NPL Report CMSC 47/04

Report to the National Measurement System Directorate, Department of Trade and Industry

From the Software Support for Metrology Programme

Exploratory Data Analysis

By
M G Cox, P M Harris, P D Kenward and I M Smith

March 2004

Exploratory Data Analysis

M G Cox, P M Harris, P D Kenward and I M Smith
Centre for Mathematics and Scientific Computing

March 2004

ABSTRACT

Least-squares data approximation or regression is a technique widely used in metrology (and other scientific disciplines) for modelling experimental data. Often, and certainly when there is no information to the contrary, the regression is undertaken under the assumption that the measurement deviations ('errors') associated with the data are independent and identically distributed. Model validation is used to establish confidence in the reliability of a least-squares fitted model and hence predictions made on the basis of the model. The application of standard statistical tests, such as the χ^2 goodness-of-fit, Kolmogorov-Smirnov and Shapiro-Wilks' tests, is considered with the aim of validating assumptions made about the statistical model for the experimental data, viz., the distribution for the measurement deviations. In cases where these tests indicate there is doubt that the assumed statistical model applies, an approach is presented to improve knowledge of the statistical model. The approach involves applying a local analysis of the data to evaluate the standard uncertainties associated with the measurements. Applications to simulated data and to real data arising in the application of thermal analysis techniques are given.

ISSN 1471–0005

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

Least-squares data approximation or regression is a technique widely used in metrology (and other scientific disciplines) for modelling experimental data. An important application of the technique is in the *calibration* of a measurement system or instrument. Here, the experimental data consists of measurements of the response of the system to a number of stimuli, and the data is modelled by a function that describes the behaviour of the system. The fitted function, or *calibration curve*, can be used subsequently to predict, for example, the response of the system to any given stimulus and, particularly, the stimulus that produces an observed response of the system.

Let x and y denote, respectively, the stimulus and response variables. Then, the ingredients for formulating a least-squares data approximation problem are

- experimental data $(x_i,\ y_i),\ i = 1, \ldots, m$, comprising measurements y_i of the response corresponding to measurements x_i of the stimulus,
- standard uncertainties $u_i,\ i = 1, \ldots, m$, that quantify the accuracy of the measurements y_i,[1] and
- a model $y = f(x, \mathbf{c})$ defined in terms of parameters $\mathbf{c}$.

In some cases, a *physical* model, based on an understanding of the measurement system or instrument or experiment giving rise to the data, will be known. In the absence of a physical model, *empirical* models are used. Important classes of empirical models are polynomial and polynomial spline curves, with the latter providing a flexible class of models that may be used to represent a wide variety of behaviour [10, 11]. The standard uncertainty u_i corresponds to the standard deviation of possible measurements at $x = x_i$ of which y_i is one realisation.[2] It is usual to assume a statistical model for the data in which the 'error' or (*measurement*) *deviation* associated with the measurement y_i is a random sample drawn from $N(0, u_i^2)$, a normal (Gaussian) distribution with mean zero and standard deviation u_i [8, 9]. In the case that the standard uncertainties u_i are *known*, a *weighted* least-squares data approximation problem is formulated and solved, with weights given by $w_i = 1/u_i,\ i = 1, \ldots, m$ [7].

In order to have confidence in the reliability of predictions made on the basis of a least-squares fitted model to experimental data it is required that

[1] In the treatment given here the x_i are taken as exact. A generalised treatment is possible, in which the x_i are also regarded as inexact. A further generalisation is possible in which mutual dependencies are permitted among the measurement data [7].

[2] It is assumed that the inexactness in the measurements $y_i,\ i = 1, \ldots, m$, arises from random effects only [9].

1. the model is capable of representing the function underlying the data, and
2. the deviations associated with the measurements satisfy the above statistical model.

Model validation [1] is the process by which such requirements are verified in particular circumstances. If it is assumed that both 1 and 2 (above) apply, then the weighted residual deviations defined by

$$\tilde{e}_i = w_i e_i, \quad e_i = y_i - f(x_i, \mathbf{c}), \quad i = 1, \ldots, m,$$

estimate the deviations associated with the measurements (scaled by their respective standard uncertainties) and can be expected to appear as a random sample from the (standard) normal distribution $N(0, 1)$. Statistical tests may be applied to test the null hypothesis that this is the case. If the null hypothesis is accepted, predictions made on the basis of the fitted model are also accepted.

It is often the case that the standard uncertainties u_i associated with the measurements are not known. Then, and certainly where there is no information to the contrary, it is usual to assume that the deviations associated with the measurements are independent and identically distributed (i.i.d.), and are drawn from $N(0, \sigma^2)$ with *unknown* standard deviation σ. In this case, an *unweighted* least-squares data approximation problem is formulated and solved,[3] and σ is estimated from the resulting fitted model. As above, model validation is necessary to verify the 'i.i.d.' assumption, and to establish confidence in both the fitted model as well as predictions made on the basis of the fitted model.

In this work polynomial spline curves are used to model the experimental data, and it is assumed that a suitable choice of function from this class of models that adequately represents the function underlying the data can be made. This is achieved by making appropriate choices for the order and interior knot positions of the spline curve. The aims of the work are then two-fold.

Firstly, consideration is given to applying standard statistical tests, such as the χ^2 goodness-of-fit, Kolmogorov-Smirnov and Shapiro-Wilks' tests, to validate assumptions made about the statistical model for the experimental data, viz., the distribution for the deviations associated with the measurements.

Secondly, in cases where these tests indicate there is evidence that the 'i.i.d.' assumption does not hold, an approach is presented to improving our knowledge of the statistical model for the experimental data. The approach involves applying a *local* analysis of the data to evaluate the standard uncertainties associated with the measurements y_i. The motivation for the approach is that, whereas the 'i.i.d' assumption may not hold *globally* across the data set, the assumption may be

[3] Equivalent to a weighted least-squares data approximation problem with weights given by $w_i = 1,\ i = 1, \ldots, m$.

applied *locally* to obtain a local estimate of the data accuracy. The local estimates are subsequently used to formulate a weighted least-squares data approximation problem for the complete data set.

The focus is on a simple modification to the statistical model for the experimental data, concerned only with removing the assumption that the deviations associated with the measurements are identically distributed. More comprehensive adjustments to achieve conformity between a model and experimental data, the latter including the uncertainties associated with the data, may be necessary [9]. In particular, the problem of achieving model and data conformity arises in the context of laboratory intercomparisons [6].

The report is organised as follows. In Section 2 a formulation is provided of the least-squares data approximation problem, together with its particularisation for the empirical models of polynomial curves and polynomial spline curves with fixed knots. In Section 3 an overview is given of a number of general techniques for testing whether a sample of observations is drawn from a specified probability distribution. A discussion is included concerning the application of these general techniques in the context of model validation, where the sample of observations consists of the (weighted) residual deviations associated with the solution to the approximation problem, and the specified probability distribution is the normal distribution. In Section 4 an approach is described to obtaining local estimates of the accuracy for a set of experimental data. Sections 5 and 6 present the results of applying the approach to, respectively, simulated data and real data arising in the application of thermal analysis techniques. Finally, Section 7 contains a summary and recommendations.

2 Least-squares data approximation

Given are data points $(x_i,\ y_i),\ i = 1, \ldots, m$, standard uncertainties $u_i,\ i = 1, \ldots, m$, associated with the y_i, and a model $y = f(x, \mathbf{c})$ defined in terms of parameters $\mathbf{c} = (c_1, \ldots, c_n)^{\mathrm{T}}$. The least-squares data approximation problem is to determine values for the parameters $\mathbf{c}$ such that the sum of squares $\tilde{S}$, where

$$\tilde{S} = \sum_{i=1}^{m} \tilde{e}_i^2, \quad \tilde{e}_i = w_i e_i, \quad w_i = \frac{1}{u_i}, \quad e_i = y_i - f(x_i, \mathbf{c}), \tag{1}$$

is minimised. The quantities e_i and $\tilde{e}_i,\ i = 1, \ldots, m$, in the above formulation are referred to, respectively, as unweighted and weighted residual deviations. The quantities $w_i,\ i = 1, \ldots, m$, are the *weights*, and are calculated in terms of the standard uncertainties $u_i,\ i = 1, \ldots, m$. The minimisation of (1) is referred to here as a *weighted* least-squares data approximation problem.

If the model f is linear in its parameters $\mathbf{c}$, i.e.,

$$f(x, \mathbf{c}) = \sum_{j=1}^{n} c_j \phi_j(x),$$

where $\{\phi_j(x) : j = 1, \ldots, n\}$ is a set of basis functions for the class of models of which f is a member, then the weighted least-squares data approximation problem has the linear algebraic formulation

$$\min_{\mathbf{c}} \ \mathbf{e}^{\mathrm{T}} V^{-1} \mathbf{e}, \quad \mathbf{e} = \mathbf{y} - \Phi \mathbf{c}, \tag{2}$$

where Φ is an $m \times n$ matrix with $\Phi_{ij} = \phi_j(x_i)$, and

$$V = \text{diag}\left\{u_1^2, \ldots, u_m^2\right\}$$

is the uncertainty matrix (covariance matrix) associated with the measurements $\mathbf{y} = (y_1, \ldots, y_m)^{\mathrm{T}}$. Formally, the solution to (2) is given by [1, 22]

$$\mathbf{c} = \left(\Phi^{\mathrm{T}} V^{-1} \Phi\right)^{-1} \Phi^{\mathrm{T}} V^{-1} \mathbf{y}, \tag{3}$$

with associated uncertainty matrix

$$V_c = \left(\Phi^{\mathrm{T}} V^{-1} \Phi\right)^{-1}. \tag{4}$$

However, for reasons of numerical stability, matrix computational methods [18] are applied to (2) to obtain the solution and the associated uncertainty matrix.

