## 3 Processing, Operations, and Simulations

**COMMENT**

Although the subject of engineering surfaces covers a large number of interdisciplinary subjects ranging from, say, economics on the one hand to chemistry on the other, there are a number of areas in which a degree of unity can be seen. One of these, perhaps the most important, is in the processing of measured information to enable it to be used most effectively for control of production or design purposes. Processing information is taken to mean making changes to the original data to isolate, separate, or identify features which may be of particular significance. Certain operations which keep on recurring will be highlighted. There are also other considerations which accompany processing such as filtering. For example, in some cases it may be necessary to construct surface profiles or areal maps from signals obtained from a surface by means other than by scanning i.e., optical diffraction or it may be necessary to investigate ways of generating surfaces in the computer data from. In this section a few of these will first be described and then illustrated with particular reference to some of the issues pertinent to the subject matter being discussed. Parts of the chapter will be tutorial in character to help ensure that a common base of understanding is available.

Using all or any combination of these “processing” methods should allow a researcher, designer, or other user not only to understand more fully the background to the subject but also to progress the technique of surface analysis and measurement further. As this part of the subject of surface metrology is continuously being advanced it will only be possible to outline lines of advance rather than to delve in great detail.

The signal being processed is that obtained from the surface by any means. This could mean digital or analog information obtained from a measuring instrument. It can, however, equally be a signal obtained by utilizing a mathematical model of the surface.

The first of these fundamental operations to be considered is that of numerical techniques, and the second is filtering in one form or another. These are singled out because of their special importance. Obviously the techniques have overlapping regions so that the discussion of filtering will necessarily involve some discrete concepts. However, the repetition of such important concepts is not a redundant exercise. Some mention will be made of splines wavelets, fractals, and various other issues roughly mirroring the more general concepts of Chapter 2.

Filtering is examined in the same sense as that used in system vibration analysis; that is to say that the frequency characteristics of the input signal are operated on by the same method whether electrical, mechanical, or computational. One reason for doing this is to introduce the important concepts of system response and convolution. These terms have wider mathematical significance than equivalent expressions used in either of the individual disciplines, for example they happen to have equivalents in random process theory. Other subjects will be included wherever deemed necessary.

The other point about the numerical analysis of random surfaces is that, to some degree, this is exploring the properties of discrete random surfaces. To this extent there is some overlap with the methods of characterizing surfaces given in Chapter 2.

### 3.1 Digital Methods

In what follows it should be assumed that digital methods are being used unless specified. Various optical processing methods will be considered in Chapter 4. Graphical methods, which are useful if the existing recorded data have to be checked and can also be used as a last resort to verify the results of simulations are mentioned.

Digital methods arise in three major areas: the first is in the analysis of the various waveforms, the second is in simulations, and the third is the use of computers to control instruments, machines, etc. Here various aspects of the analysis of waveforms using digital methods will be considered; simulations such as generation and reconstruction will be discussed where appropriate: control will be left until Chapter 4. Digital methods have virtually replaced analog methods in surface metrology because they are more flexible and more accurate than analog methods.

Three basic considerations have to be taken into account in order to get useful digital results. These are the sampling, quantization, and the numerical technique. Each is briefly considered together with typical examples. Emphasis is given to those problems unique to surface metrology. The starting point is taken to be a typical waveform which could have been obtained by any number of methods.

#### 3.1.1 Sampling

The operation of sampling both in time (or space) and frequency is shown in Figure 3.1 in which (a) represents a typical waveform \( z(t) \) to be sampled and (b) its frequency characteristic which is shown limited at a frequency \( B \). In what follows the word “time” will be synonymous with “space”: strictly speaking surfaces are analyzed in terms of spatial frequency not temporal.

The process of taking a single sample of a time signal is equivalent to multiplying the time function by a unit impulse \( \delta \) at \( t_i \) and integrating. Thus

\[
z(t_i) = \int_{-\infty}^{\infty} z(\tau)(t_i - \tau) d\tau
\]

(3.1)
where $\tau$ is a dummy variable. This operation is called the sampling property of impulses. Sampling at regular intervals “$h$” in time is equivalent to multiplying the waveform by an impulse train, where each impulse is separated by $h$, and then integrating. The equivalent operation in frequency is shown in Figure 3.1b, d, and f. The Fourier transform of an impulse train is itself an impulse train whose spacing is $1/h$ as shown in Figure 3.1d. Because time sampling involves a multiplication, the equivalent operation in frequency is a convolution. Thus Figure 3.1f shows the effect of convoluting the frequency impulse train with that of the frequency characteristic of the waveform (shown symmetrical about the zero-frequency axis).

The criterion for good sampling is that all information should be recoverable. From Figure 3.2 it is obvious that passing this sampled signal through a low-pass filter whose low-frequency cut is higher than $B$ will remove the other bands of frequency introduced by the sampling, namely $A$, $B$, $C$, etc. But this is only possible if the other bands do not encroach into the band around zero frequency and this is only

![FIGURE 3.1](image-url) Pictorial representation of Nyquist sampling theorem showing (a) signal profile, (b) spectrum of signal profile, (c) sampling points comb, (d) spectrum of sampled comb, (e) sampled signal, (f) spectrum of sampled signal, (g) sample comb with interval too wide, (h) spectrum of (g) showing spectral overlap causing aliasing.

![FIGURE 3.2](image-url) Use of filter to reduce bandwidth.
possible providing \(1/h > 2B\), otherwise the situation shown in Figure 3.1h arises in which an overlap occurs. Cutting out the high-frequency bands, even with an infinitely sharp cut filter still does not isolate the original frequency because some degree of scrambling has taken place. The extent of this scrambling can be seen by the cross-hatched area shown in Figure 3.1h. It is in effect a folding back of the frequency characteristic, on itself, about the mid-line. Problems arising from this folding will be described shortly. However, one important fact emerges: in order to be sure of preserving all the information in an analog signal of frequency \(B\) it is necessary to sample in time at a spacing which is a maximum of \(1/2B\) long. This is called the Nyquist criterion.

Unfortunately signals in the real world do not have a distinct cut-off in frequency \(B\) as shown in Figure 3.2. Various insignificant frequencies are invariably present therefore precluding satisfactory sampling and opening the way to the folding (or aliasing) effect. Consequently, it is usual to decide on the highest frequency of real interest, to filter the signal by analog means to remove higher frequencies and then to sample.

Notice that the Nyquist criterion has to be relaxed slightly depending upon how sharp the analog filter can cut. Simply sampling at \(1/2B\) will cause a fold-over at \(B\) due to the inability of filters to attenuate infinitely sharply. A guard band \(G\) of frequency is usually incorporated to cater for the finite drop-off of the filter.

\(G\) is taken to be about 0.5\(B\) so that the sample rate becomes \(\sim 1/3B\) in time. It is possible to use a digital filter for this preprocessing only if the data are previously sampled at a much higher rate than the Nyquist rate for the frequency of interest. But this is sometimes wasteful in effort. It has the advantage that artificial filters can be used (described in the section on filters).

To illustrate the sort of misleading results which can occur when there is an interaction between sampling rate and signal frequency, consider Figure 3.3. This shows that by sampling at a distance slightly different from that of the true wavelength a false wavelength appears. This is similar to “beats” between waves.

“Aliasing” is a similar effect. Apparent low frequencies are introduced by the folding of frequencies around the sampling frequency. It becomes impossible to detect whether a signal of frequency \(f\) is genuine or whether it is the result of an aliased signal \(f_2-f\), or is it simply \(f_1\) as shown in Figure 3.4.

Other forms of sampling can reduce these problems. Second-order sampling is still periodic in nature but within each period two measurements are taken, usually close together. This sort of sampling has been used on signals having a band-pass frequency characteristic. Random sampling has also been used where fewer samples need to be taken but the problem then arises of unscrambling the data afterwards.

Summarizing, samples should be taken at about \(3 \times\) the rate of the highest frequency required and known to be present. Sampling much more often than this can be wasteful and only results in highly correlated data which can give biased results and lead to numerical problems.

In surface texture measurement the signal representing the roughness waveform has already been smoothed relative to the true surface profile because of the finite size of the stylus tip, which acts as a mechanical filter, or optically by the limited resolution of the sensor so that the problem of trying to decide the highest frequencies has to some extent been solved prior to measurement. In practical instruments the tip is about 2.5 \(\mu\)m, which implies that sampling should take place every micrometer or so. Similarly the wavelength of light sets an optical limit as discussed together with the restrictions of other sensors in Chapter 4.

### 3.1.2 Quantization

This is not concerned with the way in which an analog signal is turned into a time series of data points it is concerned with the conversion of an analog waveform into a digital form.
This always means a choice between two signal levels: the separation being determined by the discrimination of the analog to digital (A/D) convertor. A point on the analog signal at \( P \) in Figure 3.5 will have to take either the value of level A or level B, whichever is the nearer. Having to take discrete values is the process of digitization.

This breaking down of the continuous signal into discrete levels can introduce errors known as quantization errors. They do not refer to instrumental accuracy and such errors are usually small. For instance, some idea can be obtained by using Sheppard’s grouped data result. With this it is easy to show that, if \( q \) is the separation of levels, then an RMS noise \( \varepsilon \) of \( q/\sqrt{12} \) will be introduced into any assessment of the RMS value of the digital signal above that of the signal itself. Normally in metrology the quantization interval, expressed as a percentage of the signal value, is about 0.1%, and hence \( \varepsilon = q/\sqrt{12} = 0.03\% \) which is negligible. It only becomes significant if the separation of levels, that is the quantization interval, becomes comparable with the signal size. In almost all practical cases in metrology the quantization intervals are equal over the whole range, but use has been made of unequal intervals in the measurement of autocorrelation, for example.

### 3.1.3 Effect of Computer Word Length

Often spurious resolution can appear to develop during a calculation, especially when converting from one type of word in the computer to another, for instance from integer to floating arithmetic. The number 10 becomes, say, 10,000; the decimal digits to the right of the decimal point are not significant in this case. Suppose the number 10 represents a profile ordinate that the A/D convertor could not resolve better than the unity digit. The signal itself may have been anywhere between 10 and 11. Certainly the probability of it being 10.000 is remote, but this does not mean that all floating point numbers derived from integer numbers are incorrect. As an example, if a profile is made up of a string of integer numbers 583, 621, 718, etc., then the low-pass-filtered profile may be expressed as 592.3 for instance, because the uncertainty in a mean (weighted mean for filtering) is less than that of individual numbers. It still does not mean, however, that all the printout of the floating point numbers is significant. As a rule of thumb, if \( q \) is the uncertainty in the individual numbers then the mean can be taken significantly to about one decimal digit further than the individual values. Note that if the A/D convertor is two digits and the word length of the computer is 16 bits (four digits), the last digit is not likely to be significant if a simple convolution type of exercise is being carried out. This has nothing to do with numerical analysis problems—it is purely the accuracy in the values themselves, whether theoretically correct or not!

Another form of number notation called the scientific or E notation appears at first sight to offer both high resolution and range that is 0.63152 E 23 would mean 0.63152 \( \times 10^{23} \). The reason for this apparent benefit is that most computers use two words or more to present a number in the scientific notation. Thus, for a 16-bit word, 32 bits are available, usually 24 bits for the mantissa and 6 bits for the exponent, so an increase in accuracy is only at the expense of store. This used to be a problem but is less so today. Whether this accuracy is real or not depends on the initial data and the arithmetic operations carried out.

Problems of this nature are always arising in curve fitting, matrix inversion, etc. Their individual solution depends on the particular problem and the type of computer. There is always an improvement if care is taken.

Some obvious checks should be made to ensure that sensible values are being obtained. One example of this is in the measurement of slope. Here the model, quantization, and sampling again conspire to confuse the issue (see Figure 3.6). This figure shows a case where all three are poor: the quantization interval is too big, the sampling is too fast, and the model (say three-point) is too restrictive. Let the sampling interval be \( q/10 \).

Then the slopes measured at points 1, 2, 3, … will all be unity until the point \( E \) is reached. Then the slope will be \( q/2 + q/10 = 5 \), giving an angle of 78%! Not only does this look ridiculous, it cannot make physical sense if the device used for obtaining the analog signal is a stylus with a semi-angle of 45°. Real slopes are greater than this cannot be seen. Also, curvature measurement as revealed digitally cannot or should not be greater than that of the stylus itself! Common-sense
3.1.4 NUMERICAL ANALYSIS—THE DIGITAL MODEL

One of the main reasons why the results obtained by different people often do not agree—even given the same data—is that not enough attention is given to numerical techniques. In this section some of the basic operations needed in analysis are examined, followed by particular problems more specific to engineering surfaces. No attempt is made to cover the subject as a whole. Suitable references enable the operations usually employed—including differentiation, integration, interpolation, extrapolation, curve fitting—to be separately studied.

3.1.4.1 Differentiation

There is a tendency among people versed in analog ways to take a very simple formula for the first differential (Figure 3.7a). Thus, the differential between points $z_1$ and $z_{-1}$ at $z_0$ is

$$\frac{z_1 - z_{-1}}{2h}.$$  (3.2)

In fact, this is only the tangent. More than just the two ordinates are needed to get a good estimate of the first differential. One usually adequate formula involves the use of seven ordinates, equally spaced by $h$. The differential

$$\frac{dz}{dx} \bigg|_{x=z_0} = \frac{1}{6h} \left[z_3 - 9z_2 + 45z_1 - 45z_{-1} + 9z_{-2} - z_{-3}\right].$$  (3.3)

The errors in this are of the order of $(l/140) \mu \delta z_0$, where $\mu$ is the averaging operator between ordinates and $\delta$ is the central differencing operator. These are very small providing that ordinates outside $z_3$ and $z_{-3}$ are well behaved. A similar error in the three-point formula given above is of the order of $(1/6) \mu \delta z_0$, which turns out to be

$$\frac{1}{3h} (z_2 - 2z_1 + 2z_{-1} - z_{-2})$$

when expressed in terms of ordinates. These error formulae show that if the ordinates outside those used are significantly different then errors can creep in Ref. [1].

The numerical Formulae 3.2 and 3.3 are examples of Lagrangian formulae.

By choosing a formula encompassing a wide range of ordinates the chances of rogue ordinates affecting the true derivatives are reduced. Hence the need for seven rather than three-point analysis. Similarly, to use

$$\frac{d^2z}{dx^2} \bigg|_{x=z_0} = \frac{1}{h^2} (z_1 + z_{-1} - 2z_0)$$  (3.4)

as a formula for the second differential is sometimes dangerous. The errors here are of the order of

$$\frac{1}{12} \delta^2 z_0$$  (3.5)

that is

$$\frac{1}{12} (z_2 - 4z_1 + 6z_0 - 4z_{-1} + z_{-2}).$$  (3.6)

An equivalent seven-point formula that reduces noise is

$$\frac{d^2z}{dx^2} \bigg|_{x=z_0} = \frac{1}{180h^2} \left[2z_3 - 27z_2 + 270z_1 - 490z_0 + 270z_{-1} - 27z_{-2} + 2z_{-3}\right]$$  (3.7)

with error $(1/560) \delta^2 z_0$.

Note the fundamental point about these formulae: it is still the central three ordinates that are dominant; the adjacent ordinates merely apply some degree of control over the value obtained should $z_1$, $z_0$, or $z_{-1}$ be freaks or in error. Similarly, the $z_3$ and $z_{-3}$ values act as constraints on $z_2$ and $z_{-2}$, and so on.

Alternative formulae to these exist. It is possible to extend the number of ordinates on either side indefinitely, their effect getting smaller and smaller. It is also possible to evaluate the differentials in terms of backward and forward differences rather than central differences. For example,

$$h \frac{dz}{dx} \bigg|_{x=z_0} = \left(\Delta + \frac{1}{2} \Delta - \frac{1}{6} \Delta^3\right) z_0.$$
where $\Delta$ is the forward difference, whence

$$\frac{dz}{dx} \Bigg|_{x=x_0} = \frac{1}{12h}(z_3 - 6z_2 + 8z_1 - 10z_0 - 3z_{-1}). \quad (3.8)$$

The derivative at $z_0$ has been evaluated by ordinates obtained later in the sequence of data points, similar formulae can be obtained for second derivatives, etc. The only usual reason for using these is because of the difficulty of using symmetrical numerical formulae at the beginning and ending of a set of data points. This enables all the waveform to be differentiated leaving a gap at the front or the end of the data rather than leaving gaps at both ends as is the case for central difference.

3.1.4.2 Integration

The tendency when attempting to integrate is to regard the sum of ordinates multiplied by their spacing as equal to the value of the integral i.e., the area under a curve. Although this is true enough when the number of terms is large it is not so when the number of ordinates is small. The numerical formulae for any function being integrated are not simply the sum of the evaluated points; the values can only be added after modification.

The simplest modification is to halve the first and last ordinates and to add the rest unchanged. This is no more than a statement of the trapezoidal rule for numerical integration and it corresponds, in fact, to joining the function values by straight lines instead of each ordinate being represented by a block. The area is made up of trapezoids rather than rectangles. It is a special case of Gregory’s formula.

Further improvement can be achieved by fitting curves between the points of evaluation. The quadratic curve gives rise to Simpson’s rule for numerical integration. Interpreted in a purely geometric way this gives the sum of the areas under second-degree parabolas. If there are $n + 1$ pairs of data points the integration formula is

$$\int_{x_0}^{x_n} bdx = \frac{h}{3}[b_0 + b_n + 4(b_1 + b_3 + ... + b_{n+1}) + 2(b_2 + b_4 + ... + b_{n-2})], \quad (3.9)$$

where the $b$’s are the ordinates of the curve. For $n + 1$ pairs if $n$ is a multiple of 3:

$$\int_{x_0}^{x_n} bdx = \frac{3h}{8}[b_0 + b_n + 3(b_1 + b_3 + b_5 + ... + b_{n-2} + b_{n-1})].$$

(3.10)

Other more elaborate formulae exist involving unequal spacing of the function values. These require fewer values for a given accuracy for the reasons given below. However, it is well known that for most engineering applications the equal-interval formulae show a surprising accuracy, especially Simpson’s rule. If, however, more accuracy is required then other techniques may have to be adopted. Lagrange interpolation formulae involve fitting a polynomial through any set of points not necessarily equally spaced. This polynomial therefore represents the function as a whole. Even when the intervals are equal, Lagrangian techniques have the advantages of permitting interpolation without the need to construct a difference table. They have the disadvantage common to all polynomial curve-fitting methods of requiring some knowledge of the degree of the polynomial needed to achieve any real accuracy.

In integration, use can also be made of the unequal interval in the Gauss’ formulae for numerical integration. Although these and other formulae exist they are most useful in those situations where a limited amount of well-understood data is present. Usually there is enough data to get the required accuracy for the basic operations using equal intervals and first- or second-degree interpolation.

3.1.4.3 Interpolation and Extrapolation

In many respects these two areas of digital analysis are the most fundamental because in the process of sampling continuous signals have been transformed into discrete values in time (space). Reconstitution of the original signal in between sampled values is of prime importance, as is the ability to project the signal outside its immediate vicinity. Fortunately in surface metrology a lot of data are usually available, and consequently the use of such techniques is rarely required. In essence a polynomial is fitted to the measured or listed points and made to conform with the fixed points. Examination of the polynomial behavior between and also outside the fixed values is then possible. In particular, the most used polynomial is called the Lagrange interpolating polynomial.

Well-known formulae exist for interpolation and extrapolation, in particular those due to Everett and Bessel [1] for equally spaced data and Lagrange for unequally spaced data. An example of the use of interpolation formulae in surface topography comes in the field of contact, where mechanical models of surface peaks are used. Using interpolation it is possible to find the position of the maximum value of a peak even if it is not touched by the sampled points. For instance, if the ordinates are $z_{-1}$ and $z_{+1}$ where $z_{-1} \neq z_{+1}$ then the apex is not at $z_0$ but at a distance $V$ from $z_0$ given by

$$V = \frac{(h/2)(z_{+1} - z_{-1})}{2z_0 - z_{-1} - z_{+1}}. \quad (3.11)$$

It is further possible to work out curvatures at the apex of this peak. Interpolation can also help to cut down the number of mean line points that need to be calculated for the mean line assessment.
3.2 DISCRETE (DIGITAL) PROPERTIES
OF RANDOM SURFACES

3.2.1 SOME NUMERICAL PROBLEMS ENCOUNTERED
IN SURFACE METROLOGY

Some examples will be given by way of illustration of some of the typical problems encountered in surface metrology. Those chosen will not be exhaustive in content but should give a fairly representative cover of the types of problem likely to be encountered.

There are a limited number of features of most interest. From the section on surface characterization much emphasis was placed on the peak and slope measurement of a profile and, in particular, in two lateral dimensions (i.e., the areal case). These features can be extended to include peak height distributions, curvature distributions, and how they vary with height, slope, and associated parameters.

It is only in the last 30 years that digital methods have become readily available to make possible the measurement of these important parameters. However, simply theorizing about parameters is not enough—they have to be measured. This apparently straightforward task is fraught with problems as Section 3.2.2 depicts. In the 1950s, parameters were restricted to those that could be measured with simple analog circuitry. This in itself imposed natural constraints upon the wavelengths which could be measured. Recorders, amplifiers, and meters have bandwidth limitations which cannot be overcome but the restriction on digital methods is less distinct; the experimenter is confronted head-on with the sampling, quantization, and numerical model problem. Sometimes these considerations are outside the experience or training of the investigator. Unfortunately, the parameters of use to the surface metrologist are just those parameters that are difficult to measure. In what follows it will become apparent that the correlation function of the surface is of fundamental importance in assessing the change in value of such parameters with sampling. Instrumental limitations such as the stylus tip or optical resolution will be incorporated more fully in Chapter 4, and parameter variability in Chapter 5.

3.2.2 DEFINITIONS OF A PEAK AND DENSITY OF PEAKS

The problem of defining peaks and assessing peak properties often arises in metrology, especially in contact theory. For instance, it is often of interest to know the density of peaks of a surface profile [2]. The question that has to be posed is how this count of peaks depends on the digitizing interval, the quantization interval, and the definition of the peak. All three can affect the count. This is one of the reasons why it is so very difficult to get agreement between researchers, even given the same data.

Take, for example, the problem of the definition. One of the most used definitions is a three-point model as shown in Figure 3.8a. If the central ordinate of three consecutive ordinates is the highest then the three together constitute a peak. An alternative one is also shown in the figure in

![Figure 3.8](image)

FIGURE 3.8 (a) Possible definition of peaks and (b) effect of quantization on peaks.
which four ordinates are used, the central two being higher than the others for a definition. Many similar possibilities exist. In any case the number of peaks counted will be different for the same data. For example, a peak counted by the three-point method could get ignored using the four or more ordinate models.

Also, differences in definition have been used within the three-point method. Some investigators have imposed a height difference constraint on the definition. For instance, the central ordinate has to be a height \( z' \) above the higher of the other two before a peak is registered, that is as in Figure 3.8a, \( z_0 - z_1 > z' \). This constraint reduces the count as before.

### 3.2.3 Effect of Quantization on Peak Parameters

The quantization interval can influence the count as is shown in Figure 3.8b. It can be seen that using exactly the same waveform, simply increasing the quantization interval by a factor of 2 means that, in the case of Figure 3.8b, the three-point peak criterion fails, whereas in the other case it does not.

So, even the A/D resolution can influence the count. In order to get some ideas of the acceptable quantization interval it should be a given ratio of the full-scale signal size, subject to the proviso that the interval chosen gives sufficient accuracy. As an example of the quantitative effect of quantization, consider a signal that has a uniform probability density. If the range of this density is split up into \( m \) levels (i.e., \( m \) blocks) then it can be shown that the ratio of peaks to ordinates is given by

\[
\text{ratio} = \frac{1}{3} - \frac{1}{2m} + \frac{1}{m^2}. \tag{3.12}
\]

This makes the assumption that the samples are independent and that the three-ordinate model is used as a definition of a peak.

Examination of the formula shows that, when \( m = 1 \), the ratio is 1/3. This makes sense because, for independent samples with no quantization restriction, one would expect one-third of all samples to be peaks. Similarly, when \( m = 1 \), the ratio is zero. Again this makes sense because no information is being conveyed.

Various other values of \( m \) are listed in Table 3.1.

<table>
<thead>
<tr>
<th>( m )</th>
<th>Ratio Peaks/Ordinates</th>
<th>Percentage Drop</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.125</td>
<td>70</td>
</tr>
<tr>
<td>3</td>
<td>0.19</td>
<td>45</td>
</tr>
<tr>
<td>4</td>
<td>0.21</td>
<td>37</td>
</tr>
<tr>
<td>5</td>
<td>0.23</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td>0.25</td>
<td>15</td>
</tr>
<tr>
<td>100</td>
<td>0.33</td>
<td>&lt; 1</td>
</tr>
</tbody>
</table>

Taking this over all possible values of interval allowed, in this case taken to be symmetrical about \( z = 0 \) and extending \( m/2x \) blocks to either extreme of \( 3\sigma \) where \( \sigma \) is the RMS value of the distribution, the probability is

\[
\int_{m/2}^{m/2} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz
\]

assuming that the \( z \) values have a zero mean and unit variance.

Taking this over all possible values of interval allowed, in this case taken to be symmetrical about \( z = 0 \) and extending \( m/2x \) blocks to either extreme of \( 3\sigma \) where \( \sigma \) is the RMS value of the distribution, the probability is

\[
\frac{1}{8} \sum_{n=-3\sigma}^{3\sigma} \left[ \text{erf}\left(\frac{n\Delta z}{\sqrt{2}}\right) \text{erf}\left(\frac{(n-1)\Delta z}{\sqrt{2}}\right) \right] + \text{erf}\left(\frac{(n-1)\Delta z}{\sqrt{2}}\right)^2
\]

which gives the following results:

- \( m = 2 \) probability = 0.125
- \( m = 4 \) probability = 0.167
- \( m = 6 \) probability = 0.211
- \( m = 12 \) probability = 0.266.

These generally follow the rectangular case except for being slightly lower. In both cases the value is asymptotic to 1/3, which is correct for the three-point analysis.

### 3.2.4 Effect of Numerical Analysis on Peak Parameters

While investigating the effect of quantization, it is informative to investigate the effect of different models. This is especially valid in peak behavior because it is in the region of a peak that changes in level tend to be small and the effect of quantization can dominate the result. To combat this predicament it is sometimes useful to use a modified three-ordinate model in which, if the next-to-central ordinate is at the same digital level as the central ordinate, the judgment of whether...
or not a peak exists is deferred for one ordinate. If this is still not conclusive it is deferred further until an adjacent ordinate is lower.

Under these circumstances one would expect rather more peaks to be revealed than when restricting the model strictly to three ordinates. The effect of this modification can be taken account of in probability merely by noting that the probability of the situation shown in Figure 3.9 is the same as for the three-point one except that the central ordinate has been repeated three times, that is the probability is \( P_1^3 P_2^2 \), where \( P_1 \) is the probability of an ordinate lying between \((n-1)\Delta z\) and \(n\Delta z\), and \( P_2 \) is the probability of an ordinate being below \((n-1)\Delta z\).

Taking all such possibilities into account gives an additive sequence of probabilities. Thus

\[
P_1 P_2^3 + P_1^2 P_2^2 + P_1^3 P_2^2 + \ldots + P_1^m P_2^2 = \frac{P_1 P_2^2}{1 - P_1}. \tag{3.15}
\]

The formulae derived previously for the three-point model definition of a peak can therefore be modified to give the following:

1. For a rectangular distribution

\[
\text{probability} = \frac{2m^2 - 2m - 1}{6m(m-1)}. \tag{3.16}
\]

2. For the Gaussian distribution

\[
\frac{1}{8} \sum_{n-m/2}^{n+m/2} \left[ 1 + \text{erf}((n-1)\Delta z/\sqrt{2}) \right]^2 - \frac{1}{2} \left[ \text{erf}(n\Delta z/\sqrt{2}) - \text{erf}((n-1)\Delta z/\sqrt{2}) \right]. \tag{3.17}
\]

Both of these formulae have the effect of considerably increasing the probability of a peak, especially at small values of \( m \) (see Figure 3.10). For both in fact, if the quantization is purely linear, that is \( 0 \pm 1 \) or \( \pm 1 \), a complete change in the count can result depending simply upon which numerical model is chosen to represent a peak—on the same original profile.

As will now be obvious, each one of the constraints of sampling, quantization, and numerical definition can aggravate the effects of the others.

Again, notice that as \( m \to \infty \) for both formulae the value of the probability tends to 1/3 because as \( m \to \infty \) the chances of more than one ordinate lying within the infinitesimal limit become increasingly remote.

The main reason for taking a transducer range of more than one decimal digit is that it is not always possible to ensure that the signal is a significant percentage of the range of the A/D convertor. For very smooth surfaces this is often the case when the part is fitted. Another reason that will be amplified later is concerned with the best use of computers in metrology. The latest trend is to let the computer take out a large part of the manual setting-up by accepting all the signal from the workpiece—including errors of form, misalignment, waviness, etc.—removing all the errors that are unwanted, digitally, and then magnifying the remainder for evaluation. This technique presupposes that the remaining signal is known with sufficient accuracy (i.e., to a sufficient resolution) to be useful. In order to ensure this, the whole of the original signal has to be digitized to a reasonably small quantization interval. This is because the process of
removing the large error signal leaves only a small percentage of the total, as seen in Figure 3.11.

### 3.2.5 Effect of Sample Interval on the Peak Density Value

The real issue here is not the physical distance between the discretely measured sampled ordinates, but how this distance relates to the autocorrelation function of the surface. In order to investigate the way in which ordinate spacing affects surface parameters it is necessary to know what constraints exist on the choice of a typical correlation function.

Perhaps the obvious choice of autocorrelation function is the exponential form shown in Figure 3.12. The argument for adopting this is that most finished surfaces exhibit such autocorrelation functions in regions other than at the origin. This is a direct result of the Poissonian nature of surface generation incurred because of the random cutting action. $\tau_{max}$ is an estimate of the correlation length (for $\rho = 0.1$) and is not to be confused with $\tau_{corr}$ which is the general term for the correlation length given by Equation 3.22.

There are, however, theoretical objections to the use of the exponential autocorrelation function, for example there are certain features of the exponential function at the origin which are undesirable. In particular, the density of peaks $D_p$ is given, for an ordinary, random process, by

$$D_p = \frac{1}{2\pi} A''(0) \left\{ A''(0) \right\}^{1/2}, \quad (3.18)$$

where $A''(0)$ is the second differential of the autocorrelation function at the origin and $A''(0)$ is the fourth differential. For the exponential function these are undefined because $A'(0)$ and $A''(0)$ are non-zero: they should be! This becomes necessary because the autocorrelation function is an even function. In practice, this is not a serious problem because there is always some degree of smoothing of the profile caused by the finite resolution of the measuring instrument. In what follows the exponential correlation will be assumed for simplicity unless specified otherwise.

Consider for example, as before, the measurement of peak density or the probability that an ordinate is a peak. Using the three-point model the latter becomes $N$, where $N$ is given by

$$N = \frac{1}{\pi} \tan^{-1} \left\{ \frac{3 - 4\rho}{1 + \rho} \right\}^{1/2}, \quad (3.19)$$

and $\rho$ is the correlation between ordinates separated by $h$. For an exponential correlation function $h = \beta \ln(1/\rho)$ as shown in Figure 3.12.

The peak density is given simply by Equation 3.20 and gives the number of peaks in the unit length in which $h$ is measured.

$$N/h.$$

In the case of a general autocorrelation function shown, for example in Figure 3.13, two correlation coefficients are needed for the three-point analysis, $\rho_1$ corresponding with the autocorrelation value at $h$ and $\rho_2$ the value at $2h$. The formula relating the peak density to the correlation function and hence to the spacing is modified somewhat to give

$$\frac{N}{h} = \frac{1}{\pi h} \tan^{-1} \left\{ \left( \frac{3 - 4\rho_1 + \rho_2}{1 - \rho_2} \right)^{1/2} \right\}.$$

How the value of peak density changes with $h$ can easily be calculated for an exponential autocorrelation function and is shown in Figure 3.14.

Varying $\rho$ from 0 to 1 gives values of $N$ from 1/3 to 1/4, respectively, in theory at least. The value of 1/3 is to be expected because when $\rho = 0$ the sampled ordinates are
independent. The separation when this occurs is usually a matter of definition. One such definition, shown in Figure 3.12, is the value of $h$ such that $\rho \approx 0.1$ to measure is $\tau_{\text{max}}$. Another definition, which is more reliable is $\tau_{\text{corr}}$, thus,

$$
\tau_{\text{corr}} = \int_0^\infty |A(\tau)| d\tau. \quad (3.22)
$$

For an exponential correlation function normalized i.e., $A(0) = 1$, $\tau_{\text{corr}} = 1$, so the value of correlation $\rho$ at this is $\exp(-1)$ which is $1/e \approx 35\%$.

When the ordinates are highly correlated for the exponential correlation the density is, in theory at least, as $\rho \to 1$, $N \to 1/4$.

When $h$ is small the practical value of peak density always falls to zero. This can be seen by postulating any correlation function having well-behaved derivatives at the origin and inserting the $\rho_1$ and $\rho_2$ values measured from it into the general formula for $N$ given above in Equation 3.21. Take for example the case when the correlation function is Gaussian.

The comparison between this and the exponential is as follows.

From the standard Gaussian form the normalized autocorrelation function is

$$
A_G(\tau) = \exp\left(-\frac{\pi \tau^2}{4}\right) \quad (3.23)
$$

and the comparison for the exponential is

$$
A_E(\tau) = \exp(-\tau). \quad (3.24)
$$

These are plotted in Figure 3.15.

Consider the formula for the probability that an ordinate is a peak given above as Equation 3.21. When $h = 1$

It can be seen from Table 3.2 that there are considerable differences between the various types of correlation function. In fact for the Gaussian correlation the relationship is very nearly linear.

