Mathematics for modern precision engineering

BY PAUL J. SCOTT1,* AND ALISTAIR B. FORBES2

1 Taylor Hobson Ltd, New Star Road, Leicester LE4 9JQ, UK
2 National Physical Laboratory, Hampton Road, Teddington TW11 0LW, UK

The aim of precision engineering is the accurate control of geometry. Precision engineering is usually characterized by the ratio of the geometrical tolerances to the size of the object. This ratio is decreasing with time as precision engineering becomes more precise. In the 1980s, this ratio was of the order of 1 part in $10^4$ [1]. Today, it is typically of the order of 1 part in $10^6$ and, for some key twenty-first century engineering projects that have been proposed, it is approaching 1 part in $10^8$. For example, in the 42 m European Extremely Large Telescope, each of the 984 hexagonal segments is of the order of 1–2 m diameter with a specified form tolerance of better than 25 nm [2].

The mathematical sciences have a long and illustrious history with geometry for precision engineering. In ancient Egypt, geometrical problems appear in both the Rhind Mathematical Papyrus (RMP) [3,4] and the Moscow Mathematical Papyrus (MMP) [5]. These examples demonstrate that the ancient Egyptians knew how to compute various geometric attributes and put this knowledge to practical utility, particularly in re-establishing field systems after the Nile flood, building pyramids, etc.

*Author for correspondence (p.j.scott@hud.ac.uk).

One contribution of 16 to a Discussion Meeting Issue ‘Ultra-precision engineering: from physics to manufacturing’.
In particular, problems 56–59 of the RMP discuss various ‘pyramid problems’ that use the *seqed* (or *seked*), the Egyptian concept for slope. The *seqed* is defined as the lateral displacement in palms for a drop of 1 cubit (seven palms). For example, in problem 56, it is required to find the *seqed* of a pyramid with a base of 360 cubits and a height of 250 cubits (figure 1). Further, in problem 14 of the MMP, the volume of a truncated pyramid (frustum) is computed. The *seqed* is an abstract mathematical concept used in the construction of the pyramids—the precision-engineered structures of the day.

2. Mathematical terms and definitions

The *seqed* is an example of a mathematical term and definition. The value of mathematical terms and definitions is to provide first a common term and then a common understanding of the concept behind the term through the definition.
As the mathematical papyri demonstrate, once a term has been defined it can be used to provide standardized calculation procedures and examples using the mathematical concepts.

The value of mathematical terms and definitions is particularly true for geometrical concepts. The great ancient work on geometry (and the ancient Greek version of elementary number theory) is Euclid’s *Elements* [6], which consists of 13 known books and includes a collection of definitions, axioms, constructions, theorems and proofs. Euclid put together the *Elements* by collecting together geometrical results developed by others and supplementing them by some original work. The axiomatic/deductive treatment of geometry has proved instrumental in the development of logic and modern science and was not surpassed until the nineteenth century.

Mathematical terms and definitions, useful to precision engineering, also have a long history, from Archimedes’ work on screws and levers [7], Descartes’ coordinate system [8], Newton’s laws of motion [9] to the current mathematization of international standards on geometrical product specification and verification. They provide an idealized conceptual framework in which to define, realize and provide calculation procedures for mathematical entities that are idealizations of real-world objects such as precision-engineered objects.

### 3. Measurement and scales

Measurement is fundamental to all science; without measurement, there can be no science. Precision-engineered objects require verification that they are manufactured to specification, and measurement procedures are the main constituent of the verification process. Although there are mathematical theories of measurement, they have not been used in precision engineering until very recently [10]. The aspect of measurement discussed in this section is the concept of a scale. The development of evermore accurate scales has been important for precision engineering from early astronomical instrumentation such as astrolabes to modern astronomical devices, through surveying instrumentation to scales on coordinate-measuring instruments, precision machine tools, etc.

A scale is a totally ordered numerical structure onto which physical quantities are mapped, with the mapping preserving the structure of the original physical quantity. The two scales of interest here are the length and angle scales, both of which have been realized using mathematical concepts. Although other scales are useful to precision engineering, e.g. hardness, temperature and other environmental scales, they are realized using concepts from physics rather than from mathematics.

Both length and angle scales are linear scales. Length requires a unit length to be defined. Early unit length standards were realized as end bars, in which the fundamental unit was the distance between the two ends of the bar. However, to define a distance using an end bar with any accuracy, it is required that the two ends be flat and parallel to each other, properties that are challenging to manufacture and difficult to maintain with use. Later, end bars were replaced by line standards in which the fundamental unit is the distance between two (parallel) lines on a standard bar, overcoming the problems associated with end bars. The SI unit of length is the metre, and between 1889 and 1960 was defined
by a platinum–iridium line standard kept at the Bureau International des Poids et Mesures in Paris, France. Angle is self-specifying with only the number of angular units in a full circle being required to be defined; for radians this is $2\pi$.

(a) Realization of the metre

In 1791, the French Academy of Sciences selected a definition of one ten-millionth of the meridian distance between the North Pole and the Equator through Paris as its definition of a unit length, which it called the metre. Subsequently, in 1792, Jean Delambre and Pierre Méchain were commissioned to lead an expedition to measure the distance between the belfry at Dunkirk, France, and Montjuïc Castle, Barcelona, Spain, to estimate the length of the meridian arc through Dunkirk. By using triangulation and accurate base lines of between 1 and 10 km, they constructed a series of imaginary triangles between Dunkirk and Barcelona to determine the required meridian distance, completing their work in 1799 [11].

However, in 1793, France adopted as its official unit of length a metre based on provisional results from the expedition. This was later determined to be short by approximately one-fifth of a millimetre because of miscalculation of the flattening of the Earth. Thus, a precision of 1 : 5000 was achieved.