In the case that the measurements are assumed to be 'i.i.d.' with $u_i = \sigma$ and σ unknown, the least-squares data approximation problem is formulated in terms of weights $w_i = 1, \ i = 1, \ldots, m$, and the problem is to determine values for the parameters $\mathbf{c}$ such that the sum of squares S, where

$$S = \sum_{i=1}^{m} e_i^2, \tag{5}$$

is minimised. The minimisation of (5) is referred to here as an *unweighted* least-squares data approximation problem. Expressions (3) and (4) apply also for the unweighted problem, but with

$$V = \text{diag}\left\{s^2, \ldots, s^2\right\} = s^2 I,$$

where I is the identitity matrix of order m, and s^2 is an estimate of σ^2 given by

$$s^2 = \frac{\sum_{i=1}^{m} e_i^2}{m - n}, \tag{6}$$

with $e_i,\ i = 1, \ldots, m$, evaluated at the solution. Note that the solution is independent of s^2 because, from (3),

$$\mathbf{c} = \left(\Phi^{\mathrm{T}} s^{-2} I \Phi\right)^{-1} \Phi^{\mathrm{T}} s^{-2} I \mathbf{y} = \left(\Phi^{\mathrm{T}} \Phi\right)^{-1} \Phi^{\mathrm{T}} \mathbf{y}.$$

In this work *empirical* models for least-squares data approximation are considered, particularly polynomial curves and polynomial spline curves with fixed knots. Both are examples of linear models. The representation of each model is described below, as is the formulation of the least-squares data approximation problem in terms of these representations.

2.1 Least-squares data approximation by polynomial curves

A polynomial curve $p(x, \mathbf{a})$ of order n (degree $n-1$) on the interval $x \in [x_{\min}, x_{\max}]$ has the representation

$$p(x, \mathbf{a}) = \frac{1}{2} a_0 T_0(X) + a_1 T_1(X) + \ldots + a_{n-1} T_{n-1}(X),$$

where $X \in [-1, +1]$ is related to x by

$$X = \frac{(x - x_{\min}) - (x_{\max} - x)}{x_{\max} - x_{\min}}$$

and $T_j(X),\ j = 0, \ldots, n-1$, are the *Chebyshev polynomials* (of the first kind) [17] defined by the recursion

$$T_0(X) = 1, \quad T_1(X) = X, \quad T_j(X) = 2X T_{j-1}(X) - T_{j-2}(X),\ j > 1.$$

The use of (2) together with the Chebyshev representation of polynomials, gives the linear algebraic formulation

$$\min_{\mathbf{a}}\ \mathbf{e}^{\mathrm{T}} V^{-1} \mathbf{e}, \quad \mathbf{e} = \mathbf{y} - \Phi \mathbf{a},$$

where $\mathbf{a} = (a_0, \ldots, a_{n-1})^{\mathrm{T}}$ and Φ is the $m \times n$ matrix with elements

$$\Phi_{i,j+1} = \begin{cases} \frac{1}{2} T_0(X_i), & j = 0, \\ T_j(X_i), & j > 0. \end{cases}$$

2.2 Least-squares data approximation by polynomial spline curves

Let $I := [x_{\min}, x_{\max}]$ be an interval of the x-axis, and

$$x_{\min} < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{N-1} \leq \lambda_N < x_{\max}$$

a partition of I. Define sub-intervals $I_j,\ j = 0, \ldots, N$, of I by

$$I_j = \begin{cases} [x_{\min}, \lambda_1)\,, & j = 0, \\ [\lambda_j, \lambda_{j+1})\,, & j = 1, \ldots, N-1, \\ [\lambda_N, x_{\max}]\,, & j = N. \end{cases}$$

Then, a polynomial spline curve s of order n (degree $n-1$) on I is a piecewise polynomial of order n on $I_j,\ j = 0, \ldots, N$. The spline is C^{n-k-1} at λ_j if $\text{card}(\lambda_\ell = \lambda_j, \ell \in \{1, \ldots, N\}) = k$.[4] So, for example, a spline of order $n = 4$ for which the λ_j are distinct is a piecewise cubic polynomial of continuity class C^2, i.e., continuous in value, first and second derivatives, at the points λ_j.

The partition points $\boldsymbol{\lambda} = \{\lambda_j\}_1^N$ are called the (interior) *knots* of s. To specify the complete set of knots needed to define s on I, the interior knots are augmented by (exterior) knots $\{\lambda_j\}_{1-n}^0$ and $\{\lambda_j\}_{N+1}^{N+n}$ satisfying

$$\lambda_{1-n} \leq \ldots \leq \lambda_0 \leq x_{\min}, \quad x_{\max} \leq \lambda_{N+1} \leq \ldots \leq \lambda_{N+n}.$$

For many purposes, a good choice of additional knots is

$$\lambda_{1-n} = \ldots = \lambda_0 = x_{\min}, \quad x_{\max} = \lambda_{N+1} = \ldots = \lambda_{N+n},$$

which readily permits derivative boundary conditions to be incorporated. On I, the spline curve s of order n with knots $\boldsymbol{\lambda}$ has the B-spline representation

$$s(x, \mathbf{c}) = \sum_{j=1}^{q} c_j N_{n,j}(\boldsymbol{\lambda}, x),$$

where $q = N + n$ and $N_{n,j}(\boldsymbol{\lambda}, x)$ is the *B-spline* of order n with knots $\{\lambda_\ell\}_{j-n}^j$ [4, 5, 16].

The linear algebraic formulation (2) of the least-squares data approximation problem applies with $\mathbf{c} = (c_1, \ldots, c_q)^{\mathrm{T}}$ and $\Phi_{ij} = N_{n,j}(\boldsymbol{\lambda}, x_i)$.

3 Tests for normality

An approach to validating the solution to the least-squares data approximation problem (Section 2) is to examine the weighted residual deviations $\tilde{e}_i,\ i = 1, \ldots, m$, evaluated at the solution. It is expected that the weighted residual deviations associated with a valid solution appear as a random sample from the standard normal distribution $N(0, 1)$. A systematic trend in the weighted residual deviations can be a symptom of a model that does not adequately represent the function underlying the data. For example, a trend will generally be observed when using

[4] $\text{card}(A)$ is used to denote the *cardinality* of the set A, i.e., the number of elements in the set.

a straight-line model to approximate experimental data that clearly describes a quadratic curve. A value for the mean-square *weighted* residual deviation, defined by

$$\tilde{s}^2 = \frac{\sum_{i=1}^{m} \tilde{e}_i^2}{m-n} \tag{7}$$

(cf. equation (6)), that is appreciably larger (or smaller) than unity can be a symptom that the standard uncertainties u_i associated with the measurements y_i are underestimated (or overestimated) [1].

Additionally, statistical tests are available to examine whether a given sample of observations is drawn from a specified probability distribution. Such tests would appear to be of value in the context of model validation, where the (weighted) residual deviations constitute the sample of observations and the probability distribution is a normal distribution: $N(0,1)$ for the weighted least-squares problem and $N(0,s^2)$ for the unweighted problem.

In the following sections four basic approaches to undertaking such tests are described, based on comparing

1. probability density functions,
2. cumulative distribution functions,
3. order statistics, and
4. moments.

Particular statistical tests that implement each approach are also indicated. The issues that arise in the application of these statistical tests to (weighted) residual deviations in the context of model validation are considered.

3.1 Comparing probability density functions

The standard test in this category is the χ^2 *goodness-of-fit test*. It is applied to data and distributions that are *binned*.

Let $x_i,\ i = 1, \ldots, m$, be a sample of observations of the random variable X, and $B_k,\ k = 1, \ldots, b$, a set of classes or bins for values of X. The χ^2 test statistic is defined by

$$\chi_0^2 = \sum_{k=1}^{b} \frac{(o_k - f_k)^2}{f_k},$$

where o_k is the *observed* frequency of observations for the kth bin, i.e.,

$$o_k = \text{card}\left(x_i \in B_k, i \in \{1, 2, \ldots, m\}\right),$$

and f_k is the *expected* frequency, calculated from

$$f_k = p_k m,$$

where

$$p_k = \text{Prob}\left(X \in B_k\right).$$

For large sample sizes m the distribution of the test statistic is approximately that of a χ^2 random variable with $b-1$ degrees of freedom [2].[5]

A large value of χ_0^2 suggests that the null hypothesis (that the observed frequencies o_k are drawn from a population represented by the expected frequencies f_k) is *not* true. The level of significance of the observed result χ_0^2 is given by

$$\text{Prob}\left(\chi^2 > \chi_0^2\right)$$

and can be used to decide whether to accept or reject the null hypothesis.

A disadvantage of the χ^2 goodness-of-fit test is that it is necessary to decide the number b of bins used as well as the 'bin-edges'. Especially for small sample sizes m the results from a χ^2 goodness-of-fit test can vary depending on the choice of b and their bin-edges. An advantage of the test is that the individual contributions to the test statistic from the bins can give some information as to the nature of the departure of the sample of observations from the specified distribution. Furthermore, the test is quick to apply, even for large samples, because the calculation of the test statistic does not require the sample to be sorted (as is necessary for other tests, see below).[6]

The NAG library [23] includes a routine (G08CGF) that provides an implementation of the χ^2 goodness-of-fit test in the case that the specified distribution takes one of the following forms: normal, uniform, exponential, χ^2, gamma, or a distribution defined by user-supplied probabilities p_k.