Summarizing, therefore,

$$
N = \frac{3 + h}{12} \quad \text{for exponential}
$$

$$
N = kh \quad \text{for Gaussian}, \quad (3.25)
$$

where $k$ has a value between $1/3$ and $1/4$ depending on the approximation. The relevant point here is that the probability of an ordinate being a peak using the three-point numerical model

![Figure 3.13 General autocorrelation function.](image1)

![Figure 3.14 Peak density-effect of correlation.](image2)

![Table 3.2 Peak Densities for Different Sample Intervals, Gaussian, and Exponential Correlation](table)

<table>
<thead>
<tr>
<th>Correlation</th>
<th>$h = 0.5$ ($\rho_1$)</th>
<th>$2h = 1$ ($\rho_2$)</th>
<th>Probability $N/h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>0.82</td>
<td>0.46</td>
<td>0.164</td>
</tr>
<tr>
<td>Exponential</td>
<td>0.61</td>
<td>0.37</td>
<td>0.280</td>
</tr>
<tr>
<td>Correlation</td>
<td>$h = 1$</td>
<td>$2h = 2$</td>
<td></td>
</tr>
<tr>
<td>Gaussian</td>
<td>0.47</td>
<td>0.01</td>
<td>0.297</td>
</tr>
<tr>
<td>Exponential</td>
<td>0.37</td>
<td>0.13</td>
<td>0.301</td>
</tr>
</tbody>
</table>

![Figure 3.15 Autocorrelation functions Gaussian vs. exponential.](image3)
is nearly constant for all sample intervals for the exponential correlation function surface but is nearly a linear function of the sample interval for surfaces having a Gaussian correlation. For small values of \( h \approx 0.1 \) many surfaces, even the exponential, will exhibit a linear relationship between probability of a peak and sample interval similar to that of the Gaussian in Equation 3.25b because of the smoothing produced by the instrument probe, whether using a stylus or any another method. Remember that for values of \( N \) the sample interval \( h \) has to be expressed as a fraction of the correlation length (which is about equal to one-third of the high-spot wavelength).

All digital features can be incorporated into the evaluation of \( N \). Thus, for example, the exponential case gives

\[
N = \frac{1}{8h} \left[ \text{erf}\left( \frac{n\Delta q}{\sqrt{2}} \sqrt{1 + \rho} \right) - \text{erf}\left( \frac{(n-1)\Delta q}{\sqrt{2}} \sqrt{1 + \rho} \right) \right]
\]

\[
\left[ 1 + \text{erf}\left( \frac{(n-1)\Delta q}{\sqrt{2}} \sqrt{1 + \rho} \right) \right]^2
\]

\[
\left[ 1 - \frac{1}{2} \text{erf}\left( \frac{n\Delta q}{\sqrt{2}} \sqrt{1 + \rho} \right) - \text{erf}\left( \frac{(n-1)\Delta q}{\sqrt{2}} \sqrt{1 + \rho} \right) \right].
\]

(3.26)

This incorporates quantization, sampling, and the numerical model and is shown in Figure 3.10 for \( \rho \) values of 0, 0.7, and 0.9, the last two being typical of those used in practice. Note that in this equation \( N \) is the peak density. Removal of the \( h \) in the RHS of the equation makes \( N \) the probability.

From what has been said it is clear that problems of surface analysis, such as peak definition, arise from three different factors: the quantization, the sampling, and the model. These three cannot be divorced from each other. Not to realize this interrelationship can be the source of much trouble when different investigators compare results. Any one of these three variables can be used as a fixed datum: the model, because it is used universally; the quantization, perhaps because of an accuracy requirement on individual measurements or subsequent analysis; and/or the sampling, because of instrumentation problems (stylus resolution) or the functional significance of the harmonic picked out.

To illustrate this interdependence, suppose that it is important to make sure that a particular numerical model is universally valid, say the simple three-point method for peaks. A criterion could be used in which the allowable loss of peak data is, for instance, 5%. If the surface is Gaussian some quantitative picture can be obtained.

From Figure 3.10 it can be seen that in order to ensure that only 5% of peak information is lost using the three-ordinate model, about 50 or 60 quantization intervals are required throughout the range of the signal. Because the ordinates in this model are Gaussian it is possible to work out a relationship between the average difference of height from ordinate to ordinate and to equate this difference of adjacent ordinate heights as

\[
E(z_1 - z_2)^2 = 2\sigma^2(1-2),
\]

where \( \sigma \) is the RMS value of the surface and \( \rho \) is the value of the correlation between sets of data separated by such an interval in space that the autocorrelation is \( \rho \).

The average separation \( \Delta z \) is \( \sqrt{2}/\pi \) the square root of the variance. Thus

\[
\Delta z = \frac{2\sigma}{\sqrt{\pi}} \sqrt{1-\rho^2} = 1.13\sigma\sqrt{1-\rho^2}.
\]

(3.27)

For \( \rho = 0 \), that is independence between ordinates, \( \Delta z = 1.13\sigma \).

If it is assumed that the range of the signal is \( \pm 3\sigma \) then the acceptable relation between \( z \) and \( q \), the quantization interval, is

\[
1.13 \times \frac{50 - 60}{6} \sim 12
\]

that is \( \Delta z \sim 12q \). (It is not just in the measurement of peaks that problems lie. There are similar problems in slope measurement and curvature, as will be seen later in Sections 3.2.6.2 and 3.2.6.3.)

Remembering that this ratio must be preserved for all other digital sample intervals in order to maintain the 95% integrity, the ratio of \( q \) to \( \sigma \) for any other interval can be determined. Take for example the case where \( \rho = 0.7 \). The value of \( \Delta z \) becomes 0.58\( \sigma \). Hence \( q = \Delta z/12 - 0.05\sigma \) Therefore, to cover the range of 6\( \sigma \) for the signal requires about 120 quantization intervals \( q \).

Thus, in order to maintain the integrity of the model from a sampling giving \( \rho = 0 \) to one giving \( \rho = 0.7 \), the number of \( q \) levels has to be increased from a nominal 60 to 120, a factor of 2. If this integrity is not preserved comparative deviations in peak count cannot be attributed to the surface profile itself. A similar quantitative result is obtained if the differences between ordinates around the peaks themselves are investigated rather than those of the profile as a whole.

Other constraints may have to be added, for instance the inability of the stylus to resolve small-wavelength undulations or the need to exclude the short undulations for functional reasons. Two things are plain: first, the three variables of numerical model, sampling, and quantization are related and, second, at least one of these three needs to be fixed, preferably by using a functional requirement—which is usually a sampling consideration. So the preferred rule is to choose the sampling to match the function and work out a suitable quantization and numerical model in terms of the sampling. In a lot of work on contact wear and friction, it is the features of long wavelength picked out by independent samples that dominate performance and therefore need to be unambiguously measured.

### 3.2.6 Digital Evaluation of Other Profile Peak Parameters

#### 3.2.6.1 Peak Height Measurement

The method adopted previously in which the exponential autocorrelation function is used as a basis will be continued
here. Expressions are derived which can easily be transformed into the general autocorrelation case, the height distribution should be nominally Gaussian.

The mean peak height \( \bar{z}_p \) as a function of correlation between ordinates of \( \rho \) is given by

\[
\bar{z}_p = \frac{\sqrt{1-\rho}}{2\pi^{1/2}} \frac{\sqrt{1+\rho}}{(1+\rho)^{1/2}} \frac{(1-\rho)\sqrt{1+\rho}}{2\sqrt{3-\rho} \tan^{-1} \left( \frac{3-\rho}{1+\rho} \right)} \left( 1 + \frac{\tan^{-1} \left( \frac{3-\rho}{1+\rho} \right)}{4(1+\rho)^{1/2}} \right) \]

(3.28)

and for the general autocorrelation by

\[
\bar{z}_p = \frac{(1-\rho)\sqrt{1+\rho}}{2N} \]

(3.29)

where \( N \) is the probability that an ordinate is a peak; \( \bar{z}_p \) is normalized with respect to \( \sigma \).

Therefore, in the simple exponential case the mean peak height expressed in terms of the RMS value of the surface migrates from \( 1.5/\sqrt{\pi} \) to zero as the correlation changes from zero to unity. This height value would have to be multiplied by \( \sqrt{\pi}/2 \) if the deviations were to be expressed in terms of the \( R_s \) value of the surface. This behavior is modified slightly in the case of the general autocorrelation because although, as one would expect, the independent result is the same, the limiting behavior as \( \rho = 1 \) and \( \rho = 0 \) tend to unity is different. There is a limiting value of mean peak height for \( \rho = 0 \), which depends on the particular type of autocorrelation function modeled. Thus, for a highly oscillatory autocorrelation function, the limiting value of \( \bar{z}_p \) could be near to or even higher than the RMS value \( R_s \) of the surface. In any event it must be positive.

The standard deviation of peak height \( \sigma_p(\rho) \) is also affected in the same way. Thus for an exponential autocorrelation function

\[
\sigma_p(\rho) = \left[ 1 + \frac{(1-\rho)(1+\rho)}{2\pi^{1/2}} \frac{\tan^{-1} \left( \frac{3-\rho}{1+\rho} \right)}{4(1+\rho)^{1/2}} \right]^{1/2} \]

(3.30)

which, for \( \rho = 0 \), gives \( \sigma_p = 0.74 \) and for \( \rho = 1 \) gives \( \sigma_p = 1 \), the same value as the profile itself. The standard deviation of mean peak height for a general correlation is

\[
\sigma_p(\rho_1,\rho_2) = \left[ 1 + \frac{(1-\rho_1)(1-\rho_2)}{2\pi N} \frac{\left( 1 - \rho_2 \right)}{3 - 4\rho_1 + \rho_2} \right]^{1/2} \]

(3.31)

where \( N \) is defined for the general case in Equation 3.22.

Some idea of the effect of quantization can be obtained by using Sheppard’s correction for grouped data. Thus if \( \sigma_p(\rho) \) is the quantized value

\[
\sigma_p^2(\rho) = \sigma_p^2(\rho) + q^2/12. \]

(3.32)

With a typical worst-case value of \( q = 0.5 \), \( R_s \) reveals that an error of a few per cent in \( \sigma_p \) will result. Obviously peak height measurements can also be in error due to numerical modeling. For example, the three-point model is only an approximation, and is in effect a vertical parabola fitted through three ordinates. What is normally assumed is that the central ordinate is at the apex of the parabola. In fact this is only true if those ordinates adjacent to the central one are of equal height; if not, the error in peak height \( \delta z_p \) is given by

\[
\delta z_p = \frac{1}{8} \frac{(z_{i+1} - z_i)^2}{(2z_0 - z_{i+1} - z_i)}. \]

(3.33)

Taking typical values of \( z_0, z_{i+1} \), and \( z_i \) to be 3, 1, and −1, respectively, an error of about 3% results. Errors are generally much smaller for highly correlated data.

### 3.2.6.2 Peak Curvature

The curvature of peaks is especially sensitive to extraneous effects. This is mainly due to the fact that curvature is defined in terms of derivatives of the signal which are notoriously prone to noise. To get an estimate of the size of digital effects, consider the basic formula for curvature:

\[
C = \frac{d^2z/dx^2}{(1+(dz/dx)^2)^{3/2}}. \]

(3.34)

Usually, because in the vicinity of a peak \( dz/dx \rightarrow 0 \) the curvature can be closely approximated by the second differential, so the curvature \( C \) can be most simply expressed by the three-ordinate Lagrangian formula

\[
C = \frac{2z_0 - z_{i+1} - z_i}{h^2} = \frac{C'}{h^2}. \]

(3.35)

where \( C' \) is curvature measured in terms of ordinate differences only.

As before, the mean peak curvature can be expressed digitally as \( \bar{C}(\rho) \) where

\[
\bar{C}(\rho) = \frac{(3-\rho)(1-\rho)^{1/2}}{2\pi^{1/2}} \frac{1}{(1+\rho)^{1/2}} \]

(3.36)

To get \( \bar{C}(\rho) \), this is simply divided by \( h^2 \).

For the general case \( \bar{C}(\rho) \) is given by

\[
\bar{C}(\rho) = \frac{3-4\rho_1 + \rho_2}{2Nh^2} \frac{R_s^2}{\pi(1-\rho_1)^{1/2}}. \]

(3.37)

Comparing, for example, \( \bar{C}(\rho) \) at \( \rho = 0.1 \) and \( \rho = 0.9 \) shows a 10–1 difference, but in practice the difference in \( \bar{C}(\rho) \) will be larger because, not only are the differences between ordinates changing as shown in \( \bar{C}(\rho) \) but the actual value of \( h \) changes and it is this which has the dominant effect. Changes in curvature of 50–1 can be obtained on ordinary surfaces by changing the sampling rates!
A simple maximum probable estimate for the errors in \( C \) due to quantization is \( \delta C \):

\[
\delta C = \frac{4q}{h^2}
\]

(3.38)

where \( h \) is, as before, the ordinate spacing and \( q \) is the quantization interval. For typical values of \( q \) and \( h \) it would appear at first sight that the error is not likely to be serious. This is not necessarily so because near to a peak the differences between ordinates is small and can even be of the same order as \( q \), with the result that very large errors can occur. Increasing the ordinate spacing \( h \) in this case not only increases the ordinate differences but reduces \( \delta C \) in the above expression.

It is possible to get a limit for \( q \) in terms of ordinate differences above which useless estimates of curvature will be obtained. The mean value of ordinate differences is given by

\[
\frac{(3-\rho)(1-\rho)^{1/2}}{4\pi}
\]

(3.39)

Hence the upper limit for \( q \) could be taken as the inequality

\[
q < \frac{(3-\rho)(1-\rho)^{1/2}}{4\pi R} \cdot \frac{\sqrt{1-\rho}}{2\sqrt{\pi}} R.
\]

(3.40)

Knowing \( q \), \( \rho \) can be found (assuming the exponential model of a given exponent), from which the minimum ordinate spacing can be worked out. For example, for \( \rho = 0.7 \) it works out that \( q = 0.17R \), which implies that there should be about 35 quantization intervals across the range of the signal.

Numerical model errors are also important in measuring curvature. The three-ordinate model used earlier is merely the simplest second-differential formula. It is only strictly valid to use it when the fourth-order central differences are small when compared with it. A factor of one-tenth is typical, that is

\[
2\zeta_0 - (z_{s+1} + z_{s-1}) = 2\zeta_0 - (z_{s+1} + z_{s-1}) + \delta^2 z_0/12.
\]

(3.41)

This in turn is most likely to be true when the correlation between ordinates is high and the profile smoothly fits between discrete sampled ordinates.

For higher accuracy better Lagrangian formulae exist; for example, the seven-point model gives

\[
C = \frac{1}{180h^2} (2y_1 - 27y_2 + 270y_1 - 490y_0 + 270y_{-1} + 27y_{-2} + 2y_{-3})
\]

(3.42)

in which case the eighth central differences have to be small.

One way to assess the suitability of a curvature formula is to find out where it fails to be a good second differential. The way to do this is to consider the formula to be a digital filter by finding the Fourier transform of the numerical sequence and thence to find the break point where the gain does not increase as \( \omega^2 \). This is because a theoretically correct second differential should have a transfer function of \( \omega^2 \) for all \( \omega \).

Hence if \( C(x) \) is the operator on \( f(x) \) to get \( f''(x) \), that is

\[
f''(x) = f(x) * C(x),
\]

(3.43)

where * denotes convolution

\[
C(x) = 2\delta \delta (x-3h) - 27\delta \delta (x-2h) + 270...
\]

Thus, if \( A(\omega) \) is the transform of \( C(x) \):

\[
A(\omega) = \frac{1}{180h^2} [2 \exp(3j\omega h) - 27 \exp(2j\omega h) + 270 \exp(j\omega h) - 490 + 270 \exp(-j\omega h)

-27 \exp(-2j\omega h) + 2 \exp(-3j\omega h)]
\]

(3.44)

which results in

\[
A(\omega) = \frac{1}{180h^2} [4\cos 3\omega h - 54\cos 2\omega h + 540\cos \omega h - 490].
\]

(3.45)

If the frequency characteristic is plotted out it will be observed that it is only when \( \omega = 2h \) that it behaves as a true second differential. For higher values of \( \omega \) the differentiating property breaks down. For the three-point differential model, \( \omega \) has to be even shorter, so even though the three-point analysis is suitable for detecting peaks without the complication of quantization, it is severely restrictive when measuring curvature and should only be used when the shortest wavelengths present are four or five times the spacing interval \( h \).

### 3.2.6.3 Profile Slopes

Exactly the same arguments can be used when attempting to measure slopes: the first-differential formulae referred to in an earlier section still hold. In this case, however, the breakdown of the formulae is that point where the frequency characteristic fails to increase as \( \omega \) and not \( \omega^2 \). A similar conclusion to that for \( C \) is reached [3].

In terms of the autocorrelation function and discrete ordinates, the mean absolute slope is given by

\[
\bar{m} = \frac{R_0}{h} \sqrt{1-\rho} \quad \text{or} \quad \Delta q = \frac{R_0}{h \sqrt{1-\rho}}
\]

(3.46)

\[
\text{from} \quad E \left[ \frac{(z_s - z_{s+1})^2}{(2h)^2} \right] \quad \bar{m} = \frac{R_0}{h \sqrt{1-\rho \pi}}
\]

for a general correlation function. As before for curvature of peaks

\[
f''(x) = M(x) * f(x).
\]

(3.47)

where \( M \) is impulse train operator not the slope \( m \) corresponding to \( C \) for curvature in Equation 3.43.
Hence
\[ M(x) = \frac{1}{12h} \left[ \delta(z(x-2h)) - 8\delta(z(x-h)) + 8\delta(z(x+h)) - \delta(z(x+2h)) \right] \] (3.48)

for the five-point model from which the spectrum \( A_m(\omega) \) is given by
\[ A_m_1(\omega) = \frac{j}{6h} [8\sin(\omega h) - \sin(2\omega h)]. \] (3.49)

Expanding this as a power series in \( \sin \omega h \):
\[ A_m_1(\omega) = \frac{j}{h} \left( \omega h - \frac{(\omega h)^3}{40} - \frac{(\omega h)^5}{252} + K \right) \] (3.50)

from which it is obvious that only the first term represents a true differentiation being proportional to \( \omega \); the others are errors due to the limitation of the five-point method.

The actual error at a specific frequency can be found by using Equation 3.49. Thus, as an example, consider a sine wave sampled at four points per cycle. It will have a wavelength of \( 4h \) and \( \omega = 2\pi 4h \), the true value is \( A_m(\omega) \):
\[ i.e., \quad A_m(\omega) = j\omega = j\frac{\pi}{2h} \] (3.51)

where as
\[ A_m(\omega) = \frac{j}{6h} \left( 8\sin \left( \frac{\pi}{2} - \sin \pi \right) \right) = j \frac{4}{3h} \] (3.52)

which is 15% down on the true value.

Note that the formula for the average values of slope curvature and other features in terms of the spacing and the correlation coefficients of the single, double, and in fact multiple spacings between ordinates can be extended.

There is a general pattern that can be used to get some idea of the applicability of surface parameters in terms of digital characteristics. This is shown in simple form in Figure 3.16.

In the figure, region \( A \) gives the wrong results because the ordinates are too highly correlated and the small differences between them (perhaps due to the finite tip of the stylus or the limited resolution of the optical device) do not show at all, because the quantization interval \( \Delta z \) does not see the differences, so region \( A \) reflects instrument resolution quantization. Region \( B \) tends to produce realistic values where the correlation is about 0.7. In this region the instrument limitations are minimal and the reasonably fine structure of the surface as well as the gross structure is picked up. In region \( C \) the fine structure is lost because \( \rho \approx 0 \), aliasing can be a problem and only the gross structure is picked up. Loss of information and misleading information are very possible. For region \( D \) the answer is completely dominated by the numerical model—the data is completely uncorrelated. The probability of an ordinate being a peak is 1/3 if the three-ordinate model is used and 1/5 if the five-ordinate model is used, so region \( D \) is numerical model limited.

The graph shown in Figure 3.16 demonstrates the limited range in which some sort of result can be obtained which is not corrupted by either the instrumentation or the digital processing. The question arises, therefore, whether the data mean anything anyway, because the answer appears to be dependent on the sampling even if no quantization effects are present! There are two issues: one is whether a digital technique would ever agree with an analog one; and the other is whether surfaces are well behaved enough in the sense of differentiability to allow a unique answer for a feature which did not continually change as the sample interval and hence “scale of size” changed. The first point will be addressed first, not because it is simpler but because it can be attempted. Even then there are two aspects to it: the profile problem and the “areal” problem. Consider the profile first and using conventional terminology.

Take for example the case of the mean number of peaks \( m_0 \) in unit distance and the mean number of crossings \( n_0 \) in unit distance in terms of \( D_s \) and \( D_z \) where
\[ D_i = \frac{d' A(h)}{dh'} \] (3.53)

These are as revealed in Chapter 2 and Section 3.3.2
\[ m_0 = \frac{1}{2\pi} \left( \frac{D_z}{-D_s} \right)^{1/2}, \quad n_0 = \frac{1}{\pi} \left( \frac{D_z}{-D_s} \right)^{1/2}. \] (3.54)

The results here can be compared with those given in digital form in Equation 3.55 for \( m_0 \) and \( n_0 \):
\[ \frac{1}{\pi} \tan^{-1} \left( \frac{3 - 4\rho_1 + \rho_2}{1 - \rho_2} \right)^{1/2} \quad \text{and} \quad \frac{1}{\pi} \cos^{-1} \rho_1 \] (3.55)

by first expressing the autocorrelation function as a Taylor expansion and then investigating the behavior as the sampling interval \( h \) is made to approach zero when \( \rho_i = \rho(h) \) and \( \rho_2 = \rho(2h) \). Thus
\[ \rho(h) = 1 - \left( -D_s \frac{h^2}{2!} + D_z \frac{h^4}{4!} + O(h^6) \right). \] (3.56)
Using this expansion and inserting the values for \( \rho_{1} \) and \( \rho_{2} \) in the equations for peak distributions and other features described above, it can be shown [4] that they become the results obtained by Rice in 1944 [5] and by Bendat [6] thereby satisfactorily allowing the conclusion that the digital method does indeed converge onto that which would have been obtained from a continuous surface with perfect instrumentation. Unfortunately this happy state of affairs is subject to two massive assumptions: that Equation 3.56 is allowable (i.e., \( D_{2} \) and \( D_{4} \) exist) and that the results obtained digitally from a profile are acceptable simplifications of the areal digital contour of a surface.

Assuming that these two conditions are acceptable—a point which will be raised later—it is possible to use the limiting case of the discrete analysis to express any parameters hitherto not obtained and yet which are of considerable importance in the continuous theory of tribology (which is in fact what counts in the macrophysical world).

Thus the joint probability density function (PDF) of the peak height and curvature is given as the probability density of an ordinate being a peak of height \( z_{0} \) and curvature \( C \) shown as \( \rho(C, z_{0}/\text{peak}) \) where

\[
\rho(C, z_{0}/\text{peak}) = \frac{C}{[2\pi D_{4}(D_{4} - D_{2}^2)]^{1/2}} \exp\left[\frac{- (D_{4} z_{0}^2 + 2D_{2} z_{0} C + C^2)}{2(D_{4} - D_{2}^2)}\right] \quad \text{for } C > 0. \quad (3.57)
\]

Also, the correlation coefficient between peak height and curvature is

\[
\text{corr}(C, z_{0}/\text{peak}) = \left(-\frac{D_{2}}{D_{4}}\right)^{1/2} \left(\frac{2 - \pi/2}{1+(1-\pi/2)D_{2}^2/D_{4}}\right)^{1/2} \quad (3.58)
\]

among others, which demonstrates that useful results can be obtained by moving between the discrete and continuous cases for surface metrology.

### 3.2.7 Summary of Profile Digital Analysis Problems

What has been indicated is how many of the most useful surface parameters of surface roughness vary in the general sense as a function of the digital (discrete) parameters. Hopefully, this gives an insight into the possible relationships that might be obtained on real surfaces. However, obviously the results in the foregoing section pose the question of whether the correlation function is measured and from this whether the measured values of the various tribological parameters of interest are predicted or should the actual parameters be measured directly from the surface using a data logging system and a computer? The former is simple and less variable because the expressions given in terms of the correlation coefficients are concerned with the average behavior of surfaces, assuming them to be nominally Gaussian in statistics. From the formulæ given, which illustrate the way in which the parameters can be estimated from the correlation coefficients, it has to be said that the specific behavior of different surfaces will depend on the actual type of correlation function. For this reason, to estimate the relationships between parameters it is necessary to assume a correlation function which approximates to the surface in question. This is a mathematical modeling problem of surfaces.

However, as a first guess the exponential correlation function is usually sufficiently adequate for values away from the origin that is \( \rho = 0.7 \). It also fits in with fractal-type surfaces. For correlation values larger than 0.7, a Gaussian correlation could be used. It is considered that in many applications of surfaces the function depends on two surfaces and not one: it is the properties of the gap between them that are important. Thus one advantage of specifying surfaces in terms of the correlation coefficients is that, at least to a first approximation, the gap properties (neglecting actual separation) can be estimated from the additive properties of the mating surfaces.

Thus if \( \sigma_{A}^{2} \) is the variance (\( R_{\text{p}}^{2} \)) value of surface \( A \) and \( \rho_{A}(h) + \rho_{A}(2h) \) are its coefficients, and similarly for surface \( B \), then

\[
\sigma_{B}^{2} = \sigma_{A}^{2} + \sigma_{B}^{2} \\
\rho_{B}(h) = [\sigma_{B}^{2} \rho_{A}(h) + \sigma_{B}^{2} \rho_{A}(2h)]/\sigma_{B}^{2} \quad (3.59)
\]

\[
\rho_{B}(2h) = [\sigma_{B}^{2} \rho_{A}(2h) + \sigma_{B}^{2} \rho_{A}(2h)]/\sigma_{B}^{2}.
\]

Use of the density of peaks and zero crossings instead of correlation coefficients is not attractive because they are not additive. See Chapter 7 for a full examination of these issues.

Using the nomenclature \( n_{q} \) and \( m_{q} \) as before, the comparison with Equation 3.59 becomes

\[
\sigma_{B}^{2} = \sigma_{A}^{2} + \sigma_{B}^{2} \\
n_{qg} = [(\sigma_{A}^{2} n_{qA}^{2} + \sigma_{B}^{2} n_{qB}^{2})/\sigma_{q}^{2}]^{1/2} \quad (3.60)
\]

\[
m_{qg} = \frac{\sigma_{A}^{2} n_{qA}^{2} m_{qA} + \sigma_{B}^{2} n_{qB}^{2} m_{qB}}{\sigma_{q}^{2} n_{qA} + \sigma_{B}^{2} n_{qB}}
\]

Any errors tend to be cumulative. (More will be said about this aspect of characterization in Chapter 7.)

The latter method of obtaining tribological data (i.e., the formal data logging method) gives better estimates of extreme properties in the sense that the odd, very high peak or valley may be picked up, but it does tend to be more time consuming to gather the necessary data. Obviously in the former method the correlation coefficients have to be determined. Another way to be discussed is concerned with estimation by other than stylus means.

### 3.2.8 Areal (3D) Filtering and Parameters

The same basic techniques can be used for areal (3D) filtering as were used in profile filtering. However, areal manipulation
offers opportunities for a better understanding of the manufacturing process and the function of the workpiece than was possible with profile information.

The starting point is the surface data \( f(x, y) \). This has a frequency (wavelength) content given by its Fourier transform

\[
F(\omega, v) = \int \int f(x, y) \exp(-j(\omega x + vy)) \, dx \, dy. \tag{3.61}
\]

\( F(w, v) \) can be modified by a filter function, say, \( H(w, v) \) so that the filtered areal data is now \( F'(w, v) \) where

\[
F'(\omega, v) = \int \int F(\omega, v) H(\omega, v) \, dx \, dy \tag{3.62}
\]

this can be inversely transformed to give the filtered areal data \( f'(x, y) \):

\[
f'(x, y) = \frac{1}{4\pi^2} \int \int F'(\omega, v) \exp(+j(\omega x + vy)) \, d\omega \, dv. \tag{3.63}
\]

The data need not be analyzed using the Fourier transform. Dong [7] has been using wavelet transforms to examine areal (3D) data.

The way to calculate the areal spectrum is by carrying out a spectrum over all the profile \( y \) values of which there may be \( N \) and finding the average and then carrying out an equivalent operation in the orthogonal direction and averaging [8]. The fast Fourier transform (FFT) is used for these calculations. The resultant output is a 2D FFT.

The same arguments apply to areal filtering as to profile filtering, namely that a conventional filter does not need to have knowledge of the surface topography in order to be applied effectively. The penalty for this freedom is losing data at the boundaries because of the finite size of the impulse response of the filter (Figure 3.17).

It is possible to play some tricks with the data in order to salvage some of the “lost” area. One is simply to fold \( A \) to \( A' \) and repeat the data (Figure 3.18).

The shaded area and the repeat areas \( A'B'C'D' \) \( A''B''C''D'' \) can now be utilized.

Unfortunately, although there is no discontinuity between the levels at \( A \) \( A' \) and \( B \) \( B'' \) and \( A \) \( A'' \) \( B \) \( B''' \), etc., the differentials at the boundaries will be discontinuous so that slope and curvature data will be corrupt. The first differentials could also be made to fit on a first order surface but higher differentials cannot.

Polynomials can be effectively fitted to the areal data providing that the longer wavelengths can be matched to the order of the polynomial which presupposes prior knowledge. Failure to do this will produce spurious results.

There has been some interest in using spline functions e.g., Ref. [9]. It is instructive to carry out the spline fit for a profile and the extend it to the area fit. Following Radhakrishnan [9] and using his nomenclature the following method can be adopted.

The equation of the spline can be written as

\[
P(x_i) = \sum_{i=1}^{n} N_k(x_i) B_i. \tag{3.64}
\]

\( N_k \) are the \( B \) spline basis functions given by

\[
N_k = \begin{cases} 1 & x_i \leq x \leq x_{i+1} \\ 0 & \text{otherwise} \end{cases}
\]

\[
N_k(x) = \frac{x - x_i}{x_{i+1} - x_i} N_{k-1}(x) + \frac{x_{i+1} - x}{x_{i+1} - x_i} N_{k+1}(x), \tag{3.65}
\]

where \( x_i \) is the “i”th knot value, \( B_i \) is the “i”th control value = values at knot, \( N \) is the number of control values, and \( k \) is order of the curve.

If \( x_i, y_i \) are data points

\[
\frac{\partial}{\partial B_i} \sum (p(x_i) - y_i)^2 = 0.
\]

The canonical equations obtained match the number of unknowns so that the \( B \)s can be found.

When simplified this becomes

\[
\sum_i N_k(x_i) \sum_{j=1}^{n} B_i N_{k}(x_j) = \sum_i y_i N_k(x_i). \tag{3.66}
\]

As \( i \) is stepped to \( n \), \( n \) equations result from which the \( n \) control values of the mean curve are obtained.
For two dimensions, Equation 3.66 can be extended to give Equation 3.67 [9, 59].

\[ P(x, y) \sum_{i} [P(x, y, Z_i)]^2 = 0 \quad (3.67) \]

\[ = \sum_{i} \sum_{j} z_{i,j} N_{i,j}(x, y) M_{i,j}(x, y) y(i). \quad (3.68) \]

Varying \( i \) from 1 to \( n \) and \( j \) from 1 to \( M \) gives a set of \( m \times n \) equations which, when solved, give the control values of the mean surface. Parameters from this mean surface are extensions of the one dimension profile case.

So,

\[ S_{x} = \frac{1}{A} \int \int \partial \varepsilon \partial \varepsilon dx dy \quad (3.69) \]

\[ = \frac{1}{2} \sum_{m} \sum_{n} \left| z(i, j) \right| \quad (3.70) \]

Other parameters can also be obtained similarly. It should be remembered that the true spline is a cubic equation and is meant to be a line of minimum potential energy. It was originally used in ship building. An elastic or bendy piece of wood was pressed against a set of \( n \) stanchions. The resulting curve made contact at \( n \) points. The curve in between restraints could be measured and drawn. The shape of the strip of wood is naturally such that it has a minimum potential energy. Mathematical splines are derived from this.

### 3.2.9 Digital Areal (3D) Measurement of Surface Roughness Parameters

#### 3.2.9.1 General

Note here that areal is taken to be the 3D picture having two independent variables, \( x \) and \( y \), the term 3D is sometimes used instead of areal but is not formally correct and is not recommended.

In the earlier sections on the characterization of surface roughness it became clear that a major objective is to measure the areal dimensional characteristics of the surface rather than to restrict the discussion to single profiles. This expansion of subject immediately brings into focus such things as specification and measurement of lay, the relationship between the moments of the spectrum of the area versus those measured from one or more profiles. Exactly the same situation exists in the digital domain. It has already been highlighted that there are differences between the theoretical values of features for profile and areal measurements according to Longuett-Higgins [10] and Nayak [11]. However, the questions emerge as to the differences which might arise between profile discrete measurement on the one hand and the convergence of areal discrete estimates of surfaces to the theoretical values on the other.

Exactly the same sort of calculation can be carried out as before, yielding some very interesting and important results.

Two attempts have been made to analyze the areal discrete problem for a random surface, the first by Whitehouse and Phillips [4] and the second by Greenwood [12]. The former results will be given here, yet the more digestible results of the latter will also be included for comparison. Both approaches can be considered to be a typology of surfaces in their own right (see Chapter 2).

So far profile information has been examined. It is understood that such information is only a convenient approximation to the real surface geometry. Exploring the areal properties on a continuous basis in the way that Longuett-Higgins and Nayak have is most informative but does not actually relate directly to what is measured. All measurements are now carried out digitally. It will be observed that areal discrete measurements cannot converge to the values for a continuous waveform because of the discontinuous spatial bandwidth of any sampling pattern. In certain circumstances, for example isotropic surfaces or when the two perpendicular components are independent, measuring the spectra either in one direction in isotropy or in two directions for independent lay. These enable Nyquist criteria to be tested if the pass conversion is possible. It is the purpose of this next section to investigate the nature of the differences that can and do occur in parameters which are functionally important, such as summit curvatures and height distributions.

Some of the theory used in investigating discrete properties overlaps the theory used for surface characterization using the discrete parameters reported in Chapter 2. The analytical results are necessarily complicated but are absolutely essential if the results for surfaces obtained by experimenters are ever to be made consistent with each other.

Also, whereas the idea of tribology parameter prediction using the correlation function for a profile might not seem very attractive, it most certainly is for areal estimation because of the very great difficulty in measuring over an area and maintaining geometric fidelity between traces. Simply taking disjointed profile tracks is not good enough—each trace has to be integrated into the overall mapping scheme in height and starting point. Another point which adds to the complexity is that, in areal measurement, there is not necessarily a straightforward grid system for data ordinates to be used. Before discussing the alternative methods, the analysis for the areal equivalent of the three-point method will be given.