The metre was realized in a physical form until 1960, when it was redefined to be $1650763.73$ wavelengths of the orange-red emission line in the electromagnetic spectrum of the krypton-86 atom in a vacuum. The current definition of the metre is in terms of the second and the speed of light [12],

The metre is the length of the path travelled by light in a vacuum during a time interval of $1/299,792,458$ of a second.

Thus, the metre, the fundamental unit of length, is now defined from concepts of physics rather than being based on the results from a mathematical survey of a geographical artefact.

(b) Dividing the fundamental unit into a scale

Having realized the fundamental unit, the next stage is to divide the fundamental unit into a scale for increased resolution. Mathematically, division is interpolation, which is the computation of new data points from known discrete points. For the length scales, this consists of dividing the unit length, defined by a line scale, into equal lengths and for angular scales dividing the circle into equal angles.

Euclid’s Elements provides various geometrical constructions to divide lengths into equal parts and constructions for particular angles and angular bisection. The method of transversals uses one geometric construction to allow finer division of a scale. Linear transversals (figure 2) consist of a series of parallel lines orthogonal to the scale to form a grid with the extended scale lines; diagonal lines (the transversals) are constructed from the uppermost corner of a column in the grid to the opposite lowest corner. Where a transversal crosses one of the parallel lines constitutes a finer graduation of the scale. Circular transversals for angular scales are also possible, but construction of the grid is considerably more complex. Linear transversals have been used to approximate finer angular scales; the finer the original angular scale, the better the approximation for the linear transversal.
The constructions require precision because any errors will remain in the final division. Iterative methods, based on finding approximate divisions (e.g. by geometrical constructions) and adjustment through comparison of lengths/angles, are straightforward to implement in practice and allow any errors to be eliminated as the iteration proceeds [13].

The Vernier scale has largely replaced the method of transversals. Here, a second sliding scale is used with slightly smaller (direct Vernier) or larger (retrograde Vernier) graduations, e.g. nine or 11 graduations on the second scale compared with 10 graduations on the main scale. From matching graduation marks on the two scales, it is possible to achieve increased resolution. The mathematics for the Vernier scale is very similar to the linear transversal with the distance between marks on a direct Vernier scale equal to the distance along the scale from the lowermost corner in a column to the intersection of the transversal in the next column in the uppermost but one parallel line.

An alternative approach for length division is through a screw. Mathematically, a screw is a helix, which links length scales to angular scales via a linear relationship. Given the pitch of the screw, an angular scale used for the rotation of the screw can be used to read linear distances. The accuracy of this approach is dependent on the precision of the screw (also the angular scale and the determination of the pitch length). Techniques to improve screw precision include using a nut with a long engagement to average out elastically (statistical averaging) local errors in a screw in the manufacture of a more precise screw. The production of scales was originally undertaken by hand but automatic ‘dividing engines’ incorporating the mathematical techniques mentioned earlier soon became the norm. A good historical review of circular and linear dividing engines is given in Evans [14].

4. Controlling geometrical variability

The aim of precision engineering is the accurate control of geometry. Euclid’s Elements deals with ideal (or nominal) geometrical entities only (i.e. perfect circles, perfect straight lines, perfect right angles, etc.). Real-world objects are typically designed to take the form of these idealized geometrical objects, but manufacturing technologies can only produce approximate idealized geometries subject to variation.
Precision-engineered artefacts can be associated with four conceptual worlds: the world of design, where the artefact is specified as imagined by the designer; the world of manufacture, where the artefact is created; the world of verification, where the artefact is measured for conformance to specification; and the world of product life, where the artefact is used.\textsuperscript{1} The control of geometrical variability must be taken into account in all four worlds, particularly when specifying a product’s geometry and managing manufacturing processes. A mechanism to distinguish between allowable and unacceptable geometries is required.

Early mechanisms to control geometrical variability include:

- The use of a master component with which the manufactured part could be compared and then adjusted, as required.
- (Early 1800s) The use of hard gauging that enabled parts to be manufactured accurately enough to be assembled interchangeably. Hard gauges are physical devices with specified features of size/length with which manufactured components can be compared in order to check the size/length of the manufactured component. (The basic mathematical theory behind hard gauging was not fully developed until Taylor’s \textsuperscript{[16]} principle in 1905.)
- (ca 1850) Dimensional drawings: gauging technology allows dimensional drawings, rather than representational drawings or models, to be used as design specifications. Although not directly concerned with geometrical variability, dimensional drawings are a mathematical nominal model of the desired geometry, a precursor to dimensioning tolerancing.
- (Early 1900s) Dimensional tolerancing: a system for specifying explicit limits on dimensional requirements in dimensional drawings in order to control workpiece variability. In 1905, William Taylor formulated the Taylor principle in Patent no. 6900 for \textit{Improvements in gauges for screws} \textsuperscript{[16]}. The most important principle associated with the design of hard gauges is as follows: \textit{a go-gauge should check the full feature (maximum material limit) while a no-go gauge should check each individual feature or dimension.}
- (1940s) Geometrical tolerancing: a system of specifying tolerance zones in engineering drawings in order to control workpiece variability. Geometrical tolerancing was developed to improve the weakness of previous tolerance systems to handle imperfect form and ambiguous references. Geometrical tolerancing is still at the core of today’s tolerancing systems: \textit{geometrical product specification and verification} is defined by the International Organization for Standardization (ISO).

There are three basic approaches to verifying that a manufactured component conforms to its specification \textsuperscript{[17]}: hard gauging; hand gauging using micrometers or similar devices; and digital metrology that uses sampled points gathered by a three-dimensional coordinate-measuring system.