3.2 Comparing cumulative distribution functions

The standard test in this category is the *Kolmogorov-Smirnov* test [20, 27, 28]. It is applied to distributions that are continuous and data that is not binned.

Let $x_i,\ i = 1,\ldots,m$, be a sample of observations of the random variable X. Suppose $F(x)$ is the expected cumulative distribution function for X, and $S(x)$ is

[5] It is assumed here that the expected frequencies sum to the sample size m, i.e., $\sum_{k=1}^{m} f_k = m$. A normalisation of the expected frequencies can be used to ensure this property. Alternatively, the bins may be chosen to ensure the probabilities sum to one, i.e., $\sum_{k=1}^{b} p_k = 1$. Note that if the model that gives the expected frequencies has free parameters that are estimated from the data, then each such parameter decreases the degrees of freedom by one.

[6] It is necessary only to decide to which a bin a sample value belongs and, generally, the number b of bins is appreciably smaller than the number m of samples.

the *sample* cumulative distribution function defined by

$$S(x) = \frac{1}{m}\text{card}\left(x_i \leq X, i \in \{1, 2, \ldots, m\}\right),$$

i.e., $S(x)$ is the fraction of the sample to the left of x. The function is constant between consecutive values $x_{(i)}$ (obtained by *sorting* the values x_i into ascending order), and jumps by the same constant $1/m$ at each $x_{(i)}$ (provided the $x_{(i)}$ are distinct).

The Kolmogorov-Smirnov test statistic is defined by

$$D = \max_x |S(x) - F(x)|,$$

i.e., the maximum value of the absolute difference between the sample and expected cumulative distribution functions. The distribution of the test statistic in the case of the null hypothesis (that the sample of observations is drawn from a distribution specified by $F(X)$) can be calculated, at least to a useful approximation, thus giving the level of significance of an observed result.

A property of the test is that it is invariant under a reparametrization of X [24]. A disadavantage is that the cumulative distribution function $F(X)$ must be fully specified, i.e., the parameters of $F(X)$, such as its expectation and variance, must be known. A variation of the Kolmogorov-Smirnov test that is applied in the case that $F(X)$ corresponds to a normal distribution whose mean and variance are estimated is the Lilliefors test [3, 21].

The NAG library [23] includes a routine (G08CBF) that provides an implementation of the Kolmogorov-Smirnov test in the case that the specified distribution takes one of the following forms: uniform, normal, gamma, beta, binomial, exponential and Poisson. In addition, NAG routine G08CCF provides an implementation of the test for a user-specified distribution.

3.3 Comparing order statistics

The most common 'test' for normality is to use a *normal probability plot*. Graph paper with scales such that a plot of the cumulative distribution function for a normal distribution takes the form of a straight-line is called 'normal probability paper'.

Let $x_i,\ i = 1, \ldots, m$, be a sample of observations of the random variable X with sample cumulative distribution $S(x)$ (see above). If X is normally distributed then the points $(x_i, S(x_i)),\ i = 1, \ldots, m$, should lie close to a straight-line when plotted on normal probability paper. A measure of the 'closeness' of the points to a straight-line is therefore a test for whether the sample of observations is drawn

from a normal distribution. This is the basis of the Shapiro-Wilks' test.[7]

The Shapiro-Wilks' test statistic [25, 26] is defined by

$$W = \frac{\left(\sum_{i=1}^{m} w_i x_{(i)}\right)^2}{\sum_{i=1}^{m}(x_i - \bar{x})^2},$$

where $x_{(i)},\ i = 1, \ldots, m$, denotes the sample values arranged in non-decreasing order, $\bar{x}$ is the sample mean, and $w_i,\ i = 1, \ldots, m$, are 'weights' whose values only depend on the sample size m. The statistic W can be interpreted as the squared correlation coefficient between the ordered sample values $x_{(i)}$ and the weights w_i that correspond to the expected values of standard normal order statistics for a sample of size m. W is a measure of the straightness of the normal probability plot.

The NAG library [23] includes a routine (G01DDF) that provides an implementation of the Shapiro-Wilk's test, returning values for the weights w_i, the test statistic W, and the significance level for W under the null hypothesis that the sample of observations is drawn from a normal distribution.

3.4 Comparing moments

The normal distribution has zero skewness (third moment) and zero (standardised) kurtosis (fourth moment). Consequently, the values of the sample skewness and (standardised) kurtosis calculated for a sample of observations $x_i,\ i = 1, \ldots, m$, can be used as the basis for testing whether the sample is drawn from a normal distribution. However, these sample statistics tend to be non-normally distributed, and so are first transformed to have an approximate normal distribution. This leads to the D'Agostino and Anscombe-Glynn tests [15]. A combined test based on comparing both skewness and kurtosis is the D'Agostino-Pearson test [14].

3.5 Application to residual deviations

The approaches and tests described above are used to test whether a sample of observations is drawn from a specified probability distribution, such as a normal distribution. It appears to be common to apply these tests without further modification to the (weighted) residual deviations obtained from fitting a model to experimental data. However, the following considerations are relevant.

[7] The normal probability plot can also be produced by plotting the ordered observations against the theoretical quantiles of a normal distribution. In the case that the observations correspond to the (weighted) residual deviations associated with a fitted model to experimental data, other 'residual plots' can also highlight non-normal behaviour, for example a plot of the weighted residual deviations against the fitted values can indicate heteroscedastic data, i.e., data for which the associated measurement deviations are independent but not identically distributed.

1. The residual deviations are not independent. This is the case even if the original observations y_i are independent. The effect of the correlation induced by the regression is usually not considered to have a significant effect.

2. The residual deviations are linear combinations of the observations, and so by the Central Limit theorem will tend to normality even if the (true) measurement deviations are not drawn from a normal distribution.

3. The rejection of the null hypothesis (of normality) may be due to an inadequate model. Consequently, tests for normality should always be used in combination with other tests for validating the suitability of the model.

It should be noted that while the theoretical framework for (least-squares) methods may be built on the assumption that the data (deviations) are normally distributed, many of the inferences made on the basis of least-squares methods are robust to that assumption. For example, if the data (deviations) are 'i.i.d.' and normally distributed, then their (sample) mean also follows a normal distribution. However, if the data is not normally distributed, then in most cases and for large samples the mean will, by the Central Limit theorem, still follow an approximately normal distribution.

A further point is that if the mean and variance of the population from which the sample is drawn are unknown, it usually requires a large number of observations to detect non-normality. Estimating the mean and variance from the sample of observations provides a 'best-fit' normal distribution for the data that will mask some of the non-normality, e.g., a large sample variance will negate some of the effect of kurtosis.

Finally, it is also necessary to consider the source of any 'non-normality'. The non-normality of a sample of observations may be due to the presence of a small number of 'odd' (outlier) values rather than a non-normal distribution for the bulk of the data. It may be sensible to remove these values before testing the bulk of the data for normality (although tests for outliers will usually assume a distribution for the data in order to assess what is to be regarded as an extreme value).

Examples of the results of applying a selection of the tests described above in the context of model validation are given in Sections 5 and 6.

4 Local analysis

In this section a procedure is presented to evaluate the standard uncertainties associated with the measurements y_i in cases where such information is not available or is shown to be invalid, for example by the tests described in Section 3. The

motivation for the approach is that, whereas the 'i.i.d.' assumption may not hold *globally* across the data set, the assumption may be applied *locally* to obtain a local estimate of the data accuracy. Polynomials are used to approximate subsets of the data, in terms of which the local data accuracy is estimated. The local estimates are subsequently used to formulate a weighted least-squares data approximation problem for the complete data set in terms of a polynomial spline curve with fixed knots.

Subsets of the complete set of data are constructed consisting of ℓ (ℓ given) successive points. Each subset X_k, where

$$X_k = \{(x_i, y_i) : \ i = k, \ldots, k + \ell - 1\}, \ k = 1, \ldots, m - \ell + 1,$$

is approximated by an (unweighted) least-squares polynomial fit of low order n (n given), and the root-mean-square residual deviation s_k associated with the residual deviations $e_{i,k}, \ i = k, \ldots, k + \ell - 1$, of the fit is computed from

$$s_k^2 = \frac{\sum_{i=k}^{k+\ell-1} e_{i,k}^2}{\ell - n}$$

(cf. equation (6)).

Suppose $\hat{x}_k$ defines the midpoint of the interval containing the x-values for the kth subset of data, viz.,

$$\hat{x}_k = \frac{x_k + x_{k+\ell-1}}{2}, \quad k = 1, \ldots, m - \ell + 1.$$

(For the particular case of uniformly-spaced data and ℓ odd, $\hat{x}_k$ will correspond to the middle x-value for the kth data set, viz., $\hat{x}_k = x_{k+(\ell-1)/2}$.) Then, s_k provides an estimate of the standard uncertainty associated with a measurement made at $\hat{x}_k$. Defining

$$(\hat{x}_0, s_0) = (x_1, s_1)$$

and

$$(\hat{x}_{m-\ell+2}, s_{m-\ell+2}) = (x_m, s_{m-\ell+1}),$$

the piecewise-linear function joining the points

$$(\hat{x}_k, s_k), \ k = 0, \ldots, m - \ell + 2,$$

provides an estimate of the data accuracy associated with measurements made anywhere in the interval $[x_1, x_m]$ containing the data.