In one dimension a continuous definition of a local maximum or peak on a profile only requires one first-order and one second-order differential, and only three ordinates are needed for the discrete definition of a peak. However, in two dimensions, the continuous definition of a local maximum (or summit, in the terminology of Nayak) requires two first-order and three second-order differentials, and a minimum of five ordinates are usually considered necessary for the discrete definition of a summit. Sayles and Thomas [13] gave two discrete definitions of a summit, one using the five nearest-neighbor ordinates, and the other nine. Here a number of possibilities will be considered; to start with, the conventional five-ordinate definition will be used. The definition states that an ordinate at height \( z_0 \) above an arbitrary datum is a five-point summit if it is higher...
than the four neighboring ordinates, which are at a distance $h$ from it, on the two orthogonal Cartesian axes. If the differences between $z_0$ and each of the four neighboring ordinates are denoted by $s_1, s_2, s_3$, and $s_4$, then the condition for the ordinate $z_0$ to be a summit is that $s_1, s_2, s_3,$ and $s_4$ are all positive.

The summit density is the proportion of ordinates that are summits. The summit height is the height $z_0$ of an ordinate that is a summit.

It seems natural to extend to two independent dimensions the discrete definition of Whitehouse and Archard [14] for peak curvatures, which was given by

$$C_h = \frac{(s_1 + s_2)}{h}$$

(3.71)

where $s_1$ and $s_2$ are defined later. This extension uses the average orthogonal peak curvatures, and gives the discrete definition of five-point summit curvature as

$$C_{h^2} = \frac{1}{2}(s_1 + s_2 + s_3 + s_4)/h^2$$

(3.72)

which again is a linear combination of the ordinates.

### 3.2.9.2 The Expected Summit Density and the Distributions of Summit Height and Curvature

The distributions of peak height and curvature for a profile in one dimension were obtained by Whitehouse and Phillips [4] for a surface with ordinates from a Gaussian (normal) distribution. (A multivariate normal distribution (MND) for a vector $Z$ will be denoted by $Z \sim N(\mu, V)$) They obtained these results from the theory of truncated random variables. This was because the peak height distribution is the conditional distribution of $z_0$ the profile height, given that $s_1$ and $s_2$ are positive, or, in other words, the distribution of $z_0$ conditional on $s_1$ and $s_2$ being positive, where

$$s_1 = z_0 - z_{-1} \quad s_2 = z_0 - z_1$$

(3.73)

and remembering that $z_{-1}$ and $z_1$ are the preceding to $z_0$ and succeeding ordinate values, respectively, on the profile. Similarly the distribution of peak curvature is the conditional distribution of $C_h$ given that $s_1$ and $s_2$ are positive. Hence the results of Whitehouse and Phillips [4] can be obtained by using the results for truncated random variables, with $m = 2$.

$$Y_0 = Z_0, X = (2 - 2\rho_1)^{-1/2}(s_1, s_2)^T \quad d = (1/2 - \rho_1)^{1/2}$$

and $V$ obtained from $V_\delta$ by removing the first row and column (and denoted by $V_\delta$) so that

$$\lambda = 1 + a + 2b.$$ 

(3.80)

Then the PDF of the five-point summit height distribution is given by

$$p(z_0/Y_\delta) = \phi(\lambda)\left(z_0\left([1-\rho_1]/(1+\rho_1)\right)^{1/2}; V_\delta\right) \phi(z_0)/\phi(\lambda)(0, V_\delta)$$

(3.81)

where $V_\delta$ is given by

$$V_\delta = \begin{bmatrix}
1 & a_{c} & b_{c} & b_{c} \\
 a_{c} & 1 & b_{c} & b_{c} \\
b_{c} & b_{c} & 1 & a_{c} \\
b_{c} & b_{c} & a_{c} & 1
\end{bmatrix}$$

(3.82)

and

$$a_{c} = (\rho_2 - \rho_2^1)/(1 - \rho_2^1) \quad b_{c} = (\rho_3 - \rho_3^1)/(1 - \rho_3^1).$$

(3.83)
The expected (average or mean) five-point summit height is given by \( E[z_0 \mid Y_3] \), where \( E \) denotes the statistical expectation:

\[
E[z_0 \mid Y_3] = \frac{2(1 - \rho/\pi)^{\gamma/2} \Phi^{(3)}(0; B_4)}{\Phi^{(4)}(0; Y_3)},
\]

(3.84)

where

\[
B_4 = \begin{pmatrix}
1 - b^2 & a - b^2 & b(1 - a) \\
(a - b^2) & 1 - b^2 & b(1 - a) \\
(b(1 - a)) & (b(1 - a)) & 1 - a^2
\end{pmatrix}.
\]

(3.85)

From Equation 3.82 it can be seen that the expected five-point summit height depends on two orthant probabilities \( \Phi^{(3)}(0; B_4) \) and \( \Phi^{(4)}(0; Y_3) \), which have to be evaluated. From Whitehouse and Archard [14], \( \Phi^{(3)}(0; B_4) \) is given by

\[
\Phi^{(3)}(0; B_4) = \frac{1}{2} - (4\pi)^{-1} \cos^{-1}[(a - b^2)/(1 - b^2)]

- (2\pi)^{-1} \cos^{-1}[(b(1 - a)(1 - b^2))^{1/2}].
\]

(3.86)

Cheng [15] has evaluated \( \Phi^{(i)}(0; Y_3) \), so using this result the expected five-point summit height can be found. It is only necessary to evaluate Plackett’s [16] method. As this orthant probability only depends on the parameters \( a \) and \( b \), it will be denoted by \( \Phi^{(i)}[a, b] \). Then

\[
pr(Y_3) = \Phi^{(i)}[a, b] = \Phi^{(i)}[a, 0] + \int_0^b \frac{\partial \Phi^{(i)}[a, t]}{\partial t} \, dt
\]

\[
= \left[ \frac{1}{4} + (2\pi)^{-1} \sin^{-1}a \right]^2 + (2\pi)^{-1} \sin^{-1}b

+ \pi^{-2} \int_0^b (1-t^2)^{-1/2} \sin^{-1} \left[ t(1-a)(1+a-2t)^{-1} \right] \, dt.
\]

(3.87)

This result was given by Cheng [15] in terms of the dilogarithm function.

The distribution of a five-point summit having a height \( z_0 \) conditional on a curvature \( C^{(2)}_h \) is normal and is given by

\[
(Z_0 \mid C^{(2)}_h) \sim N[h^2(1+a+2b)^{-1}c^{(2)}_h, 1-2(1-\rho)/(1+a+2b)].
\]

(3.88)

This is henceforth called the conditional distribution of summit height given curvature. Thus the expected five-point summit curvature is given by

\[
E[C^{(2)}_h \mid Y_3] = h^{-2}(1+a+2b)E[Z_0 \mid Y_3].
\]

(3.89)

Hence, by the application of the theory of Gaussian (normal) truncated random variables, it has been possible to obtain the expectations of the five-point summit height, curvature, and density in terms of the correlation coefficients. These are the basic tribological parameters required. The results have been arrived at by using a five-point model for summits (called the tetragonal model) and this implies a rectangular grid of sampled data. These results therefore can be regarded as the tribological parameters of a discrete random surface with values at the intersections of a rectangular grid. As long as the correlation is \( \rho_1 \) and \( \rho_2 \) at \( h \) and \( 2h \) in all four directions or scaled accordingly the results are exact. A wide variety of surfaces can be modeled by allowing \( \rho_1 \) and \( \rho_2 \) to vary.

### 3.2.9.3 The Effect of the Sample Interval and Limiting Results

The distributions of five-point summit height and curvature have been derived in terms of the correlation coefficients between ordinates. These correlation coefficients are \( \rho_1 \) for ordinates a distance \( h \) apart, \( \rho_2 \) for ordinates \( 2h \) apart, and \( \rho_3 \) for ordinates \( \sqrt{2}h \) apart. If the surface is isotropic then

\[
\rho_1 = \rho(h) \quad \rho_2 = \rho(2h) \quad \text{and} \quad \rho_3 = \rho(h\sqrt{2}).
\]

(3.90)

So \( \rho_1, \rho_2, \) and \( \rho_3 \) will vary as \( h \) varies, depending on the shape of the autocorrelation function of the surface.

As \( h \) approaches zero

\[
\lim_{h \to 0} \rho_1(h) = \lim_{h \to 0} \rho_2(h) = \lim_{h \to 0} \rho_3(h) = 1
\]

(3.91)

and as \( h \) approaches infinity

\[
\lim_{h \to \infty} \rho_1(h) = \lim_{h \to \infty} \rho_2(h) = \lim_{h \to \infty} \rho_3(h) = 0.
\]

(3.92)

If \( \rho_1, \rho_2, \) and \( \rho_3 \) are plotted in three dimensions then, as \( h \) varies, the curve will start at (1,1,1) for \( h = 0 \) and end at (0,0,0) for \( h = +\infty \). In order that the matrix \( V_1 = dJ \) is positive definite it is necessary for this curve to lie in the region bounded by \( \rho_2 < 1 \) and \( 1/(1+\rho_2) + 2\rho_1^2 < \rho_3 < 1/(1+\rho_2) \).

Results for the summit height have been obtained by Nayak [11] for the continuous surface, so it is possible to compare his results with those obtained for the discrete results of this chapter as the sampling interval \( h \) approaches zero. The expected summit height depends on \( \rho_1, \rho_2, \) and \( \rho_3 \) through \( a \) and \( b \), and

\[
\lim_{h \to 0} a = -1 \quad \text{and} \quad \lim_{h \to 0} b = 0.
\]

The autocorrelation function of the surface can be approximated by Taylor’s expansion

\[
\rho(h) = 1 + D_2 h^2/2! + D_4 h^4/4! + O(h^6),
\]

where \( D_2 \) and \( D_4 \) are the second and fourth derivatives of the autocorrelation function at the origin and if

\[
\eta = -D_2 (D_4)^{1/2} < \sqrt{2}.
\]

(3.93)
The limiting value for the expected five-point summit height is given by

$$\lim E[Z_0 | Y_k] = \frac{16}{\pi + 2 \sin^{-1}(1/3) + 4\sqrt{2}} \left(\frac{1}{2} \pi\right)^{1/2} \eta \tag{3.94}$$

$$= 1.688 \left(\frac{1}{2} \pi\right)^{1/2} \eta.$$ \tag{3.95}

Nayak [11] showed that the expected continuous summit height for the areal case was

$$E[z | Y_k] = (4\sqrt{2}\pi) \left(\frac{1}{2} \pi\right)^{1/2} \eta = 1.801 \left(\frac{1}{2} \pi\right)^{1/2} \eta, \tag{3.96}$$

which is comparable with the expected continuous peak height for the profile

$$E[z | Y_k] = \left(\frac{1}{2} \pi\right)^{1/2} \eta \tag{3.97}$$

a result given by Whitehouse and Phillips [4]. Then it can be seen that the limit of the expected five-point summit height (Equation 3.94) is 69% larger than the expected peak limit (Equation 3.96) as opposed to 80% larger than the expectation of the distribution of summit height for the continuous definition of Nayak [11] (Equation 3.94). However, this is only an overall reduction of about 6% and suggests that the discrete five-point definition may be adequate. Compare this with the seven-point result, Equations 3.128 and 3.132 below.

It is possible to obtain $\Phi^{(2)}(z_0/[1-p_0]/[1 + p_0])^{1/2}$, $V$ by the methods of [15] and [16], and hence to obtain the PDF of the limiting distribution of the five-point summit height as $h$ converges to zero. This is given by

$$\lim_{h \to 0} p(z_0 | Y_k) = 12 \pi \varphi(z_0) \left(\pi + 2 \sin^{-1} \frac{1}{3} + 4\sqrt{2}\right)^{-1}$$

$$\times \{1 - \eta^2\} \varphi(w) + w \Phi \varphi(w)}(w - \frac{1}{3} - \eta^2)(1 - \eta^2) - (1 - \eta^2) \Phi^{(2)}(w; 0) + 4(1 - \eta^2) T^{(2)}(w, 2 - 3 \eta^2 - 1/2), \tag{3.97}$$

where

$$T^{(2)}(w, u) = \frac{1}{2\pi} \int_0^u (1 + x^2)^{-2} \exp\left(-\frac{w^2}{2} - (1 + x^2)^2\right) dx \tag{3.98}$$

and

$$w = \eta(1 - \eta^2)^{-1/2} z_0. \tag{3.99}$$

This PDF is compared with the PDF of continuous summit height, given by

$$p(z | Y_k) = 3 \eta(2 - 3 \eta^2)^{1/2} \varphi(0) \varphi\left(1 - \frac{3}{2} \eta^2\right)^{-1/2} z$$

$$+ 3\sqrt{3} \eta^2 (z^2 - 1) \Phi \left[\frac{3}{2} \left(1 - \frac{3}{2} \eta^2\right)^{1/2} z\right] \Phi(z)$$

$$+ 2(1 - \eta^2)^{-1/2} \left(\frac{1}{2} \left[(1 - \eta^2)(1 - \frac{3}{2} \eta^2)\right]^{1/2} \eta\right)$$

$$\times \varphi((1 - \eta^2)^{-1/2} z) \tag{3.100}$$

which was obtained by Nayak [11] for three values $\eta = 0$, $\sqrt{1/3}$, and $\sqrt{2/3}$, as shown in Figure 3.19. For $\eta = 0$ both distributions are the standardized normal distribution. When $\eta = \sqrt{2/3}$, the PDF of the continuous summit height is

$$p(z | Y_k) = 2 \sqrt{3} (z^2 - 1 + 2\pi [\varphi(z)]^2) \varphi(z) \quad \text{for } z > 0$$

$$= 0 \quad \text{for } z \leq 0 \tag{3.101}$$

while the limiting distribution of the five-point summit height is given by

$$\lim_{h \to 0} p(z | Y_k) = \frac{8\pi \varphi(z_0) \left[(2z_0^2 - 1) \Phi(\sqrt{2}z_0) - \frac{1}{2}\right]}{\pi + 2 \sin^{-1} \frac{1}{3} + 4\sqrt{2}} \quad \text{for } z_0 > 0$$

$$= 0 \quad \text{for } z_0 \leq 0 \tag{3.102}$$

**FIGURE 3.19** The probability density function (PDF) of the distribution of summit height (full line) and the limiting distribution of the five-point summit height (broken line) for $\eta = 0$, square root 1/3, and square root 2/3.
Nayak [11] used the definition of mean summit curvature $K_m$ given by Sokolnikoff as minus the average of the second partial derivatives in orthogonal directions.

$$K_m = -\frac{1}{2} \left[ \frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} \right]$$  \hspace{1cm} (3.103)

With this definition the conditional distribution of continuous summit height given the curvature is a normal distribution with

$$E[z/K_m, Y_N] = \frac{3}{2} \eta K_m (D_s)^{-1/2}$$  \hspace{1cm} (3.104)

and

$$\text{var}(z/K_m, Y_N) = 1 - \frac{3}{2} \eta^2.$$  \hspace{1cm} (3.105)

This is also the limit of the conditional distribution of $z_0$ given $C_{0h}$. Hence, the limit as $h$ tends to zero of the expected five-point summit curvature will be 6% smaller than the expectation of the continuous summit curvature. Thus the operational procedure of sampling in a plane and taking the sample interval to zero gives different results from the continuous values for the surface!

Greenwood [12] approached the problem of profile measurement in the same basic way using the multinormal distribution from the joint distribution of height $z$, slope $m$, and curvature $k$:

$$p(z, m, k) = \frac{1}{2\pi^{3/2}\sqrt{1-r^2}} \times \exp\left(-\frac{1}{2(1-r^2)}(\xi^2 + r^2 + 2\eta^2)\right) \exp\left(-\frac{r^2}{2}\right),$$  \hspace{1cm} (3.106)

where $\xi = z/\sigma_z$, $s = m/\sigma_m$, and $t = -k/\sigma_k$, and $\xi$, $s$, and $t$ are the standardized height, slope, and curvature; $z$, $m$, and $k$ normalized to their standard deviations $\sigma_z$, $\sigma_m$, and $\sigma_k$, respectively. In Equation 3.103 $r$ is $\sigma_z^2/\sigma_m \sigma_k$ and corresponds to the variable $\alpha = m_r m / \sigma_m^2$ as $t^2$.

Greenwood introduces another variable, $\theta$ to replace in part the sampling interval defined by

$$\sin\theta = h \sigma_s / 2 \sigma_m$$  \hspace{1cm} (3.107)

where $\sigma_m$ is the standard deviation of the slope and $\sigma_s$ that of curvature. If $h$ is the sampling interval, this transcription from $h$ and the correlation coefficients $\rho(0)$, $\rho(h)$, and $\rho(2h)$ preferred by Whitehouse and Phillips for a profile makes some of the subsequent equations simpler in form. However, the concept of sampling interval is masked, which is a disadvantage to the investigator.

Formulae for the joint distributions of height and curvature are found, as are those for summit curvature distributions (Figure 3.20).

The reader is requested to consult the paper by Greenwood [12] to decide the simplest approach to the areal problem. However, it is gratifying to note that the results and predictions of both methods are compatible.

### 3.2.10 Patterns of Sampling and Their Effect on Discrete Properties (Comparison of Three-, Four-, Five- and Seven-Point Analysis of Surfaces)

An investigation of the various sampling pattern options open to a researcher will be given here to see if there is any advantage to be had by changing from the conventional rectilinear sampling pattern. The real issue is to find what sampling pattern best covers the area and picks out summits. Obviously, there are instrumental factors that have to be taken into account as well as purely theoretical ones. One such factor is the ease with which a complex sampling pattern can be achieved by hardware comprising a specimen table driven by two orthogonal, motorized slides.

One thing that emerges from these sampling procedures is that the properties so revealed are substantially different from each other.

#### 3.2.10.1 Four-Point Sampling Scheme in a Plane

A comparison of possible sampling schemes for sampling in an areal plane will now be presented. They are illustrated in Figures 3.21 and 4.73. In all cases the distance between measurements is $h$.

First, there is sampling along a straight line (from a profile of the surface). This sampling scheme only takes measurements in one dimension of the plane. This has been presented for completeness and because it was the first case considered by Whitehouse and Archard [9] and Whitehouse and Phillips [4]. In Figure 3.21a it is illustrated with $k = 2$.

Second, a sampling scheme could be used that would take measurements at the vertices of a hexagonal grid. The summit properties could be defined using four ordinates, that is the measurement at a vertex and the three adjacent ordinates...
at a distance \( h \) from the chosen vertex. This is the case when \( k = 3 \) and is referred to as the trigonal symmetry case.

In order to produce such a hexagonal grid pattern on a surface, it would be necessary to sample along parallel lines separated alternately by a distance of \( \sqrt{3}h \) and \( h \). The spacing between ordinates along a line would be \( \sqrt{3}h \) but the position at which the first ordinate is measured would be 0 for the \((4j-3)\)rd and \(4j\)th lines and \( \sin \theta = h \sigma_z / 2 \sigma_h \sqrt{3}h \) for the \((4j-2)\)nd and \((4j-1)\)st lines for \( j > 1 \). This is illustrated in Figure 3.21b. Alternatively, it would be possible to sample along parallel lines a distance of \( 1/2 \). \( \sqrt{3}h \) apart but this would involve a different position for the first ordinate and the spacing between ordinates would alternatively be \( h \) and \( 2h \).

Third, there is sampling on a square grid. This was considered by Whitehouse and Phillips [4] and Greenwood [5] and will be referred to as the tetragonal symmetry case. It is illustrated in Figure 3.21c with \( k = 4 \). The sampling scheme requires sampling along parallel lines separated by a distance \( h \) and with a spacing between ordinates along a line of \( h \).

Results will also be given for the hexagonal grid or trigonal symmetry and the hexagonal case with \( k = 6 \). For purposes of comparison, the cases when \( k = 2, 3, \) and \( 4 \) will also be considered from earlier. The notation will be used below. If the \( m \) random variables \( x = (x_1, x_2, ..., x_m) \) have a joint multivariable Gaussian (normal) distribution with mean \( \mu \) and variance–covariance matrix \( V \) then this is denoted by \( x \sim N(\mu, V) \). Also the convention of using an upper-case letter for a random variable and a lower-case letter for a realization of the random variable will be followed, as in the previous section.

### 3.2.10.2 The Hexagonal Grid in the Trigonal Symmetry Case

Results have been obtained for the PDF and expectation (mean) of peak (or summit) height, the density of summits and the expected peak (or summit) curvature in the cases when \( k = 2 \) by Whitehouse and Phillips [17] and when \( k = 4 \) by Whitehouse and Phillips [4] and Greenwood [12]. The results for the hexagonal grid (\( k = 3 \)) in the trigonal symmetry and the hexagonal case \( k = 6 \) will now be given [14]. These can be obtained from the general results of truncated random variables in the appendix of Whitehouse and Phillips [4].

For measurements with four ordinates let \( z_0 \) be the height of the central ordinate and \( s_1, s_2, \) and \( s_3 \) be the differences between this ordinate and the three adjacent ordinates at a distance \( h \). The ordinate \( z_0 \) will be defined to be a four-point summit if \( s_1, s_2, \) and \( s_3 \) are all positive. By analogy with the three-point and five-point definitions of curvature the discrete definition of four-point curvature is

\[
C = \frac{2(s_1 + s_2 + s_3)}{3h^2}. \tag{3.108}
\]

Assuming that the surface height measurements have a multivariate Gaussian distribution and that the surface is isotropic then

\[
(z_0, 2 - 2 \rho_1)^{1/2}(s_1, s_2, s_3) \sim N[0; V_4]. \tag{3.109}
\]

The probability that \( s_1, s_2, \) and \( s_3 \) are all positive gives the probability that an ordinate is a four-point summit. Thus, again using the nomenclature of Cheng [15],

\[
V_4 = \begin{pmatrix}
1 & d & d & d \\
d & 1 & a & a \\
d & a & 1 & a \\
d & a & a & 1 \\
\end{pmatrix}
\tag{3.110}
\]

with

\[
d = \left( \frac{1}{2} - \frac{1}{2} \rho_1 \right)^{1/2} \quad \text{and} \quad a = \frac{1}{2} (1 - 2 \rho_1 + \rho_{z1})(1 - \rho_1), \tag{3.111}
\]

where \( \rho_1 = \rho(h) \) and \( \rho_{z1} = \rho \left( \sqrt{3}h \right) \).

If \( Y_z \) is the event \( (s_1 > 0, s_2 > 0, s_3 > 0) \) then the distribution of four-point summit height is the conditional distribution of \( z_0 \) given that \( Y_z \) has occurred. This can be obtained using the results of Whitehouse and Phillips [4] with \( m = 3 \):

\[
X = (2 - 2 \rho_1)^{-1/2}(s_1, s_2, s_3) \tag{3.112}
\]

\[
d = \left( \frac{1}{2} - \frac{1}{2} \rho_1 \right)^{1/2}. \tag{3.113}
\]
The distribution of the height $z_0$ of a four-point summit conditional on curvature $C$ is Gaussian with an expectation given by
\[ E[z_0 | C = c, Y_4] = \frac{3h^2(1-\rho_i)c}{4(2-3\rho_i+\rho_{\xi\xi})} \] (3.123)
and variance given by
\[ \text{var } [z_0 | Y_4] = \frac{1-3\rho_i + 2\rho_{\xi\xi}}{2(2-3\rho_i+\rho_{\xi\xi})}. \] (3.124)

This is the same as the distribution of the height $z_0$ of an ordinate conditional on the four-point curvature but not conditional on the ordinate being a summit, and is a result which holds for the three values of $k = 2, 3, \text{ and } 4$. This is because $V_k$ is of the form
\[ V_k = \begin{pmatrix} 1 & d' \\ d & \varphi \end{pmatrix} \] (3.125)
where $V$ is a correlation matrix with a constant row (column) sum ensured by cofactors $d'$ and $d''$. This result enables the expected $(k + 1)$-point peak (or summit) curvature to be obtained from the expected $(k + 1)$-point peak (or summit) height.

Hence the expected four-point summit curvature is given by
\[ E[C | Y_4] = \frac{4(2-3\rho_i+\rho_{\xi\xi})}{3h^2(1-\rho_i)} E[Z_0 | T_k] \]
\[ = \frac{1}{2} \left[ \frac{1}{4} + \frac{1}{2\pi} \sin^{-1} \left( \frac{1-2\rho_i + \rho_{\xi\xi}}{3-4\rho_i+\rho_{\xi\xi}} \right) \right] \]
\[ \left\{ h^2 [\pi(1-\rho_i)]^{1/2} \left[ \frac{1}{2} - \frac{3}{4\pi} \cos^{-1} \left( \frac{1-2\rho_i + \rho_{\xi\xi}}{2-2\rho_i} \right) \right] \right\}. \] (3.126)

It is also possible to obtain the following simple connection between the variances of the $(k + 1)$-point peak (or summit) height and curvature:
\[ \text{var } [Z_0 | T_{k+1}] = \text{var } (Z_0 | C, Y_{k+1}) + \left[ \frac{E[Z_0 | C = C, T_{k+1}]}{C} \right]^2 \text{var } (C, T_{k+1}). \] (3.127)

This relation was shown by Whitehouse and Phillips [4] for $k = 1$ (with $z_0 \equiv y_0$) and by Greenwood [12] for $k = 4$.

So, by the application of the theory of Gaussian truncated random variables, it has been possible to obtain connections between the expectations and variances of four-point summit and curvature. Also, it is straightforward to show, without getting involved in the calculation that the probability
TABLE 3.3
Expected Summit (Peak) Density

<table>
<thead>
<tr>
<th>Model</th>
<th>k</th>
<th>pr(T_k, 1) Limit as h→0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Three points</td>
<td>2</td>
<td>[ \frac{1}{2\pi} \left( \frac{D_1}{D_2} \right)^{\frac{1}{2}} ]</td>
</tr>
<tr>
<td>Four points</td>
<td>3</td>
<td>[ \frac{1}{6\pi} \left( \frac{D_1}{D_2} \right) h^2 = 0.0531 \left( \frac{D_1}{D_2} \right) h^2 ]</td>
</tr>
<tr>
<td>Five points</td>
<td>4</td>
<td>[ \frac{\pi + 2\sin^{-1} \left( \frac{1}{3} \right) + 4\sqrt{2} \sqrt{D_1}}{24\pi^2} \left( \frac{D_1}{D_2} \right) h^2 = 0.400 \left( \frac{D_1}{D_2} \right) h^2 ]</td>
</tr>
<tr>
<td>Seven points</td>
<td>6</td>
<td>[ \frac{\pi + 6(\sqrt{3} - 1)}{24\pi^2} \left( \frac{D_1}{D_2} \right) h^2 = 0.0318 \left( \frac{D_1}{D_2} \right) h^2 ]</td>
</tr>
</tbody>
</table>

From this the limiting values as h→0 can be found and inserted in Table 3.3.

It can be seen from Figure 3.20 that the values of the summit properties approach that of the continuous case as k increases—a not unexpected result. It has also been shown for the hexagonal case k = 6 that the data storage is 13% less than for the rectangular case and in many instances the processing is quicker (see Ref. [52]). Also, if faults are present in the surface they are more easily detected. This result is shown by Whitehouse and Phillips [18].

The most important point to emerge is that the best results are those where the sampling pattern follows most closely the areal bandwidth pattern of the surface. For example, the hexagonal case k = 6 is most suitable for isotropic surfaces which have circular symmetry about the origin in frequency (and wavelength) [19]. In the case of anisotropic surfaces a suitably scaled sample interval in both directions with a rectangular grid for the tetragonal case will probably be best.

So simply applying a sampling procedure to a surface is not good enough. The sampling pattern should, whenever possible, image the surface properties. There is obviously another good case here for looking at the surface before trying to evaluate it.

3.2.10.3 The Effect of Sampling Interval and Limiting Results on Sample Patterns

It is important to investigate the variation of parameters with h because it is due to the large number of possible differences in sampling interval that the scatter of measured values of parameters occurs between investigators [15].

The distributions of four-point summit height and curvature have been derived in terms of correlation coefficients between ordinates. These two correlation coefficients are \( \rho_i \) for ordinates a distance \( h \) apart, and \( \rho_{ij} \) for ordinates a distance \( \sqrt{3}h \) apart. If the surface is isotropic and the autocorrelation function is \( \rho(x) \), then \( \rho_i = \rho(h) \), and \( \rho_{ij} = \rho(\sqrt{3}h) \). So \( \rho_i \) and \( \rho_j \) will vary as \( h \) varies, depending on the shape of the autocorrelation function of the surface.

\[ E[z | Y_i] = \frac{3\sqrt{(1 - \rho_i)}\pi \cdot \Phi^6(0, B_e)}{pr(Y_i)}, \]

where \( B_e \) is the variance-covariance matrix of the conditional distribution of the differences \( s_1, s_2, ..., s_5 \) given \( s_6 \), from which [20]

\[ E[z | Y_i] = \frac{3\sqrt{(1 - \rho_i)} h}{8\sqrt{3} \pi^2} \left[ 6\pi - 12\tan^{-1}(1/\sqrt{2}) - \sqrt{2}\pi \right] \Delta \left( \frac{D_1}{D_2} \right)^{\frac{1}{2}}. \]
Results for the summit height have been obtained for the continuous surface, so it is possible to compare these results with those obtained earlier for the discrete results as the sampling interval \( h \) converges to zero.

To do this it is necessary, as before, to make assumptions about the behavior of the autocorrelation function \( \rho(h) \) near the origin. It will be assumed as before that

\[
\rho(h) = 1 + D_2 h^2/2! + D_4 h^4/4! + O(h^4) \quad (3.133)
\]

where

\[
D_2 < 0 \quad D_4 > 0 \quad \text{and} \quad \eta = -D_2(D_4)^{-1/2} < \frac{\sqrt{3}}{6}. \quad (3.134)
\]

\( D_2 \) and \( D_4 \) are the second and fourth derivatives of the autocorrelation function at the origin.

Comparison will be made for the estimates of parameters measuring peak and summit properties of the surface. This will be done for the four cases of three-, four-, five-, and seven-point estimates.

The first parameter that will be considered is the density of peaks or summits. These parameters are known for a continuous random Gaussian surface and were given for peaks as

\[
D_{\text{peak}} = \frac{1}{2\pi} \sqrt{\frac{D_4}{-D_2}}. \quad (3.135)
\]

by Rice [5] and for summits as

\[
D_{\text{sum}} = \frac{1}{6\pi^3} \left( \frac{D_1}{-D_2} \right) \quad (3.136)
\]


The density of the peaks or summits is the number of peaks per unit length or summits per unit area, using the \((k + 1)\)-point definition of peak for \( k = 2 \) and summit for \( k = 3, 4, \) and 6. The expected density of peaks or summits is given by the product of \( \text{pr}(Y_{k+1}) \) and the density of ordinates, where \( T_{k+1} \) is the event \((s_1 > 0, \ldots, s_k > 0)\) and \( s_1, s_k \) are the differences between the central ordinate and the \( k \) adjacent ordinates at a distance \( h \).

The limiting behavior of \( \text{pr}(Y_{k+1}) \) as \( h \) tends to zero, the density of ordinates and the limit of the expected density of peaks (or summits) are given in Table 3.3. The limits are given in terms of the limiting results for a continuous surface given by Refs. [4, 11]. It can be seen that the density of peaks (when \( k = 2 \)) converges to the continuous limit. This is not the case for summits (when \( k = 3, 4, \) and 6). In these cases the density would be overestimated by 73%, 31%, and 4%, respectively (see Figure 3.20).

The second parameter that will be considered is the average peak (or summit) height. The results are known for a continuous random Gaussian surface and were given for peaks as

\[
E[Z \mid \text{continuous peak}] = \frac{\pi}{\sqrt{2}} \eta \quad (3.137)
\]

by Rice [5] and Whitehouse and Phillips [4] and for summits as

\[
E[Z \mid \text{continuous summit}] = 1.801 \left( \frac{\pi}{\sqrt{2}} \right)^{1/2} \eta \quad (3.138)
\]

by Nayak [11]. Hence the average summit height is 80% higher than the average peak height.

Again the expected height of peaks (when \( k = 2 \)) converges to the continuous limit for peaks on a profile, but this is not the case for summits (when \( k = 3 \) and 6) as is seen in Table 3.4. In these cases the expected summit height is underestimated by

\[
\text{TABLE 3.4}
\]

Expected Summit (Peak) Height \( E[Z_{1/2Y_{k+1}}] \)

<table>
<thead>
<tr>
<th>Model</th>
<th>( k )</th>
<th>Limit as ( h \to 0 )</th>
<th>Limit as ( h \to \infty )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Three points</td>
<td>2</td>
<td>( \left( \frac{\pi}{2} \right)^{1/2} \eta = 0.555 \frac{4}{\sqrt{\pi}} \eta )</td>
<td>0.846</td>
</tr>
<tr>
<td>Four points</td>
<td>3</td>
<td>( 2 \left( \frac{3}{\pi} \right)^{1/2} \eta = 1.559 \left( \frac{\pi}{2\sqrt{2}} \right)^{1/2} \eta )</td>
<td>1.029</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( = 0.866 \frac{4}{\sqrt{\pi}} \eta )</td>
<td></td>
</tr>
<tr>
<td>Five points</td>
<td>4</td>
<td>( \frac{8\sqrt{2}\pi}{\pi + 2\sin^{-1} \left( \frac{1}{3} \right) + 4\sqrt{2}} \eta = 1.688 \left( \frac{\pi}{2} \right)^{1/2} \eta )</td>
<td>1.163</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( = 0.938 \frac{4}{\sqrt{\pi}} \eta )</td>
<td></td>
</tr>
<tr>
<td>Seven points</td>
<td>6</td>
<td>( \frac{3\sqrt{\pi} \left[ 6\pi - 12 \tan^{-1} \left( \frac{1}{\sqrt{2}} \right) - \sqrt{2}\pi \right]}{\sqrt{\pi} \left[ \pi + 6 (\sqrt{3} - 1) \right]} \eta = 1.779 \left( \frac{\pi}{2} \right)^{1/2} \eta )</td>
<td>1.352</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( = 0.938 \frac{4}{\sqrt{\pi}} \eta )</td>
<td></td>
</tr>
</tbody>
</table>
13% for the four-point case, by 6% for the five-point case, and by only 1% for the hexagonal case.

Because the conditional distribution of height given curvature is Gaussian with a mean which is a linear function of curvature, for all values of \( k \), the expected summit curvature will converge in the same manner as the expected summit height.

