profiles.

From this type of simulation, it is possible to

- Quantify the bias in the estimate of form error due to undersampling.

	$\sigma = 1e-7$		$\sigma = 0.0005$	
m	$\bar{e}$	$\bar{u}$	$\bar{e}$	$\bar{u}$
4	0.0023	0.0010	0.0023	0.0011
5	0.0037	0.0016	0.0038	0.0010
14	0.0084	0.0011	0.0085	0.0012
19	0.0091	0.0008	0.0091	0.0010
25	0.0095	0.0005	0.0096	0.0007
31	0.0096	0.0003	0.0098	0.0007
38	0.0098	0.0002	0.0100	0.0006
47	0.0098	0.0001	0.0101	0.0006
57	0.0099	0.0001	0.0104	0.0006
68	0.0099	0.0001	0.0104	0.0005

Table 6.15: Mean estimates of the maximum form error $\bar{e}$ and their standard deviations $\bar{u}$ for different numbers of data points m and noise levels σ.

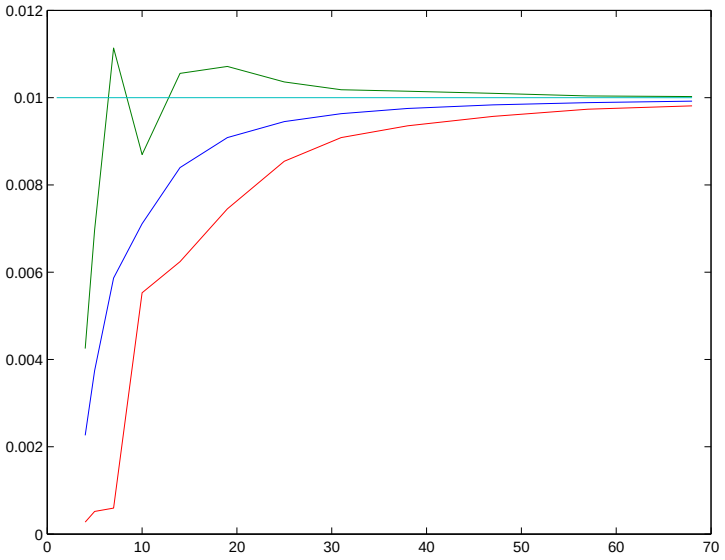

Figure 6.24: Estimates of the maximum for error $\bar{e}$ and $\bar{e} \pm 2\bar{u}$ where $\bar{u}$ is the standard deviation of the estimates determined from 1000 Monte Carlo simulations for accurate measurement data, $\sigma = 10^{-7}$.

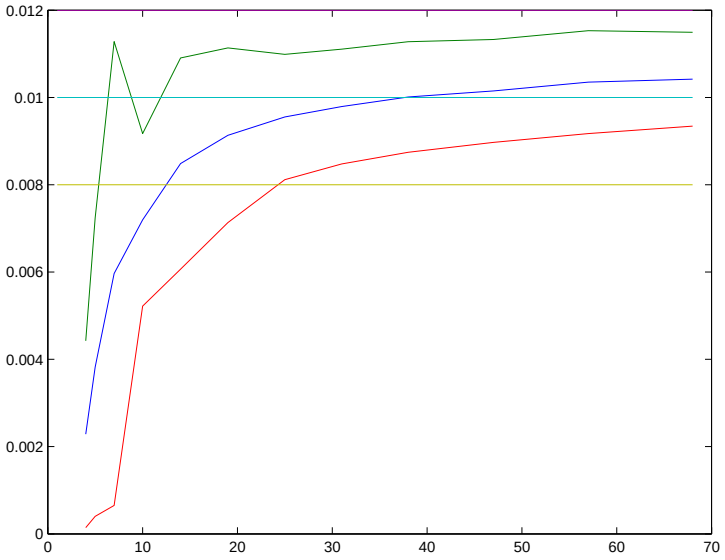

Figure 6.25: Estimates of the maximum for error $\bar{e}$ and $\bar{e} \pm 2\bar{u}$ where $\bar{u}$ is the standard deviation of the estimates determined from 1000 Monte Carlo simulations for measurement data with $\sigma = 0.0005$.