To use this procedure it is necessary to choose values for ℓ and n. Clearly, ℓ must exceed n, for otherwise each subset is interpolated by a polynomial, and nothing can be learnt about the data accuracy. On the other hand, if ℓ is large compared with n, the polynomial models may consistently underfit the data giving overestimates of the uncertainties associated with the measurements. We have found that the

choice $\ell = 5$ and $n = 3$, i.e., quadratic polynomial fits to subsets of five points, gives acceptable results in many cases. By repeating the analysis using other values for ℓ and n, the estimates obtained by this procedure may be validated: in some cases the value of s_k may be sensitive to the choice of ℓ and n.

5 Application to simulated data

In this section simulated data is used to illustrate the application of the procedure described in Section 4 to evaluate the standard uncertainties associated with measurements y_i.

Figure 1 shows an example of some simulated data.[8] The function underlying the data is a cubic (order $n = 4$) spline curve with $N = 6$ knots. The measurement deviation associated with the measurement $y_i,\ i = 1, \ldots, m$, is a random sample from $N(0, \sigma_i^2)$ with σ_i^2 varying as a quadratic function of x_i. (Thus, the measurement deviations are mutually independent but *not* identically distributed.) The 'standard uncertainty profile' for the data, corresponding to the piecewise-linear function joining the points $(x_i, \sigma_i),\ i = 1, \ldots, m$, is shown in Figure 2.

Figure 3 shows the weighted residual deviations $\tilde{e}_i,\ i = 1, \ldots, m$, associated with a weighted least-squares spline fit to the data using (correct) weights $w_i = 1/\sigma_i,\ i = 1, \ldots, m$.[9] Figure 4 shows the weighted residual deviations as a frequency distribution (histogram) scaled to have an area of unity[10] together with the probability density function for $N(0, \tilde{s}^2)$ with $\tilde{s}$ the weighted root-mean-square residual deviation evaluated from

$$\tilde{s}^2 = \frac{\sum_{i=1}^{m} \tilde{e}_i^2}{m - (n + N)}. \tag{8}$$

The first row of numerical values in Table 1 contains the value of $\tilde{s}^2$ and the results of applying three statistical tests, viz., χ^2 goodness-of-fit,[11] Kolmogorov-Smirnov and Shapiro-Wilks', to test the null hypothesis that the weighted residual deviations form a sample of observations of a random variable with distribution $N(0, \tilde{s}^2)$.[12] The table contains, for each test, the 'p–value' for the observed value of the test

[8] In this example, there are $m = 1001$ measurements uniformly spaced between $x_{\min} = 100$ and $x_{\max} = 110$.

[9] Throughout this example, the fitted model is a cubic spline curve with the same (six) knots as the curve used to simulate the data. Consequently, the model is capable of representing the function underlying the data (Section 1).

[10] This scaled frequency distribution provides an approximation to the probability density function for the distribution of weighted residual deviations.

[11] With $b = 50$ bins.

[12] The expected value for $\tilde{s}^2$ is unity. We choose here to test whether the weighted residuals are samples from *a* normal distribution, rather than from the *particular* normal distribution $N(0, 1)$ (Section 3).

statistic for the test, viz., the probability of observing a similar or more extreme value for the test statistic given the null hypothesis is true. A p–value that is small (for example, less than 0.05) implies a significant result and evidence to reject the null hypothesis. The other rows in this table show the results for other data sets, simulated on the basis of the same underlying functional and statistical models. The results of Figures 3 and 4, and Table 1, are obtained on the basis of the 'correct' statistical model for the data. It is concluded that the null hypothesis about the statistical model is accepted, and (usually) with a high level of confidence.[13]

Figures 5 and 6 show the residual deviations $e_i,\ i = 1, \ldots, m$, associated with an unweighted least-squares spline fit to the data of Figure 1. In Figure 6 the probability density function for $N(0, s^2)$ is also shown,[14] with s the root-mean-square residual deviation evaluated from

$$s^2 = \frac{\sum_{i=1}^{m} e_i^2}{m - (n + N)}. \tag{9}$$

Table 2 contains the value of s^2 and the results of testing the null hypothesis that the residual deviations form a sample of observations of a random variable with distribution $N(0, s^2)$. The results of Figures 5 and 6, and Table 2, are obtained on the basis of an 'incorrect' statistical model for the data. The inhomogeneity of the residual deviations is clear in Figure 5, and to a lesser extent in Figure 6. The p–values given in Table 2 are generally smaller than the corresponding values given in Table 1.[15] On the basis of the χ^2 goodness-of-fit test, the null hypothesis would be rejected for all four data sets.

Figure 7 shows the standard uncertainties $s_k,\ k = 0, \ldots, m-\ell+2$, associated with the data y_i obtained using the local analysis procedure described in Section 4.[16] The standard uncertainty profile for the data (Figure 2) is shown as the solid curve. It is clear that the values s_k describe the correct 'shape' of standard uncertainty profile for the data, but exhibit considerable scatter about that profile. Figure 8 shows estimates of the profile obtained by 'smoothing' the standard uncertainty data $(\hat{x}_k, s_k),\ k = 0, \ldots, m - \ell + 2$, using polynomials of degrees 2, 3, 4 and 5. Again, the standard uncertainty profile for the data is shown as the black solid curve. The reason for replacing the individual estimates s_k by 'smoothed' values is that each estimate, being derived from a small number of measurements, might not be expected to be very reliable. The results show that the 'smoothed' values are not unduly sensitive to the degree of the approximating polynomial.

[13] The exception is the χ^2 goodness-of-fit test applied to data set 2 and (to a lesser extent) data set 4.

[14] It is interesting that the scaled frequency distribution shows a 'sharper peak' than the probability distribution function. This appears to be the case for the other results presented in this section and Section 6.

[15] The exception is the Shapiro-Wilks' test applied to data set 1.

[16] Based on using quadratic polynomial fits to subsets of $\ell = 5$ points (Section 4).

Finally, Figures 9 and 10 show the weighted residual deviations associated with a weighted least-squares spline fit using estimated weights. The weights are calculated in terms of standard uncertainties obtained by evaluating the degree five polynomial approximation to the standard uncertainty data shown in Figure 7. In Figure 10 the probability density function for $N(0, \tilde{s}^2)$ is also shown. Table 3 contains the value of $\tilde{s}^2$ and the results of testing the null hypothesis that the weighted residual deviations form a sample of observations of a random variable with distribution $N(0, \tilde{s}^2)$. The results in Figures 9 and 10, and Table 3, are obtained on the basis of an 'estimated' statistical model for the data. The weighted residual deviations appear 'random' and homogenous (compare with Figure 5). The p–values given in Table 3 are generally greater than the corresponding values given in Table 2 and comparable to those given in Table 1.[17] There is evidence to accept the null hypothesis about the statistical model, and (usually) with a high level of confidence.[18]

6 Application to the analysis of thermophysical data

The example considered here concerns the application of thermal analysis techniques, including differential scanning calorimetry (DSC), dynamic mechanical analysis (DMA) and deflection temperature under load (DTUL), for the assessment of the processing and performance of polymer composites and adhesives. DSC [19] is a thermal analysis technique in which the difference between the heat flux (power) into a test specimen and that into a reference specimen is measured as a function of temperature (and/or time) while the specimens are subjected to a controlled temperature programme.

A study of a data set arising from the application of DSC is available [12]. The aim of the study is to 'extract features' from the data, in particular the positions of peaks and troughs. The features provide information about the characteristic temperatures that correspond to, for example, the crytallization and melting transitions of the material of which the test specimen is made. Here, we use a part of that data set to illustrate the application of the procedure described in Section 4 to evaluate the standard uncertainties associated with the measurements of heat flow. The part of the data set considered comprises $m = 361$ measurements between 200 °C and 300 °C, and is shown in Figure 11. Throughout this section, the x–axis represents temperature (in °C) and the y–axis heat flow (in mW).

[17] The exceptions are the χ^2 goodness-of-fit test applied to data sets 2 and 3, and the Shapiro-Wilks' test applied to data set 1.

[18] The large range of p–values observed in Tables 1 and 3 for the χ^2 goodness-of-fit test is somewhat surprising. When applying any of the tests considered to normally distributed data one would expect, by chance, to classify some as non-normal. Here, the χ^2 goodness-of-fit test appears to be rejecting (although for a small number of cases) a larger than expected proportion. This aspect requires further investigation.

	$\tilde{s}^2$	p–value		
		χ^2	K–S	S–W
Data set 1	1.03	0.63	1.00	0.23
Data set 2	0.99	0.05	0.76	0.70
Data set 3	1.00	0.74	1.00	0.91
Data set 4	0.98	0.16	1.00	0.26

Table 1: Results of applying χ^2 goodness-of-fit, Kolmogorov-Smirnov (K–S) and Shapiro-Wilks' (S–W) tests of the null hypothesis that the weighted residual deviations associated with a weighted least-squares spline fit using correct weights form a sample of observations of a random variable with distribution $N(0, \tilde{s}^2)$ with $\tilde{s}$ the weighted root-mean-square residual deviation. Data set 1 corresponds to the simulated data illustrated in Figures 1–10.

	s^2	p–value		
		χ^2	K–S	S–W
Data set 1	0.009 3	0.00	0.19	0.25
Data set 2	0.009 1	0.00	0.02	0.21
Data set 3	0.009 1	0.00	0.16	0.22
Data set 4	0.008 9	0.00	0.33	0.03

Table 2: Results of applying statistical tests of the null hypothesis that the residual deviations associated with an unweighted least-squares spline fit form a sample of observations of a random variable with distribution $N(0, s^2)$ with s the root-mean-square residual deviation.