In 1988, Richard Walker of Westinghouse Corporation issued a US GIDEP alert \textsuperscript{[18]} on the method divergence between digital metrology and traditional metrology approaches based on hard and hand gauging. This alert has led ultimately to the process of redefining geometrical tolerancing concepts in terms

\textsuperscript{1}The first three worlds are taken from the International Organization for Standardization \textsuperscript{[15]}.
of well-defined mathematics, i.e. to produce mathematics-based national and international standards. This process demonstrated that the mathematics of geometrical tolerancing was not complete, with many concepts requiring precise mathematical definitions to avoid definition uncertainty [19]. For example, the recently published ISO standard on linear sizes [20] has expanded one ambiguous definition of size to a precise definition of what constitutes a ‘feature’ of size, together with an unambiguous toolbox of techniques to define different types of size, depending on the functional requirements for that geometry. For example, a UK 50 pence coin has constant two-point diameter and thus has a well-defined two-point size, which ensures that it can roll down a slot, even though it is far from being round (figure 3).

The challenge with the ‘mathematization’ of geometrical product standards is to merge the traditional with the digital using well-defined mathematics that allows utility for function, design, metrology and product lifetime management. This is an exciting time for geometrical product standardization with many philosophical discussions still to come over exactly what a particular concept really means and how it can be represented mathematically. The outputs of these discussions will have consequences for precision engineering.

5. The digital paradigm

Traditional metrology is essentially analogue, using hard and hand gauges to verify that an artefact is to specification. The mathematics is encapsulated within the mechanical set-up (jigs and fixtures) and the relationships of the gauges (figure 4). Post-measurement calculation is relatively simple and, traditionally, carried out by hand. Traditional metrology was restricted mainly to simple geometrical shapes (planes, spheres, cylinders, cones, etc.) that described the idealized geometry of artefacts of the day [21].

Digital metrology, through sampling in a coordinate system, provides a digital representation of an artefact, enabling computational techniques to be applied (figure 5). The mathematics is encapsulated in the post-measurement calculations through software implementing algorithms that are becoming increasingly complex. When digital metrology first appeared in the 1960s, standards and methodologies still reflected the language and concepts of traditional metrology. But digital metrology extends much beyond the traditional metrology domain.

In many technologically advanced industries, such as optics, the shapes of the surfaces of manufactured objects are becoming much more complicated than simple geometrical shapes. Unlike conventional surfaces, these advanced surfaces
have no axes of rotation or other symmetries and in the future could have almost any designed shape. These latter geometrical surface shapes are called freeform surfaces [22]. Over the past decade, the design and manufacture of advanced optics has begun to include freeform optical elements and micro-structured optical surfaces, which are critical components in high added-value products such as mobile phone cameras, laser printers, flat bed scanners, displays,
telecommunications hardware and photonics hardware, e.g. broadband optical fibre connectors. Freeform surfaces allow these optical components to contain less ‘glass’, making them lighter, cheaper to manufacture and, in most cases, perform better [23,24].

The rapidly increasing use of ultra-precision freeform surfaces is, however, not limited to the optics field: bioimplants such as knee prostheses use freeform surfaces as the bearing components. One of the primary advances in the biomedical field is freeform hard-bearing couples, requiring micrometre-scale form control with nanometric surface topography control [25,26].

In the four worlds of a precision artefact, because the world of design comes first, specification sets the pattern that the other three worlds follow. The specification for a precision artefact consists of many independent call-outs, each of which requires three elements to be defined: the surface features that constitute the measurands, the characteristics to define the desired geometrical property of the features, and the comparison rule to determine acceptable/unacceptable limits of the value of the characteristic.

Surface features are defined by a series of four basic operations carried out on a surface: sampling, moving from a continuous object to a discrete object, reconstruction, moving from a discrete object to a continuous object, decomposition of a surface into different features and composition, building a surface up from different features. For decomposition, there are many different subtypes, including: partitioning an artefact into different functional surfaces; filtering a surface into features of different sizes; association, i.e. determining the best-fit-specified nominal geometrical surface to the sampled points, etc. The main mathematical challenges are: to determine suitable subtypes of the four operations that have functional utility; to determine efficient sampling strategies for a geometrical surface to obtain sufficient points to carry out subsequent operations in order to verify that it conforms to its specification (a complication is that it is not possible to mathematically sample uniformly on most geometrical surfaces; see Berger [27]); to determine suitable ‘best-fit’ methodologies for different geometrical surface types; the specification of different geometrical surface types (this is discussed in detail in the case study in §5a), to determine filtration for all types of geometrical surfaces.

There are two main types of characteristic: intrinsic (property of a single feature) and relational (property of the relationship between two or more features). The main mathematical challenge is to determine stable and robust characteristics for functional utility. Here, a stable characteristic means that ‘small changes’ in the value of a characteristic imply ‘small changes’ in the surface (see Scott [10] for a full mathematical description) and a robust characteristic means that outlier events (e.g. spikes, scratches, steps, continuity breaks) do not lead to ‘large changes’ in the value of the characteristic.

Comparison rules are the part of the call-out that limits the geometrical variability. There are three ways to specify the permissible limits: by dimension, by zone and by statistics. The mathematical challenge is to incorporate uncertainty (see §8) into the comparison rule.

One very important current challenge is the specification of freeform and deterministic patterned (e.g. structured) geometries. Some initial work has been carried out [28,29] but it is still early days.
(a) Case study: specification of geometric surfaces

The specification of the ideal shape of a surface is typically defined in terms of:

— geometric elements: planes, spheres, cylinders, cones and tori,
— (more general) surfaces of revolution: aspherics, paraboloids, hyperboloids, and
— freeform parametric surfaces: parametric splines, non-uniform rational B-splines (NURBSs).