To study the qualitative effect of the change of the sampling interval \( h \) on the digital measurements of an isotropic surface it is necessary to specify a model for the autocorrelation function of the surface. Note that any autocorrelation could be used. For the model to fit in with observed autocorrelation functions of surfaces it would be desirable to have a negative exponential function with a multiplicative periodic function. Whitehouse and Phillips [4] “smoothed” the exponential cosine function to give a function that was smooth at the origin. This alternative approach replaced the negative exponential function by another function which is smooth at the origin but behaves like the negative exponential function for large lag values. Lukyanov [21] has used such models extensively. Both of these are close to the autocorrelation functions of many typical practical surfaces. Note, however, that the results are quite general and that the specific models are only used to give an idea of actual scale. This model autocorrelation function is given by

\[
\rho(h) = \text{sech}\left(\frac{1}{2} \pi h A(\theta)\right) \cos(2\pi h A(\theta)) \tag{3.139}
\]

where

\[
A(\theta) = \text{sech}(2\pi \theta) + 2 \sum_{r=0}^{\infty} \frac{(-1)^r \theta}{2\pi \theta^2 + \left[1/4 \cdot (2r + 1)^2\right]} \cdot \frac{\sinh(\pi(2r + 1)/8\theta)}{\sinh(\pi/8\theta)} \tag{3.140}
\]

where \( h \) is the sampling interval. The values of \( \theta \) used are 0 and 1/2. For this autocorrelation function

\[
D_2 = -(1/2 \pi)^2 [1 + (4\theta)^2] [A(\theta)]^2 \tag{3.141}
\]

and

\[
D_4 = +(1/2 \pi)^4 [5 + 6(4\theta)^2 + (4\theta)^4] [A(\theta)]^4. \tag{3.142}
\]

A point needs to be made here. The rather complicated correlation function (Equation 3.139) is not the only function for which the analysis will work. It will work for any correlation function having \( \rho_1, \rho_2 \) etc., as values at the given spacings. This is wide enough to cover most, if not all, reasonable surfaces. The reason for this complex correlation function is simply to ensure that it has near perfect properties at the origin and elsewhere just to forestall any criticism. For practical cases the exponential, Gaussian, Lorentzian or whatever correlation function could have been picked to give an idea of the quantitative effects of sampling on the tribological properties of random surfaces.

What these results show is that by measuring the correlation values on a surface at spacings \( h, 2h, \sqrt{3}h, \) etc., the tribological summit parameters can be found simply by inserting the values of \( \rho_1, \rho_2, \) etc., into the derived formulae. There is no need to measure the surface itself to get curvatures etc., providing that it is reasonably Gaussian (to within a skew of ±1, see Stauffert [22]). The very tricky parameters can be simply calculated by knowing the correlation values.

For theoretical purposes a model of the surface can be devised as above and the values of \( \rho_1, \rho_2 \) etc., calculated. The tribological parameters can then be similarly evaluated.

The expected density of summits is given in Figures 3.22 and 3.23 and the expected height of peaks or summits is given...
in sampling schemes is to try and see whether taking measurements using a non-conventional sampling scheme would produce any advantages to outweigh the disadvantage of complexity. The advantages are less information to collect, easier analytical derivation of theoretical results and simpler numerical methods.

The sampling schemes that were considered all had the property that the information could be collected by sampling along parallel straight lines with a fixed sampling interval. (It might be necessary, however, to have a variable starting point, though this would follow a regular pattern.)

This ensured that if a measurement (ordinate) was chosen when using a particular scheme it would always have the same number of adjacent ordinates at a distance \( h \) (the chosen sampling interval), provided the chosen ordinate is not on the boundary.

From the point of view of simplicity of sampling mechanism the square grid \((k = 4)\) in the tetragonal case is the best. In this case the spacing between the lines is constant and equal to the sampling interval \( h \) along the line. Also the starting points for the sampling all lie along a straight line. However, other schemes do have advantages to offset their complexity.

The trigonal \((k = 3)\) case has the advantage that measurements of slope can be taken in three directions as opposed to two for the tetragonal \((k = 4)\) case. Though the theoretical results have been restricted to the consideration of isotropic surfaces it may still be of practical value to be able to check the assumption of isotropicity in more than two directions.

The trigonal \((k = 3)\) case can be obtained by an alternative sampling method but this involves alternating the sampling interval from \( h \) to \( 2h \). This alternative method is equivalent to rotating the grid through \( \pi/6 \). For \( k = 6 \) the sampling is very straightforward, as can be seen from Figure 3.21.

From the point of view of collecting digital information, the trigonal \((k = 3)\) case is preferable as “less” information is collected. The density of ordinates is \((4/3\sqrt{3})h^2 \) compared with \( 1/h^2 \) for the square grid \((k = 4)\), so in the same area 23% fewer ordinates would be needed. The advantage of this would need to be weighed against the disadvantages.

Another advantage of the trigonal \((k = 3)\) case is that fewer ordinates are used when defining the properties of the extremities. To check the definition of a four-point summit only three conditions have to be obeyed, as opposed to four conditions for the five-point summit and six for the seven point. It should also be noted that some properties of the discrete defined random variables, such as the limiting values of the \((k + 1)\)-point summit (or peak) height as the sampling interval tends to infinity, are simply a function of the numerical definition and are independent of the surface being measured. The problem is that the surface has to be measured to get any values for the properties.

Any discrete measurement of a surface must lose information compared with a complete “map” of the surface. This is inevitable! However, ideally, any discrete measurement...
should produce results that converge to the results for the continuous surface as the sampling interval \( h \) tends to zero.

For sampling along a straight line \( (k = 2) \) it has been seen that the discrete results do converge to those for the continuous profile. They do not, however, converge to the results of the areal measurement. For example, \( D_{\text{peak}}^2 = 0.83 \ D_{\text{sum}} \), so that assuming independent measurements at right angles would produce a limit that is 17\% too small.

For areal measurements when sampling with \( k = 3, 4, \) or 6, the limiting results for expected summit density and expected summit height do not converge to the results for the continuous surface. In the case of expected summit density the limit is 73\% too large for \( k = 3 \), 31\% too large for \( k = 4 \) and 4\% for \( k = 6 \). Again for expected summit height, the case \( k = 3 \) is worse than for \( k = 4 \) and \( k = 6 \) but the differences are not so large. This suggests that some surface parameters may be estimated by discrete methods fairly well but others may not.

In order to consider the merits of sampling schemes it is necessary to study their theoretical properties. By doing this it is possible to obtain new insights into the general problem, but only by using models which lead to tractable mathematics. The three sampling schemes with \( k = 2, \) 3, and 4 considered here have been chosen because they have a common property that enables them to be investigated using the analytical results of theoretical statistics. Using the trigonal \( (k = 3) \) symmetry case leads to a simpler mathematical model than for the tetragonal \( (k = 4) \) symmetry case, as this reduces the dimension by one. However, taken as a whole it may be that a hexagonal sampling plan where \( k = 6 \) offers the maximum benefit in terms of the three criteria mentioned above. One message which has emerged is that the conventional grid pattern method of sampling is not necessarily the best. The implications of this in general pattern recognition and image analysis scanning systems could be significant.

The result given here has been concerned primarily with the effect of sampling patterns on the values of parameters obtained from measured surfaces. It has not therefore been aimed at investigating the actual nature of surfaces in general. Well-behaved correlation functions have been assumed and certain specific examples have been used to give the researcher an idea of the value of parameter changes that might be expected to occur on typical measured surfaces. This has been justified by the fact that to some extent all instruments used for obtaining the data have a short-wavelength filter incorporated, whether it is a stylus or a light spot, which tends to force the correlation at the origin to be smooth. However, there can be no denying that the very nature of random manufacture encourages the presence of exponential and other misbehaved characteristics in the correlation function. The effect of sampling patterns on such fractal (multiscale) surfaces will be the subject of further comment in Chapter 7.

Rosen [62] has made some interesting comments concerning sampling strategies. He proposes that the sampling rate should be flexible during a measurement: the rate depending upon the nature of the surface. Alternative “sparse matrices” with variable sampling distances and sampling patterns which include star and cross as well as the conventional orthogonal sampling are suggested. Also “adaptive sampling” which uses the surface “as found” as the basis for determining the spacing, as an alternative to fixed sampling. He maintains that the level of confidence of a reading should match the importance of the feature being measured. His philosophy is that the measurement strategy should be related to the variation of the topography over the whole product and not just one part of it. A good way to approach this problem, he says is functionally: a functional element like a piston ring sliding along a surface will “feel” the variation of the cylinder liner roughness and will react accordingly regarding the oil film formation and the sealing properties, which will strongly affect the life, energy consumption, emissions, and power loss of the engine.

He splits the problem into two: one tackling the local variations and the other the global variations.

The technique proposes the production of parameter variation maps, both globally and locally, to make sure that critical measurements are taken where needed most and that only enough measurements are taken to ensure a specific level of confidence and that once achieved the measurements stop. This way the dual objectives are met namely that the measurements are significant and that the time used is a minimum. The criteria for this sampling plan are admirable but the problem is that there would have to be as many plans as functions and consequently would be too flexible to be practical.

In the above sections much emphasis has been placed on statistical parameters, distributions and functions. Some effort has to be expended to make sure that the values obtained are realistic. The next section therefore will revert back to a simple discussion of the processing of these parameters. However, because of its central importance in random process theory, the role and evaluation of Fourier transforms will be described here.

### 3.3 Digital Form of Statistical Analysis Parameters

#### 3.3.1 Amplitude Probability Density Function

This contains the height information in the profile. It has been used extensively in one form or another in surface analysis. Various terms are used to describe the function, its derivatives and its integral.

In Figure 3.26 the value for the amplitude probability density function (APDF) at a given height is called \( p(z) \) and is the number of profile ordinates having a numerical value between \( z \) and \( z + \delta z \) divided by the total number of profile ordinates. There are two basic ways of evaluating the APDF digitally. The first and most obvious way is to select a height interval, say between \( z \) and \( z + \delta z \), and to scan through the whole profile data, counting how many ordinates lie within
the height range. The height interval is changed and the operation is repeated. This is carried on until all the vertical range of the profile has been covered. In this method the height interval is selected before the counting is started. The other method involves examining every profile ordinate in sequence, finding to which height interval it belongs and registering a count in the store position corresponding to this level. The advantage of the last method is that the whole of the amplitude distribution (APDF) is found after just one traverse and the only store requirement is that of the number of height intervals to cover the height range; the profile itself need not be stored. Measurement of the central moments, skew and kurtosis can be made from the APDF directly.

**FLOW CHART 3.1** Global sampling strategy. (From Rosen, B-G., XII Int. Colloquium on Surfaces, Chemnitz, January 28–29, 2008. With permission.)

**FIGURE 3.26** Statistical height distributions.
For instance, the RMS value $R_q$ is given by

$$R_q = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (i\Delta z)^2 p_i}$$

(3.143)

where $\Delta z$ is the quantization interval of the pockets and where the mean

$$\bar{z} = \frac{1}{m} \sum_{i=1}^{m} (i\Delta z) p_i$$

(3.144)

and there are $N$ levels of $p_i$ in the APDF. Notice that the $R_q$ value is subject to an error of $\Delta z/\sqrt{2}$ which is a penalty for being in discrete rather than continuous form. These moments can be measured more accurately without recourse to increasing the store as running summations in terms of each other. It is not necessary to determine $p_i$ as such, the natural quantization interval of the incoming data $q$ being the only limitation in accuracy. Thus

$$R_q = \left(\frac{1}{m} \sum_{i=1}^{m} z_i^2 - \left(\frac{1}{m} \sum_{i=1}^{m} z_i\right)^2\right)^{1/2}$$

(3.145)

where the number of ordinates is $m$ and the $z$ values are measured from an arbitrary datum.

### 3.3.2 Moments of the Amplitude Probability Density Function

**Skew**

$$skew = \frac{1}{m} \sum_{i=1}^{m} z_i^3 - \frac{3}{m^2} \sum_{i=1}^{m} z_i \sum_{i=1}^{m} z_i^2 + \frac{2}{m^3} \left(\sum_{i=1}^{m} z_i \right)^3$$

$$\left[\frac{1}{m} \sum_{i=1}^{m} z_i^2 - \left(\frac{1}{m} \sum_{i=1}^{m} z_i\right)^2\right]^{3/2}$$

(3.146)

**Kurtosis**

$$kurtosis = \frac{1}{m} \sum_{i=1}^{m} z_i^4 - \frac{4}{m^2} \sum_{i=1}^{m} z_i \sum_{i=1}^{m} z_i^3 + \frac{6}{m^3} \left(\sum_{i=1}^{m} z_i \right)^4 - 3$$

$$\left[\frac{1}{m} \sum_{i=1}^{m} z_i^2 - \left(\frac{1}{m} \sum_{i=1}^{m} z_i\right)^2\right]^2$$

(3.147)

In other words the central moments can all be expressed in terms of the moments about an arbitrary level. It is the central moments which contain the information in surface metrology.

All the important moments can be found directly from the APDF, even the $R_q$ value. It is given by

$$R_q = 2 \left( \sum_{\tau < \bar{z}} \sum_{i=1}^{N} p_i - \sum_{\tau < \bar{z}} \Delta \bar{z}_i p_i \right)$$

(3.148)

so that, providing the APDF is measured as the input signal enters, all the moments, peak values (subject to a $\Delta \bar{z}$ limit on accuracy) and $R_q$ can be measured without actually storing the data!

The distribution function is merely the cumulative sum of $p_i$, up to a given level. Thus

$$P_j = \sum_{i=1}^{j} p_i.$$ 

(3.149)

The bearing ratio, material ratio, or Abbott–Firestone curve often used in surface metrology are all $1 - P_j$.

One of the problems in assessing the profile parameters without evaluating $p_j$, is that the number of ordinates can be so great that the summations may overload the computational word length.

These techniques are not restricted to the measurement of a surface profile; any waveform can be analyzed in the same way whether it is derived from the profile itself, the slope of the profile or whatever. There is a problem, however, in isolating the different components of, say, texture. This is because, although these quick and economic techniques for statistical moment evaluation do not require knowledge of the mean level, they do assume that the signal follows the general direction of the surface. In other words, it is not possible to eliminate the need for a reference line but it is possible to relax the stringent positioning of it relative to the profile. Fortunately this is not a problem in measuring slope or curvature; the general trend of the waveform will be zero because of the differentiation that takes place.

### 3.3.3 Autocorrelation Function

There are a number of ways in which the autocorrelation function can be evaluated; which one is chosen depends on time and space factors. Most often the temporal (serial) form is used for convenience although the ensemble (parallel) method is strictly the correct method. The most obvious way for the temporal method is to store all the initial filtered data of the profile, say $m$ points, $z_1, z_2, ..., z_m$, and then to evaluate, step by step, the correlation coefficient corresponding to each lag point, making sure that the average values are taken. Thus for a lag $j\Delta x$ the correlation array

$$A(j\Delta x) = \frac{1}{m-j+1} \sum_{i=1}^{m-j+1} z_i z_{i+j} \left/ \frac{1}{m} \sum_{i=1}^{m} z_i^2 \right.$$ 

(3.150)

This method requires that all the data are stored, which may well not be convenient. Under these circumstances an alternative way is possible which, although taking longer, requires considerably less storage. The data are sifted and operated on as it becomes available from the input; the correlation array is built up in parallel rather than serial as in the first method.

Organizationally the operation is as shown in Figure 3.27. After the $m$ ordinates have been operated on the array $A(1(r))$ is normalized with respect to the variance of $z_i$ with
the mean value removed and the result in $A(1(r))$ is the autocorrelation function. Only $j$ data storage locations have been required as opposed to $m$ for the previous method. Because $j$ is of the order of $m/10$ a considerable gain in space can be achieved. This method is used when storage is at a premium. Fortunately, with the spectacular increase in storage capability schemes like this are becoming redundant.

By far the most often used method is that involving the FFT simply because of the time saving. The method is different from the others because it involves getting the power spectral density first and then deriving the autocorrelation from it.

### 3.3.4 Autocorrelation Measurement Using the Fast Fourier Transform

Schematically the calculation follows the steps below:

1. Profile $z(x) \rightarrow \xi(k\Delta x)$ (sampled profile) fast Fourier transform
2. Discrete Fourier transform using fast Fourier transform
3. $P(h\Delta f)$ power spectral density
4. Discrete inverse Fourier transform using the fast Fourier transform
5. $A(\Delta \tau)$ autocorrelation function

In this technique the basic FFT operation is performed twice, once to get the power spectrum and once to get the autocorrelation function from it. This is the reverse of the old method in which the power spectrum is obtained from the autocorrelation function. Note that $P$ is not the distribution function of equation 3.149.

Apart from the obvious time advantage there is a storage advantage. For autocorrelation this involves a realization that two steps are required and, furthermore, because of the nature both of the data and of the transformation some economy can be made. In this case the original data—the profile, say—is real. This means that the discrete Fourier transform (DFT) will be conjugate even (the amplitude terms have even symmetry about the frequency origin while the phases have odd) and consequently only half the transform needs to be evaluated—the other half can be deduced and the storage allocated to these points can be used elsewhere. The original data points are completely replaced in the store by an equal number of transformed points in the FFT as described in Section 3.8.2. Furthermore, in undergoing the inverse discrete Fourier transform a further gain can be effected because, although the data (in this case the Fourier transform of the original profile) is conjugate even, only the cosine transformation needs to be used—the phase is lost in the autocorrelation function. Hence space which would normally be used to house the imaginary part of the complex numbers (resulting from the FFT) can be used for other purposes.

All these methods use the “time average” or “temporal” version of the autocorrelation function for evaluation rather than the ensemble average where it is understood in the case of surfaces that the term temporal really means spatial average because it is the surface geometry which is being used and not a voltage or other time function.

$$A(\tau) = E[z(\tau)|z(\tau + \tau)].$$  \hspace{1cm} (3.151)

This choice is for purely practical reasons. It is easier to evaluate $A(\tau)$ from one record than to measure it from a number of records in parallel. It does assume ergodicity, however.

A similar function known as the structure function $S(\tau)$ can be obtained by measuring the expected variance between ordinates $z_1$ and $z_2$ separated by $\tau$

$$S(\tau) = E[(z_1 - z_2)^2] = 2\sigma^2[1 - A(\tau)].$$  \hspace{1cm} (3.152)

The structure function is a form of cumulative autocorrelation function. It has the advantage that any slope on the waveform is eliminated and the value $z_1 \rightarrow z_2$ can be evaluated directly by means of a skid-stylus combination without the need to have two independent styluses or without having to store any data (providing that $\sigma$ is known). If $\sigma$ is not known, a digital filter can be used to preprocess the data in real time using very little storage. From the point of view of classification it is not necessary that $\sigma$ be known because it is a constant unless fractal—the essential variation of the structure with the lag $\tau$ is the same as the autocorrelation function, that is

$$\frac{dS(\tau)}{d\tau} = -k dA(\tau).$$  \hspace{1cm} (3.153)

### 3.3.5 Power Spectral Density

The power spectral density can be evaluated from the data either via the autocorrelation function or the Fourier transform of the data. The former method is possible because of the Wiener–Khinchine relationship linking the two together. Thus in continuous form

$$P(\omega) = 2 \int_0^\infty A(\tau) \cos(\omega \tau) d\tau.$$  \hspace{1cm} (3.154)
In practice the record from which the autocorrelation function has been measured is finite, of length \( L \). Thus

\[
A(\tau) = \frac{1}{L - \tau} \int_{0}^{L-\tau} f(x) f(x + \tau) \, dx. \tag{3.155}
\]

This limits the available length of autocorrelation that can be used to get a power spectral density. This truncation of the autocorrelation causes some problems in the frequency domain. These problems are twofold: one is that the statistical reliability of the information is limited, and the other is that the shape of the truncation causes spurious side effects in the spectrum. Truncation in the extent of the record is equivalent to multiplying the data waveform by a box function. This box function when transposed into frequency is highly resonant; it produces “ringing” around simple peaks in the spectrum, for instance frequencies corresponding to the feed of a tool. The transform of the box function is a sine function, which has a considerable lobbing up to 25% of the peak value. To reduce this in practice a weighting function can be applied to the autocorrelation function prior to transformation. A criterion for the shape of this is that ringing does not occur, or at least to only a few per cent.

The box function is an example of what is in effect a “lag window.” This lag window would normally be a box function, but it can be shaped in the correlation domain to have a minimum effect in the frequency domain. The frequency equivalent of the lag window is called a spectral window. The criterion for a well-behaved spectral window is that it should have a highly concentrated central lobe with side lobes as small and rapidly decaying as possible [24].

The most used lag window \( W_L(\tau) \) is that due to Hanning, which has the formula

\[
W_L(\tau) = 0.5 + 0.5 \cos(\pi \tau / \tau_{\text{max}}) \tag{3.156}
\]

where \( \tau_{\text{max}} \) is the maximum autocorrelation lag allowed by reliability considerations. An alternative is also due to Hamming

\[
W_L(\tau) = 0.54 + 0.46 \cos(\pi \tau / \tau_{\text{max}}). \tag{3.157}
\]

Some other examples are shown in Figure 3.28, which also illustrates the well-used Gaussian window. This is unique in that the spectral window is the same shape as the lag window.

Assuming that the Hanning window is chosen the digital equivalent formula for the power spectrum is given by

\[
P(\omega) = 2 \Delta \sum_{k = 0}^{N} W(k \Delta \tau) A(k \Delta \tau) \cos(\omega k \Delta \tau). \tag{3.158}
\]

For the power spectrum the sample points are taken \( \Delta \tau \) apart, usually equal to \( \Delta \tau \), the spacing of the ordinates or a multiple of it.

Another way to measure the power spectrum is directly from the data using the FFT routine described earlier. First

\[
\text{FIGURE 3.28} \quad \text{Windows for the correlation function.}
\]

the periodogram \( |F(\omega)|^2 \) is obtained by transforming the real data. This will yield \( N \) transform points corresponding to \( N \) real data points. To get the PSD from this it is necessary to apply the spectral window corresponding to the Hanning (or other) lag window. Now this is operated on the frequency data by means of convolution. Thus

\[
P(\omega) = 0.25(F(\omega_{i+1}))^2 + 0.5(F(\omega_0))^2 + 0.25(F(\omega_{i-1}))^2 \tag{3.159}
\]

for the Hanning window, where \( \omega_{i+1}, \omega_0, \) and \( \omega_i \) are adjacent frequencies in the array. For the special case where \( \omega = 0 \):

\[
P(0) = 0.5(F(0))^2 + 0.5(F(\omega_{i}))^2 \tag{3.160}
\]

and for the Hamming window

\[
P(\omega) = 0.23(F(\omega_{i+1}))^2 + 0.54(F(\omega_0))^2 + 0.23(F(\omega_{i-1}))^2. \tag{3.161}
\]

3.4 DIGITAL ESTIMATION OF REFERENCE LINES FOR SURFACE METROLOGY

3.4.1 GENERAL

Filtering methods are the natural way to isolate specific bands of information of the surface. Obviously, originally the most usual way to apply filtering techniques was to pass the signal through a passive or active electrical network so
that the breakdown of the signal occurred in frequency. This is not the natural domain of the surface, which is spatial, so the recent introduction of digital methods has been an important step in surface analysis. Each digital measurement can now be referred more easily to a point on the surface. Fundamentally the great advantage of filters, as mentioned in the previous Chapter, is that they take the signal as it is and do not assume any particular waveform for the surface. They take the waveform “as received” and operate on it, unlike best-fit polynomial curves which can completely distort the remnant signal if the order of the polynomial is mismatched compared with the general shape of the surface.

### 3.4.2 Convolution Filtering

The first step is to work out the impulse response of the filter and express it in digital form or sequence. Thus, let \( h(n) \) be the digital sequence representing the impulse response and \( z(n) \) be the profile signal. Then the output from the filter \( g(n) \) becomes

\[
g(n) = \sum_{m=0}^{n} h(m) z(n-m)
\]  

where \( h(m) \) and \( z(n-m) \) are zero outside the limits.

The relationship between the weighting function and the impulse response has been explained in Section 3.2.2.

<table>
<thead>
<tr>
<th>TABLE 3.5</th>
<th>Ordinate Spacing and Normalizing Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No of Ordinates</strong></td>
<td><strong>per Cut-off</strong></td>
</tr>
<tr>
<td><strong>Cut-off</strong></td>
<td><strong>Cut-off</strong></td>
</tr>
<tr>
<td>50</td>
<td>0.02</td>
</tr>
<tr>
<td>150</td>
<td>0.0067</td>
</tr>
</tbody>
</table>

The working formula is

\[
g = K(a_0b_0 + a_1b_1 + \ldots + a_nb_n).
\]  

Usually \( a_i, a_2 \) are equally spaced ordinates, but not necessarily; \( K \) is a constant equal to \( 1/\text{the number of samples per cut-off length} \). The \( b \) terms are the digitized values of the weighting function (reversed impulse response of the filter).

The required spacings, and the associated value of \( K \), are shown in Table 3.5.

How the weighting function appears digitally with respect to the profile (shown with the stylus) is illustrated in Figure 3.29. Here the shape of the weighting function is that of the 2CR filter Figure 3.29a and b for the Gaussian case.

A pictorial demonstration of how the weighting function is constructed can be seen in Figure 3.30. The 2CR filter is shown because it is asymmetrical, i.e., causal, and shows clearly the fundamental difference between the weighting function and the impulse response, but once the weighting function has been determined the principle is exactly the same for all other weighting functions. Basically the weighting function for the standard type of filter is the time (space) inverted version of the impulse response, but in the case of phase-corrected filters, the weighting function is the same as the impulse response because it is an even function in \( x \). In all roughness filters there is an impulse which corresponds to the profile and a time dependent portion from which the weighting function is obtained and which is used to obtain the mean line.

To get the mean line value multiply each weighting factor by the profile ordinate in line with it, add the products and divide by the normalizing factor, which is \( 1/K \). This gives the height of the mean line above the same datum and in the same units used for the profile ordinates.

Each mean line ordinate that has been calculated in this way refers to that profile ordinate which is directly in line with the zero time value of the weighting function or its center of symmetry. The weighting factor corresponding to
the maximum $b_0$ of the weighting function corresponds to the first term in a causal system: the $b_0$ value for example in the standard 2CR filter.

A comparison of Figure 3.29a and b immediately illustrates the difference in the shape of the two weighting functions and, more importantly, it shows the fact that the mean line position acts in a different place relative to the weighting function for a phase-corrected weighting function than it does for a weighting function which is not phase-corrected.

Subtraction of the mean line value from the profile ordinate to which it refers gives the filtered profile at that point. Taking the average of the differences (without taking account of sign) over a number of cut-off lengths, usually two, gives the $R_a$ parameter.

3.4.2.1 Repeated Convolutions

An interesting property of the Gaussian function can be used in one application in surface metrology, and this is the use of a Gaussian filter to exclude waviness. The property is that if a number of windows are convoluted together they will always eventually produce an effect which is Gaussian, irrespective of the shape of the window. A good example is the box function (Figure 3.33). Three convolutions of the box function will produce an equivalent weighting function window which is already very close to Gaussian (Figure 3.31).

Some instrument manufacturers advocate the Gaussian filter simply because it has got this property. It means that if a running-average procedure is repeated three times, the end result is as if the profile signal had been subjected to a low-pass Gaussian filter. The very simplicity of the technique makes it fast and inexpensive.

In this form the weighting function has a low-pass characteristic that is the waviness profile. To get the surface roughness value of the surface data corresponding to the mid-point of the weighting function should be selected and the waviness line, as found by this Gaussian weighting function taken from it (Figure 3.32).

The Gaussian shape has some advantages. Probably the most important is that it is recognizable in industry, being the basis for the acceptance curve used in statistical process control. This means that production engineers are comfortable working with it. Also the curve is always positive and it falls off rather faster than the 2CR transmission curve. Large components of waviness therefore do not impinge on the roughness.
The weighting function is given by

\[ h(x) = \frac{1}{\alpha \lambda \alpha} \exp\left(-\pi \left(\frac{x}{\lambda \alpha}\right)^2\right) \] (3.164)

and its transform giving the transmission curve of Equation 3.165 where \( x \) is measured from the axis of symmetry.

\[ H\left(\frac{1}{x}\right) = \exp\left(-\pi \left(\frac{\lambda \alpha}{x}\right)^2\right) \] (3.165)

Where \( \alpha = \sqrt{\ln(2)} = 0.4697 \).

Another more practical reason for using Gaussian type filters is that they minimize the RMS duration of the product of \( h(x) \) and \( H(\omega) \) where \( \omega = 2\pi/x \).

Bodschwinna [26] proposed a modified form of the Gaussian filter incorporating a second order polynomial to reduce the influence of end effects caused by the finite duration of the weighting function. The combination of least-square polynomial as well as filtering is somewhat messy. However, he suggests a form \( h(x) \) given by

\[ h(x) = \frac{1}{\lambda \alpha} \sqrt{\frac{\pi}{C}} \left[ 1.5 - \frac{\pi^2}{C \lambda^2 x^2} \right] \exp\left(-\frac{\pi^2}{C \lambda^2} x^2\right) \] (3.166)

which has a transmission

\[ H\left(\frac{1}{x}\right) = \left[ 1 + C \left(\frac{\lambda \alpha}{x}\right)^2\right] \exp\left(-C \left(\frac{\lambda \alpha}{x}\right)^2\right). \] (3.167)

Krystek [60] uses a spline to get rid of form and waviness and he produces a complicated form for the filter. It is questionable whether the extra complication is needed.

Wavelets have also been proposed as a means of filtering. The ability to vary the resolution and range—usually in octave bands [25] makes them useful in fractal analysis [27]. The term “mathematical zoom lens” has been used for wavelet analysis. Apart from the standard filtering use of wavelets, which has a questionable credibility it seems that defect detection using “raised wavelets” has some benefits [28].

### 3.4.3 Box Functions

While on the subject of box functions, it should be recognized that, when convoluted with the profile signal, a box function is a simple running average, the extent of the average being the length of the box. This average is taken to be at the midpoint of the box (Figure 3.33).

The process of using the averaging procedure was advocated by Reason [29] as a means of producing a mean line of the surface texture. It was then called the “mid-point locus” line. Unfortunately it has a poor frequency response (Figure 3.34). That is,

\[ B\left(\frac{x}{\lambda \alpha}\right) \Rightarrow \frac{\sin(2\pi(\lambda \alpha/x))}{2\pi(\lambda \alpha/x)}. \] (3.168)

However, the “averaging” procedure produced by a running box function is a convolution and, therefore, the whole process acts as a filter. Furthermore, it is in effect a phase-corrected filter providing that the output is taken to act at the center of the box. It therefore represents the first attempt at phase-corrected (or linear phase) working in surface metrology.

It is obvious that in order to get no distortion of the signal but just an attenuation at certain frequencies a phase-corrected characteristic should be used. Another way to get phase-corrected characteristics, which is simple, is to use a “double-pass” method. This is useful when only analog methods are available, but the technique has also been used digitally. To see how this technique works imagine a filter has an impulse response \( h(t) \) whose Fourier transform is \( H(\omega) \). Then if the signal is put through this filter and the output reversed and put through the same filter again the final output is

\[ |H(\omega)|^2 \exp(-j\omega T). \] (3.169)
This is achieved because turning the output round is equivalent to conjugation, that is

\[ h(T - t) \leftrightarrow \exp(-j\omega T) h^*(\omega), \]  

where \( T \) is the time taken to reverse the output signal and enter it into the filter again. So the first pass is \( H(\omega) \). The second pass (in the reversed mode) is simply

\[ \exp(-j\omega T) H^*(\omega) \]  

yielding overall \( |H(\omega)|^2 \exp(-j\omega T) \)

(3.171)

This equation has the linear phase term \( \exp(-j\omega T) \) in it. This is phase corrected about the same time \( T \).

Physically this means that to enable the signal (suitably filtered) to take on phase-corrected characteristics it has to be delayed by time \( T \). It is impossible to get phase-corrected properties without the delay because an impulse response cannot be an even function about \( t = 0 \) and all realizable systems are causal (i.e., do not exist before \( t = 0 \)). This argument is the same in spatial terms.

To use this technique it is essential that the square root of the intended final transmission can be taken, that is \( H(\omega) \) must exist. This means that any characteristic, if capable of being expressed as \( H(\omega)^2 \), can be made into a phase-corrected filter by using this reverse technique. Once this is done the calculation is very fast and the storage small.

### 3.4.4 Effect of Truncation

The finite length of the impulse response may call for comment. Many impulse responses have infinite extent (IIR) (infinite impulse response). Practically, in communication theory they are difficult to use because of the need to store masses of data points. What happens is that the impulse response is curtailed as in the case when only 100 ordinates are used. This results in truncation errors. These can be quantified easily enough because, instead of the frequency output being \( G(\omega) = H(\omega)Z(\omega) \) it becomes

\[ G'(\omega) = H(\omega) * B(\omega)Z(\omega) \]  

(3.172)

where \( B(\omega) \) is the frequency spectrum of the truncating function, usually a box function. This shows the other operation performed by the box function, a multiplication rather than a convolution in averaging.

The effect of truncation is shown in Figure 3.35.

The amount of truncation allowable is usually decided by the percentage error produced on one of the averaging parameters, such as the \( R_p \) value. It has been found that making the weighting function between two and three cut-offs long suffices for an error of less than 5%. Even this can be greatly reduced by increasing (or decreasing) the weighting factor ordinates so that, if the truncation length is \( a \), then

\[ \int_{-a/2}^{a/2} h(x) dx = 1 \]  

(3.173)

If this crude compensation is used then at least the average frequencies are correct.

Later in this chapter recursive filters will be discussed. Here it will be found that they have less of a truncation problem than the convolution methods because the equivalent weighting function builds up as the number of evaluated points increases.

Errors in parameters resulting from truncated weighting functions are easy to evaluate for \( R_p \) and, hence, assuming a correction factor, for \( R_n \). But for peak parameters the problem is much more complicated.

However, it can be argued that communication criteria should not be dominant in surface metrology. Two questions have to be asked. The first is whether an IIR means anything functionally when two surfaces make contact. The second is whether the requirements for filtering should be different for the functional applications of surfaces and for the control of manufacture.