	$\tilde{s}^2$	p–value		
		χ^2	K–S	S–W
Data set 1	1.34	0.48	1.00	0.15
Data set 2	1.30	0.00	1.00	0.84
Data set 3	1.16	0.06	1.00	0.96
Data set 4	1.25	0.99	0.95	0.19

Table 3: Results of applying statistical tests of the null hypothesis that the weighted residual deviations associated with a weighted least-squares spline fit using estimated weights form a sample of observations of a random variable with distribution $N(0, \tilde{s}^2)$.

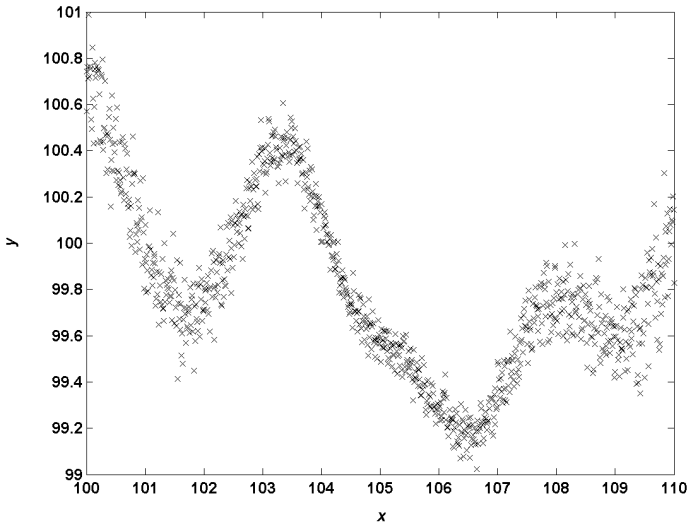

Figure 1: Simulated measurements. The function underlying the data is a cubic (order $n = 4$) spline curve with $N = 6$ interior knots. The measurement deviation associated with y_i is a random sample from $N(0, \sigma_i^2)$ with σ_i^2 varying (quadratically) with x_i: see Figure 2.

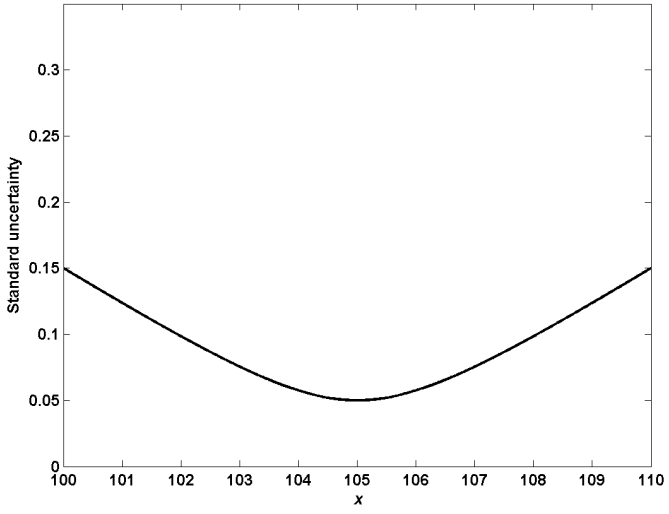

Figure 2: Standard uncertainty profile for the simulated measurements of Figure 1. The profile is shown as the piecewise-linear function joining the points $(x_i, \sigma_i),\ i = 1, \ldots, m$.

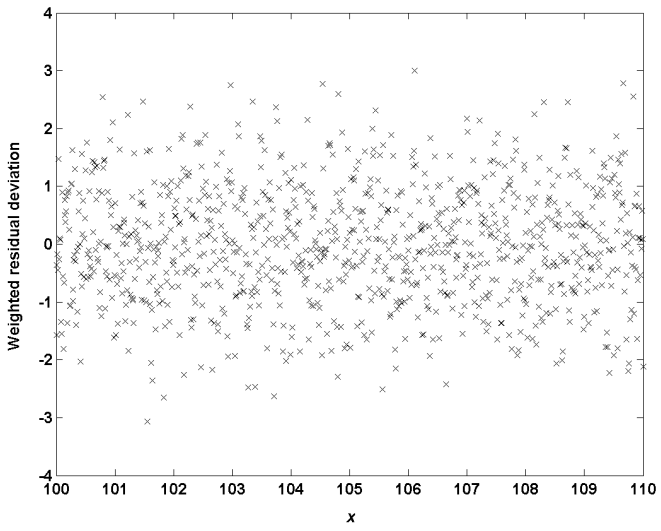

Figure 3: Weighted residual deviations associated with a weighted least-squares spline fit to the data shown in Figure 1, with weights $w_i = 1/\sigma_i,\ i = 1, \ldots, m$.

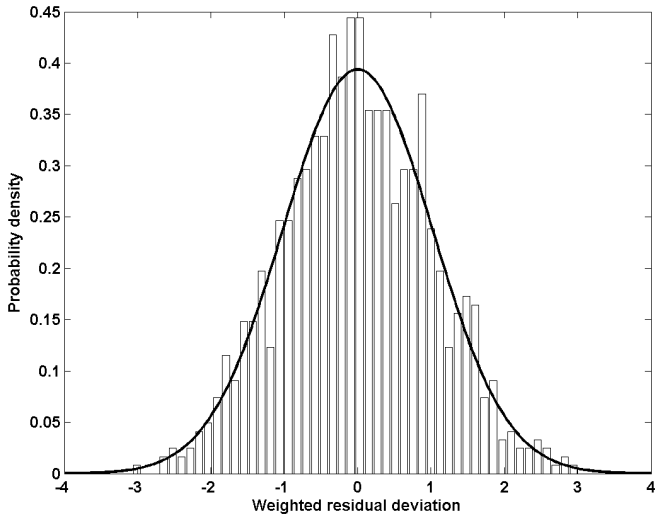

Figure 4: Scaled frequency distribution for the weighted residual deviations shown in Figure 3. The probability density function for $N(0, \tilde{s}^2)$ is also shown with $\tilde{s}$ the weighted root-mean-square residual deviation.

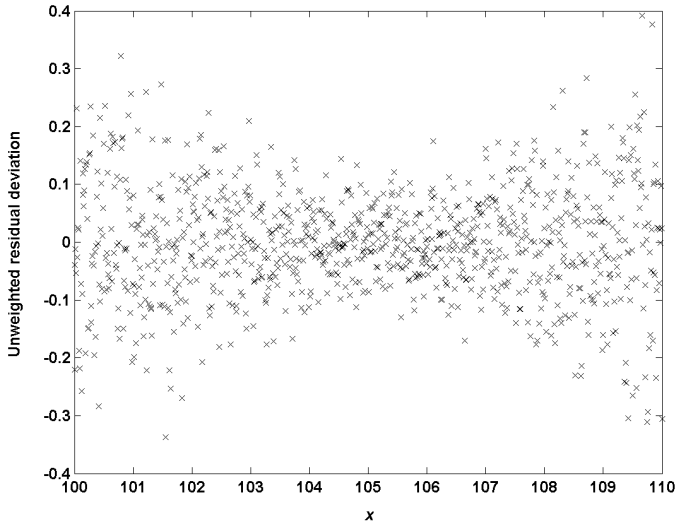

Figure 5: Residual deviations associated with an unweighted least-squares spline fit to the data shown in Figure 1.

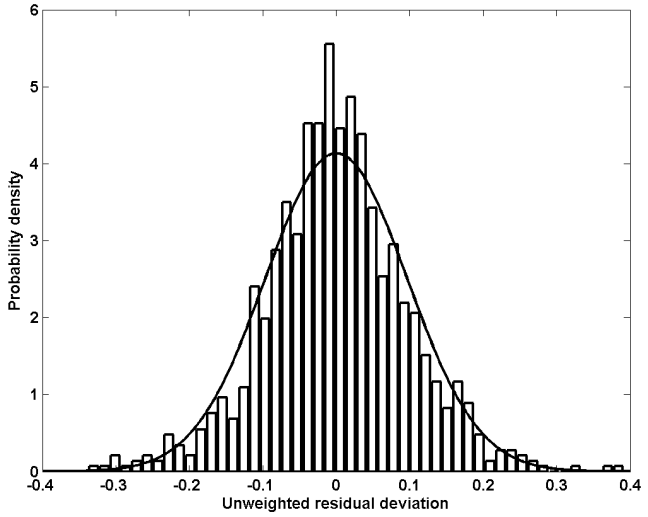

Figure 6: Scaled frequency distribution for the residual deviations shown in Figure 5. The probability density function for $N(0, s^2)$ is also shown with s the root-mean-square residual deviation.