In general, the geometry can be defined as a mathematical surface \( \mathbf{u} \rightarrow \mathbf{f}(\mathbf{u}, \mathbf{b}) \), where \( \mathbf{u} \) are the patch or footpoint parameters and \( \mathbf{b} \) are parameters that define the position in some fixed frame of reference, size and shape of the surface. It is also possible to define a surface implicitly in the form \( \mathbf{f}(\mathbf{x}, \mathbf{b}) = 0 \), where again \( \mathbf{b} \) specify the position and shape of the surfaces. Even for the case of standard geometric elements such as spheres and cylinders, element parametrization is not straightforward. We let \( E \) be the space of geometric elements, e.g. cylinders. A parametrization \( \mathbf{b} \rightarrow \{ \mathbf{u} \rightarrow \mathbf{f}(\mathbf{u}, \mathbf{b}) \} \) is a locally one-to-one and onto mapping \( \mathbb{R}^n \rightarrow E \).

Parametrizations are not necessarily globally one-to-one. For example, the cylinder with axis normal \( \mathbf{n} \) is the same as that defined by \( -\mathbf{n} \). Parametrizations are not unique. For example, a cone is defined by six parameters. Two standard parametrizations of a cone whose axis is approximately aligned with the \( z \)-axis are (i) cone vertex (three parameters), direction of the axis (two), e.g. angles of rotation about the \( x \)- and \( y \)-axes, cone angle (one), i.e. the angle between the cone generator and its axis; and (ii) intersection of the cone axis with the plane \( z = 0 \) (two), direction of the axis (two), radii (two) at two distances \( h_1 \) and \( h_2 \) along the cone axis from the point of intersection with \( z = 0 \). These two parametrizations are not equivalent; the first parametrization breaks down when the cone angle approaches zero, while the second parametrization does not.

The condition that a parametrization is locally onto cannot, in general, be strengthened to being globally onto. The reason for this is that the topology of \( E \) need not be flat like \( \mathbb{R}^n \). For example, the space \( N \) of cylinder axis direction vectors \( \mathbf{n} \) is a sphere in \( \mathbb{R}^3 \) with \( \mathbf{n} \) identified with \( -\mathbf{n} \) and has a non-flat topology. For element spaces with non-flat topologies, any parametrization \( \mathbb{R}^n \rightarrow E \) has at least one singularity. This has implications for developers of element-fitting algorithms because the algorithms may need to change the parametrization as the optimization algorithm progresses. Any implementation that uses only one parametrization will (probably) break down for data representing an element at (or close to) a singularity for that particular parametrization.

The problem of parametrization becomes more difficult for freeform surfaces defined by parametric B-splines or the more general NURBSs [30]. A (cubic) spline curve is made by joining sections of (cubic) polynomials together. The locations of the joining points are called knots. Fitting spline curves to data is very efficient because the matrices involved are banded and this structure can be exploited. As with computations involving polynomials [31], it was only with the development of numerically stable algorithms [32,33] using the so-called B-spline basis functions \( N_j(x) \) that the use of cubic splines became practical. In a
parametric B-spline surface, each coordinate is represented as a tensor product spline on the same knot set:

\[
f(u, b) = \sum_{i=1}^{n} \sum_{j=1}^{n} N_i(u) N_j(v) p_{ij},
\]

where \( p_{ij} = (p_{ij}, q_{ij}, r_{ij})^T \) are control points, the coefficients of the basis functions, and \( b \) is the vector of control points. The control points \( p_{ij} \) form a discrete approximation to the surface and changing \( p_{ij} \) influences the shape of the surface local to \( p_{ij} \). NURBSs, the ratio of two tensor product splines, generalize parametric splines so that standard geometric elements can be represented also by NURBSs.

For parametric spline surfaces, there is a strong geometrical relationship between the control points and the surface itself. For example, applying a rigid body transformation to the control points results in the same transformation of the surface; scaling the control points scales the surface by the same amount. While the control points \( b \) (along with weights, if we are dealing with a NURBS surface) specify the surface, the vector \( b \) does not in general represent a parametrization as defined above because a change in the shape of the control points does not necessarily mean a change in the shape of the associated surface. A simple example is given by parametric spline curves of the form \( f(u, b) = (f(u, p), g(u, q)) \) on the same knot set. If \( p = q \), the parametric spline curve specifies a section of the line \( y = x \); changing the coefficients associated with the interior knots but keeping the relationship \( p = q \) does not change the line segment, only the speed with respect to \( u \) the point \( f(u, b) \) moves along the line segment. This behaviour suggests that we allow only those changes in the shape of the control points that correspond to changes in the shape of the surface.

\( b \) Case study: Gaussian processes and form errors

While parametric surfaces such as NURBS are very flexible tools for representing ideal geometries, an actual manufactured part will depart from the ideal geometry in a way that is difficult to predict. An estimate of how much the actual surface departs from ideal geometry can be determined by measuring the part and fitting the ideal geometry to the measured coordinates. The residual distances \( d(x_i, b) \) can be used to give a measure of the departure from the ideal geometry. However, it is clear that such a measure will depend on the points sampled, and selecting a different set of points will probably give a somewhat different measure. The difficulty is that evaluating the form error involves the complete surface, whereas the data represent only a discrete sampling of the surface.