Referring to the first question, it could well be that the effective impulse response could use, as a weighting function, the pressure ellipse of the Hertzian contact zone [30] (Figure 3.36). The term functional filtering has been used [31, 32]. The advantage of such an approach is that it more nearly represents what goes on in a functional situation. The disadvantage is that the instrument maker cannot provide every possibility: the obvious answer is to let the customer choose the filter nearest to his/her requirement.
The advantage of the convolution method, using a weighting function representing a functional condition has many advantages. One is that it is versatile. The filter characteristics can be chosen to fit the function, within reason; the function can effectively be shaped to fit. The disadvantage is that the arithmetic operation is expensive on storage and takes a long time to perform. Simple tricks like the equal-weight method outlined below can be used. There are, however, other methods which preserve the equal-spacing criteria of ordinates.

### 3.4.5 Alternative Methods of Computation

#### 3.4.5.1 Overlap Methods

The first technique is called the overlap-add method. If the profile is long \( N_1 \) and the weighting function \( h(n) \) short \( N_2 \), the profile is split up into samples of \( N_3 \) ordinates. Convenient values of \( N_2 \) would correspond to the sampling length \( \left\lfloor \frac{z}{z} \right\rfloor \).

The convolution becomes

\[
g(n) = \sum_{m=0}^{N_2} h(m) \sum_{i=0}^{N_1} z_i(n-m) = \sum_{i=0}^{N_1} \sum_{m=0}^{N_2} h(m)z_i(n-m).
\]

The duration of each of the convolutions of Equation 3.176 is \((N_1 + N_2 - 1)\) samples — so there is a region of \((N_2 - 1)\) samples over which the \(i\)th convolution overlaps the \((k + 1)\) convolution and the outputs from each therefore overlap and so have to be added; hence the name overlap-add when referred to the outputs.

The second method to be used, which achieves an output reasonably quickly, is called the overlap-save method. This differs from the previous one in that it involves overlapping input sections rather than output sections and is used most often in the filtering of periodic signals, as in the case of roundness. This technique makes use of circular convolution.

In this method the input is sectioned into overlapping sections of length \( L \) and overlap \( M-1 \). Then each of the input sections is convoluted with the filter and the resultant outputs are added together, but the first \( M-1 \) outputs from each section are discarded because these will have been the last \( M-1 \) values of the previous section. These methods are valuable in metrology when using conventional computers because they do allow some output to be obtained very quickly rather than to have to wait until all the input data is stored. This time factor is not essential for post-process inspection but can be serious when in-process measurement is envisaged.

Different methods can also be used to speed up the processing time, one of which follows.

The evaluation of filtered results in the frequency domain can obviously be done by taking the data, doing a FFT on it, and then multiplying the spectrum by the desired filter characteristic. To get the filtered profile all that is needed is to carry out an inverse FFT. This is much quicker because all convolution operations are replaced by multiplications. However, if functional filtering is the objective, for example in contact situations, the spatial convolution method is more suitable because the weighting function can be shaped (as, for example, to the shape of a pressure distribution) to run across the surface (Figure 3.36).

#### 3.4.5.2 Equal-Weight Methods

In the general convolution form for filtering shown earlier, the output at time \( t \) is given by \( g(t) \) where

\[
g(t) = \text{constant} \sum_{i=0}^{n} a_i b_i.
\]

The constant here is a constant of proportionality dependent on the ordinate spacing \( h \). In this equation the \( b_i \) represent weighting function ordinates derived from the impulse response of the filter. Letting the constant be \( h' \) (equal to the reciprocal of the density of ordinates in a sampling length), the equation can be written in a different form which makes better use of the data. The equation shows that each ordinate

![Figure 3.36](image-url)
weighting function. The only disadvantage is that the \( a_i \) values are not equally spaced: their location is inversely proportional to the height of the weighting function. Their location addresses have to be stored in much the same way that the weighting function ordinates in the conventional form have to be stored. However, the look-up process is much faster than the multiplication process and gains in speed of an order of magnitude are possible. This numerical procedure for equal weight can also be applied to a graphical method—which needs to be possible so as to enable calibration routines to be established for people without computers. So, instead of multiplying equally spaced ordinates each by a different weighting factor, a useful approximation can be arrived at by multiplying unequally spaced ordinates by a constant factor, which reduces to adding up the graph ordinates at these spacings and multiplying by the factor (or dividing by its reciprocal if that is more convenient). The use of 26 ordinates was found to give reasonable accuracy and a convenient divisor of 20. A 26-ordinate template designed for 0.8 mm cut-off and a horizontal magnification of 100 has been used with surprisingly accurate results ~5% for the \( R_a \) value.

### 3.4.6 Recursive Filters

The convolution method requires a lot of storage and computation. The output depends on previous and present inputs. Since a filter effectively contains a memory of past inputs, it might be expected that considerable savings in total effort could be obtained by calculating the past and present inputs and past outputs.

This technique is analogous to feedback in a linear control system and for this reason is called recursive. With conventional filtering methods such as that involved in the 2CR filter, the savings in time and storage over the convolution method are substantial. For instance, the example given earlier, which involved 100 weighting factors requires 100 multiplications and additions for each output point. The recursive approach reduces this to the order of only four each.

To appreciate this method (which is similar in every way to the autoregressive moving average equation (ARMA) models used for surface characterization examined in Chapter 2) an example will be given.

#### 3.4.6.1 The Discrete Transfer Function

A convenient notation for the discussion of sampled systems is the \( z \) transform in which the term \( z^{-1} \) is the unit-delay operator. Then a sequence of sampled values can be represented as the coefficients of a power series in \( z^{-1} \). For instance, the infinite sequence \( X = 1, B, B^2, B^3, \) can be written as

\[
X(z) = 1 + Bz^{-1} + B^2z^{-2} + \ldots
\]

\[
= \frac{1}{1 - Bz^{-1}} \quad \text{for } B < 1.
\]

The unit-delay operator is related to the \( z \) transform operator as

\[
z^{-k} \Rightarrow \delta(n-k)
\]

There are several ways of converting the continuous filter into a digital one. The standard method is that known as the impulse-invariant technique in which the discrete impulse response is identical to the sampled impulse response of the continuous filter. In this the continuous function is expanded into partial fractions and then transformed by

\[
\frac{1}{p+a} \rightarrow \frac{1}{1 - \exp(-aT)z^{-1}}.
\]

where \( T \) is the sampling interval so that although digital filters can be directly realized, it is more usual to approach the problem from a linear analysis. The Laplace operator is \( p \).

Tables exist for transformations of this type. The difficulty of this approach is that the gain of the digital filter is proportional to the sampling frequency, which may cause problems if this frequency is not always the same value. Also, it may be noted that the continuous transfer function may not easily resolve into a form suitable for transformation.

A way of avoiding the need to expand the transfer function is to map the \( p \) plane onto the \( z \) plane.

A suitable way of doing this is to use the bilinear substitution:

\[
p = \frac{2}{T} \left( \frac{1 - z^{-1}}{1 + z^{-1}} \right).
\]

One of the main difficulties of this method may be shown by considering the unit circle in the \( z \) plane. Now

\[
z = \exp(-j\omega_0T)
\]

\[
p = \frac{z^{-1}}{z+1}
\]
becomes

\[ p = j \tan \left( \frac{\omega_s T}{2} \right) \]  

(3.184)

so that, as required, the imaginary axis of the \( p \) plane is mapped onto the unit circle of the \( z \) plane. However, it also follows that

\[ \omega_s = \tan \left( \frac{\omega_s T}{2} \right). \]  

(3.185)

Thus the method causes a non-linear warping of the frequency scale. It would, therefore, generally be necessary to precompensate the continuous function.

A third method uses an approximation of the convolution integral on a continuous transfer function that has been either factorized or resolved into partial fractions. This method, of which the impulse-variant approach is a special case, has a gain that is independent of sampling rate and is more readily applied to a high-pass filter than the impulse-invariant transformation.

As computational requirements suggested the use of a cascade approach and the 2CR filter causes no factoring problems, the third method of transformation can be used and this will be discussed in more detail.

For convenience in the following analysis, the single-stage filter will be considered throughout. The two-stage filter is derived directly from it by squaring the transfer function. The basic form of filter section is the low-pass and such sections in cascade have their transfer functions multiplied. So let the general low-pass filter have the form

\[ H(p) = k \prod_{i=1}^{N} \frac{1}{p + a_i}. \]  

(3.186)

Considering a single stage

\[ H(p) = \frac{Z(p)}{X(p)} = \frac{1}{p + a} \]  

(3.187)

and expressing its time domain response by the convolution integral, assuming the filter to be initially relaxed gives, for the output \( z(t) \) and input \( x(t) \),

\[ z(t) = \int_{0}^{\infty} \exp[-a(t-\tau)] x(\tau) d\tau ] \]  

(3.188)

\[ z(t) = \exp(-at) \int_{0}^{\infty} \exp(a\tau) x(\tau) d\tau. \]

In a digital system, time is necessarily discrete (it is no longer “real time”) and Equation 3.188 may be set to successive values \( nT \) and \( nT-\tau \) where \( n \) is an integer:

\[ z(nT) = \exp(-anT) \int_{0}^{\infty} \exp(a\tau) x(\tau) d\tau \]  

(3.189)

Multiplying Equation 3.190 by \( \exp(-aT) \) and subtracting from Equation 3.189 gives

\[ z(nT) = \exp(-anT) z(nT-T) \]

\[ + \exp(-anT) \int_{0}^{\infty} \exp(a\tau) x(\tau) d\tau. \]  

(3.191)

This integral requires that an analytic expression for \( x(\tau) \) exists within the interval \((nT-T)\) to \( nT \). An approximate solution can be obtained by applying the constraint, which is in any case normally imposed by the A/D conversion, that \( x(\tau) \) is constant in the interval \( nT-T \leq \tau < nT \). Then Equation 3.191 becomes

\[ z(nT) = \exp(-anT) z(nT-T) \]

\[ + \exp(-anT) x(nT-T) \frac{\exp(anT)}{a} [1 - \exp(-aT)] \]

\[ z(nT) = \exp(-anT) z(nT-T) + \frac{[1 - \exp(-aT)]}{a} x(nT-T). \]  

(3.192)

Since time \((nT-T)\) is one unit delay from time \( nT \) the discrete transfer function can be found directly to be

\[ H(z^{-1}) = \frac{Z(z^{-1})}{X(z^{-1})} = \frac{1}{a} \frac{[1 - \exp(-aT)]z^{-1}}{1 - \exp(-aT)z^{-1}}. \]  

(3.193)

Having established the transfer function for the low-pass section, the other elementary sections can be derived. For instance, the bandpass can be expressed as the difference of two low-pass sections. Of particular interest is the high-pass section, which has a transfer function

\[ H(p) = \frac{p}{p + a} = 1 - \frac{a}{p + a}. \]  

(3.194)

Thus the high-pass is realized as a direct link of the profile in parallel with and opposite to a low-pass section. However, in the discrete system, if this approach were used, the fact that time is not continuous means that the signal applied to the low-pass section could not affect the output until a finite time later and so would not combine properly with the direct-link signal. The filter response would thus deteriorate, so it is necessary, for best performance, to introduce a compensating delay in the direct-link path. To find the required delay, it is necessary to estimate the phase angle of the discrete low-pass section. Using Equations 3.192 and 3.194 and grouping the constants gives
\[ H(p) = \frac{A \exp(-pT)}{1 - B \exp(-pT)}. \]  

(3.195)

Substituting \( p = j \) and applying de Moivre's theorem gives

\[ H(j\omega) = \frac{A(\cos \omega T - j \sin \omega T)}{1 - B \cos \omega T + j B \sin \omega T} \]  

(3.196)

so that the phase angle is

\[ \varphi = -\omega T - \beta - \tan^{-1}\left( \frac{B \sin \omega T}{1 - B \cos \omega T} \right). \]  

(3.197)

where \( \beta \) is a constant.

For \( B \) close to unity, which is equivalent to the product \( |aT| \) being much less than unity, Equation 3.197 simplifies to

\[ \varphi = -\omega T - \beta - \tan^{-1}\left( \cot \frac{\omega T}{2} \right). \]  

(3.198)

From this, the group delay of the section is

\[ -\frac{d\varphi}{d\omega} = \frac{T}{2} \]  

(3.199)

so that a delay of half a sampling period should be introduced into the direct path. The high-pass section now has the form

\[ h_{hp}(z^{-1}) = z^{1/2} - \frac{A z^{-1}}{1 - B z^{-1}} \]  

(3.200)

which is not a realizable linear digital filter since its transfer function is not rational in \( z^{-1} \). A further transformation is therefore required and a suitable one is \( z^{-1} \rightarrow z^{-2} \) which in the continuous filter would be \( p \rightarrow 2p \). Thus, to retain the same frequency response under this transformation, the cut-off frequency must be doubled and the fully realized high-pass section is

\[ H_{hp}(z^{-1}) = z^{-1} - \frac{[1 - \exp(-2aT)] z^{-2}}{1 - \exp(-2aT) z^{-2}}. \]  

(3.201)

Thus the required two-stage high-pass filter has the transfer function \[30\]

\[ H(z^{-1}) = \left| z^{-1} - \frac{[1 - \exp(-2aT)] z^{-2}}{1 - \exp(-2aT) z^{-2}} \right|. \]  

(3.202)

The discrete transfer function of Equation 3.201 has a pole near the Nyquist frequency (\( 1/2T \) Hz) which is caused by the frequency-doubling procedure of the foregoing analysis and so will not be usable near this frequency.

It is the correction for group delay used here that gives the improved high-pass characteristics of the convolution integral approach over that of the impulse-invariant method. If the analysis embodied in Equations 3.196 through 3.200 were carried out on the impulse-invariant-derived transfer function, it would be found that the group delay was \( T/2 \), that is, a half time-period advance would be needed as compensation. Although this could be arranged when processing previously stored data, it is not a physically realizable system since it requires information about the future. It may be noted that this problem is similar to that encountered when a phase-corrected filter is being realized, because a true linear phase method requires both linear and group delay correction.

The single-stage low-pass section of Equation 3.193 can be rearranged to give

\[ z = AXz^{-1} + BZz^{-1} \]  

(3.203)

where \( X \) and \( Z \) are the input and output sequences, respectively.

The high-pass section described by Equation 3.201 can be implemented as it stands or multiplied out to give

\[ H_{hp}(z^{-1}) = \frac{z^{-1} - Az^{-2} - Bz^{-3}}{1 - Bz^{-2}}. \]  

(3.204)

It will be noted that the parallel representation, Equation 3.193, requires less arithmetic than the canonical form. In general, therefore, this approach should be both faster and more accurate.

### 3.4.6.2 An Example

A particular application of the recursive 2CR filter is in the fast assessment of profiles. With the limited memory size available at the time, the saving of the need to store a weighting function is significant and, for on-line work, the filter should preferably work in “real time.” In this sense, “real time” means that the time taken to process one sample in the filter is less than the smallest sampling period used. For the reasons discussed in the previous sections, the chosen method of implementation could be a cascade implementation of the “convolution-approximation” section.

The requirements of other programs with which the filter has to be used demands that, initially, there should be 400 samples per cut-off. (This requirement is to ensure that a reasonably well-defined 100:1 bandwidth for average wavelength measurements could be produced.) However, at small cut-offs (sampling lengths) this represents samples considerably less than one stylus width apart and requires high sampling rates. (It is likely that, in general, less severe restrictions will be imposed.) As an example, for the two-stage filter to demonstrate 50% transmission at the cut-off it can be shown that for each stage

\[ a = \frac{2\pi}{k\lambda_x} \]  

(3.205)

or, if there are \( N \) ordinates per cut-off, \( a = 2\pi/NT \).

Thus for an \( N \) of 400, the \( \exp(-2aT) \) will be very close to unity and, consequently, \( 1 - \exp(-2aT) \) will be near zero. The multipliers that are used in the filter are fractional and
so cannot be handled by the integer arithmetic. There are methods of pre- and post-scaling coefficients and coordinates which allow fractional quantities to be handled but in this case the near zero value is likely to introduce considerable errors. In the interests of accuracy it may be better to use floating point arithmetic although, since this depends on computer software, it is slower in operation.

Recursive methods can only be used in cases where the discrete transfer function can be written in terms of a rational ratio of polynomials such as

\[ H(z^{-1}) = \frac{a_0 + a_1 z^{-1} + a_2 z^{-2} \ldots}{1 + b_1 z^{-1} + b_2 z^{-2} \ldots} \quad (3.206) \]

The question arises of whether this can be suitable for linear phase (phase-corrected) filters. Another point is whether the poles and zeros of Equation 3.206 are realizable. In practice the latter is decided by the transmission characteristics of the filter and whether or not it suffers discontinuities. From the earlier discussion on optimized filters it is clear that the 3:1 and 2:1 filters are not well behaved in so far as they have discontinuous slope in the transform and so cannot be directly implemented in a recursive mode. However, the fact that these filters cannot be implemented does not mean that linear phase filters as a whole cannot. One simple way of implementing filters of the kind shown above relies on the ability to describe the transfer function in a square root mode.

Remember that, if the impulse response is \( h(t) \) whose transform is \( H(\omega) \) then reversing the output can be shown to be equivalent to conjugation, with an additional linear phase term; that is, if \( h(t) \Rightarrow H(\omega) \) then

\[ h(T-t) \Leftrightarrow \exp(-j\omega T)H^*(\omega) \quad (3.207) \]

where \( H^*(\omega) \) is the complex conjugate of \( H(\omega) \). Hence, passing the signal through a filter, storing the output, reversing it and passing it back through the filter yields

\[ H(\omega)H^*(\omega)\exp(j\omega T) = |H(\omega)|^2 \exp(-j\omega T). \quad (3.208) \]

If a recursive method is possible then this method is very fast.

### 3.4.7 USE OF THE FAST FOURIER TRANSFORM IN SURFACE METROLOGY FILTERING

#### 3.4.7.1 Areal Case

The FFT routine can be used very effectively in surface metrology because it is very easy to implement any desired filter characteristic. Because the input data (i.e., the surface) is real, the storage is reduced by a factor of 2 relative to conventional complex data. Once transformed the desired amplitude characteristic is simply a multiplication of the transform data by the required characteristic and then this is followed by an inverse FFT to reconstitute the filtered signals. Phase effects are invariant if the linear phase filter is being simulated.

As shown in the earlier section on FFTs there is a very big speed advantage of filtering in this way. However, there is the problem that all the data has to be stored before filtering can take place. This is not the case in the direct convolution method, so depending upon the storage capability either one or the other can be used. It is invariably a compromise between storage, speed, and accuracy. From the latter point of view the effective weighting function length is, as in the recursive method, equal to the length of the data.

Whatever is true for profile filtering is even more true of 2D filtering. One method due to Rabiner and Gold [8] allows a 2D (i.e., areal) filtering to take place in terms of profile filtering methods. If \( x, y \) are integer numbers in the \( xy \) plane and \( v, \omega \) in the frequency domain, then the areal transform of \( z(x,y) \) is \( Z(v,\omega) \) where

\[
Z(v,\omega) = \sum_{x=0}^{N_x-1} \sum_{y=0}^{N_y-1} z(x,y) \exp \left[ -j \left( \frac{2\pi \omega}{N_1} x + \frac{2\pi \nu}{N_2} y \right) \right]
\]

or

\[
Z(v,\omega) = \sum_{x=0}^{N_x-1} \exp \left[ -j \left( \frac{2\pi \omega}{N_1} x \right) \right] \left\{ \sum_{y=0}^{N_y-1} z(x,y) \exp \left[ -j \left( \frac{2\pi \nu}{N_2} y \right) \right] \right\}.
\]

The term in braces in Equation 3.210 is a series of \( N_1 \) 1D DFTs obtained by varying \( x \) from 0 to \( N_1 - 1 \) as shown in Figure 3.37.

So the transforms for each profile can be taken with \( x \) fixed for each profile, yielding \( N_1 \) transforms. These give the transform for a given \( \omega \) given \( Z(v,\omega) \). Once \( Z(v,\omega) \) is mapped it can be multiplied (because it is in the frequency domain) by whatever 2D filtering characteristic is required, so fast convolution using the FFT is probably the most important means for realizing 2D filters. Note, however, that every one of the profile graphs shown in the figure has to have mechanical fidelity relative to each other. It is no good measuring each profile using a skid instrument because this will lose the reference level for each profile and the 2D spectrum will have little macro-geometry spectrum detail.

The filtered characteristic would become

\[
Z_f(v,\omega) = Z(v,\omega)W(v,\omega)
\]

where \( W \) is the amplitude characteristic of the required filter in two dimensions. From this the filtered surface \( z_f(x,y) \) emerges as
This is much more efficient than the direct convolution. A typical number of points in a surface map is 250,000. To give some idea of the advantage in time of this method, although the ratio of data points (assuming 250 per profile) is 250:1 the ratio of the FFT method to the direct method is about 50:1 for a profile and about 14,000:1 for the surface area, so the gain in time is tremendous. The problem for surface metrology is visualization; a true contact acting over an area is easy to imagine and is a functional effect. Unfortunately, because its effect is non-linear it cannot readily be transformed into the frequency domain. It can therefore not be advantageous to develop convolution-type operations in the spatial domain in order to simulate contact and other functional situations.

3.5 EXAMPLES OF NUMERICAL PROBLEMS IN STRAIGHTNESS AND FLATNESS

The best-fit line has a slope \( m \) given by

\[
m = \frac{\sum x \sum z - N \sum xz}{(\sum x)^2 - N \sum x^2}
\]  
(3.213)

and \( m \) is relatively small.

If the spacings of measurement in the \( x \) direction are taken to be of equal increments then the slope becomes

\[
m = \frac{12 \sum z_i \sum i z_i - 6(N + 1) \sum z_i^2}{N(N^2 - 1)}
\]  
(3.214)

or

\[
m = \sum z_i \left( \frac{12i}{N(N^2 - 1)} - \frac{6(N + 1)}{N(N^2 - 1)} \right)
\]

where

\[
k_1 = \frac{12}{N(N^2 - 1)} \quad \text{and} \quad k_2 = \frac{6}{N(N - 1)}
\]  
(3.216)

So, from the computational point of view, each \( z \) can be multiplied by a weighting factor which, for a given fixed \( N \), can always be the same.

Similarly for flatness, the slopes become

\[
M_1 = \frac{12 \sum j = \frac{N}{2} \sum z_j}{MN(N + 1)(N + 2)}
\]  
(3.217)

\[
M_2 = \frac{12 \sum j = \frac{M}{2} \sum z_j}{MN(M + 1)(M + 2)}
\]  
(3.218)

assuming that the measurements are being taken from the center of the coordinate system.

Typical of the dilemmas that can arise is the play-off between the quantization and the sampling of a roughness waveform to determine the best-fit least-squares line. Usually one requires a small quantization interval and high sampling to get good results. But if one of these is not available it may not be advantageous to hold to the other; a mutual relaxation may sometimes be necessary. In this example, the least-squares mean line slope \( m \) as above is given by

\[
m = \frac{12 \sum N \sum z_j^2 - 6(N + 1) \sum N \sum z_j^2}{N(N^2 - 1)}
\]  
(3.219)

where the \( z \) values are profile ordinates taken at \( N \) equal unit intervals along the surface. The essential numerical point to notice is that the numerator is made up of the difference between two summations. Two quantization factors influence the value of \( m \) that is obtained; first, the resolution of the digital measurement of the analog signal, and second the resolution in the computer, that is the word length. The best way to show these effects is by a simple example.

Suppose that five measurements are taken on a surface using a very high-resolution A/D converter and let these be processed in a computer of a long word length. Let the numbers
as measured be 10.000, 10.000, 10.000, 10.000, 10.1999. Substituting these into the equation for \( m \) yields

\[
m = \frac{1811.94 - 1807.164}{120} = 0.0398 = 2.2790°.
\]

If now the same measurements had been taken with an A/D convertor capable of seeing only three decimal digits, the numbers would be 10.0, 10.0, 10.0, 10.0, 10.2 giving

\[
m = \frac{1806 - 1803.6}{120} = 0.0200 = 1.1458°.
\]

Again if a high-resolution A/D convertor was used and the computer could only work to four decimal digits (13 bits), then

\[
m = \frac{1811 - 1807}{120} = 0.0333 = 1.909°.
\]

Finally if the A/D convertor measures three digits and the computer four, then

\[
m = \frac{1806 - 1803}{120} = 0.025 = 1.432°.
\]

Four different answers have been obtained (2.279°, 1.1458°, 1.909°, and 1.432°) using what were intended to be the same data from the surface profile—nearly 100% variation!

Part of the discrepancy is due to the small variations between the numbers being suppressed by the limited resolution of the A/D convertor and part due to the small difference between the two summations in the numerator being suppressed or modified by the limited word length. Both have entered into it. Yet another factor has to be taken into account and this is the sampling rate. If there were three times as many numbers, each of the originals counted three times, what effect does this have? If the A/D convertor has three digits and the computer four, the answer should be 0.025 = 1.432° in the five-ordinate case, but it is 0.0089 = 0.511° for the 15 ordinates 10.0, 10.0, 10.0, 10.0, 10.0, 10.0, 10.0, 10.0, 10.0, 10.1, 10.1, 10.1, 10.1, which contain exactly the same information. The reason for this further discrepancy is that merely adding more numbers into each of the summations of the numerator eventually leads to the least significant numbers (which contain the vital information) being rounded off or truncated. This is due to the limited word length. Again, as before, too much sampling aggravates the problems associated with quantization.

What is to be done? A number of possibilities are open. The first is only to use independent samples; do not use redundant information by sampling too fast. Next make sure that the real variations are preserved in the computer. This can be achieved in the example above by removing an estimate of the mean value of the data from each measurement before working out the numerator. Under these circumstances, the data becomes 0, 0, 0, 0, 0.1 and the danger of the summation overflowing the word length limitation is considerably reduced. In addition to these measures, it is always possible to use double word length arithmetic if the computer allows it, and it is also possible to increase the resolution of the A/D convertor at the expense of extra work and slower data capture rates. In general it is more advantageous to anyone working on the numerical analysis problem than it is on validating the data itself. The latter is of limited use without the former.

### 3.6 ALGORITHMS

#### 3.6.1 DIFFERENCES BETWEEN SURFACE AND DIMENSIONAL METROLOGY AND RELATED SUBJECTS

#### 3.6.1.1 Least-Squares Evaluation of Geometric Elements

In assessing the departure from straightness, roundness, flatness, sphericity, cylindricity, conality, and other geometric forms the method of least-squares occupies a dominant place, being the preferred method often included in international and national standards. Although this would appear to be straightforward, it is not so because of the many algorithms that could be used to achieve the assessment. Depending on which is used the result could be obtained quickly and accurately or, sometimes, not at all. Also, another problem that arises is very often the parameter is not linear, in which case, prior to applying least-squares some linearization has to take place, from which an approximate solution is obtained and used in the next iteration.

In the text so far the least-squares method has been used in a number of situations. There is, however, one general approach that might prove to be useful. This is based on the technique developed at NPL by Forbes [33] for use in coordinate-measuring problems. The technique is valid for surface metrology problems providing linearization is taken into account. The algorithms employ a common approach which has stable parameterization of the elements. What has to be watched carefully is the nature of the data obtained from the measuring instrument. Surface metrology instruments are different from dimensional measuring machines. In measuring an arc, for example, both can produce different forms of data; one can usually easily be linearized, the other cannot! Very unstable results can be produced if one sort of algorithm is applied to the wrong sort of data.

The actual shapes of elements are those already considered in Chapter 2 and which are specified in BS 7172 [34].

The rationale is as follows. It concerns parameterization, which in turn concerns the way in which the problem of solving for the best-fit solution is posed.

As an example, the best-fit plane will be discussed. This plane or any other can be specified by a point in it, say \((x_0, y_0, z_0)\), and the direction cosines of the normal perpendicularly such that \(a^2 + b^2 + c^2 = 1\). Thus, any plane can be defined by six parameters which are not all independent of each other.

If a plane exists as above and it is required to see how a set of points fit to it the intuitive thing to do is to measure the
distance of each point from it. For a point \(x, y, z\) the distance \(d\) is given by

\[d = a(x - x_o) + b(y - y_o) + c(z - z_o). \tag{3.220}\]

The \(i\)th point has a distance \(d_i:\)

\[d_i = a(x_i - x_o) + b(y_i - y_o) + c(z_i - z_o) \tag{3.221}\]

and the one way to estimate the goodness of fit is to look at

\[S = \sum d^2_i. \tag{3.224}\]

This sum \(S\) depends on the parameters of the plane \(x_o, y_o, z_o\) and \(a, b, c\). These parameters have to be chosen to minimize \(S\).

The steps for a least-squares fit are therefore as follows.

1. Choose parameters to describe the position, shape and sometimes also the orientation and size of the geometrical part.
2. Derive a formula for the distance of a point from this geometrical object.
3. Express the sum of squares \(S\) in terms of a set of data points and their distances from the geometrical object.
4. Develop a stable algorithm for determining the parameters such as \(a, b, c, x_o, y_o, z_o\).

A precursor to evaluating the best-fit least-squares to improve the numerical accuracy is to find the centroid of the data points –

\[\text{centroid} = \frac{1}{n} \sum \text{point}.\]

This is a set of linear equations which can be put in the matrix form \(Au = b:\)

\[
\begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m1} & \cdots & \cdots & a_{mn}
\end{pmatrix}
\begin{pmatrix}
u_1 \\
u_2 \\
\vdots \\
u_m
\end{pmatrix}
=
\begin{pmatrix}
b_1 \\
b_2 \\
\vdots \\
b_n
\end{pmatrix}. \tag{3.224}
\]

In general \(m > n\) so that all the equations will not be satisfied simultaneously. Reliable and efficient algorithms do exist to solve Equation 3.285 in the least-squares sense:

\[A^T A u = A^T b. \tag{3.225}\]

These are the “normal equations.”

In Chapter 2 it has already been shown that, in surface metrology, this approach is really sufficient, because in the case of a circle and sphere, for example, the limaçon approach basically changes the equation problem from being that of a quadratic to that of being linear. It is this feature which differentiates the coordinate measurement from that of the surface metrology approach. Rarely is it impossible to linearize the system from the point of view of surface metrology. This is because the “surface skin,” of whatever shape, is in practice so very much smaller than the dimension or position of the geometric object. Surface metrology instruments are built to see only the former; coordinate-measuring machines have to cater for the latter.

However, because the two methods are converging, as the scale of size reduces as discussed in Chapter 8, it is informative to take the coordinate-measuring approach because ultimately surface metrology will be incorporated.

### 3.6.1.4 Eigenvectors and Singular Value Decomposition

For linear systems a good algorithm follows.

Given a square matrix \(B\), an eigenvector \(u\) of \(B\) is such that

\[Bu = \lambda u \tag{3.226}\]

for some eigenvalue \(\lambda\). The case in question is such that

\[B = A^T A, \tag{3.227}\]

as in Equation 3.226, for some \(m \times n\) rectangular matrix \(A\) where \(m > n\). In this situation a stable numerical solution is obtained by finding a “singular value decomposition” (SVD) of the matrix \(A\). In this \(A\) can be written as a product

\[A = U S V^T \tag{3.228}\]

with \(U\) and \(V\) orthogonal matrices and \(S\) a diagonal matrix containing the singular values of \(A\). If \(i\) is as in Equation 3.227 the squares of the diagonal elements of \(S\) are the eigenvalues of \(B\) and the columns of \(V\) are the corresponding eigenvectors. These are usually produced as standard output from most software implementations. SVD is now standard for many solutions. See, for example Ref. [44].

### 3.6.2 Best-Fit Shapes

#### 3.6.2.1 Planes

The steps are as follows.

1. Specify point \(x_o, y_o, z_o\) on a plane
2. Evaluate direction cosines \((a, b, c)\) of a normal to a plane; note that any point on a plane satisfies...
\begin{align*}
a(x-x_0) + b(y-y_0) + c(z-z_0) &= 0 \\
3. \text{ Evaluate distance from plane} \quad d_i &= a(x_i-x_0) + b(y_i-y_0) + c(z_i-z_0) \\
4. \text{ Describe the algorithm}
\end{align*}

The best-fit plane \( P \) passes through the centroid \( \bar{x}, \bar{y}, \bar{z} \) and this specifies a point in the plane \( P \). It is required to find the direction cosines of \( P \). For this \( (a, b, c) \) is the eigenvector associated with the smallest eigenvalue of

\[ B = A^T A \] (3.229)

where \( A \) is the \( m \times 3 \) matrix whose \( i \)th row is \( (x_i-x, y_i-y, z_i-z) \); alternatively \( (a, b, c) \) is the singular vector associated with the smallest singular value of \( A \). Thus, an algorithm to find the best-fit line in 3D is:

1. Calculate the centroid \( \bar{x}, \bar{y}, \bar{z} \)
2. Form matrix \( A \) from the data points and \( \bar{x}, \bar{y}, \bar{z} \)
3. Find the SVD (singular value decomposition) of \( A \)
   and choose the singular vector \( (a, b, c) \) corresponding to the smallest singular value. The best-fit plane
   is therefore \( \bar{x}, \bar{y}, \bar{z}, a, b, c \).

Similarly for the best-fit line to data in two dimensions.

### 3.6.2.2 Circles and Spheres

These shapes involving an axis of revolution are usually evaluated by linearization of the basic equations mechanically as stated by the process of radius suppression by mechanically shifting the instrument reference to a position near to the surface skin of the geometric element being measured. Failing this an iterative method has to be used. The Gauss–Newton iterative method can be used when the relationship between the distances \( d_i \) and the parameters \( u_i \) is non-linear. Hence an iterative scheme has to be used. This is similar to the Deming method given in Chapter 2. The situation is shown in Figure 3.38.

One iteration of the Newton algorithm for computing the zero of a function is as follows.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{gauß-newton.png}
\caption{Gauss–Newton method.}
\end{figure}

Suppose that there is a first estimate \( u_0 \) of where the function \( u \) crosses the \( u \) axis. Then:

1. Evaluate \( f(u_0) \)
2. Form a tangent to the graph at \( (u_0, f(u_0)) \)
3. Find \( u_1 \) where the tangent crosses the \( u \) axis

Then

\[ u_0 = u_0 + \frac{f(u_0)}{f'(u_0)} = u_0 + p. \] (3.230)

\( u_1 \) is now the new estimate of where \( f(u) \) crosses the \( u \) axis. This is repeated until the result is close enough to \( u^* \).