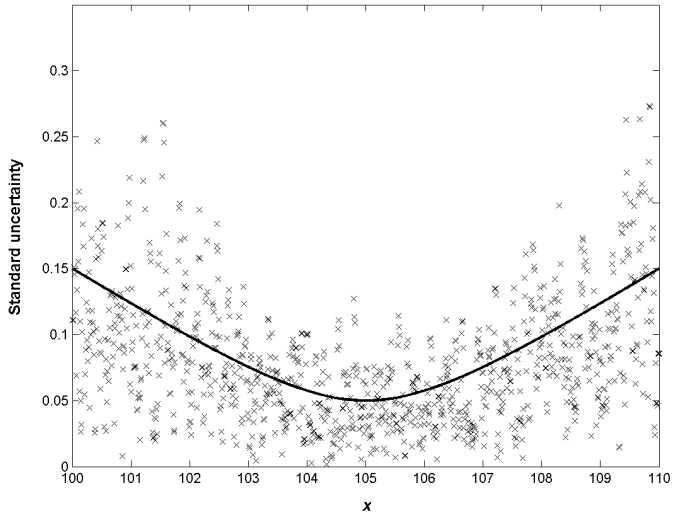

Figure 7: Standard uncertainties associated with the measurements y_i for the data shown in Figure 1 obtained using the local analysis procedure described in Section 4. The solid line shows the standard uncertainty profile for the data.

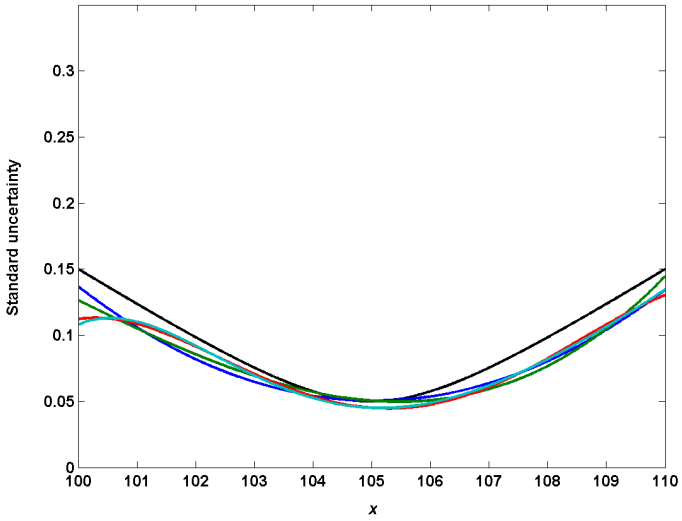

Figure 8: Estimated standard uncertainty profiles obtained by 'smoothing' the standard uncertainties shown in Figure 7. Smoothing is based on polynomials of degrees 2, 3, 4 and 5. The black solid line shows the (correct) standard uncertainty profile for the simulated data.

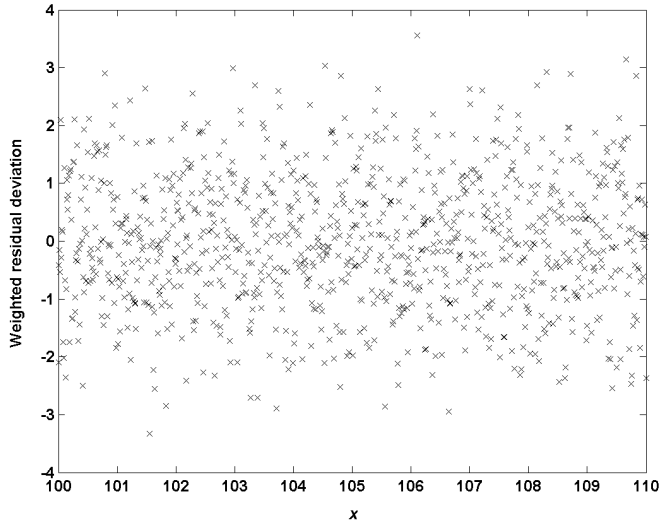

Figure 9: Weighted residual deviations associated with a weighted least-squares spline fit to the data shown in Figure 1, with estimated weights. The weights are based on standard uncertainties obtained by evaluating a degree five polynomial approximation to the standard uncertainty data shown in Figure 7.

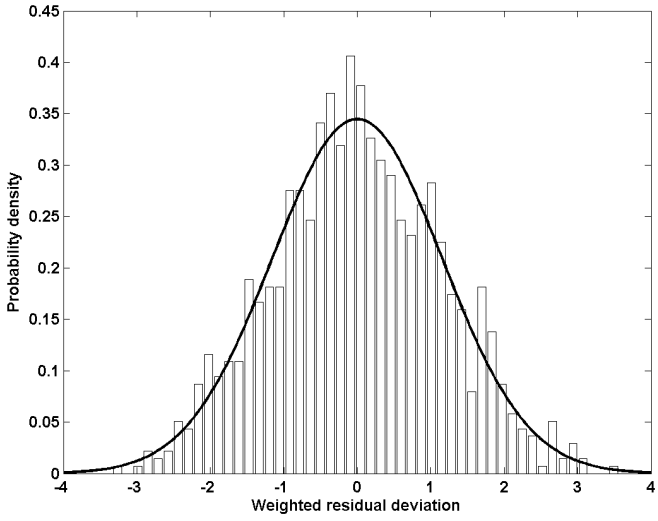

Figure 10: Scaled frequency distribution for the weighted residual deviations shown in Figure 9. The probability density function for $N(0, \tilde{s}^2)$ is also shown.

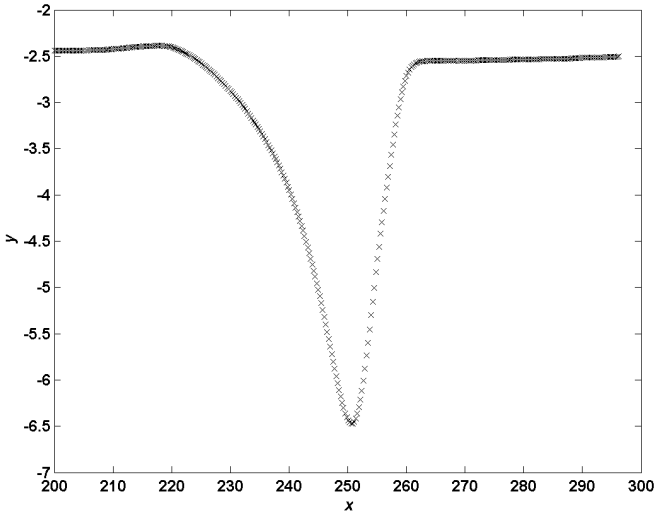

Figure 11: Part of a data set obtained from the application of the thermal analysis technique of differential scanning calorimetry.

No (quantitative) information concerning the standard uncertainties associated with the measurements is provided.[19] Figure 12 shows the residual deviations $e_i,\ i = 1, \ldots, m$, associated with an unweighted least-squares spline fit to the data. Figure 13 shows the residual deviations as a scaled frequency distribution together with the probability density function for $N(0, s^2)$, where s is the root-mean-square residual deviation given by (9). The fitted function is chosen to be a cubic (order $n = 4$) spline with $N = 80$ knots.[20] The first row of numerical values in Table 4 contains the value of s^2 and the results of applying the three statistical tests to test the null hypothesis that the residual deviations form a sample of obervations of a random variable with distribution $N(0, s^2)$.

It is clear from these results that it is inappropriate to assume that the deviations associated with the measurements are identically distributed. In particular, it would appear from Figure 12 that the uncertainty associated with measurements of heat flow made at temperatures less than (approximately) 210 °C and greater

[19] However, it is assumed that the standard uncertainties associated with measurements of temperature are small compared with those associated with measurements of heat flow. Consequently, for the purposes of this study, the measurements of temperature are regarded as exact.

[20] Although not presented here, a simple validation of this choice of fitting function was undertaken. The validation involved calculating the root-mean-square residual deviations associated with unweighted least-squares cubic spline fits to the data with an increasing number N of knots from $N = 0$ to $N = 100$. In each case the knots are chosen so that there is an approximately equal number of data points in each knot interval, except that in the first and last knot intervals there are approximately twice as many [10, 11]. A number of knots is chosen to correspond to the smallest number for which the root-mean-square residual deviation saturates to an (essentially) constant value.

than (approximately) 260 °C is smaller than that associated with measurements made between these temperatures.

	$\tilde{s}^2$	p–value		
		χ^2	K–S	S–W
(a) Unweighted	2.47×10^{-7}	0.00	0.00	0.00
(b) Weighted, estimated	3.74	0.00	0.03	0.76
(c) Weighted, smoothed	2.86	0.00	0.01	0.17

Table 4: Results of the analysis of the data shown in Figure 11. The residual deviations associated with least-squares spline fits to the data are tested for normality. The fits are obtained on the basis of weights that are assigned (a) to be unity (unweighted), (b) on the basis of a local analysis of the data (weighted, estimated), and (c) by smoothing the results from a local analysis of the data (weighted, smoothed).

Figure 14 shows the standard uncertainties $s_k,\ k = 0, \ldots, m - \ell + 2$, associated with the data obtained using the local analysis procedure described in Section 4.[21] The overall trend in the standard uncertainty values shown in this figure is consistent with the observations inferred (above) from Figures 12 and 13 about the inhomogeneity of the measurements. However, some additional information is apparent, viz., that the standard uncertainty associated with measurements of heat flow would appear generally to increase for temperatures below (approximately) 255 °C and, thereafter, generally to decrease before becoming essentially constant for temperatures above 265 °C.

Also shown in Figure 14 is a standard uncertainty profile for the data obtained by 'smoothing' the standard uncertainty values obtained from the local analysis. A function is chosen for the smoothing, viz., a piecewise-linear function,[22] to represent the 'underlying trends' apparent in the standard uncertainty data. This is in contrast to the approach taken for the simulated data considered in Section 5 for which a 'smooth' (polynomial) function was chosen to represent the standard uncertainty profile.