Gaussian processes [34] give an approach to generating estimates of quantities defined over a continuum from discrete data. The traditional approach to solving this problem is to assume that the quantity is a member of a particular model space, e.g. polynomials, splines, NURBSs, and then use the measured data to select the member from the model space that optimally matches the observed data. However, the validity of the approach depends on defining the correct model space. If the actual behaviour of the system is not covered by the selected model space, then inferences based on the assumed model space may be flawed.
Gaussian process models are specified in terms of correlation behaviour rather than functional model spaces. For example, for form error, we may expect that form error at nearby locations will be similar and the closer the locations, the stronger the correlation (but not exactly the same owing to surface roughness effects). For example, this correlation structure could be modelled as

\[
\text{cov}(\delta_i, \delta_j) = k(x_i, x_j) = \sigma_F^2 \exp(-\alpha_F \Delta_{ij}^2 - \beta_F),
\]

where \( \delta_i \) and \( \delta_j \) are the form error at \( x_i \) and \( x_j \), respectively, measured normal to the nominal surface, \( \Delta_{ij} = \|x_i - x_j\| \), and \( \sigma_F, \alpha_F \) and \( \beta_F \) control the variances and length scales over which the correlation operates. Such a correlation structure can be used to predict the form error at any point from a discrete set of measurements. One approach, sometimes referred to as universal kriging, can be described in terms of linear models. Suppose

\[
\begin{bmatrix} \eta \\ \xi \end{bmatrix} | \alpha \sim N \left( \begin{bmatrix} C \\ D \end{bmatrix} \alpha, V \right), \quad V = \begin{bmatrix} V_{11} & V_{12}^T \\ V_{21} & V_{22} \end{bmatrix} = L L^T, \quad L = \begin{bmatrix} L_{11} \\ L_{21} \\ L_{22} \end{bmatrix},
\]

or, equivalently,

\[
\eta | \alpha = C \alpha + L_{11} e_1, \quad \xi | \alpha = D \alpha + L_{21} e_1 + L_{22} e_2, \quad e_1 \sim N(0, I), \quad e_2 \sim N(0, I).
\]

Suppose observations \( y \) of \( \eta \) are made. What can be said about \( \alpha \) and \( \xi \)? The best estimate \( \hat{\alpha} \) of \( \alpha \) is given by \( \hat{\alpha} = (C^T V_{11}^{-1} C)^{-1} C^T V_{11}^{-1} y \), with associated variance matrix \( V_\alpha = (C^T V_{11}^{-1} C)^{-1} \), the best estimate \( e_1 \) of \( e_1 \) is \( e_1 = L_{11}^{-1} (y - C \alpha) \), and the best estimate \( z \) of \( \xi \) is

\[
z = D \alpha + L_{21} e_1 = D \alpha + V_{21} V_{11}^{-1} (y - C \alpha) = V_{21} V_{11}^{-1} y + E \alpha, \quad E = D - V_{21} V_{11}^{-1} C,
\]

with associated variance matrix \( V_z = V_{22} - V_{21} V_{11}^{-1} V_{21}^T + E V_\alpha E^T \). On the basis of the information supplied by the observations \( y \), the state of knowledge distribution for \( \xi \) is now \( \xi | y \sim N(z, V_z) \). In practice, universal kriging is used to take measurements \( y \) at a set of discrete locations specified by observation matrix \( C \) to predict what the model values are at locations specified by observation matrix \( D \).

For example, suppose we are interested in assessing the form error of a circle, modelled as

\[
\text{cov}(\delta_i, \delta_j) = k(\theta_i, \theta_j) = \sigma_M^2 \exp(-\alpha_M \Delta_{ij}^2 - \beta_M), \quad \Delta_{ij} = |\theta_i - \theta_j|.
\]

We assume that a measurement system is used to sample points at angles \( \theta_i \) around the circle, with random noise with standard deviation \( \sigma_M \) and that we wish to predict the form error at angles \( \phi_j \). Thus

\[
V_{11}(i, i) = \sigma_M^2 + k(\theta_i, \theta_i), \quad V_{11}(i, j) = k(\theta_i, \theta_j), \quad i \neq j,
\]

and

\[
V_{22}(i, j) = k(\phi, \phi_j), \quad V_{12}(i, j) = k(\theta, \phi_j).
\]

Figure 6 shows the best estimate of the form deviation \( z \) evaluated at 100 points (solid curve) derived from seven measurements \( y \). The figure also shows two
Figure 6. Estimated form deviation at 100 points (solid) curve derived from seven measurements (circles), along with two potential realizations (dots and crosses) of form error. (Online version in colour.)

samples $z_1, z_2 \in N(z, V_z)$ (dots and crosses) representing potential realizations of the form error. The calculations take into account the measurement uncertainties associated with $y$.

6. Mathematical models

All mathematics used within precision engineering, from the Egyptian seqed to modern complex computer algorithms, has an underpinning mathematical model. It is now not possible to develop any complex device or precise artefact without first having to simulate or build a mathematical model of that entity in order to test out ideas of its effectiveness.

The choice of mathematical model and decisions of the type of model is critical to being able to answer the questions being asked of it. Should it be an analytical or numerical model, discrete or continuous model, stochastic or deterministic model, etc.? There are no easy answers to these questions. The choice of model depends on the mathematical tools available, the physics of the problem, etc. It is not possible to give justice to the full range of nuances of mathematical modelling within the space of this paper; so a case study from precision engineering will be used to illustrate some of the concepts behind mathematical modelling. A fuller discussion of these issues within mathematical models can be found in Gershenfeld [35].

One very important use of mathematical models is to optimize the design of machine tools through modelling vibration, thermal and other environmental effects. Finite-element analysis (FEA) is a very common technique used to model
these effects. An example output of an FEA is given in figure 7: figure 7a shows the temperature distribution, and figure 7b shows the distortions in a machining structure due to thermal distribution (example from Mian et al. [36]).

(a) Case study: Gaussian processes and machine errors

Gaussian processes can also be used to model measuring system errors; for example, the kinematic errors associated with a coordinate-measuring machine (CMM). Usually, these kinematic errors are modelled as polynomial or spline functions. However, models based on a correlation structure could also be of
value. A very simple correlation model has been used for the calibration of regular features such as hole plates, ball plates, ball bars and two-dimensional target arrays for vision systems. Owing to the regular symmetries associated with these artefacts, it is possible to translate or rotate the artefacts, so that the features (hole or ball centres) are located approximately at the same nominal points $x_l$ in the working volume.