- Quantify the bias in the estimate of form error due to random measurement error.
- Provide valid estimates of the uncertainty in form error estimates.
- Select a measurement strategy that provides adequate estimates of form error.

Simulation is the only practical way of arriving at this information.

While this case study has been concerned with the effect of form error on roundness assessment, the underlying features are quite generic. Many experiments in metrology are influenced by small systematic effects that are incompletely characterised. For example, properties of materials are not known exactly and materials are not perfectly homogeneous. The departure from homogeneity can be modelled using empirical models and the effect of inhomogeneity examined using simulations, both in terms of parameter estimates and estimates of their uncertainties.

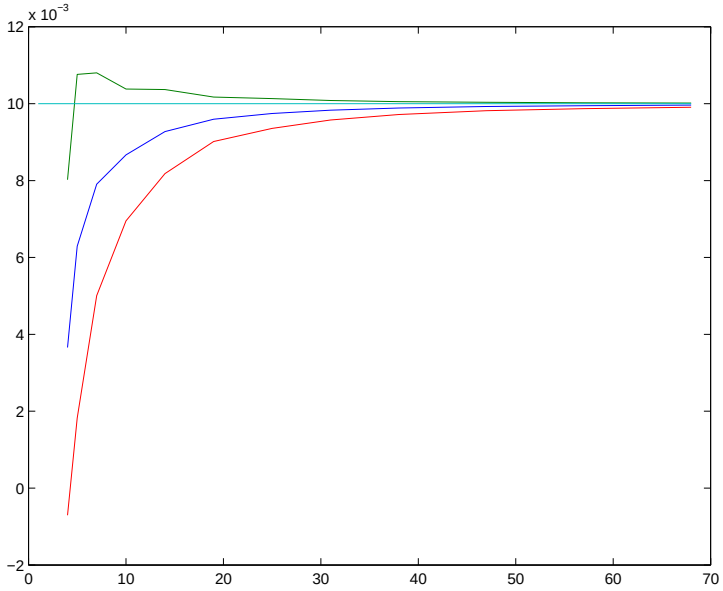

Figure 6.26: Estimates of the maximum for error $\bar{e}$ and $\bar{e} \pm 2\bar{u}$ where $\bar{u}$ is the standard deviation of the estimates determined from 1000 Monte Carlo simulations for accurate measurement data, $\sigma = 10^{-7}$. Each simulation involved a profile represented by a Fourier series with 20 harmonics generated at random.

Chapter 7

Best-practice in discrete model validation: a summary

This Best Practice Guide has been concerned with validation of the following components of discrete modelling:

Functional model consisting of:

- Problem variables representing all the quantities that are known or measured.
- Problem parameters representing the quantities that have to be determined from the measurement experiment. The problem parameters describe the possible behaviour of the system.
- The functional relationship between the variables and parameters.

Statistical model for the measurement errors consisting of:

- The error structure describing which variables are known accurately and which are subject to significant measurement error.
- The description of how the measurement errors are expected to behave.

Estimator describing a method of determining estimates of the problem parameters from the measurement data. Good estimators are unbiased, efficient and consistent.

Experimental design and measurement strategy specifying how the data is to be gathered.

The following summarises the validation techniques that can be applied to validate each of these components.

Function model validation:

- Conduct design review by an expert in the metrology field to examine the choice of model variables and check assumptions.
- Perform numerical simulations to compare the behaviour of comprehensive models with simpler models.
- Conduct design review by an expert in the metrology area to check the modelling of the underlying physics.
- Conduct design review by a modelling expert to check the mathematical derivation of equations.
- Perform numerical simulations to check the effect of approximations, simplifications, linearisations, etc., on the model values (relative to the likely measurement error).
- Evaluate the model at variable/parameter values for which the physical response is known accurately.
- Perform numerical simulations to check the qualitative behaviour of the model against expected behaviour.
- Conduct design review by a metrology and/or modelling expert to check an empirical model is appropriate for the expected type of behaviour.
- Perform numerical simulations to check the qualitative behaviour of the model against expected behaviour.