Figures 15 and 16 show the weighted residual deviations $\tilde{e}_i,\ i = 1, \ldots, m$, associated with a weighted least-squares spline fit using estimated weights. The weights are calculated in terms of the standard uncertainties obtained using the local analysis procedure *without* applying any smoothing to the standard uncertainty data. Figures 17 and 18 show the results when the weights are calculated in terms of standard uncertainties obtained by evaluating the smoothed uncertainty profile shown in Figure 14. In Figures 16 and 18 the probability density function for $N(0, \tilde{s}^2)$ is also shown, where $\tilde{s}^2$ is the weighted root-mean-square residual

[21] Based on quadratic polynomial fits to subsets of $\ell = 5$ points (Section 4).

[22] In fact, a linear (order $n = 2$) spline with knots, chosen by inspection, at 210 °C, 255 °C and 265 °C. In Figure 14 the values of the gradient of the first and second straight-line pieces of the piecewise-linear function are approximately equal and, consequently, appear as a single piece.

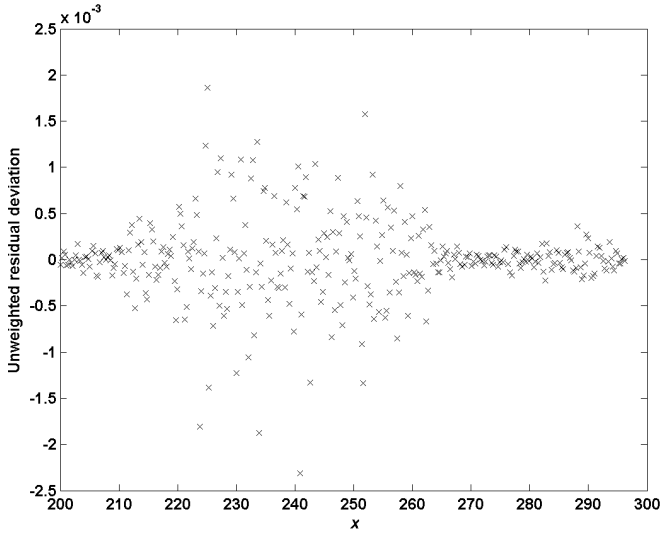

Figure 12: Residual deviations associated with an unweighted least-squares spline fit to the data shown in Figure 11.

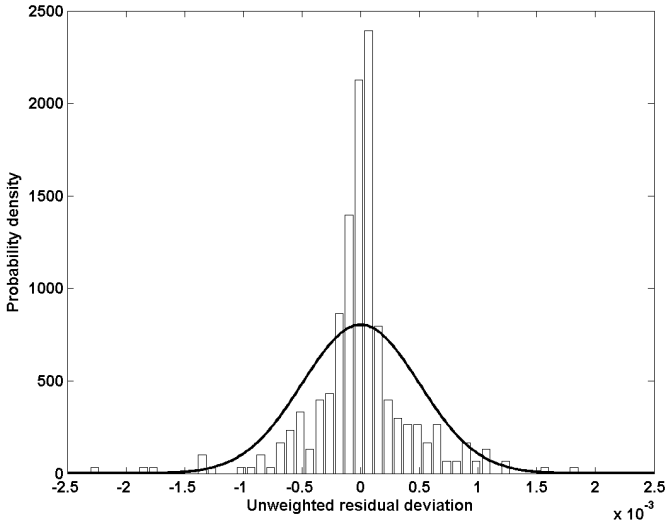

Figure 13: Scaled frequency distribution for the residual deviations shown in Figure 12. The probability density function for $N(0, s^2)$ is also shown, where s is the root-mean-square residual deviation.

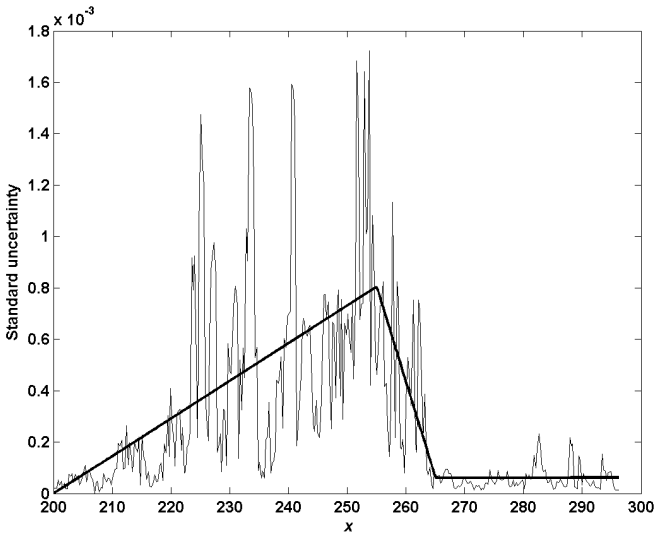

Figure 14: Standard uncertainties associated with the measurements y_i for the data shown in Figure 11 obtained using the local analysis procedure described in Section 4. A 'smoothed' standard uncertainty is also shown, which takes the form a piecewise linear function composed of four straight-line pieces.

deviation given by (8). The second and third rows of numerical values in Table 4 contain for the two cases (of, respectively, 'unsmoothed' and 'smoothed' standard uncertainty profiles) the values of $\tilde{s}^2$ and the results of applying statistical tests to test the null hypothesis that the weighted residual deviations form a sample of observations of a random variable with distribution $N(0, \tilde{s}^2)$.

In both cases, much of the inhomogeneity apparent in the residual deviations (of Figure 12) has been removed. The results of the statistical tests are somewhat inconclusive, but nevertheless suggest that the statistical model for the measurements inferred from the local analysis is an improvement over one based on the assumption of equal weighting of the data. We would expect to have more confidence in predictions made on the basis of weighted least-squares spline fits to the data using weights obtained from such an analysis.

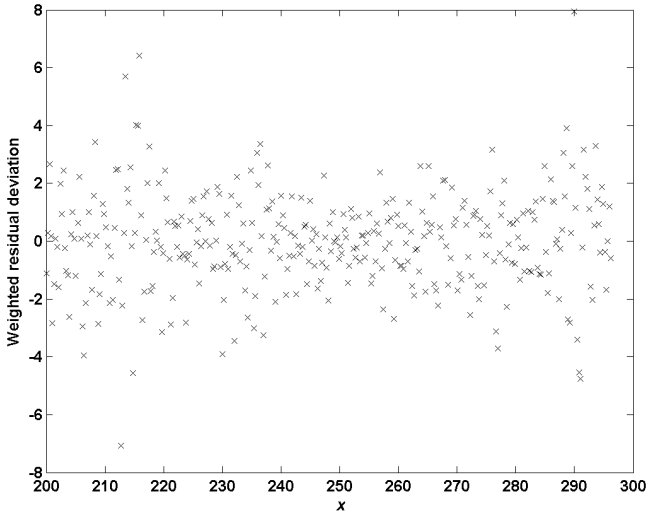

Figure 15: Weighted residual deviations associated with a weighted least-squares spline fit to the data shown in Figure 11, with estimated weights. The weights are based on standard uncertainties obtained using the local analysis procedure described in Section 4. No smoothing is applied to the standard uncertainties provided by that procedure.

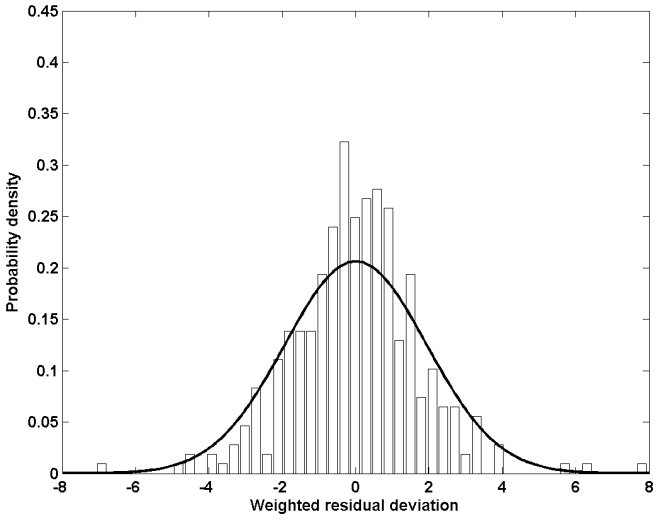

Figure 16: Scaled frequency distribution for the weighted residual deviations shown in Figure 15. The probability density function for $N(0,1)$ is also shown.

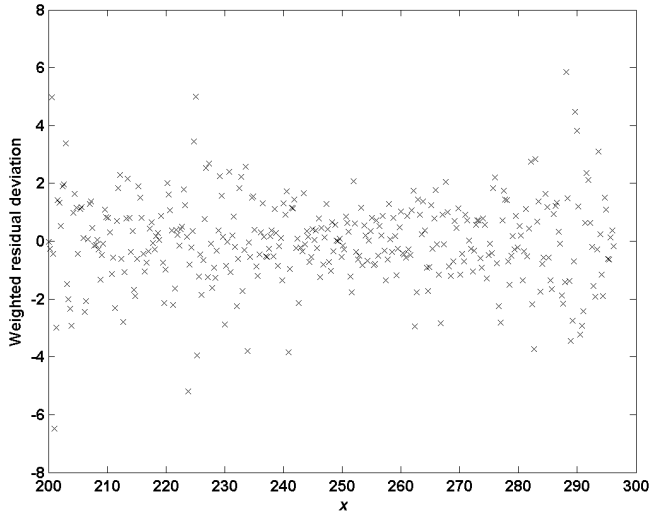

Figure 17: Weighted residual deviations associated with a weighted least-squares spline fit to the data shown in Figure 11, with estimated weights. The weights are based on standard uncertainties obtained by smoothing the standard uncertainties provided by the local analysis procedure described in Section 4.