This feature gives rise to observation equations of the form $x_i + e_l = y_{j,k} + \varepsilon_i$, where $x_i$ are the measurement coordinates close to $x_l$, $e_l$ is the fixed, systematic error at $x_l$, $y_{j,k}$ is the location of the $j$th feature $y_j$ in the $k$th position of the artefact and $\varepsilon_i$ represents random effects associated with $x_i$. Thus, $e_l$ presents the machine error that is taken to be constant, i.e. perfectly correlated, within a small neighbourhood of $x_l$. With appropriate measurement strategies [37] and additional scale-setting information, it is possible to calibrate both the artefact features $y_j$ and the machine errors $e_l$ simultaneously, by solving a large nonlinear least-squares problem. Because the observation equations involve at most one $y_j$ and one $e_l$, the associated Jacobian matrix is extremely sparse, and, for large artefacts with many features, sparse matrix techniques have to be used in order to make the computations practical [38]. Gaussian processes can be used to extend this approach to tasks in which there is approximate symmetry associated only with the artefact.

7. Inverse problems

In forward problems, input data are taken through a system to produce output data. In inverse problems, the output is known but with errors, and the problem is to compute either the input or the system, given the other [39].

Frequently, it is required to relate the parameters in a mathematical model to the observed–measured data. This is a classic inverse problem [40]. Inverse problems are ubiquitous within precision engineering and particularly in metrology, from the interpretation of measured results, through calibration and adjustment of instruments, fitting surfaces to measured data (see the case study in §7a), to correction of a probe’s geometry. Concepts such as resolution, uncertainty, Ockham’s razor, etc. are automatically induced as part of the detailed framework of inverse problems. Inverse problems are an important challenge for the mathematics for precision engineering (and science as a whole).

It is stated in Hadamard [41] that there are three issues in inverse problems: existence (no model exactly fits the observed data); uniqueness (there are potentially many exact solutions); and stability of the solution (inverse solutions can be extremely unstable). All three problems make interesting mathematical challenges for inverse problems in precision engineering. There is an extremely large amount of literature on inverse problems with linear models being particularly developed, including techniques to allow stable approximate solutions to ill-posed inverse problems (e.g. regularization algorithms such as spectral and Tikhonov regularization) [40]. It is with nonlinear inverse problems where the real challenges lie.

(a) Case study: fitting geometric surfaces to data

A key activity in precision engineering relates to assessing how close a part matches its design intent; nowadays, specified in terms of geometric elements,
NURBSs or similar. In traditional engineering, this assessment was achieved using hard gauges to check the various dimensions of a part. Since the introduction of the CMM in the 1960s, increasingly, part-checking is achieved by matching a discrete representation of the part through a set of coordinates $x_i = \{x_i : i \in I\}$, $I = \{1, \ldots, m\}$, gathered by a coordinate-measuring system, to a mathematical representation $u \rightarrow f(u, b)$ of the ideal surface. The fitting process usually involves calculating or estimating the distance from a point to a surface. Let $x$ be a datum point reasonably close to the surface $u \rightarrow f(u, b)$, and let $u^*$ be the solution to the ‘footpoint’ problem $\min_u (x - f(u, b))^T (x - f(u, b))$ so that $u^* = u^*(b)$ specifies the point on the surface closest to $x$. Let $n = n(b)$ be the normal to the surface at $f(u^*, b)$, likewise a function of $b$, and set

$$d(x, b) = (x - f(u^*, b))^T n.$$  

Then, $d(x, b)$ is the signed orthogonal distance of $x$ from $f(u, b)$ and, furthermore,

$$\frac{\partial d}{\partial b_j} = -\left( \frac{\partial f}{\partial b_j} \right)^T n,$$

where all terms on the left-hand side are evaluated at $u^*$. For standard geometric elements, the distance function $d(x, b)$ can be defined as an explicit function of the parameters but for freeform surfaces the optimal footpoint parameters $u^*$ have to be determined using numerical techniques [42–44].

In least-squares orthogonal distance regression (LSODR), the best-fit surface $u \rightarrow f(u, b)$ to a set of data $x_I = \{x_i : i \in I\}$ is that which minimizes the sum of squares of the orthogonal distances, i.e. solves $\min_b \sum_{i \in I} d^2(x_i, b)$. The optimization problem may be restricted to adjusting the position and scale of the surface but keeping the shape constant. The Gauss–Newton algorithm is usually employed to perform the optimization. Given an estimate of $b$, an updated estimate is given by

$$b := b + p, \quad p = -(J^T J)^{-1} J^T d, \quad d = (d_1, \ldots, d_m)^T, \quad J_{ij} = \frac{\partial d}{\partial b_j}(x_i, b),$$

involving the $m \times n$ Jacobian matrix $J$.

The LSODR problem can also be posed as

$$\min_{u, b} \sum_{i \in I} (x_i - f(u_i, b))^T (x_i - f(u_i, b))$$

explicitly involving the footpoint parameters $u_I = \{u_i : i \in I\}$. The approach can be made efficient by exploiting the block–angular structure of the Jacobian matrix,

$$J = \begin{bmatrix} J_1 & J_{0,1} \\ J_2 & J_{0,2} \\ \vdots & \vdots \\ J_m & J_{0,m} \end{bmatrix},$$  

where the diagonal blocks store the derivatives with respect to the point coordinates and the border blocks those with respect to the surface parameters.
8. Uncertainty

When reporting the result of a measurement, it is obligatory that some quantitative indication of the quality of the result is given so that those who use it can assess its reliability. Without such an indication, measurement results cannot be compared, either among themselves or with reference values given in a specification or standard. ISO/IEC Guide 98-3:2008 (GUM) [45] provides a set of procedures for characterizing the quality of a result of measurement called uncertainty.

Further, GUM distinguishes between the measurement result and the characteristics of the measurement equipment. Uncertainty is used only for characterizing the quality of measuring results. Maximum permissible errors are used for characterizing the quality of attributes of the measuring equipment.