Statistical model validation

- Conduct design review by an expert in the metrology field to examine the statistical model for the measurement data and check assumptions.
- Conduct design review by modelling expert to check the statistical models for derived quantities.
- Perform numerical simulation to check the effect of approximations, simplifications, linearisations, etc., associated with the statistical model.
- Perform Monte Carlo simulations to check the variation in derived quantities against the predicted variation.

Estimator validation

- Perform Monte Carlo simulations to examine the bias and variation of the solution estimates on datasets generated according to the statistical model.

- Perform Monte Carlo simulations to compare the predicted variation of parameter estimates with the actual variation on datasets generated according to the statistical model.
- Apply the estimator to datasets for which the estimates provided by an optimal estimator are known.
- Compare the actual variation of parameter estimates on datasets generated according to the statistical model with the predicted variation for an optimal estimator.
- Compare the actual statistical model with the statistical model for which the estimator is known to perform well.

Validation of the model solution

- Examine the goodness of fit in terms of the size of the residual errors.
- Compare of the size of the residual errors with the statistical model for the measurement errors.
- Plot the residual errors to check for random/systematic behaviour.
- Plot the root-mean-square residual error for a number of model fits to select an appropriate model (e.g., the polynomial of appropriate degree).
- Calculate and check the covariance matrix for the fitted parameters against requirements and/or expected behaviour.
- Calculate and check the uncertainty associated with the model predictions againsts requirements and/or expected behaviour.

Validation of the experimental design and measurement strategy

- Examine the singular values of the matrix defining the solution for exact data. If the matrix is full rank, the measurements are sufficient to determine all the parameters. If the matrix is rank deficient the right singular vectors contain information on what degrees of freedom are left unresolved.
- Calculate the covariance matrix of the fitted parameters for data generated according to the proposed experimental design. Check that the uncertainties in the fitted parameters are sufficient to meet the required uncertainty targets.

Bibliography

[1] G. E. P. Box and G. M. Jenkins. *Time Series Analysis*. Holden-Day, San Francisco, 1976. 2.2

[2] B. P. Butler, M. G. Cox, A. B. Forbes, P. M. Harris, and G. J. Lord. Model validation in the context of metrology: a survey. Technical Report CISE 19/99, National Physical Laboratory, Teddington, 1999. 4.1.1

[3] M. G. Cox, M. P. Dainton, and P. M. Harris. Software Support for Metrology Best Practice Guide No. 6: Uncertainty and statistical modelling. Technical report, National Physical Laboratory, Teddington, 2001. 1.2, 2.2, 3.2, 3.2

[4] M. G. Cox, A. B. Forbes, and P. M. Harris. Software Support for Metrology Best Practice Guide 4: Modelling Discrete Data. Technical report, National Physical Laboratory, Teddington, 2000. 1.1, 3.2.1, 3.3.1, 3.3.2, 3.4.1, 5.1, 6.1, 6.2, 6.3.2, 6.3.2, 6.3.2, 6.5, 6.5, 6.5

[5] R. S. Dadson, S. L. Lewis, and G. N. Peggs. *The Pressure Balance*. HMSO, London, 1982. 6.4.1

[6] A. B. Forbes and P. M. Harris. Estimation algorithms in the calculation of the effective area of pressure balances. *Metrologia*, 36(6):689–692, 1999. 6.4.2, 6.4.3

[7] A. B. Forbes and P. M. Harris. Simulated instruments and uncertainty estimation. Technical Report CMSC 01/00, National Physical Laboratory, 2000. 2.1.1

[8] G. H. Golub and C. F. Van Loan. *Matrix Computations*. John Hopkins University Press, Baltimore, third edition, 1996. 3.3.1, 5.1

[9] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, London, 1979. 3.4.1

[10] J. A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, CA, second edition, 1995. 2.2

[11] B. A. Wichmann. Software Support for Metrology Best Practice Guide No. 1: Measurement system validation. Technical report, National Physical Laboratory, Teddington, 2000. 1.3