Basically the Gauss–Newton method is as follows:

Suppose there is a first estimate \( u \) of \( u^* \). Then solve the linear least-squares system

\[ Jp = -d \] (3.231)

where \( J \) is the \( m \times n \) Jacobean matrix whose \( i \)th row is the gradient of \( d_i \) with respect to \( u \), that is

\[ J_{ij} = \frac{\partial d_i}{\partial u_j}. \] (3.232)

This is evaluated at \( u \) and the \( i \)th component of \( d \) is \( d_i(u) \). Finally, the estimate of the solution is

\[ u := u + p \quad (:= \text{ means update}). \]

These steps are repeated until \( u \) is close enough to \( u^* \). Ideally, changes in the iteration should be small for this method to be quick in convergence and stable.

For example, for the best-fit circle:

1. Specify circle center \( x_0, y_0 \), radius \( r \). Note that \( (x-x_0)^2 + (y-y_0)^2 = r^2 \)
2. Obtain distance from the circle point:

\[ d_i = r_i - r \]
\[ r_i = [(x_i-x_0)^2 + (y_i-y_0)^2]^{1/2} \] (3.233)

3. The elements of the Jacobean \( J \) are

\[ \frac{\partial d_i}{\partial x_0} = -(x_i-x_0)/r_i \]
\[ \frac{\partial d_i}{\partial y_0} = -(y_i-y_0)/r_i \] (3.234)
\[ \frac{\partial d_i}{\partial r} = -1. \]
Processing, Operations, and Simulations

4. Algorithm: knowing $x_0, y_0$ and $r$ for the circle center and radius estimates, use them in a Gauss–Newton iteration. Form $J \ p = -d$ from the $d$ of Equation 3.233 and the $J$ of Equation 3.234

5. Solve

$$J \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} = -d$$

(3.235)

for $p$

6. Update the $x_0, y_0, r$ according to

$$x_0 := x_0 + p_{x0}$$
$$y_0 := y_0 + p_{y0}$$

(3.236)

$$r := r + r_i.$$  

Carry on until successful and the algorithm has converged.

The linear best-fit circle can be evaluated by an approximation described earlier used by Scott [61] (Chapter 2, Ref. [172]). In this model

$$S = \sum f_i^2$$

(3.237)

is minimized, where $f_i = r_i^2 - r^2$ rather than $r_i - r$ as in the linear case—the trick is to make the $f_i$ linear.

By changing the parameters $f$ can be made into a linear function of $x_0, y_0$, and $p = x_0^2 + y_0^2$. $r$.

$$f_i = (x_i - x_0)^2 + (y_i - y_0)^2 - r^2$$

$$= -2x_i x_0 - 2y_i y_0 - (x_0^2 + y_0^2 - r^2) + (x_i^2 + y_i^2).$$

(3.238)

Thus

$$A \begin{pmatrix} x_0 \\ y_0 \\ p \end{pmatrix} = b$$

(3.239)

(from which $x_0, y_0$, and $p$ are found) where the elements of the $i$th row of $A$ are the coefficients $(2x_i^2y_i - 1)$ and the $i$th element of $b$ is $x_i^2 + y_i^2$.

An estimate of $r$ is

$$\sqrt{x_0^2 + y_0^2} - p.$$  

(3.240)

This can be used to get a first estimate of the parameter for the non-linear method if required.

Both the linear and non-linear methods described above can be used for spheres.

### 3.6.2.3 Cylinders and Cones

It has been suggested [33] that a modified Gauss–Newton iterative routine should be used in the case of cylinders because one of the parameters is the direction of a line that is the axis of the cylinder. Such a line $x_0, y_0, z_0, a, b, c$ in 3D can be specified by four parameters together with two rules to allow the other two to be obtained:

Rule 1: represent a direction $(a, b, 1)$

Rule 2: given the direction above, ensure $z_0 = -ax_0 - by_0$

For nearly vertical lines these two rules give stable parameterization for $a, b, x_0$, and $y_0$. The problem of finding the distance of a data point to an axis is quite complicated. The following strategy is therefore followed based on the fact that for axes which are vertical and pass through the origin, $a = b = x_0 = y_0 = 0$ and all expressions become simple.

The strategy is as follows:

1. Iterate as usual but, at the beginning of each iteration, translate and rotate the data so that the trial best-fit cylinder (corresponding to the current estimates of the parameters) has a vertical axis passing through the origin.

2. This means that when it is time to evaluate the Jacobean matrix the special orientation can be used to simplify the calculations. At the end of the iteration use the inverse rotation and translation to update the parameterizing vectors $x_0, y_0, z_0, a, b, c$, and thereby determine the new positions and orientation of the axis.

#### Note:

To rotate a point $(x, y, z)$ apply a $3 \times 3$ matrix $U$ to the vector $(x, y, z)^T$; the inverse rotation can be achieved by using the transpose of $U$:

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = U \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$  

(3.241)

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = U^T \begin{pmatrix} u \\ v \\ w \end{pmatrix}.$$  

(3.241)

A simple way to construct a rotation matrix $U$ to rotate a point so that it lies on the $z$ axis is to have $U$ of the form

$$U = \begin{pmatrix} C_2 & 0 & S_2 \\ 0 & 1 & 0 \\ -S_2 & 0 & C_2 \end{pmatrix}.$$  

(3.242)

where $C_i = \cos \theta_i$ and $S_i = \sin \theta_i$, $i = 1, 2$. So if it is required to rotate $(a, b, c)$ to a point on the $z$ axis, choose $\theta_i$ so that

$$bC_1 + cS_1 = 0 \quad \text{and} \quad \theta_2 = aC_2 + (cC_1 - bS_1)S_2 = 0.$$  

These notes are only suggestions. There are other methods that can be used but these are most relevant to geometric parts like cylinders, which in surface metrology can usually be oriented to be in reasonable positions that is for a cylinder nearly vertical.
Care should be taken to make sure that the algorithm is still stable if reasonable positions for the part cannot be guaranteed.

**Cylinder**

1. Specify a point \( x_0, y_0, z_0 \) on its origin, a vector \( a, b, c \) pointing along the axis and radius \( r \).
2. Choose a point on the axis. For nearly vertical cylinders
   \[
   z_0 = -ax_0 - by_0 \quad c = 1.
   \]

3. Distance of the chosen point to cylinder is
   \[
   d_i = r_i - r \quad r_i = \frac{(u_i^2 + v_i^2 + w_i^2)^{1/2}}{(a^2 + b^2 + c^2)^{1/2}}
   \]
   where
   \[
   u_i = c(y_i - y_0) - b(z_i - z_0) \\
   v_i = a(z_i - z_0) - c(x_i - x_0) \\
   w_i = b(x_i - x_0) - a(y_i - y_0).
   \]

To implement the Gauss–Newton algorithm to minimize the sum of the square distances the partial deviation needs to be obtained with the five parameters \( x_0, y_0, a, b, r \) (the five independent variables for a cylinder). These are complicated unless

\[
x_0 = y_0 = a = b = 0 \quad \text{in which case} \quad r_i = \sqrt{u_i^2 + v_i^2}\]

\[
\frac{\partial d_i}{\partial x_0} = -x_i/r_i \\
\frac{\partial d_i}{\partial y_0} = -y_i/r_i \\
\frac{\partial d_i}{\partial a} = -x_iz_i/r_i \\
\frac{\partial d_i}{\partial b} = -y_iz_i/r_i \\
\frac{\partial d_i}{\partial r} = -1.
\]

**Algorithm Operation**

1. Translate data so that the point on the axis lies at the origin:
   \[
   (x_i, y_i, z_i) := (x_i, y_i, z_i) - (x_0, y_0, z_0).
   \]

2. Transform the data by a rotation matrix \( U \) which rotates \( a, b, c \) to a point on the \( z \) axis:
   \[
   \begin{pmatrix}
   x_i \\
   y_i \\
   z_i
   \end{pmatrix} := U
   \begin{pmatrix}
   x_i \\
   y_i \\
   z_i
   \end{pmatrix}.
   \]

3. Form the right-hand side vector \( d \) and Jacobean according to Expressions 3.244, 3.246, and 3.247.
4. Solve the linear least-squares system for \( P_{a0} \) etc.

5. Update the parameter estimates to
   \[
   \begin{pmatrix}
   a \\
   b \\
   c
   \end{pmatrix} := U^T \begin{pmatrix}
   P_a \\
   P_b \\
   1
   \end{pmatrix}.
   \]

These steps are repeated until the algorithm has converged. In Step 1 always start with (a copy of) the original data set rather than a transformed set from the previous iteration.

If it is required to have \( (x_0, y_0, z_0) \) representing the point on the line nearest to the origin, then one further step is put in:

\[
\begin{pmatrix}
   x_0 \\
   y_0 \\
   z_0
   \end{pmatrix} := U^T \begin{pmatrix}
   a \\
   b \\
   c
   \end{pmatrix}.
   \]

(See Forbes [33] for those situations where no estimates are available.) Luckily, in surface metrology, these iterative routines are rarely needed. Also it is not yet clear what will be the proportion of workpieces in the miniature domain that will have axes of centro-symmetry. Until now, in micro-dynamics the roughness of rotors and stators has made it difficult to measure shape. However, there is no doubt that shape will soon be a major factor and then calculations such as the one above will be necessary.

**Cones**

These can be tackled in the same way except that there are now six independent parameters from \( (x_0, y_0, z_0), (a, b, c)\),
\[ \phi, \quad t, \quad \text{where } t \text{ is shown in Figure 3.39}. \]

\[ z_0 = S_0 - ax_0 - by_0 \]  

(3.252)

for some constraint \( S_0 \), which is position sensitive and the choice of which has to be determined or specified with care depending on whether the cone is narrow angle or wide angle.

The distance of the point from the cone, \( d_i \), is given by

\[ d_i = e_i \cos(\phi/2) + f_i \sin(\phi/2) - t, \]  

(3.253)

where \( e_i \) is the distance from \( x_i, y_i, z_i \) to the line specified by \((x_0, y_0, z_0)\) and \((a, b, c)\).

Again, as for the cylinder, making \( x_0 = y_0 = a = b = 0 \):

\[ e_i = r_i = \sqrt{x_i^2 + y_i^2} \]  

(3.254)

\[ f_i = z_i - S_0 \]  

and

\[ \frac{\delta d_i}{\delta x_0} = -x_i \cos(\phi/2)/r_i \]

\[ \frac{\delta d_i}{\delta y_0} = -y_i \cos(\phi/2)/r_i \]

\[ \frac{\delta d_i}{\delta x_i} = -x_i w_i/r_i \]  

(3.255)

\[ \frac{\delta d_i}{\delta y_i} = -y_i w_i/r_i \]

\[ \frac{\delta d_i}{\delta \phi} = +w_i/2 \]

\[ \frac{\delta d_i}{\delta t} = -1 \]

with

\[ w_i = (z_i - S_0) \cos(\phi/2) - t \sin(\phi/2). \]  

(3.256)

For most cones \( S_0 \) is chosen such that

\[ S_0 = \overline{z} - \overline{r} \tan(\phi/2). \]  

(3.257)

**Algorithm Description**

For cones with moderate apex angle \((< 0.9\pi)\) let \( S_0 = 0 \):

1. Translate data so that the point on the axis lies at the origin:

\[ (x_i, y_i, z_i) := (x_i, y_i, z_i) - (x_0, y_0, z_0). \]

2. Transform the data by a rotation matrix \( U \) that rotates \((a, b, c)\) to a point on the \( z \) axis:

\[ \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} := U \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix}. \]  

(3.258)

3. Form RHS vector \( d \) and \( J \) according to the above with \( S_0 = 0 \)

4. Solve the linear least-squares system

\[ \begin{pmatrix} P_0 \\ P_y \\ P_z \end{pmatrix} = -d. \]  

(3.259)

5. Update the parameter estimate

\[ \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix} := \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix} + U^T \begin{pmatrix} P_0 \\ P_y \\ -P_y P_0 - P_z P_z \end{pmatrix} \]  

(3.260)

\[ \begin{pmatrix} a \\ b \\ c \end{pmatrix} := U^T \begin{pmatrix} P_0 \\ P_y \\ 1 \end{pmatrix} \]  

(3.261)

\[ \phi := \phi + P_y \]

\[ t = t + P_z. \]  

(3.262)

6. This step is the same as for the cylinder

For the special case when the periphery of the component is incomplete, the best-fit are has to be generated in order to provide a reference. This has been developed earlier [36]. It suffices here to repeat the formula. The formula used is the
simplified version where the angular reference is taken to be
the bisector of the arc angle $\theta_0$.

Thus the center of curvature and radius are given by

\[
\begin{align*}
\bar{x} &= \left( \int_0^\theta r \cos\theta d\theta - \frac{\sin\theta}{\theta} \left( \int_0^\theta r d\theta \right) \right) \left( \theta_2 + \frac{\sin\theta_0}{\theta_2} \right) \\
\bar{y} &= \left( \int_0^\theta r \sin\theta d\theta \right) \left( \theta_2 - \frac{1}{2} \sin2\theta \right) \\
R &= \frac{1}{2\theta} \left( \int_0^\theta r d\theta - 2\bar{x}\sin\theta \right).
\end{align*}
\]

(3.263)

These results have already been given in Chapter 2 in the
roundness section. They can obviously be extended to the
3D case.

Equation 3.263 gives the best-fit conditions for a partial
arc which can enclose any amount of the full circle. Often it
is necessary to find the unique best-fit center to a concentric
pair of circular arcs. This involves minimizing the total sum
of squares of the deviations. Arc 1 has sum of squares $S_1$ and
arc $S_2$.

\[
S_1 = \sum_{i=1}^M (r_{ij} - R_1 - \bar{x}\cos\theta_j - \bar{y}\sin\theta_j)^2
\]

(3.264)

\[
S_2 = \sum_{j=1}^N (r_{ij} - R_2 - \bar{x}\cos\theta_j - \bar{y}\sin\theta_j)^2.
\]

Minimizing $S_1 + S_2$ and differentiating these polar equations
with respect to $\bar{x}$, $\bar{y}$, $R_1$ and $R_2$ gives

\[
\begin{align*}
\sum \cos^2\theta_i + \sum \cos^2\theta_j &= \sum \sin\theta_i \cos\theta_j + \sum \sin\theta_j \cos\theta_i \\
\sum \sin\theta_i \cos\theta_j + \sum \sin\theta_j \cos\theta_i &= \sum \sin^2\theta_i + \sum \sin^2\theta_j \\
\sum \cos\theta_i &= \sum \cos\theta_j \\
\sum \cos\theta_j &= \sum \cos\theta_i \\
\sum r_{ij}\cos\theta_i + \sum r_{ij}\cos\theta_j &= \sum r_{ij}\sin\theta_i + \sum r_{ij}\sin\theta_j \\
\sum r_{ij}\sin\theta_i + \sum r_{ij}\sin\theta_j &= \sum r_{ij}\cos\theta_i + \sum r_{ij}\cos\theta_j.
\end{align*}
\]

These equations are useful when data are available in the
polar form. But when data are available in the Cartesian form
the other criterion, namely minimizing the deviation from
the property of the conic, is useful as described below. In this
case, the equations of the arcs are written as

\[
x^2 + y^2 - ux - vy - D_1 = 0
\]

\[
x^2 + y^2 - ux - vy - D_2 = 0
\]

(3.266)

and the total sum of the squares of the deviations from the
property of the arc/conic are defined as

\[
E_S = \sum (x_i^2 + y_i^2 - u x_i - v y_i - D_1)^2
\]

\[
+ \sum (x_j^2 + y_j^2 - u x_j - v y_j - D_2)^2
\]

(3.267)

where $D_i$ are dummy radii. Differentiating partially with
respect to $u$, $v$, $D_1$, $D_2$, the equations in matrix form to find
the solution of $u$, $v$, $D_1$, $D_2$ are given by

\[
\begin{bmatrix}
\sum x_i^2 + \sum x_j^2 & \sum x_i x_j + \sum x_j x_i & \sum x_i & \sum x_j & 1 & 1 & M & 0 & D_1 \\
\sum x_i y_i + \sum x_j y_j & \sum y_i^2 + \sum y_j^2 & \sum y_i & \sum y_j & 1 & 1 & M & 0 & D_2 \\
\sum x_i & \sum y_i & M & 0 & 1 & 1 & N & \bar{x} \\
\sum x_j & \sum y_j & M & 0 & 1 & 1 & N & \bar{y}
\end{bmatrix}
\]

(3.268)

\[
\begin{bmatrix}
\sum (x_i^2 + y_i^2) x_i + \sum (x_j^2 + y_j^2) x_j \\
\sum (x_i^2 + y_i^2) y_i + \sum (x_j^2 + y_j^2) y_j \\
\sum x_i^2 + \sum y_i^2 \\
\sum x_j^2 + \sum y_j^2
\end{bmatrix}
\]

(3.265)

Then

\[
\bar{x} = u / 2 \quad \bar{y} = v / 2
\]

\[
R_1 = \sqrt{D_1 + (u^2 + v^2) / 4}
\]

\[
R_2 = \sqrt{D_2 + (u^2 + v^2) / 4}.
\]
Obviously the key to solving these sorts of problems is how to make the equations linear enough for simple solution. This is usually done automatically by the choice of instrument used to obtain the data. The fact that a roundness instrument has been used means that the center \( a, b \) is not far from the axis of rotation—which allows the limaçon approximation to be valid. If a Coordinate measuring machine (CMM) had been used this would not be the case unless the center portions were carefully arranged.

The best-fit methods above have hinged on the best-fit limaçon technique because this is the natural way in which a roundness instrument sees the signal. Should the data be obtained with the radius not suppressed it can be treated as a circle. The equation for minimization then becomes

\[
k(x^2 + y^2) + ux + vy - 1 = 0
\]

and the sum of errors \( S \) is

\[
S = \sum [k(x_i^2 + y_i^2) + ux_i + vy_j - 1]^2.
\]

Differentiating partially with respect to \( k, u, \) and \( v \) to minimize \( S \), the sum of squared residuals gives the matrix form

\[
\begin{bmatrix}
\sum (x_i^2 + y_i^2)^2 & \sum x_i (x_i^2 + y_i^2) & \sum y_i (x_i^2 + y_i^2) \\
\sum x_i (x_i^2 + y_i^2) & \sum x_i^2 & \sum x_i y_i \\
\sum y_i (x_i^2 + y_i^2) & \sum x_i y_i & \sum y_i^2
\end{bmatrix}
\begin{bmatrix}
k \\
u \\
v
\end{bmatrix}
= \begin{bmatrix}
\sum (x_i^2 + y_i^2) \\
\sum x_i \\
\sum y_i
\end{bmatrix}, \tag{3.270}
\]

Then the unknowns \( \bar{x}, \bar{y}, \) and \( R \) are given by

\[
\bar{x} = u/(2k) \quad \bar{y} = v/(2k) \quad R = (\sqrt{\bar{x}^2 + \bar{y}^2} + 1)/k.
\]

An alternative least-squares approach will be given at the end of the chapter for all shapes.

### 3.6.3 Other Methods

#### 3.6.3.1 Minimum Zone Method

The other basic method adopted in most standards is the minimum zone method. This is found by an iterative method based on the simplex method.

This method is a search routine (of which the Steifel exchange is just a subset) which is designed to climb mathematical maxima or minima. The simplex figure is obtained from the geometrical figure used in the search process. In 2D it is an equilateral triangle and in 3D it is a triangular pyramid or tetrahedron.

The basic principle is as follows:

1. Express the problem of minimization in mathematical terms:

\[
R_i = [(x_i - x)^2 + (y_i - y)^2].
\]

The objective or object function is to get a function, say \( B \), minimized, that is

\[
B = \min[\max R_i - \min R_i].
\]

2. Select a set of feasible values for the independent variables and use this as a starting point. This starting point is invariably chosen as the best-fit center, that is

\[
\bar{x} = 2 \sum x_i / N \quad \bar{y} = 2 \sum y_i / N.
\]

3. Evaluate the objective function from this center, that is the \( B \) value from \( \bar{x}, \bar{y} \)

4. Choose, by means of the simplex figure, a second location near to \( \bar{x}, \bar{y} \)

5. Compare the objective. If it is smaller move to this new point; if not try again

Rules for minimization:

1. Reject highest point of \( B \)
2. Do not return to original point
3. Terminate after a given number of iterations

This is one of many hill-climbing techniques but it is probably the simplest.

#### 3.6.3.2 Minimax Methods—Constrained Optimization

The criteria for parameters encountered in surface metrology fall into two categories: those based on least squares and those based on peaks and valleys or extrema.

Consider, for example, roundness. Typical references are the best-fit circle, the minimum circumscribing circle, the maximum inscribed circle and the minimum zone. The zonal methods, unlike the best-fit methods, rely for their basis on a limited number of high peaks and valleys rather than all of the data, as is the case with the best-fit methods. The problem in the extreme cases is, fundamentally, one of some sort of maximization or minimization, subject to a number of constraints. Thus, for example, the requirement for finding the plug gauge circle is to maximize the radius of a circle subject to the constraint that all data points lie on or outside the circle. In other words, the requirement for finding the plug gauge circle (or limaçon) is to find the maximum “radius” for which a limaçon may be constructed such that the limaçon lies completely inside the data representing the data being measured.
A number of methods are available for finding the optimum radius or zone subject to such constraints. Originally the assessment was based on a trial-and-error estimate. For roundness measurement, this involved drawing many circles on a chart in order to get the center which maximized, minimized, or optimized the circle or zone.

### 3.6.3.3 Simplex Methods

This technique, although crude, could give answers accurate to a few per cent providing that the workpiece was reasonably centered—in which case it is acceptable to use compasses and to draw circles on the chart.

There is a simple method which gives a good approximation to the true method. As an example of this consider the roundness profile shown in Figure 3.40. Imagine that the minimum circumscribing circle (ring gauge method) is being used.

In all problems like this there are a number of ways of tackling the situation. Often there is a problem between technological relevance and mathematical elegance. In metrology the former has to take precedence. It has been pointed out many times by Lotze [37] that agreement on terminology and methodology should pre-empt any agreement on which algorithms to use for the calculation of any of the parameters in dimensional as well as in surface metrology. This is very true because as yet there is no agreement as to the procedures, let alone algorithms, for use in minimax routes (for example [38]). This has caused a lot of confusion between surface metrologists and coordinate-measuring machine metrologists.

To illustrate this, a simple approach will be given to show that it is easily possible to get to within a reasonable range of the true value in a search routine. This approach will be followed by demonstrating that it is one application of linear programing. The basic elements of linear programing in metrology will be reviewed together with an indication of the concept of the elegant dual linear programing methods for use in metrology. Finally, an example of an algorithmic approach will be given following Dhanish and Shunmugam Ref. [39] which could, in principle, be applied to dimensional and surface metrology, with the proviso that great care in setting up the components is taken.

Select a point in Figure 3.40 at O₁. This can be arbitrary or it can be based on a least-squares center [36]. Draw a circle from it with a minimum radius to touch just one point P₁. Move in the direction of OP' at each step sweeping a radius around and reducing it until two points are at the same minimum radius. Then move the center along the bisector again reducing the radius until a further point P₁ is touched. The center position O₁ is then very often the exact center frame on which the ring gauge circle is centered. This is because only three points are needed to define a circle, the center coordinates and the radius, so three contacting points are required. The same is true for the maximum inscribed circle. Four points are needed for the minimum zone circle: two for the center and two for the two radii. A sphere needs four, a cylinder five, etc.

The method described above is not necessarily unique, even for the circumscribing circle, which should, in principle, give a unique center. The plug gauge method does not, neither does the minimum zone.

It should be pointed out here that in the case of roundness measurement it should be limacons and not circles that are being described so that the evaluation should be of the least circumscribing limacon, etc.

Although the method described above works in almost all cases it is not formal. The best method for constrained optimization is to use linear programing techniques. The simple method outlined above contains the basic elements of the technique, so for completeness the basic method will be outlined below and then followed by some examples. Although the basic idea for this optimizing technique has been used in the past quite independently [40], the formal derivation has been achieved by Chetwynd [41] and his terminology will be followed here. However, other attempts have also been used. It will become obvious that the subject is quite complicated and that a number of different approaches to get a workable, cheap, and fast solution are possible.

Linear programing implies constrained optimization involving either minimization or maximizing a function (called an objective function) while keeping other relationships within predefined bounds. If these relationships are linear, then they can be expressed as a set of linear parameters and the optimization becomes linear. This so-called linear programing is fundamental to the understanding and operation of the “exchange”-type algorithms which are now used extensively in metrology. See, for example, Figure 3.42.

Take the measurement of straightness, for example. The criteria, expressed in instrument coordinates, are, given a sequence of Cartesian datum points \((x_i, y_i)\), and the minimum zone value \(Z\)

\[
\text{Minimize } Z = h \quad \text{subject to } mx_i + c + h \geq y_i \\
mx_i + c - h \leq y_i
\]

(3.271)

**FIGURE 3.40** Simple algorithm for minimum circumscribing circle.
for all \((x, y)\) simultaneously. This illustrates a convenient parameterization, namely a single line (slope \(m\), intercept \(c\)) together with a zone of acceptability of width \(2h\) set symmetrically about it. Equation 3.271 is a linear program in \((m, c, h)\). (It is also a simple form of the minimax polynomial fit for which the so-called Steifel exchange algorithm offers an efficient solution. This may be derived from, and owes its efficiency to the properties of, the associated linear program, Figure 3.42.)

Standards present a method, originally for calculating the least-squares parameters, which has been extensively studied and is known as the “limaçon approximation” for roundness measurement. Referring to the notation adopted, the eccentric circle is reduced, providing \(R\), to

\[
\rho = a \cos \theta + b \sin \theta + R.
\]

This is a linearization of the parameters about the origin, which is produced mechanically by the instrument. Linearization about any other point involves considerable extra complexity of the coefficients. Whenever the limaçon approximation is valid in surface metrology the calculation of limiting reference circles becomes a linear program. For example, the minimum circumscribing figure to a set of data points \((r, \theta)\) is expressible as

\[
\text{minimize } Z = e \quad \text{subject to } a \cos \theta + b \sin \theta + R \geq r \quad (3.273)
\]

for all \(i\). Others may be expressed in the form of either Equation 3.272 or Equation 3.273.

Before proceeding to develop algorithms from these formulations, it is useful to establish a practical context and a mathematical notation by first illustrating the earlier work on reference circles and reviewing, extremely briefly, the main points of linear programing theory following Chetwynd [41].

3.7 BASIC CONCEPTS IN LINEAR PROGRAMING

3.7.1 GENERAL

A linear program is an optimization in which the objective function (e.g., minimizing a zone or maximizing a radius) and all the constraints are linear in the parameters. Using vector notation for brevity, it can be expressed as

\[
\text{maximize } Z = c^T x \quad \text{subject to } Ax \leq b
\]

where, for \(m\) positive parameters, \(x\), and \(n\) constraints, \(c\) is an \(m\) vector, \(b\) an \(n\) vector and \(A\) an \(m \times n\) matrix.

It is known (there is extensive literature on this subject) that the optimum solution occurs when each of the constraints \((c)\) is satisfied to its limit by one of the parameters. Hence only certain combinations of parameter values need be examined. An orderly search through these is obtained by using the simplex method in which iterations involve only elementary row operations on the matrix–vector representation. Simplex organizes these vectors as a partitioned matrix (a tableau). This has the form

\[
K \begin{bmatrix} c \\ b^T \end{bmatrix} Z
\]

where \(K\) is \(A\) augmented by an \(n \times n\) identity matrix and \(c\) is correspondingly extended by \(n\) zero elements. This adds \(n\) “slack variables” to the original parameters. If the \(i\)th parameter is limiting a particular constraint, the column \(K_i\) in \(K\) will have value 1 in the row corresponding to that constraint and zero in all other elements. The set of defining parameters so identified form the “basis.” Initially the basis is the \(n\) slack variables. Iterations attempt to match parameters to constraints in such a way that \(Z\) is rapidly maximized. It is usual always to maintain the feasibility of the current iteration by ensuring that no constraint is ever violated that is that no element of \(b\) becomes negative. This is one of the problems not easily addressed in the simplex method. The prime indicates the vector that currently occupies the position originally occupied by \(b\). At each iteration, the largest positive element of \(c^T\) is chosen and its column brought actively into the solution (this is the strategy of “steepest descent”). When no positive elements remain in \(c^T\), optimality has been achieved and the solution values are readily interpreted from the tableau.

At any iteration, the columns which originally consisted of the identity matrix carry a complete and interpretable record of the row transformations carried out on the tableau. Likewise, the columns of the current basis carry the same information in the inverse of their original form. The computationally efficient method of revised simplex does not update the full tableau but merely notes what would have been done at each iteration. The work required relates to that of inverting \(n \times n\) matrices. It may, therefore, be advantageous to use a dual program. For any \(m \times n\) linear program (termed the primal), an \(n \times m\) dual can be defined as

\[
K \begin{bmatrix} c \\ b^T \end{bmatrix} Z
\]

where \(K\) is now the augmented form of \(A^T\) (compare Equation 3.275) and the optimization has changed from minimization to maximization or vice versa. It contains exactly the same information as before, subject to the correct relative interpretation of specific elements.

3.7.2 DUAL LINEAR PROGRAMS IN SURFACE METROLOGY

Straightness, flatness, and all routine roundness measurements involve reference fittings which appear naturally as linear programs. For more complex geometries, the errors inherent in parameter linearization may be judged acceptable when weighed against the computational efficiency of simplex. All the resulting formulations essentially have features in common indicating that the dual program will offer the most efficient solutions.
The sign of the parameters required for simplex cannot be guaranteed with metrological data and so each parameter is replaced by a pair having equal magnitude but opposite sign. Even then the number of constraints usually dominates the number of parameters. Thus, a circumscribing limaçon fit involves six parameters and the minimum zone seven, but typical measurements involve several hundred profile points each generating a constraint; each generates two in the case of the minimum zone because it may contribute to the inner or outer circles of the zone. The sources of the difficulties encountered with early attempts at circle fittings are now apparent. They did not exploit the simplex method of searching only certain basic solutions and, furthermore, they worked with a primal or conventional formulation involving, say, six parameters and 500 constraints or points, rather than dual which, while having 500 parameters, has only six constraints. This makes the computing long-winded, so, in moving from the primal to the dual, the roles of vectors $b$ and $c$ are interchanged. If at any iteration the dual is maintained in a feasible condition (all elements of $c$ positive), the corresponding primal would be interpreted as being in an optimal, but generally infeasible, condition. The implications of dual feasibility are critical to what is to follow. Consider a physical interpretation for the case of a circumscribing limaçon (or circle). The primal feasibility condition amounts to starting with a figure which is too large but which certainly encloses the profile and then shrinking it to the smallest radius that still closes the profile. Dual feasibility would entail initially choosing a figure which is the smallest to enclose some of the data points and then expanding it as little as possible so as to include all the data—the same problem looked at a different way.

**Note:**

If a primal has three parameters, the dual has three constraints. The corresponding geometric observation is that a circle is defined by exactly three contacts with the data, which makes physical sense.

### 3.7.2.1 Minimum Radius Circumscribing Limaçon

Transferring the primal, geometrical statement of the minimum radius circumscribing limaçon to the dual, the initial tableau can be written as a minimization:

$$
\begin{bmatrix}
\cos \theta_1 & \ldots & \cos \theta_j & \ldots & \cos \theta_n & 0 \\
\sin \theta_1 & \ldots & \sin \theta_j & \ldots & \sin \theta_n & 0 \\
1 & \ldots & 1 & \ldots & 1 & 1 \\
-r_1 & \ldots & -r_j & \ldots & -r_n & 1
\end{bmatrix}
$$

At any iteration giving a feasible solution, the basis will be formed from three of these columns, so taking three general contact points at $\theta_i$, $\theta_j$, and $\theta_k$ means that the basis $\beta$ is given by

$$
\beta^{-1} = \begin{pmatrix}
\cos \theta_i & \cos \theta_j & \cos \theta_k \\
\sin \theta_i & \sin \theta_j & \sin \theta_k \\
1 & 1 & 1
\end{pmatrix}.
$$

No significance (such as $\theta_i < \theta_j$, for example) can be read into this matrix; the relative positioning of columns depends upon the workings of revised simplex in previous iterations. The determinant of $\beta^{-1}$ is given by the sum of the cofactors of its third row, that is by the same cofactors which identify the elements of the third column of $\beta$. The non-negativity of the elements of the third column of $\beta$, thus, requires that these cofactors, $\Delta_j$, $\Delta_k$, $\Delta_l$, must have the same sign where

$$
\Delta_j = \begin{vmatrix}
\cos \theta_i & \cos \theta_j \\
\sin \theta_i & \sin \theta_j \\
1 & 1
\end{vmatrix} = \sin(\theta_j - \theta_i)
$$

and similarly for the others. Using Cartesian coordinates, the cofactor can be expressed as

$$
\Delta_j = \frac{1}{r_j} \begin{vmatrix}
x_i & x_j \\
y_i & y_j
\end{vmatrix}
$$

and related to this is a function

$$
\Delta_j = \frac{1}{r_j} \begin{vmatrix}
x_i & x_j \\
y_i & y_j
\end{vmatrix} = 0
$$

which (apart from an indeterminacy at the origin, of little importance here) is a straight line passing through $(x_i, y_i)$ and $(0, 0)$ and dividing the $xy$ plane into the two areas where $\Delta_j > 0$ and where $\Delta_j < 0$. The line is also the locus of all points having $\theta_j$ as their argument. Noting the order of indices, dual feasibility requires that $\Delta_j$ and $\Delta_k$ have opposite sign and so lie on opposite sides of the line. An exactly similar argument applies to the other points and $\Delta = 0$ or $\Delta = 0$. If point $k$ is to lie on the opposite side of $\Delta_j = 0$ from point $j$ and on the opposite side of $\Delta_j = 0$ from point $i$, it can only occupy the sector shown in Figure 3.41. As it is only in this geometry that $\Delta_k$ and $\Delta_j$ will have opposite signs, as required for dual feasibility, the following theorem, termed here “the 180° rule”, is proved. (A point known for a long time but proved by Chetwynd [43].)

**FIGURE 3.41** The 180° rule in simplex iteration.
The 180° Rule

A circumscribing limaçon on a given origin to a set of points is the minimin radius circumscribing limaçon to those points if it is in contact with three of them such that no two adjacent contact points subtend an angle at the origin of more than 180°, where the term “adjacent” implies that the angle to be measured is that of the sector not including the third contact point.