Uncertainty of measurement is defined as (see ISO [19]):

Parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand.

GUM characterizes this dispersion in terms of standard deviations. Thus, all error and uncertainty components (figure 8) are estimated as standard deviations as per their influence on the final result of the measuring process. The standard uncertainty of the result of a measurement when there are a number of uncertainty components is calculated as a combined standard uncertainty and is equal to the positive square root of the sum of terms, involving the variances (squares of the standard deviations) and covariances (measures of association between pairs of variables) of the individual uncertainty contributors and the sensitivity coefficients of the output with respect to the inputs.

When reporting uncertainty, it is usual to state the expanded uncertainty, which is the combined standard uncertainty multiplied by a numerical factor, called the coverage factor, which is typically in the range of 2–3 (the default value is 2 and corresponds, for a normally distributed output, to a 95% level of confidence). The expanded uncertainty may be expected to encompass a large fraction of the distribution of values that could reasonably be attributed to the measured value.

When reporting an uncertainty analysis, the measurand shall be specified; its value and the expanded uncertainty shall be stated, together with the coverage factor, if this differs from 2. The standard uncertainties of each of the individual uncertainty components are also reported, together with their type of evaluation (A or B; see GUM [45]) and any assumptions made in estimating the individual uncertainty components.

GUM uses concepts based on observable quantities. The definition of uncertainty of measurement is an operational one that focuses on the measurement result and its evaluated uncertainty.

Error budgeting, using uncertainty, is one of the main approaches to improve machine tool/measuring instrument accuracies. Identification of the main sources of error allows either error reduction (isolation of the error source and elimination of that source) or error correction (calibration and adjustment through hardware or software correctors). When a machine tool/measuring instrument has been optimized, the error budget typically has several (four to
Figure 8. Some typical sources of error for an error budget. Reproduced from ISO [46]. (Online version in colour).

six) main error sources that contribute equally to the error budget. Improvement in accuracy requires all the main error sources to be reduced together, which may not be economic.

(a) Case study: uncertainties associated with the fitted parameters

Suppose measured coordinates \( x_I = \{ x_i : i \in I \} \) are perturbations of true data \( x_I^* = \{ x_i^* : i \in I \} \) lying exactly on \( u \mapsto f(u, b^*) \) for some \( b^* \) according to \( x_I \in N(x_I^*, V) \), i.e. the perturbations are drawn from a multivariate Gaussian distribution with variance matrix \( V \). Then, to first order, the fitted parameters are a draw

\[
\mathbf{b} \in N(\mathbf{b}^*, V_b), \quad V_b = S V S^T, \quad S = (J^T J)^{-1} J^T N,
\]

where \( N \) is the \( m \times 3m \) block–diagonal, orthogonal matrix with normal vector \( \mathbf{n}_i^T \) in its \( i \)th row. The matrix \( S \) describes the sensitivity of the fitted parameters with respect to the random perturbations. If \( V = \sigma^2 I \), indicating that the noise associated with the measured data is independently distributed with standard deviation \( \sigma \), then \( V_b = \sigma^2 (J^T J)^{-1} \). In this case, the LSODR estimate corresponds to the maximum-likelihood estimate and represents statistically efficient use of the measurement data to estimate the surface parameters.
For many measurement systems, the assumption of independent Gaussian noise is an oversimplification. For example, it is known that a CMM will generally have kinematic errors relating to scale, squareness, roll, pitch and yaw. These systematic effects can be modelled by polynomial or spline functions depending on parameters \( \mathbf{a} \) to be determined and corrected for on the basis of observation. Because these corrections are not exact, there will be residual uncertainties associated with them, encoded by a variance matrix \( V_a \), that are propagated through to the measured coordinates \( \mathbf{x}_I = \{ \mathbf{x}_i : i \in I \} \). In this case, the variance matrix \( V \) associated with the measured coordinates will have the form

\[
V = S_{11} S_{11}^T + S_{12} S_{12}^T,
\]

(8.1)

with \( S_{11} = \sigma I \), so that \( V_b = SS_{11} S_{11}^T S_{12} S_{12}^T \), decomposing the variance associated with the fitted parameters into random and systematic components.

For general variance matrices, the LSODR estimator is not the maximum-likelihood estimator and a statistically more efficient approach is to determine the estimate that solves

\[
\min_{\mathbf{u}_I, \mathbf{b}} \left( \mathbf{x}_I - \mathbf{f}_I(\mathbf{u}_I, \mathbf{b}) \right)^T V^{-1} \left( \mathbf{x}_I - \mathbf{f}_I(\mathbf{u}_I, \mathbf{b}) \right),
\]

(8.2)

involving the \( 2m \) footpoint parameters \( \mathbf{u}_I = \{ \mathbf{u}_i : i \in I \} \) as well as the surface parameters \( \mathbf{b} \). For large datasets, this optimization problem becomes computationally expensive. However, for many coordinate measurement systems, the variance matrix associated with the coordinates can be expressed as in (8.1), with \( S_{11} \) a block–diagonal matrix with \( 3 \times 3 \) blocks \( S_i \) along the diagonal and \( S_{12} \) a full matrix with \( 3 \times n S_{0,1} \) blocks corresponding to each point \( \mathbf{x}_i \). In this case, (8.2) can be reformulated as

\[
\min_{\mathbf{u}_I, \mathbf{e}_I, \mathbf{e}_0, \mathbf{b}} \left\{ \mathbf{e}_0^T \mathbf{e}_0 + \mathbf{e}_I^T \mathbf{e}_I \right\} \quad \text{subject to} \quad \mathbf{x}_I = \mathbf{f}_I(\mathbf{u}_I, \mathbf{b}) + S_{11} \mathbf{e}_I + S_{12} \mathbf{e}_0
\]

(8.3)

or

\[
\min_{\mathbf{e}_0, \mathbf{b}} \left\{ \mathbf{e}_0^T \mathbf{e}_0 + d^2(\mathbf{x}_i, \mathbf{b}, \mathbf{e}_0) \right\},
\]

where \( d(\mathbf{x}_i, \mathbf{b}, \mathbf{e}_0) \) is the distance from the point \( S_i^{-1} \mathbf{x}_i \) to the surface \( \mathbf{u} \rightarrow \mathbf{f}_i(\mathbf{u}, \mathbf{b}, \mathbf{e}_0) = S_i^{-1} \mathbf{f}_i(\mathbf{u}, \mathbf{b}) + S_i^{-1} S_{0,1} \mathbf{e}_0 \). In other words, (8.2) can be solved as a LSODR problem.