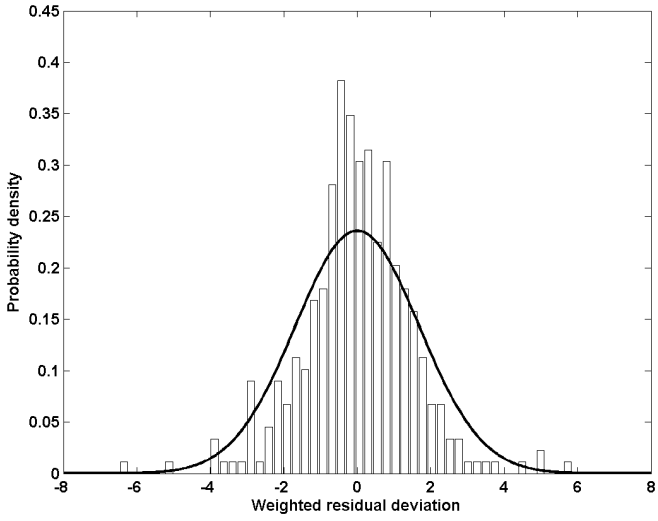

Figure 18: Scaled frequency distribution for the weighted residual deviations shown in Figure 17. The probability density function for $N(0,1)$ is also shown.

7 Summary and recommendations

This report has considered least-squares data approximation or regression for modelling experimental data as can arise, for example, in the calibration of a measurement system or instrument. Often, and certainly when there is no information to the contrary, the regression is undertaken under the assumption that the measurement deviations ('errors') associated with the data are samples of independent and identically distributed random variables described by a normal distribution. In many circumstances such as assumption will be valid, and predictions made on the basis of the fitted model can be regarded as reliable.

Consideration has been given to the application of standard statistical tests, such as the χ^2 goodness-of-fit, Kolmogorov-Smirnov and Shapiro-Wilks' tests, to validate assumptions made about the statistical model for the experimental data, viz., about the distribution for the measurement deviations. Alone, these tests (at least for the examples considered here) can be misleading and (together) inconclusive. It is therefore strongly recommended that the tests are not used in isolation, but in combination with other indicators, particularly plots of the residual deviations associated with the fitted model.

In cases where there is doubt that the statistical model applies, an approach is presented to modify the model. The modification is in terms of relaxing the assumption that the measurement deviations are samples of random variables that are identically distributed. The approach involves applying a local analysis of the data to evaluate the standard uncertainties associated with the measurements, and representing the derived uncertainties by a 'smooth' standard uncertainty profile.

Applications to simulated and real data have been given. The focus has been on using empirical models, in particular polynomial spline curves, to represent the measured data. However, the ideas would apply also for other model functions. Indeed, the local analysis procedure has been used in the context of modelling spectral irradiance data for which a hybrid model is used that takes the form of a product of a Planck function and a polynomial [13].

Particular recommendations are as follows:

- It is difficult to draw conclusions about the application of the statistical tests considered within the context of this work. The χ^2 goodness-of-fit test behaved 'as expected' for simulated data (although it gave a large range of p–values) but not for real (thermophysical) data. It is possible that the apparently good performance of the χ^2 goodness-of-fit test at identifying non-normality when 'incorrect' weights are used (Table 2) is a consequence of the test being sensitive to small deviations from non-normality, and this makes it a poor test in practical situations (Table 4). On the other hand, the Kolgomorov-Smirnov and Shapiro-Wilks' tests appeared to be more 'dis-

cerning' of the statistical model for real data but less so for simulated data.

- Although the data sets analysed here were (reasonably) large (hundreds of data points), the interpretation of the results from the tests was not straightforward. It is likely to be (even) less straightforward for smaller sets.

- Visual aids, including residuals plots, frequency distributions and normal probability paper, are an important and useful tool for making decisions about the statistical model, and should always be used.

- To derive a standard uncertainty profile using the local analysis approach described it is desirable to have a large quantity of data.

- Further work is required regarding the application of statistical tests (of normality) to the residual deviations associated with a fitted model to experimental data in the context of validating the statistical model assigned to the data. In particular, it is recommended that an investigation in terms of (a large number of) simulated data sets is undertaken.

8 Acknowledgements

The work described here was supported by the National Measurement System Directorate of the UK Department of Trade and Industry as part of its NMS Software Support for Metrology programme.

Geoff Morgan (formerly of the Numerical Algorithms Group Ltd.) contributed material for the section on tests for normality, and Martyn Byng (Numerical Algorithms Group Ltd.) commented on a draft of this report. Our colleagues Graham Sims and David Mulligan (NPL Materials Centre) provided the thermophysical data.

References

[1] R. M. Barker, M. G. Cox, A. B. Forbes, and P. M. Harris. Best Practice Guide No. 4. Discrete modelling and experimental data analysis. Technical report, National Physical Laboratory, Teddington, UK, 2004. Available for download from the SS*f*M website at www.npl.co.uk/ssfm/download/bpg.html.

[2] C. Chatfield. *Statistics for technology*. Chapman and Hall, third edition, 1983.

[3] W. J. Conover. *Practical Nonparametric Statistics*. New York, Wiley, 1980.

[4] M. G. Cox. The numerical evaluation of B-splines. *J. Inst. Math. Appl.*, 10:134–149, 1972.

[5] M. G. Cox. Practical spline approximation. In P. R. Turner, editor, *Notes in Mathematics 965: Topics in Numerical Analysis*, pages 79–112, Berlin, 1982. Springer-Verlag.

[6] M. G. Cox, A. B. Forbes, J. L. Flowers, and P. M. Harris. Least squares adjustment in the presence of discrepant data. To be published.

[7] M. G. Cox, A. B. Forbes, P. M. Harris, and I. M. Smith. The classification and solution of regression problems for calibration. Technical Report CMSC 24/03, National Physical Laboratory, Teddington, UK, 2003.

[8] M. G. Cox and P. M. Harris. Best Practice Guide No. 6. Uncertainty evaluation. Technical report, National Physical Laboratory, Teddington, UK, 2004. Available for download from the SS*f*M website at www.npl.co.uk/ssfm/download/bpg.html.

[9] M. G. Cox and P. M. Harris. Statistical error modelling. Technical Report CMSC 45/04, National Physical Laboratory, Teddington, UK, 2004.

[10] M. G. Cox, P. M. Harris, and P. D. Kenward. Fixed- and free-knot univariate least-squares data approximation by polynomial splines. Technical Report CMSC 13/02, National Physical Laboratory, Teddington, UK, 2002.

[11] M. G. Cox, P. M. Harris, and P. D. Kenward. Fixed- and free-knot univariate least-squares data approximation by polynomial splines. In I. J. Anderson J. Levesley and J. C. Mason, editors, *Algorithms for Approximation IV*, pages 330–345, Huddersfield, UK, 2002. University of Huddersfield.

[12] M. G. Cox, P. M. Harris, P. D. Kenward, and I. M. Smith. Extracting features from experimental data. Technical Report CMSC 22/03, National Physical Laboratory, Teddington, UK, 2003.

[13] M. G. Cox, P. M. Harris, P. D. Kenward, and Emma Woolliams. Spectral characteristic modelling. Technical Report CMSC 27/03, National Physical Laboratory, Teddington, UK, 2003.

[14] R. D'Agostino, A. Belanger, and R. D'Agostino Jr. A suggestion for powerful and informative tests of normality. *The American Statistician*, 44(4):316–321, 1990.

[15] R. B. D'Agostino and M. A. Stephens, editors. *Goodness-of-fit Techniques*. Dekker, New York, 1986.

[16] C. de Boor. On calculating with B-splines. *J. Approx. Theory*, 6:50–62, 1972.

[17] L. Fox and I. B. Parker. *Chebyshev polynomials in numerical analysis*. Oxford University press, 1968.

[18] G. H. Golub and C. F. van Loan. *Matrix Computations*. John Hopkins University Press, Baltimore, 1996. third edition.

[19] ISO. ISO 11357. Plastics – differential scanning calorimetry (DSC) – Part 1: General principles, 1997.

[20] A. N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933.

[21] H. Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62:399–402, 1967.

[22] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, London, 1979.

[23] Numerical Algorithms Group Limited. *The NAG Fortran Library (Mark 20)*. Copyright ©2001.

[24] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in FORTRAN: The Art of Scientific Computing*. Cambridge University Press, second edition, 1992.

[25] J. P. Royston. Algorithm AS181: The W test for normality. *Appl. Statist.*, 31:176–180, 1982.

[26] J. P. Royston. An extension of Shapiro and Wilks W test for normality to large samples. *Appl. Statist.*, 31:115–124, 1982.

[27] N. Smirnov. Estimate of deviation between empirical distribution functions in two independent samples. *Bull. Moscow Univ.*, 2:3–16, 1933.

[28] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Statist.*, 19:279–281, 1948.