A complete simplex iteration for the minimum radius circumscribing limaçon in the dual consists of selecting any point which violates the reference (conventionally, the point giving the largest violation is chosen) and substituting it for one of the points defining the reference in such a way that dual feasibility is maintained. The 180° rule allows the general iteration to be simplified to the following exchange algorithm:

1. Choose any three data points such that no two adjacent ones subtend an angle at the origin of more than 180°.
2. Construct a reference limaçon through these three points.
3. If no data points lie outside this limaçon the solution is found, otherwise choose the point which violates the reference by the largest amount.
4. Replace one of the reference points by this new point such that the 180° rule is still obeyed and go back to Step 2.

The exchange between any new point and the contacts is always unique.

An exchange algorithm depends upon the iterations moving monotonically toward an optimum solution in order to guarantee that cyclical exchanges do not occur. Here this is the case, because, as the exchange is unique at each iteration, it must be identical to the variable change at the simplex iteration of the linear program and that is known to converge monotonically.

A similar procedure needs to be followed to get the formal linear programing matrix for the minimum zone but essentially the result is as follows.

The conditions for the optimum solution to the minimum radial zone limaçons give rise to the following geometric interpretation. Expressing the zone as a band of width 2h placed symmetrically about a single limaçon:

1. All data points must lie not more than a distance $h$, measured radially from the origin, from the limaçon
2. There must be four data points all lying exactly $h$ from the limaçon such that they lie with increasing angle $\theta$, alternately inside and outside the limaçon

It may be noted that this alternation property is not unique to this problem; it occurs, for instance, in the Steifel exchange algorithm for best-fit polynomials, which may also be derived by a linear programing method [42]. These rules may be used to formulate an exchange algorithm. Thus:

1. Choose arbitrarily four data points.
2. Fit to these a reference limaçon such that they are radially equidistant from it and lie alternately on either side of it with increasing angle.
3. If no other points are further from the reference the solution is found.
4. Otherwise substitute the point which lies furthest from the reference for one of the four defining points such that the new set of points lie alternately on either side of the reference and return to Step 2.

3.7.3 Minimum Zone, Straight Lines, and Planes

The minimum separation parallel and straight lines belong to the well-documented class of minimax polynomials, that is curves having the smallest possible maximum divergence from the data. The condition for this to occur is that, relative to an $n$th-order polynomial, the data must have $n + 2$ maxima and minima, all of equal magnitude. The solution can be found by the Steifel exchange algorithm, which proceeds by fitting the polygonal according to this condition to $n + 2$ points and then exchanging points further away from it with the defining set, while maintaining the condition. In terms of the minimum zone straight lines there will be three points, two contacting one line and one the other in an alternate sequence, which are iterated by exchanges (see Figure 3.42).

The minimum zone planes can be expressed, in instrument coordinates, as

$$\minimize Z - h \text{ subject to } ax_i + by_i + c + h \geq z_i$$

$$ax_i + by_i + c - h \leq z_i$$  \hspace{1cm} (3.282)

for all data points $(x_i, y_i, z_i), a, b, c$ are sign unrestricted and $h \leq 0$. Noting that $h = 0$ is valid only for the trivial

![Figure 3.42 Steifel exchange mechanism.](image-url)
condition that all points are coplanar, then it may be asserted that four points will be represented in the basis of the dual, which can be expressed as

$$\beta^{-1} = \begin{bmatrix} S_i x_i & S_j x_j & S_k x_k & S_l x_l \\
S_i y_i & S_j y_j & S_k y_k & S_l y_l \\
S_i z_i & S_j z_j & S_k z_k & S_l z_l \\
1 & 1 & 1 & 1 \end{bmatrix}$$

(3.283)

where \( S_i \) etc., take values of +1 or −1 according to whether \((x_i, y_i, z_i)\) contacts the upper or lower of the minimum zone planes. As before, dual feasibility is guaranteed if all terms in the final column of \( \beta \) are positive, which will be true providing that the cofactors of the final row of \( \beta^{-1} \) all have the same sign. These cofactors are

$$-S_l S_k S_i \Delta_{jkl} - S_l S_j S_i \Delta_{ikl} - S_k S_j S_i \Delta_{lil} - S_l S_j S_k \Delta_{lij}.$$

(3.284)

They must all have the same sign. Consider the determinant equation representing the boundary between positive and negative regions of \( \Delta_{ij} \):

$$\Delta_{ij} = \begin{vmatrix} x_j & x_k & x_l \\
y_j & y_k & y_l \\
1 & 1 & 1 \end{vmatrix} = 0.$$

(3.285)

It is a plane parallel to the \( z \) axis (since it is independent of \( z \), passing through points \((z_j, y_j)\) and \((x_j, y_k)\). Dual feasibility requires that if \( S_i = S_j \) (contacts with the same place) \( \Delta_{jkl} \) and \( \Delta_{jki} \) must have different signs and \textit{vice versa}. So if the \( i \)-th and \( j \)-th contacts are with the same plane they lie on opposite sides of \( \Delta_{jkl} = 0 \), but if they contact different planes they both lie on the same side of \( \Delta_{jkl} = 0 \). A parallel argument shows that the same is true for all pairs of points.

These relationships show the relative positions of contacts that give dual feasibility (Figure 3.43). There can be two contacts with each of the minimum zone planes, in which case the plan of lines joining the alternate types must form a convex quadrilateral or a 3:1 split, in which case the single contact must lie in the plan of the triangle formed by the other three contacts.

Even with this most simple of 3D zone fits, the advantage of using specific exchange algorithms rather than a general revised simplex solution in an automatic system is becoming questionable—hence the difficulty of standardization.

**Note:**

It may, at first glance, seem surprising that limaçon fitting rather than the apparently simpler case of flat surfaces has been used as the primary example. This section, being primarily concerned with linear programming, does not report detailed comparisons between algorithms but some comments are needed. The following remarks are based on practical tests made by Chetwynd [41].

A typical roundness “profile” would have 512 equally spaced radial ordinates each resolved over a 10- or 12-bit working range. Exchange algorithm systems have now been working with data of this type in both industrial and research environments for several years and their robustness has been established. Even with an arbitrary choice of points for the initial basis, the exchange algorithm virtually always solves for the minimum circumscribing limaçon in five or less iterations, while the minimum zone only occasionally needs more than five on real engineering profiles.

The earlier (primal-based) algorithms were run with poorly defined end conditions, typically making 32 relatively coarse-stepped iterations and then 32 finer steps, after which the process was terminated with a result assumed to be close to the desired optimum. The dual techniques yield at least a 10:1 saving in the number of iterations as well as giving a fully determined convergence and so better accuracy. With both algorithms the iteration is dominated by the almost identical computation and checking of the updated figure, so the program size and the cycle times are closely similar on similar machines.

The direct use of revised simplex on dual programs representing limaçon fitting has been studied using a specially developed package containing only the subroutines essential for solving this class of problem. Memory requirements are only slightly larger than those of exchange algorithms and execution is typically about 20% slower. This is due to the simple way artificial variables are treated. This difference can be removed at the cost of extra program length.

The limaçon fits to roundness measurement have simple exchange rules that can be expressed in a few numerical comparisons and logic operations. Thus in a specialized system both a size reduction and a speed increase would be obtained by replacing the direct use of revised simplex on the dual by an exchange algorithm. However, the exchange logic is specific, so if several different references are to be implemented there will be less shared code. With more complex geometries it is of even greater importance that the efficiency of dual-based methods is obtained. Yet, even with the simplest 3D geometry, the exchange rules are becoming quite complicated. Using a conclusion made by Chetwynd, duality theory has shown that the “obvious” geometrical method is rarely the best approach.

**FIGURE 3.43** Contacts which give dual feasibility.
The study of exchange algorithms gives a very clear insight into the geometrical implications of reference fitting. This is important for metrology. Measurement should always be based on engineering relevance rather than a mathematically convenient abstraction. The exchange algorithm also provides a good method for solution by hand should it be necessary. Relatively flexible measurement systems are likely to use a more general implementation of a revised simplex algorithm. This is no cause for concern: both are firmly based on the same theoretical foundation.

**Note:**
Even these considerations should show that the use of dual and exchange methods require a considerable knowledge of the subject so that the parameters to be measured and the constraints to be applied can be formulated properly. The gains are massive if the mathematical background is sound, but it should never be forgotten that these methods rely on a relatively small number of geometrical points to satisfy the ultimate optimum solution. This is a weakness because of the fleeting reliability of such extrema. If this is a consideration in one dimension such as in roundness and straightness, it is even more so in the 2D cases of sphericity, cylindricity, etc. In these cases, especially that of cylindricity, the whole problem becomes very complicated because the straightforward requirement of five constraints becomes masked in the data produced by a surface metrology instrument due to tilt and eccentricity effects. For a more detailed discussion on this subject, see Ref. [43].

### 3.7.4 Minimax Problems

#### 3.7.4.1 General Algorithmic Approach

Because there is no absolutely correct algorithm or even a fully agreed algorithm for the minimax problems in surface and dimensional metrology it is informative to give an example here. The algorithm chosen is to determine the minimax zone value of a number of different forms commonly found in industry, ranging from the flat to the cylindrical in parallel with the best-fit least-squares procedures given earlier.

The reason why the minimum zone has been selected is because, next to least squares, it is often the preferred way of specifying the tolerance on a form. In fact the ISO standards refer to it often but do not attempt to indicate how to obtain it! This is not surprising because nobody has yet agreed how to do it. The optimum obviously depends on a number of things, not the least being simplicity and technological relevance rather than mathematical elegance.

Such a straightforward approach has been suggested by Dhanish and Shunmugam [39]. This approach assumes that the form parameters have already been linearized, which is consistent with the constraints of most surface metrology instruments and potentially with coordinate-measuring machines providing that the part is very well aligned. Problems such as tilt, eccentricity, and direction of measurement have been ignored. The basic problem with tilt and eccentricity has only really been attempted properly by Chetwynd.

#### 3.7.4.2 Definitions

Let the function of the surface be \( b \), and \( \phi \) the function which best fits the surface. Assuming that the form shape has been linearized in surface metrology

\[
\phi_k = a_{i_1} u_{i_1} + \ldots + a_{i_m} u_{i_m},
\]

(3.286)

where \( n \) is the number of variables, \( a_{ij} \) denotes the value of the \( j \)th variable at the \( i \)th place, and \( u_j \) is the coefficient of the \( j \)th variable.

As in least squares the deviation \( e_i = \phi_i - b_i \) is

\[
e_i = a_{i_1} u_{i_1} + \ldots + a_{i_m} u_{i_m} - b_i.
\]

(3.287)

The form error is computed as

\[
h_i = |e_{max}| + |e_{min}|
\]

(3.288)

where \( e_{max} \) and \( e_{min} \) are the maximum and minimum errors.

The general problem for the minimum zone is, given \( a_{ij} \) and \( b_i \) find \( u_i \) of \( \phi \) such that \( \max |e_j| \) is a minimum (following Dhanish and Shunmugam’s method below). The steps are as follows:

1. Choose a set of \( m + 1 \) points as a reference set. It is usual to use a set based on the least-squares approximation, usually the points giving maximum deviation from the least squares. Call these points \( v_{ij} \) so that \( v_{ij} = a_{ij} \); \( t \) takes the values in the reference set. The steps from 2 onwards operate only on the reference set.

2. Find \( \lambda_k \), \( k = 1 \ldots m + 1 \), solutions of the equations

\[
\sum_{k=1}^{m+1} \lambda_k v_{ij} = 0.
\]

(3.289)

Since these are \( m \) homogeneous equations with \( m + 1 \) unknown the solution in determinant form is

\[
\lambda_k = (-1)^{i+j} |v_{ij}| \quad j = 1 \ldots m, \quad i = 1 \ldots m + 1 (i \neq k).
\]

(3.290)

3. Calculate the reference deviation \( d \) where

\[
d = \sum_{k=1}^{m+1} \lambda_k b_k \left/ \left( -\sum_{k=1}^{m+1} \lambda_k \right) \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right. \right.
If $\lambda_0 = 0$ then Haar’s condition is violated and the situation is ambiguous with respect to sign, as mentioned earlier, so both the positive and negative $d$ have to be tried. For both $+$ and $-$ values continue to Step 6 with that sign which gives the lowest value of error. Also this sign shows the direction to go.

5. Find the leveled reference by interpolation:

$$\varphi_2 = \sum_{j=1}^{m} u_j v_{kj} \quad k = 1 \ldots m + 1 \quad (3.293)$$

so $m + 1$ equations exist for $m$ coefficients $u_1 u_2 \ldots$ so that any $m$ equations will suffice.

6. Find the value $d_i$ of this reference function at all the given points. The value of $d_i$ whose absolute value is the highest is referred to as the supreme error $e^*$ if

$$|e^*| \leq d_i. \quad (3.294)$$

Then the criterion of optimality is satisfied and the search is stopped. Then the function $\varphi$ is the Chebychev approximation sought and the form error is calculated based on Equation 3.288 for $h_i$.

Otherwise the point $i^*$ corresponding to the supreme error $e^*$ enters the reference set and the search is continued.

7. Find the point to be discarded from the reference set by the Steifel exchange mechanism used by Chetwynd. Solve for $\mu_k$ in the equations

$$\sum_{k=1}^{m+1} \mu_k v_{kj} + a_q = 0 \quad j = 1 \ldots m \quad (3.295)$$

Since there are $m + 1$ variables and only $m$ equations a suitable value for one of the $\mu$ could be assumed; then calculate $q_k$ where

$$q_k = \mu_k / \lambda_k \quad k = 1 \ldots m + 1. \quad (3.296)$$

If $\text{sgn}(d)e^* > 0$ the point with minimum $q_k$ should be removed from the reference set, otherwise the point with maximum $q_k$ is discarded. (If any $\lambda_k = 0$ a “trial exchange” has to be tried. Each point in the reference set is replaced by the point of supreme error one at a time and the trial reference deviation is computed for each set. The combination leading to the largest value of $|d|l$ is taken as the next reference step.) With this reference set, go back to Step 2. Note that the trial reference is not the exchange therein but alternative to it in the special case $\lambda_k = 0$ where the exchange will not work.

Any example taken from Ref. [39] shows how this algorithm works. There are worked examples given for flatness, roundness, cylindricity, and sphericity.

However, even though the parameterization has been linearized there is still quite a number of special situations which have to be taken into account for obtaining a general algorithm. This detail is much greater when the system has to be linearized by iteration and when the effects of misalignment and eccentricity are taken into account. It seems that the more versatile and wide ranging the instrumentation becomes, the more difficult it will be to finalize a simple general algorithm. It may be that the best that will be done is to agree on a procedure whereby the algorithms will be able to be well behaved. For the fine detail of algorithms for working out minimax problems see books on computational geometry (e.g., Ref. [44]).

### 3.8 FOURIER TRANSFORMS AND THE FAST FOURIER TRANSFORM

#### 3.8.1 General Properties

The Fourier transform may be of use whenever frequency, either in reciprocal time or wavenumber, is of importance. This is often the case in engineering, in a temporal sense in manufacturing systems and spatially in metrology. Examples of its use in correlation and filtering have already been mentioned as well as its use in random process theory.

The general form of the equation relating a time function $f(t)$ to a frequency function $F(\omega)$ is

$$F(\omega) = \int_{-\infty}^{\infty} f(t) \exp(-j\omega t) dt. \quad (3.297)$$

Alternatively this can be written with the factor $1/2\pi$ in front of the integral. The corresponding equation connecting $f(t)$ with $F(\omega)$ is

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) \exp(j\omega t) d\omega. \quad (3.298)$$

As long as there is a $2\pi$ relationship between these two equations it does not matter in which domain it occurs. Some authors use a factor of $1/\sqrt{2\pi}$ outside both integrals to provide some symmetry. The first equation is referred to as the Fourier integral. As the transforms are usually considered to be time functions in behavior they will be designated $f$ rather than $z$ which is most often used in surface geometry. Before embarking upon the FFT algorithm there will be a short listing of some of the essential properties of the Fourier integral as follows:

1. If $f(t)$ is real, as is usually the case in metrology, then the real and imaginary parts of the $F(\omega)$ namely $R(\omega)$ and $X(\omega)$ are given by

$$R(\omega) = \int_{-\infty}^{\infty} f(t) \cos \omega t \ dt \quad (3.299)$$
$$X(\omega) = -\int_{-\infty}^{\infty} f(t) \sin \omega t \ dt \quad (3.300)$$

that is $R(\omega)$ is an even function and $X(\omega)$ is odd. Thus $R(-\omega) = R(\omega)$ and $X(-\omega) = -X(\omega)$.
2. If \( f(t) \) is even, that is \( f(-t) = f(t) \),
\[
R(\omega) = 2 \int_{-\infty}^{\infty} f(t) \cos(\omega t) \, dt
\]
\[
X(\omega) = 0
\]
which is particularly relevant when \( f(t) \) takes the form of an autocorrelation function.

3. If \( f(t) \) is odd, that is \( f(-t) = -f(t) \),
\[
R(\omega) = 0 \quad X(\omega) = -2 \int_{-\infty}^{\infty} f(t) \sin(\omega t) \, dt
\]  
(3.301)

4. Linearity: if \( F_1(\omega) \) and \( F_2(\omega) \) are the Fourier integrals of time functions \( f_1(t) \) and \( f_2(t) \) then
\[
f_1(t) + f_2(t) \leftrightarrow F_1(\omega) + F_2(\omega)
\]  
(3.302)
where the symbol \( \leftrightarrow \) indicates Fourier transformation.

5. Scaling:
\[
f(Kt) = \frac{1}{K} F\left(\frac{\omega}{K}\right).
\]  
(3.303)

6. Shifting in time by \( t_0 \):
\[
f(t-t_0) \leftrightarrow \exp(-j\omega t_0) F(\omega)
\]
a property which is useful when applying the theory of the phase-corrected (linear phase) filter as seen earlier. A similar relationship occurs in frequency. Thus
\[
F(\omega - \omega_0) \leftrightarrow \exp(j\omega_0 t) F(t).
\]  
(3.304)

7. Differentiation:
   a. Time:
\[
\frac{d^n f}{dt^n} \leftrightarrow (j\omega)^n F(\omega)
\]  
(3.305)
a theorem which is useful for evaluating the frequency characteristic of a complicated impulse response. A similar theorem exists which enables the reverse to be done in frequency differentiation.
   b. Frequency:
\[
\frac{d^n F}{d\omega^n} \leftrightarrow (-j\omega)^n f(t).
\]  
(3.306)

8. Convolution: this has already been used in filtering techniques. In condensed form, if \( f_1(t) \leftrightarrow F_1(\omega) \), \( f_2(t) \leftrightarrow F_2(\omega) \) then written
\[
\int_{-\infty}^{\infty} f_1(\tau) f_2(t-\tau) d\tau \leftrightarrow F_1(\omega) \times F_2(\omega)
\]  
(3.307)
where the symbol \( \times \) denotes convolution. These relationships will be used extensively in Chapter 4.

The DFT may be written
\[
F(k) = \frac{1}{N} \sum_{n=0}^{N-1} f(n) \exp\left(-2\pi j nk / N\right) \quad n = 0, \ldots, N-1.
\]  
(3.308)

### 3.8.2 Fast Fourier Transform

The problem with the DFT as it stands is that it takes quite a long time to compute. This situation has been changed by the rediscovery of the FFT algorithm. This is not a special transform, it is a generic term enfolding a whole set of algorithms for computing the DFT. Some algorithms are matched to specific types of data etc. Generally, using the FFT, a reduction in the number of operations by a factor of \( \log_2 N \) is possible, where “operation” is usually taken to mean a complex multiplication and an addition. Thus the larger the number of data points \( N \) the greater the advantage. The FFT is also efficient in terms of storage because intermediate results in the calculation are stored in the original data places so that no extra storage is required for those extras beyond that of the data. Yet another benefit derives directly from the computation reduction: less computation means less numerical errors. Thus the FFT has three remarkable advantages: it is faster, more accurate and requires minimal storage. The disadvantage is that there have to be as many spectral points evaluated as there are data points, which may be more than is needed.

#### 3.8.2.1 Analytic Form

This is a fast algorithm of efficiently computing \( F(k) \)
\[
F[k] = \sum_{n=0}^{N-1} f(n) \exp\left(-j2\pi kn / N\right).
\]  
(3.309)
For \( N = 2^r \) there are many variants.

Let \( k \) and \( n \) be represented in binary form as
\[
k = k_{r-1} 2^{r-1} + k_{r-2} 2^{r-2} + \ldots + k_0
\]
\[
n = n_{r-1} 2^{r-1} + \ldots + n_0
\]  
(3.310)
where \( W = \exp(j2\pi/N) \).
Then
\[
F(k_{r-1}, k_{r-2}, \ldots, k_{0}) = \sum_{n_{r-1}=0}^{1} \sum_{n_{r-2}=0}^{1} \ldots \sum_{n_{0}=0}^{1} f(n_{r-1}, n_{r-2}, \ldots, n_{0})
\times W^{k_{r-1}2^{r-1}} \times W^{(k_{l}^{2^{l}+k_{0})n_{l}2^{l-2}}}
\times \ldots \times W^{(k_{r-1}2^{r-1}+k_{r-2}2^{r-2}+\ldots+k_{0})n_{0}}.
\]
(3.311)

Performing each of the summations separately and labeling the intermediate results the following is obtained:
\[
F_{0}[n_{r-1}, n_{r-2}, \ldots, n_{0}] = f[n_{r-1}, n_{r-2}, \ldots, n_{0}]
\]
\[
F_{1}[k_{0}n_{r-2}, \ldots, n_{0}] = \sum_{n_{r-2}=0}^{1} F_{0}[n_{r-1}, n_{r-2}, \ldots, n_{0}] W^{k_{0}2^{r-2}}
\]
\[
F_{2}[k_{0}, k_{1}, \ldots, n_{0}] = \sum_{n_{2}=0}^{1} F_{1}[k_{0}, n_{r-2}, \ldots, n_{0}] W^{k_{2}2^{r-2}}
\]
\[
F_{r-1}[k_{0}, k_{1}, \ldots, k_{r-1}] = \sum_{n_{r-1}=0}^{1} F_{r-2}[k_{0}, k_{1}, \ldots, n_{0}] W^{k_{r-1}2^{r-1}+k_{r-2}2^{r-2}+\ldots+k_{0}n_{0}}
\]
\[
F[k_{r-1}, k_{r-2}, \ldots, k_{0}] = F_{r}[k_{0}, k_{1}, \ldots, k_{r-1}].
\]
(3.312)

This operation is performed in \( N \log N \) complex multiplications through the indirect method above. If \( F[k] \) is evaluated directly as
\[
\sum_{n=0}^{N-1} f(n) \exp \left(-j \frac{2\pi n}{N} \right) \quad (3.313)
\]
it takes the order of \( N^2 \) complex multiplications, so the FFT is obviously faster. If \( N \) is large, as it usually is for the data set in any surface metrology, the gain in time is very great. The fact that there has to be as much store for the spectrum as there is for the data are no longer a disadvantage, although at one time when storage was at a premium in computers this was.

The only better way of achieving further gains is to use one or other of the clipped transforms, such as the Walsh transform [23]. As far as speed is concerned, some instrumental techniques can dispense altogether with digital methods and use coherent optical methods, as will be seen in the section on Fourier diffraction techniques in Chapter 4. In this case the transform is produced at the speed of light. A further advantage is that because the light is shone onto the surface, no digitization of the surface is required so that optical methods are very valid in surface metrology. The possibility of an optical method that is practical for evaluating the ambiguity function is also feasible because it is a real function.

In what follows, some conventional digital ways of filtering will be given with particular emphasis on the types of filter used in surface metrology.

Some of the comparative properties of the Fourier transform, ambiguity function, and Wigner distribution are given in Table 3.6.

As previously mentioned the transform pair involves a pair of complementary transforms, one with positive exponents and one with negative exponents, the pair taken together having a normalization factor of \( 1/2\pi \) or \( 1/N \) for the DFT. This choice is arbitrary. The equation with positive exponential

<table>
<thead>
<tr>
<th>Complex Valued ( f(x) )</th>
<th>Complex Valued ( F(\omega) )</th>
<th>Complex Valued ( \Lambda(\omega, \chi) )</th>
<th>Real Valued ( W(x, \omega) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial shifting ( f(x-x_0) )</td>
<td>( F(\omega) \exp(-j\omega x_0) )</td>
<td>( A_{1}(\sigma, \chi) \exp(-j\omega \sigma_0) )</td>
<td>( W_{1}(x-x_0, \omega) )</td>
</tr>
<tr>
<td>Frequency shifting ( f_{\omega}(x) \exp(j\omega x_0) )</td>
<td>( F(\omega - \omega_0) )</td>
<td>( A_{1}^{\omega}(\sigma, \chi) \exp(-j\omega \sigma_0) )</td>
<td>( W_{1}(x, \omega - \omega_0) )</td>
</tr>
<tr>
<td>Spatial limited ( [x_1, x_2] ) for ( x )</td>
<td>( -[\infty, \infty] ) for ( \omega )</td>
<td>(-[x_1, x_2], [-\infty, \infty] ) for ( \chi )</td>
<td>([-x_1, x_2], [-\infty, \infty] ) for ( \omega )</td>
</tr>
<tr>
<td>Frequency limited ( [\omega_1, \omega_2] ) for ( x )</td>
<td>( [\omega_1, \omega_2] ) for ( \omega )</td>
<td>(-[\omega_1, \omega_2], [\omega_1, \omega_2] ) for ( \sigma_0 )</td>
<td>(-[\omega_1, \omega_2], [\omega_1, \omega_2] ) for ( \omega )</td>
</tr>
<tr>
<td>Convolution ( \int \int f(x)h(x)dx )</td>
<td>( F(\omega)H(\omega) )</td>
<td>( \int \int \Lambda_{0}(\sigma, x)\Lambda_{0}(\sigma, \chi - x)dx )</td>
<td>( \int \int W_{0}(x, \omega)W_{0}(x, \chi - \omega)dx )</td>
</tr>
<tr>
<td>Modulation ( f(x)\exp(j\omega x) )</td>
<td>( \frac{1}{2\pi} \int F(\omega)M(\omega - \sigma)\omega d\omega )</td>
<td>( \frac{1}{2\pi} \int \Lambda_{0}(\sigma, x)\Lambda_{0}(\sigma - \omega, \chi - x)dx )</td>
<td>( \frac{1}{2\pi} \int W_{0}(x, \omega)W_{0}(x, \omega - \sigma)\omega d\omega )</td>
</tr>
<tr>
<td>Typical stationary ( A\exp(j\omega_{0}x) )</td>
<td>( A_{2}\pi\delta(\omega - \omega_{0}) )</td>
<td>(</td>
<td></td>
</tr>
<tr>
<td>Typical non-stationary ( A\exp(j\chi) )</td>
<td>( A_{2}\pi\exp(j\pi/4)\exp(-j\omega_{0}x) )</td>
<td>(</td>
<td></td>
</tr>
</tbody>
</table>
is usually called the direct transform and the negative exponential the inverse transform (IDFT):

$$f(t) = \frac{1}{2\pi} \int \int F(\omega)\exp(j\omega)\,d\omega.$$  (3.314)

Sometimes both are referred to as the DFT and sometimes the IDFT is called the Fourier series, as in most textbooks.

To see how the discrete form of the FFT works consider the discrete form of $f(t)$: $f(n)$

$$f(n) = \sum_{k=0}^{N-1} F(k) \exp\left(\frac{2\pi i}{N}jn\right) \quad n = 0, \ldots, N - 1. \quad (3.315)$$

Equation 3.315 can be written

$$f(n) = \sum_{k=0}^{N-1} F(k)W^{nk} \quad \text{where } W = \exp\left(\frac{2\pi i}{N}\right) \quad (3.316)$$

and $W^{nk} = \exp[(2\pi j/n)nk]$ which has a period of $N$.

The basic method used in the FFT algorithm is to make use of the symmetry of the exponential function.

$W^{nk}$ is a sinusoid of period $N$ and it also displays odd symmetry about $k = N/2$ for fixed $n$, etc. Thus $W^{nk} = -W^{n(k+N/2)}$.

This cyclic property is of fundamental importance. It can be shown that the efficiency of the FFT routine depends on the data sequence length. In fact it is optimized for the case where $N = 3^m$, where $m$ is an integer. However, for digital computer operation it is more convenient to use $N = 2^a$ or $4^a$ without significant loss in efficiency. Obviously base 2 is more suitable for computational work than base 3.

To see how it is possible to derive a base 2 FFT without recourse to the general theory, consider Equation 3.315. If $N$ is divisible by 2 the summation can be split into two smaller sums involving odd and even indices. Thus

$$f(n) = \sum_{k=0}^{N/2-1} F(2k)W^{2nk} + \sum_{k=0}^{N/2-1} F(2k + L)W^{2(k+1)} \quad (3.317)$$

so that the first sum uses values of $F(0)$ to $F(N-2)$ and the second sum uses values of $F(1)$ to $F(N-1)$. Hence $f(n)$ may be written in terms of odd and even indices. Thus

$$f(n) = E(n) + W^nO(n) \quad \text{where } W^n = \exp\left(\frac{2\pi jn}{N}\right) \quad (3.318)$$

and

$$E(n) = \sum_{k=0}^{N/2-1} F(2k)W^{2nk} \quad \text{and}$$

$$O(n) = \sum_{k=0}^{N/2-1} F(2k + 1)W^{2nk}. \quad (3.319)$$

$W^n$ is a constant for constant $n$.

The two series in Equation 3.319 will be recognized as discrete transforms of the odd and even-indexed terms of the original series in Equation 3.317. Note that the 2 in the index of $W$ is appropriate because these two smaller series only have a length of $N/2$ and not $N$. It is this factor that is decisive in cutting the number of operations. To see this, consider Equation 3.317. To evaluate one $f(n)$ requires $N$ operations and to evaluate $N$ values of $f(n)$ therefore requires $N^2$ operations where operation refers to a complex multiplication and addition. However, the situation has changed in Equation 3.318 because there are only $N/2$ terms in the even series and $N/2$ terms in the odd series. To evaluate all the even components of all the $F(n)$ therefore requires $(N/2)^2$ operations and to work out all the odd components of the $F(n)$ also requires $(N/2)^2$. The two, together with the $N$ multiplications of $W^n$ and $N$ additions of the $E(n)$ and $O(n)$, give $(N^2/2 + N)$ operations. For $N$ large this tends to $N^2/2$, a gain over the direct method of 50% in computation. Similarly, if $N/2$ is also divisible by 2 a gain may be made by breaking down the two series. This reduction process can be continued $M$ times until finally no further decomposition can take place. The overall number of operations carried out in the entire operation is $N \log_2 N$ as distinct from $N^2$ in the direct approach, the ratio being $\log_2 N/N$. Thus for $N = 256$ the computational time is only 3% of that used in the direct method.

This iterative process is shown in Figure 3.44, which illustrates how an eight-point transform can be produced by three iterations. Notice that the input points are not arranged in their natural order. The reason for this becomes clear from the figure; at each stage a pair of points depends only on the pair previously residing at the same locations. Thus the results of each iterative step can overwrite the operands without subsequently affecting future calculations. The computation can proceed “in place”; no extra storage is needed for intermediate results. The order into which the original sequence has to be arranged is successively determined by separating out the odd from the even indices from the progressively shorter sequences. It can be shown that each element of the sequence should be placed at an address given by reversing the order of the bits in the binary representation of its original address.

Because of the increasing use of the FFT routine some consideration will be given here actually to implementing the routine on a small computer, and one or two special tricks associated with metrology will be indicated. This will be by no means an exhaustive account but it should serve to illustrate some of the features of the technique. It will be followed by some general applications to metrology problems on surfaces.

### 3.8.2.2 Practical Realization

The term practical realization will be taken here to mean a form of the algorithm suitable for use with a general purpose digital computer which can be efficiently applied to typical transform problems. It is widely recognized that the optimum computational efficiency for a specific problem is obtained only by tailoring the program to its requirements. It follows from this that a general purpose routine will need to
compromise between the sometimes conflicting needs of different problems. An obvious example is that if large amounts of data are to be handled, computer storage may be at a premium and it may be preferable to make the FFT routine small at the expense of reducing its speed. For a “real-time” problem, the opposite approach of using more storage to increase the computational speed would probably be necessary.

The first compromise is to use a base 2 FFT algorithm. The effect of this is to get good computational efficiency and very little “housekeeping” so that the actual FFT program may be both fast and physically small. To offset this, the data must be equal to \( 2^m \) points, where \( m \) is a positive integer.

Having decided on the base 2 FFT, the actual form of the derivation used is of little consequence. For the further discussion here the approach described earlier will be used, that is that based on Equations 3.317 and 3.318.

Some general considerations will now be given to the practical realization. In the first place the general situation, which involves complex data, will be considered. This will be followed by some details of how these can be recovered in the face of the less restrictive data often found in surface metrology.

3.8.3 General Considerations of Properties

Remembering that all sequences in the DFT are complex, and that, therefore, each data point will occupy more than one store address, three distinct sources of potential inefficiency can be identified. These are the complex multiply and add operations at the heart of each calculation, the addressing of the data points at each stage of the iteration and the generation of the complex exponents.

There is little than can be done about the actual operation of multiplication without the use of special hardware, particularly if floating point operation is used. In fixed point operations, where scaling between operations is used, some optimization is possible since the nature of the multiplier (the exponential term) is known in advance. Prechecking each operand for zero and unity value can save time by eliminating unnecessary multiplications, but in general the overall time of the complete evaluation may not be reduced since extra time is required before every multiplication in order to perform this test.

Each iteration of the FFT requires at least one access to each point in the data array for reading and one for writing. The order in which the points are accessed will change for each iteration, since the point pairs move their relative positions, so that the data array must be treated as random access rather than cyclic. The accessing of array elements can introduce significant overheads in terms of the speed of operation and so should be kept to a minimum. While at least two accesses are needed to each pair of locations, these always occur as a group before the locations are addressed. Thus the addresses need only be calculated once per iteration, being stored for

![FIGURE 3.44](a) Eight-point fast Fourier transform decimation-in-frequency branch transmissions in power of \( W_8 \), output shuffled and (b) eight-point fast Fourier transform decimation-in-tune branch transmissions in powers of \( W_8 \), input shuffled.
use during accessing, without significantly increasing the total store requirements. If the exponentials are stored in a look-up table, the amount of addressing to that can be minimized by ordering the calculations so that those using a particular exponent occur consecutively during each iteration.