The advantage of solving the problem as formulated in (8.3) as opposed to (8.2) is that the solution values for \( \mathbf{e}_0 \) can be used to update the estimates of the systematic effects associated with the measuring system. For example, measuring a ring gauge with a low form error will provide information about the squareness errors associated with a CMM. This functionality is also important in comparing the measurements of the same set of points using two systems. Suppose two systems produce estimates and variance matrices

\[
\mathbf{x}_I = \{ \mathbf{x}_i : i \in I \}, \quad V = S_{11} S_{11}^T + S_{12} S_{12}^T, \quad \mathbf{y}_I = \{ \mathbf{y}_i : i \in I \},
\]

\[
W = T_{11} T_{11}^T + T_{12} T_{12}^T,
\]

associated with a set of points \( \mathbf{z}_I = \{ \mathbf{z}_i : i \in I \} \). Let \( \hat{\mathbf{z}}_I = \hat{\mathbf{z}}_I(\mathbf{z}_I, \mathbf{t}) \) be a rigid body transformation of \( \mathbf{z}_I \) depending on transformation parameters \( \mathbf{t} \). The combined
estimate of $z_I$ is found by solving

$$\min_{t,x_I,f_0,g_0} \left\{ f_0^T f_0 + g_0^T g_0 + f^T f + g^T g \right\},$$

with

$$f = S_{11}^{-1} (x_I - \hat{z}_I(t_I) - \hat{S}_{I} f_0), \quad g = T_{11}^{-1} (y_I - z_I - T_{I} g_0).$$

The solution values for $f_0$ and $g_0$ can be used to update the estimate of the systematic effects associated with both systems. Because any element of the functions $f$ and $g$ involves only one $z_i$, the associated Jacobian matrix is block angular, as in (7.1), and the computation can be made very efficient.

9. Future challenges

Precision artefacts are becoming more complex, more precise and will have a larger range of functional requirements with increased performance. They are also required to be manufactured at a lower unit cost. All of this will reflect in the future mathematical challenges within the four worlds that precision artefacts live.

Specification will become more flexible to cover the range of geometries and functional requirements. The toolbox concept for specification (through specific tools for features, characteristics and conditions) allows for this flexibility and can be designed to emulate specific functional requirements. The concepts of ‘scale-limited surfaces’ is becoming increasingly important in specification. There are many different types of scale spaces (spectral, morphological, segmentation, PDE-smoothing, etc.), each of which has properties useful for different surface functions. The three ways to limit a scale space, useful for specification, are: removal of the smaller scales only; removal of smaller scales together with form removal; and removal of both the smaller and larger scales (including form) but retaining the middle scales. The three types of scale-limited surface, together with the different functional scale spaces, are helping to solve one important challenge in making specification a unified mathematical system that covers all scales (size, shape and texture) and all geometries.

Manufacture will become more reliant on simulation and mathematical modelling of the manufacturing processes to ensure consistent and more accurate results. The mathematical modelling of vibration, thermal and environmental effects, etc. will become more detailed as computers possess more power, because the power of today’s supercomputers becomes tomorrow’s power for desktops. These computational approaches will be used in the design of improved manufacturing processes, particularly machine tools. The models can also be used to calibrate and adjust the control software to correct for any predicted errors.

Verification is undergoing a revolution in moving to the digital paradigm. Improved scales such as holographic gratings or direct interferometric scales for metrology frames allow for the increased accuracy of the coordinate system. The specification of scale-limited surfaces with a smallest surface scale will allow various sampling theorems to apply when establishing a sampling procedure. It is also possible to data-fuse sensor information to improve range and resolution of the measurement and to adjust the measured values resulting from the sensor to correct for errors. But the main advantage of the digital paradigm is the

*Phil. Trans. R. Soc. A* (2012)
use of powerful mathematical algorithms capable of mirroring the specification operations, particularly form fitting, and being able to calculate almost any desired characteristic.

Product life challenges are mainly in the design of the precision artefact to achieve the desired quality and utility over the life of the artefact. This includes functional efficiency, increased life of the artefact, cost of using the artefact, etc. Simulation and mathematical modelling will play their full role in optimizing product life quality and utility.

As a final thought, the importance of mathematical models for the future is illustrated by a quote from William H. Press of Harvard University:

Simulation and mathematical modeling will power the twenty-first century the way steam powered the nineteenth [35].

This work was partly funded by the UK’s National Measurement System Mathematics and Modelling for Metrology programme. We thank Dr Ian Smith, NPL, for comments on a draft of this paper. We also thank the directors of Taylor Hobson Ltd for permission to publish.

References

1 Jones, R. V. 1979 Precision engineering; from supertankers to integrated circuits (comment). Prec. Eng. 1, 3. (doi:10.1016/0141-6359(79)90069-2)
3 Ahmose. ca 1557 BC The Rhind mathematical papyrus, year 33 of the reign of King Auserre (Apophis), Fifteenth Dynasty, second intermediate period Ancient Egypt.

Phil. Trans. R. Soc. A (2012)