The generation of complex exponentials is almost universally performed using de Moivre’s theorem:

\[ \exp(j\theta) = \cos\theta + j \sin\theta. \]  

By this method it is only necessary to calculate either the sine or cosine, since they are simply related and the calculation time is relatively small. However, compared with the other operations within the FFT the time to calculate \( \sin\theta \) is large. Since the calculation would also require quite a lot of programming, it is more usual to use some form of look-up table for the values. Only the first two quadrants need to be stored. The form which a look-up table takes depends largely on the compromise between speed and storage. The table may be made progressively shorter by increasing the amount of calculation required to access a particular value. The fastest access is by simply storing the \( N/2 \) complex exponential values. The store may be reduced with little overhead by having two separate arrays of \( N/2 \) points for sine and cosine, which may, of course, be overlapped by one quadrant, and combining them into a complex number at access time. An interesting method of retaining fairly fast access with only small storage is to use the approximate relationship that, for small \( B \),

\[ \sin(A + B) = \sin A + \cos A \sin B. \]  

Here only a short table containing widely spaced \( A \) values and another short table of \( B \) values are needed.

Also, during the second iteration \( W \) may take only the values \(+1, -1, j, -j\). By programing the first two iterations separately outside the iterative loop, no multiplications are needed during them.

Because of the ready availability of FFT subroutines, they will not be described here.

### 3.8.3.1 Fourier Series of Real Data

By taking into account symmetries and identities readily identifiable it is possible to take great advantage of the FFT routine. Furthermore, in the case of evaluating the autocorrelation function from the former spectral density and vice versa, additional benefits can be achieved because the data is not only real but also even (i.e., symmetrical about the origin). Under these circumstances use can be made of the cosine transform rather than the full Fourier transform.

Instead of grouping the data in such a way as to appear complex it is grouped to appear as a conjugate, even series. Thus \( z(n) \) may be described as

\[ f_1(n) = z(2n) \]
\[ f_2(n) = z(2n + 1) - z(2n - 1) \]
\[ f(n) = f_1(n) + j f_2(n) \quad \text{for } n = 0, 1, \ldots, N - 1. \]  

Since \( z(n) \) is real and even and \( f(n) \) is conjugate even, the sequence \( f(n) \) need only be calculated for \( n = 0, \ldots, N/2 \); only the \( z(n) \) values of \( n = 0, 1, \ldots, N \) are needed for this. Therefore only \( N/2 + 1 \) complex storage locations are needed for processing \( 2N \) real, even data.

Using this technique only \( N + 1 \) real or \( N/2 + 1 \) complex numbers need to be stored, giving a saving of four times in store over the \( 2N \) complex data series storage requirement.

Similar statements can be made about the sine transform. It is obvious that the FFT routine has many uses. Some will be outlined here.

### 3.8.4 Applications of Fourier Transforms in Surface Metrology

Main areas of use are:

1. Digital filtering
2. Power spectral analysis
3. Correlation
4. Other convolutions
5. Interpolation
6. Other analysis, for example Cepstrum
7. Roundness and other shapes—least-squares

### 3.8.4.1 Fourier Transform for Non-Recursive Filtering

This is basically a convolution filter using the impulse response of the filter \( \omega \) as a window for the data. Thus the output \( g \) of a filter with input \( f \) is \( g = w * f \). Using the transform it is preferable to work in the frequency domain; thus

\[ G_i = W \times F_i \]  

since convolution becomes multiplication in the transform domain. So filtering becomes:

1. Transform data
2. Multiply transformed data by frequency weighting function
3. Invert transformed and modified data

Note that it sometimes beneficial to change the spectrum obtained from 1 above to an analytical function. This involves making \( \omega < 0 = 0 \). This operation involves taking the Hilbert transform (see Section 3.10.4.2).

Whether this method is quicker to compute will depend on the lengths of the data sequence and weighting function. Recently this technique has become progressively more attractive as the data lengths have increased. Because of the reciprocal nature of the transform domains, an added advantage of the transform technique may occur: a long, spatial weighting function will produce a short-frequency weighting function so that even less operations are required.

In the frequency domain, the amplitude and phase characteristics of the filter are separated into the real and imaginary
3.8.4.2 Power Spectral Analysis

The measurement of power spectral density is the most obvious application of the Fourier transform. Put simply, a data sequence from the surface is Fourier transformed and a power spectrum obtained by squaring each term independently. This gives the periodogram, which is an estimate of the power spectrum.

The periodogram is a rather unstable estimate and various smoothing methods may be applied.

3.8.4.3 Correlation

The calculation of auto and cross-correlation functions may be undertaken by transform methods since the correlation function and power spectrum form a transform pair.

Autocorrelation is performed by transforming the surface data, squaring term by term and then retransforming back to the spatial domain. For cross-correlation both data sequences must be transformed and a cross-power spectrum obtained so that more computational effort is needed. However, for raw data, the system of performing two transforms at once using the FFT can reduce the total time taken.

The transform technique of obtaining correlation functions always gives all possible lag values up to the total data points, whereas, particularly with autocorrelation, often only lags up to a small proportion of the data sequence length are needed. Under these circumstances it may be computationally quicker to use traditional lag calculation for the autocorrelation function, especially if the total data length is fairly small.

The DFT operates by imposing a periodicity of wavelength equal to the data sequence length in that the Fourier integral is defined for an infinite waveform made up of repeats of the actual data. In forming the correlation function it is these constructed infinite sequences that are, in effect, lagged so that points lagged beyond the end of the sequence do not “drop-off” but match up with points on the start of the first repeat. This looks like a circular correlation where the data is placed in a complete, closed circle and the lag produced by rotating this.

This effect can be useful in roundness measurement, for instance when comparing two randomly oriented profiles. However, it can lead to erroneous values in linear situations. To overcome this, a sequence of zero-value points of length up to at least the maximum desired lag should be appended to the data sequence. The action of these is made clear in Figure 3.45.

3.8.4.4 Other Convolutions

The main use of convolution is covered under filter and correlation. The heading here just stresses that any convolution in space may be handled as multiplication in frequency. Beware of the circular convolution of the DFT “described” above.

3.8.4.5 Interpolation

It is easily possible to produce an interpolation of a data sequence in which the harmonic quality of the original data is retained in the expanded sequence.

The discrete transform pair always has the same number of points in the sequence in either domain. Thus a short data sequence may be transferred and the frequency domain sequence expanded by adding zero-value elements at the high-frequency position (the center, since the transform is symmetrical). Inverse transformation will produce a longer data sequence with harmonic interpolation, since no non-zero-frequency components have been added.

3.8.4.6 Other Analysis in Roughness

Various other signal-processing techniques use Fourier transforms. Discrete transforms can be used to produce a form of discrete Laplace transform analysis.

Another possibility is Cepstrum analysis. The Cepstrum is defined as the transform of the logarithm of the power spectrum. As such it highlights periodicities of the power spectrum, which are, of course, harmonic sequences in the original data. Should the power spectrum of the data consist of two or more harmonic sequences multiplied together in the Cepstrum (log spectrum) they will be added. The power spectrum of this will clearly separate these different added harmonic sequences.

3.8.4.7 Roundness Analysis

The Fourier transform will yield roundness reference lines from full circles since the first term of the Fourier series is the amplitude of the least-squares limaçon.
Similarly approximation to reference ellipses, and so on, can be obtained. Because of the computation involved, it is unlikely that the DFT would be used just to determine the reference circle.

Direct evaluation of the harmonic series (since the circle is periodic) also seems useful for assessing the performance of bearings.

Perhaps more could be done using the DFT to produce circular cross-correlations, for instance to allow the comparison of cam-forms with stored masters without the need for accurate orientation. Specific applications of Fourier transforms will be dealt with in many places within this book.

3.9 TRANSFORMATIONS IN SURFACE METROLOGY

3.9.1 General

The application of random process analysis has had a profound influence on the understanding of surface texture from the point of view of generation and of function. Random process analysis is concerned with the Fourier transform and its inverse. However, the success of the Fourier transform has been so great that it has brought on an appetite for more and better transforms. Do they exist?

Are there transforms which have simpler or faster properties? Does another transform tell more about the surface?

These questions have been raised many times in recent years and have produced a plethora of transforms in the literature almost equivalent to the “parameter rash” reported earlier [45].

While there is an undeniable excitement in introducing a new transform into the field, care has to be exercised that it does not produce an over-reaction. What usually happens is that a transform may give a benefit in one area, but not overall.

The transforms introduced fall into two main categories:

1. Faster and simpler ways of producing comparable results
2. Transforms giving more information

In the former class there are the Walsh, Hartley, Hadamard, BFORE, and Haar transforms (all orthogonal transforms) and in the latter case the Wigner, ambiguity, wavelet, and Gabor transforms (these being space–frequency transforms).

Another factor that enters into the question of the use of transforms different from the Fourier is that the old criteria of calculation are no longer as pressing as they once were, for example storage size and speed of implementation. Nowadays storage is no problem and the speed of modern, even small, computers is such that real-time operations are possible, so the benefit obtained by using other transforms has to be significant to warrant their use. Also there is the point that the Fourier transform is very well known and understood and is standard in many existing instruments. To displace it would pose many educational and usage problems.

The basic question in surface metrology is this: how relevant to surface behavior are new transforms that have usually been devised because of new demands in the subjects of communications and coding? Fundamentally the new transforms are devised for temporal properties. They all need to be transformed in one way or another to be useful in surface metrology. A case in point could be the curvature of an areal summit. It is hard to see how this could be important in communications yet it is very important in tribology. Unless or until transforms are developed which deal with the vertical (or top-down) properties of spatial features directly and not via the Fourier transform and the multinormal distribution, it seems that it is probably best to stick to the Fourier approach. However, because of their existence in the literature, mention of them will be made here. Some transforms are very close to the Fourier transform, for example the Hartley transform [46].

3.9.2 Hartley Transform

The Hartley transform \(H(u, v)\) is the difference of the real and imaginary parts of the Fourier transform \(F(u, v)\). Thus if

\[
F(u, v) = F_{\text{real}}(u, v) + F_{\text{imag}}(u, v) \quad (3.324)
\]

then

\[
H(u, v) = F_{\text{real}}(u, v) - F_{\text{imag}}(u, v)
\]

\[
jF_{\text{imag}}(u, v) = \frac{1}{2}[F(u, v) - F(-u, -v)] \quad (3.325)
\]

A coherently illuminated object in the front focal plane of an optical system produces the Fourier transform in the back focal plane. Unfortunately this is not the case in the Hartley transform, but there are some features of the Fourier transform that are inconvenient. For example, the phase of the field contains substantial information but available optical sensing elements respond to intensity and not to phase. A photographic record of a Fourier transform plane abandons phase information which can be very important. On the other hand, the direct recording of the squared modulus of the Hartley transform would provide much more information. In cases where the transform does not go negative \(|H(u, v)|^2\) suffices to recover \(f(x, y)\) in full. In other cases, where the Hartley transform does go negative, knowledge of \(|H(u, v)|^2\) does not by itself determine the sign; however, sign ambiguity is a much less serious defect than the complete absence of phase knowledge and sign knowledge can often be inferred.

The Hartley transform does not inherently use complex notation:

\[
H(u) = \int_0^l f(x) \cos(kux) dx, \quad (3.326)
\]
where the symbol “cas” means cosine and sine:

\[ \text{cas} \theta = \cos \theta + \sin \theta. \]  \hspace{1cm} (3.327)

The discrete 1D Hartley transform offers speed gains over the FFT for numerical spectral analysis and therefore it has great potential in communications, but there has been difficulty in carrying the advantage over to more than one dimension. This is because, whereas \( \exp[-jk(ux+vy)] \) is easily separable into two factors, the function \( \text{cas}[k(ux+vy)] \) is not. Until this becomes straightforward it seems that the Hartley transform will be of use only in profile evaluation [35].

### 3.9.3 Walsh Functions—Square Wave Functions—Hadamard

The most popular of the square wave functions is the Walsh function. It is similar to the Fourier transform except that the sinusoidal function is replaced by an on-off signal, that is, a pseudo-square-wave signal.

Because it is not a sinusoidal form the frequency axis has been renamed “sequency”. This is a measure of “half the number of zero crossings per unit time or length”, so any signal can be represented by a Walsh spectrum as opposed to a Fourier spectrum. Thus

\[
f(\theta) = c_0 + c_1 \text{Wal}(1, \theta) + c_2 \text{Wal}(2, \theta) = a_0 + a_1 \text{Cal}(1, \theta) + a_2 \text{Cal}(2, \theta) + b_1 \text{Sal}(1, \theta) + b_2 \text{Sal}(2, \theta)
\]

where \( \text{Cal} = \cos \) Walsh and \( \text{Sal} = \sin \) Walsh.

The Walsh functions, originally defined in 1923, take only the values \( 1, 0, -1 \) and consequently they are very fast to compute.

If the Walsh function is factorized using the Cooley–Tukey FFT routine then the speed advantage over the FFT is considerable because the usual routine of complex multiplication is replaced by one of multiplying simply by \( \pm 1 \) (similar to the Stietjes technique for correlation).

Which of the two methods, Fourier or Walsh, is best? The answer is that it depends on the use. Common-sense metrology would say if the surface geometry is angular or discontinuous in any way, use the Walsh; if it is continuous and smooth use the Fourier, the argument being that the metrology should follow the function wherever possible. This is illustrated by the fact that a square wave has a Fourier transform of an infinite number of sinusoidal components, and one Walsh component. A sine wave has infinite Walsh coefficients, yet only one Fourier coefficient [47].

It is also beneficial that the surface can be reconstituted from the Walsh spectrum albeit with many coefficients.

The whole of the Walsh philosophy is discontinuous rather than continuous and as such its operation and properties have been examined by Smith and Walmley [47] using dyadic calculus. In this Nayak's results have been shown to have an equivalent derived via the Walsh spectrum rather than the Fourier. Because the Walsh spectra \( P_w \) are most suited to discrete operation they appear in the literature in this form:

\[ F_w(\theta) = \frac{1}{N} \sum_{n=0}^{N-1} f(n) \text{Wal}(n \theta) \quad \theta = 0, 1, 2, \ldots, N-1 \]

and the Walsh power spectrum by

\[ P_w(\theta) = P_w^2(\theta) \quad \theta = 1, 2, \ldots, N/2-1 \]

Unlike the power spectra of the Fourier transform the Walsh transform is not invariant to circular time shifts. This has led to the development of phase-invariant square wave transforms, in particular the Walsh phase-shift-invariant transform which is obtained from the autocorrelation function by means of a series of translation matrices. This has the effect of summing and averaging all the possible Walsh transforms of the time-shifted versions of the data—not very elegant but effective!

Another possibility is the BIFORE transformation \( P_b \), which is obtained from the Walsh transforms (BIFORE = binary Fourier representation):

\[ P_b(\theta) = \frac{1}{2^{N/2}} \sum_{N=1}^{N/2} P_w(2^n \theta - 2^{n-1}) \quad \theta = 1, 2, \ldots, \log N. \]

This has only \( 1 + \log_2 N \) spectral values spaced logarithmically in sequence.

Mulvaney [48] compared some of these transforms and came to the general conclusion that, after all, the Fourier transform is best for surface characterization. Although the orthogonal binary transforms such as the Walsh, Haar, etc., were faster they did not give spectral estimates which converged rapidly as a function of record length. The phase-invariant ones, Fourier and BIFORE, were slower to evaluate yet converged quickly.

A fundamental problem keeps on emerging when considering the use of binary-type transforms for use in surface analysis and this is the difficulty in identifying strongly periodic continuous characteristics such as may well result from a surface analysis. Transforms like the Walsh do not respond well to such features; other ones are the Haar. Consequently many features that reflect machine tool misbehavior, such as chatter, would be missed at worst or degenerated at best. This seriously limits the use of such transforms. Other orthogonal transforms such as the Hadamard, like the Hartley, have been used in optics and in general signal processing. The Hadamard, which is another binary-type transform based on
the Walsh functions, gives an order of magnitude increase in speed over the Fourier transform yet inherits the difficulties of application of the Walsh for basically continuous functions. In the past few years another variant of the Walsh-like transforms has emerged. This is called the wavelet transform [49]. This, like the others above, is binary and is especially useful for the compression of data containing many edges, such as fingerprints and in some cases TV images. It is not essentially suitable for surface analysis, which is basically continuous, but it may be that with the growing importance of fractal-like characteristics the wavelet transform may find a place.

It should be noted that these classes of transform based on the binary signal were originally devised to reduce the bandwidth of a signal being transmitted from one place to another, so they can hardly be expected to provide substantial advantages in their application to surface metrology.

3.10 SPACE–FREQUENCY FUNCTIONS

3.10.1 General

There is little doubt that the most important tool of characterization has probably been that of random process analysis in terms of spectral analysis and the correlation functions. Random process analysis is a very powerful tool for extracting the most significant information from a very noisy background. So far it has been shown that it can be used for characterizing both deterministic and random surfaces. It will be shown in Chapters 6 and 7 how it can be used with great effect to give information on manufacture and performance. Some examples of this have already been hinted at.

However, there is scope for improvement. Random process analysis, as used here, is best fitted to examine stationary random surfaces and their statistical properties. It tends to bring out average statistics. But in many cases the very thing that causes failure of a component is the non-uniformity of a part, not its uniformity—corrosion patches and fatigue cracks are just two functional cases. Tool wear is one example in manufacture, neither correlation nor spectra are well placed to look for non-stationarity because they are basically simple averages in space or frequency and, as a result, tend to integrate out changes from position to position or frequency to frequency. Ideally a function should be used that has the equal capability of characterizing random and periodic signals and their non-stationary embodiments. This implies using a function which has two arguments one spatial and the other in frequency. There is a class of function which fits this description called space–frequency functions. Note, not time-frequency functions because surface metrology is in space not time. Among these there are two obvious ones: one is called the Wigner distribution, which derives from quantum mechanics [50] and the other is the ambiguity function used in radar [51]. Both have been mentioned earlier as adjuncts to the Fourier transform.

The ambiguity function $A(\chi, \sigma)$ and the Wigner distribution $W(x, \omega)$ are defined as follows

$$A(\chi, \sigma) = \int \int F(f + \frac{\chi}{2}) F^*(f - \frac{\chi}{2}) \exp(-j\omega x) dx$$

$$W(x, \omega) = \int \int W(x, \omega) \exp(-j(\sigma x - \omega\chi)) dx d\omega$$

(3.331)

(3.332)

Notice that there are two formulae for each function one spatial and the other in frequency.

They are related by the following formulae

$$A(\chi, \sigma) = \frac{1}{2\pi} \int \int W(x, \omega) \exp(-j(\sigma x - \omega\chi)) dx d\omega$$

(3.333)

$$W(x, \omega) = \frac{1}{2\pi} \int \int A(\chi, \sigma) \exp(j(\sigma x - \omega\chi)) d\chi d\sigma$$

(3.334)

Both the ambiguity function and the Wigner function are space–frequency functions occupying neither space nor frequency but both. They can be thought of as being in between the two. This makes them suitable for easy access to either domain.

The basic difference between the two is that the ambiguity function is basically a correlation (because it integrates over the variable $x$ leaving shift $\chi$ constant) and the Wigner distribution is basically a convolution (because it retains $x$ and integrates over $\chi$ which is in effect the dummy argument in the convolution). It is this latter property—the retention of the value of $x$, the position in space or position in frequency—which makes the Wigner distribution the most useful of the two in engineering, although the ambiguity function has uses in optics.

Some properties will be given for comparison with the Fourier transform without proof.

3.10.2 Ambiguity Function

This is

$$A_r(\chi, \sigma) = \int \int f\left(x + \frac{\chi}{2}\right) f^*\left(x - \frac{\chi}{2}\right) \exp(-j\sigma x) dx$$

(3.335)
and

\[ A_f(\chi, \omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(\chi, \omega) \exp(-j\omega x) dx \]  

Figure 3.46 summarizes the relationship.

\[ f(x-x_0) \xrightarrow{A} A_f(\chi, \omega) \exp(-j\omega x_0) \]  

where \( \xrightarrow{A} \) denotes “ambiguity function of”.

**Spatial shift**

\[ f(x) \exp(j\omega_0 x) \xrightarrow{A} A_f(\chi, \omega) \exp(j\omega_0 x). \]  

**Frequency shift**

\[ f(x) \exp(j\omega_0 x) \xrightarrow{A} A_f(\chi, \omega) \exp(j\omega_0 x). \]  

**Convolution**

If \( g(x) = f(x) * h(x) \), then

\[ A_g(\omega, \chi) = \int_{-\infty}^{\infty} A_f(\chi, \omega) A_h(x-\chi, \omega) dx. \]  

**Modulation**

If \( g(x) = f(x) h(x) \), then

\[ A_g(\chi, \omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} A_f(\chi, \omega) A_h(\chi, \omega-\omega) d\omega. \]  

### 3.10.3 Discrete Ambiguity Function (DAF)

For the discrete form of a complex signal \( f(n) = 2, 1, 0, 1 \), the discrete ambiguity function (DAF) \( A_f[n, \eta] \) is

\[ A_f[n, \eta] = \sum_{n=0}^{\infty} 2f[n + \eta] f^*[n - \eta] \exp(-j2\eta n) \]  

or

\[ A_f[n, \eta] = \frac{1}{\pi} \int_{-\pi}^{\pi} F(\theta + \nu) F^*(\theta - \nu) \exp(j2\eta \theta) d\theta \]  

Note that the function is discrete in the spatial domain and continuous in the frequency domain. Note also the frequency doubling.

As in the continuous case

\[ f[n] \xrightarrow{A} A_f[n, \nu] \]  

where \( A \) is the ambiguity transformation.

### 3.10.3.1 Discrete Ambiguity Function Computation

**Computation** (where \( n \) is restricted to 0, \( N-1 \), and \( N \) is even)

1. Compute \( r[n, \eta] = 2f[n + \eta]^*[n - \eta] \) for \( n = 0, 1, ..., N-1, \eta = -M, ..., 0, ..., M \) where \( M = -N/2 \).
2. Using the DFT compute

\[ \sum_{n=0}^{N-1} r[n, \eta] \exp(-j\frac{k\pi n}{N}) \]  

for \( \eta = -M, ..., 0, ..., M \).
3. \( A_f[n, \eta] \) is then obtained for \( \omega = k\pi N \) where \( k = 0, 1, ..., N-1 \) and \( \eta = -M, ..., 0, ..., M \). Because \( A_f[n, \omega] \) is restricted in \( \eta \) and periodic in \( \omega \), so \( A_f[m, k] = A_f[n, k\pi N] \) is obtained for \( k, \eta = -2, -1, 0, 1, 2 \); for odd \( N \) it is similar provided that it can be represented by \( N = N_1, N_2, ..., N_1 \Delta \).

The point to notice about the discrete ambiguity function is that the variable is the spatial \( x \) and spectral \( \omega \) shown in Figure 3.47. It covers an area in \( x, \omega \) space of area \( \chi \omega \); for this reason features of the order \( \chi \omega \) will be highlighted in the space–frequency domain—hence the comment that the ambiguity function is good for identifying the presence or absence of features \( (\chi, \omega) \). However, it is not good for positioning absolutely in space or frequency; in other words, it is not very useful for finding the non-stationarity of a signal but good for detecting it!.

The box \( \chi, \omega \) scanning the space–frequency domain acts as a correlation with \( \chi, \omega \) being fixed for all \( x \) and \( \omega \) and acting as space–frequency lags. The instantaneous values extend by \( x' \leq \chi/2 \) and \( \omega' \leq \omega/2 \), where \( x', \omega \) are the positions at the center of the domain in \( x', \omega \) of extent \( \chi \) and \( \omega \), as seen in Figure 3.32.
### 3.10.4 Wigner Distribution Function $W(x, \omega)$

\[
W_f(x, \omega) = \int_{-\infty}^{\infty} f(x + \frac{\omega}{2}) f^*(x - \frac{\omega}{2}) \exp(-j\omega \chi) \, d\chi
\]

\[
= \frac{1}{2\pi} \int \overline{F(\omega + \frac{\omega_0}{2}) F^*(\omega - \frac{\omega_0}{2})} \exp(j\omega_0 \chi) \, d\omega_0.
\]

(3.344)

#### 3.10.4.1 Properties

\[f(x) \longleftrightarrow W \rightarrow W_f(x, \omega)\]  \hspace{1cm} (3.345)

$W$ is Wigner transformation.

**Symmetry**

\[W_f(x, \omega) = W_f(x, -\omega).\]  \hspace{1cm} (3.346)

**Reality**

For any complex-valued signal the Wigner distribution is always real:

\[W_f(x, \omega) = (W_f(x, \omega))^*.\]  \hspace{1cm} (3.347)

**Spatial shift**

\[f(x + x_0) \longleftrightarrow W \rightarrow W_f(x + x_0, \omega).\]  \hspace{1cm} (3.348)

**Frequency shift**

\[f(x) \exp(j\omega_0 \chi) \longleftrightarrow W \rightarrow W_f(x, \omega - \omega_0).\]  \hspace{1cm} (3.349)

**Spatial energy**

\[\frac{1}{2\pi} \int \overline{W_f(x, \omega)} \, d\omega = \|f(x)\|^2\]  \hspace{1cm} (3.350)

**Frequency energy**

\[\int \overline{W_f(x, \omega)} \, dx = \|F(\omega)\|^2\]  \hspace{1cm} (3.351)

**Total energy**

\[\frac{1}{2\pi} \int \int \overline{W_f(x, \omega)} \, dx \, d\omega = \|f(x)\|^2\]  \hspace{1cm} (3.352)

The integral of the Wigner distribution over the whole plane ($\omega_0$) is the total energy of the signal $f(x)$.

**Convolution**

If $g(x) = f(x)^* h(x)$, then

\[W_g(x, \omega) = \int \overline{W_f(\chi, \omega)} W_h(x - \chi, \omega) \, d\chi.\]  \hspace{1cm} (3.353)

#### Modulation

If $g(x) = f(x)^* h(x)$ then

\[W_g(x, \omega) = \frac{1}{2\pi} \int \overline{W_f(x, \omega)} W_h(x, \omega - \omega) \, d\omega.\]  \hspace{1cm} (3.354)

#### 3.10.4.2 Analytic Signals

**Note:**

For real-valued signals mainly found in surface texture, in order for the Wigner distribution to be applied the signal should be converted to the analytic form (whose frequency content is $0$, $\omega < 0$). Thus, taking the Hilbert transform in order to get the analytic signal yields

\[\hat{f}(x) = \int \frac{f(\chi)}{x - \chi} \, d\chi\]  \hspace{1cm} (3.355)

so that $f_a = f(x) + j\hat{f}(x)$. Rendering the signal analytic reduces the risk of aliasing because it reduces the artifact of frequencies less than zero.

It is possible that both the Wigner and the ambiguity functions have many useful properties in engineering (see Refs. [52, 53]). In this context it will suffice to derive the digital form and the moments of the Wigner function and then to consider surface data $z(n)$.

In discrete form

\[W_f(n, \theta) = \sum_{\eta=-\infty}^{\infty} 2 f(n + \eta) f^*(n - \eta) \exp(-j2\theta)\eta\]  \hspace{1cm} (3.356)

where $F(\eta)$ is the DFT of $f(n)$ and

\[f_a(n) = f(n) + j\hat{f}(n)\]

where $\hat{f}(n)$ is the discrete Hilbert transform of $f[n]$, $H_a f(n)$:

\[f(n) = (H_a f)[n] = \sum_{\eta=-\infty}^{\infty} f[\eta] \frac{\sin^2(\pi(n-\eta)/2)}{(n-\eta)}\]  \hspace{1cm} (3.357)

#### 3.10.4.3 Moments

The discrete moments are important in surface analysis: the zeroth in frequency is

\[P_f(n) = \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} W_f(n, \theta) \, d\theta.\]  \hspace{1cm} (3.358)

Because the argument in the Wigner distribution has two variables, moments exist in both space and frequency as well as global moments involving both.
The first order, $\Gamma_j(n)$, is

$$\Gamma_j(n) = \frac{1}{2} \arg \left( \int_{-\pi/2}^{\pi/2} \exp(j2\theta) W_j(n, \theta) d\theta \right). \quad (3.359)$$

Then

$$\Gamma_j(n) = \frac{1}{2} \arg (f(n + 1) f^*(n - 1)). \quad (3.360)$$

If $f(n) = v[n] \exp(j\phi(n))$, then

$$\Gamma_j(n) = \frac{\varphi(n + 1) - \varphi(n - 1)}{2} \mod \pi \quad (3.361)$$

that is, the first-order moment in frequency represents the instantaneous frequency, being the differential of the phase $\phi$ over the region.

The second-order moment in frequency, $M_j(n)$, is

$$M_j(n) = \left\{ \begin{array}{ll}
P_j(n) & \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} \exp(j2\theta) W_j(n, \theta) d\theta \\
- \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} \exp(j2\theta) W_j(n, \theta) d\theta 
\end{array} \right\} \quad (3.362)$$

$$M_j(n) = \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} \exp(j2\theta) W_j(n, \theta) d\theta \quad (3.363)$$

The zeroth-order moment in space, $P_j(\theta)$ is

$$P_j(\theta) = \sum_{n=-\infty}^{\infty} W_j(n, \theta) [1 F(\theta) |^2 + 1 F(\theta + \pi) |^2]. \quad (3.364)$$

The first-order moment $X_j(\theta)$ is

$$X_j(\theta) = \sum_{n=-\infty}^{\infty} W_j(n, \theta) n / P_j(\theta). \quad (3.365)$$

For analytic signals $X_j$ can be expressed as

$$X_j(\theta) = -\text{Im} \left( \frac{d \ln F(\theta)}{d\theta} \right) = -\text{Im} \left( \frac{F'(\theta)}{F(\theta)} \right) \quad (3.366)$$

and similarly the second order as

$$M_j(\theta) = -\frac{1}{2} \text{Re} \left( \frac{d \ln F(\theta)}{d\theta} \right) \quad (3.367)$$

(see Figure 3.48).

The Wigner distribution (Figure 3.48) centers on specific $x$ and $\omega$ and scans via $\chi$ and $\Theta$ over all the area. It is in effect a 2D convolution with $\chi$ and $\Theta$ as dummy variables. The 2D integral gives complete information about the spot $x'$, $\omega'$.

It was shown in Chapter 2 how the moments can be used to extract information about amplitude modulation, frequency modulation and phase modulation effects as well as chirp. Also, the spatial moments pick out the position and width of pulses. The dual character of this function enables the extraction of the most relevant information from the correlation technique and the spectral technique with the benefit of having to use only one function.

### 3.10.4.4 Digital Wigner Distribution

**Applied to Surfaces**

For practical applications the digital form of the Wigner distribution is used. For this the signal is $f(n)$ where $n = -2, -1, 0, 1, 2$, and $W(x, \omega)$ becomes the discrete Wigner distribution (DWD) $W(n, \theta)$ where

$$W(n, \theta) = \sum_{\zeta=-\infty}^{\infty} 2[z[n + \zeta] z^* [n - \zeta] \exp(-2j\theta\zeta) \quad (3.368)$$

or equivalently

$$W(n, \theta) = \frac{1}{n} \int F(\theta + v) F^*(\theta - v) \exp(2j\theta n) dv$$

where $F(\nu)$ is the DFT of $z[n]$. The digital form of the moments become zeroth order

$$P(n) = \frac{1}{2\pi} \int W(n, \theta) d\theta \quad (3.369)$$

The first-order moment $\Gamma(n)$ is

$$\Gamma(n) = \frac{1}{2} \arg \left( \int_{-\pi/2}^{\pi/2} \exp(2j\theta) W(n, \theta) d\theta \right) \quad (3.370)$$

$$= \frac{1}{2} \arg \{ z(n + 1) z^* (n - 1) \}$$
This is the instantaneous frequency; for example, for \( z(n) = v(n) \exp(j\phi(n)) \)
\[
\Gamma[n] = \frac{\phi(n+1) - \phi(n-1)}{2} \quad \text{mod} \quad \pi = \varphi'(n) \tag{3.371}
\]
and the second-order moment is
\[
m(n) = \frac{p(n) - \left(1/2\pi\right) \int_{\pi}^{\pi} \exp(2j\theta) W(n, \theta) \, d\theta}{p(n) + \left(1/2\pi\right) \int_{-\pi}^{\pi} \exp(2j\theta) W(n, \theta) \, d\theta} \tag{3.372}
\]
\[
m[n] = \frac{|f(n)|^2 - |f(n+1)|^2 f^*(n-1)}{|f(n)|^2 + |f(n+1)|^2 f^*(n-1)}. \tag{3.373}
\]
Because the Wigner function has been applied usually to roundness the data points are restricted to \((0 \rightarrow N-1)\) where \( N \) is a binary number (e.g., 256):
\[
W(n, \theta) = \sum_{m=-\infty}^{\infty} 2z(n+m)z^*(n-m) \exp(2j\theta m). \tag{3.374}
\]
Letting \( M = N/2 - 1 \)
\[
W(n, \theta) = \sum_{m=-N/2}^{N/2-1} 2z(n+m)z^*(n-m) \exp[-2j(m-M)\theta]. \tag{3.375}
\]
Letting \( \theta = k\pi/N \),
\[
W\left( n, \frac{k\pi}{N} \right) = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} z(n+m-M)z^*(n-m+M)\exp\left(-2jm\pi k/N\right) \exp\left(2jM\pi k/N\right). \tag{3.376}
\]
Therefore, to compute the WD it is only necessary to compute:
1. \( 2z[n+m\, M]z^*[n-m+M] \) where \( n, m = 0, 1, 2, \ldots, N-1 \).
2. \( (1/\sqrt{N}) \sum_{m=0}^{N-1} 2z[n+m-M]z^*[n-m+M] \exp j2mk\pi k/N \) where \( n, k = 0, 1, 2, \ldots, N-1 \).
3. \( W(n, k\pi/N) = (1/\sqrt{N}) \sum_{m=0}^{N-1} (2z[n+m-M]z^*[n-m+M] \exp j2mk\pi k/N) \) \( n, k = 0, 1, 2, \ldots, N-1 \).

Equation 3.385 represents the WD computed for \( n, k = 0, 1, 2, \ldots, N-1 \). The local moments in frequency become
\[
\text{zero order } P(n) = \frac{1}{2N} \sum_{k=0}^{N-1} W\left( n, \frac{k\pi}{N} \right) \tag{3.377}
\]
\[
\text{first order } \Gamma(n) = \frac{1}{2} \arg \left[ \sum_{k=0}^{N-1} W\left( n, \frac{k\pi}{N} \right) \exp\left(2j\pi k/N\right) \right]. \tag{3.378}
\]
How these moments can be applied is shown in an example in Chapter 6, in the detection of waviness in the form of chatter.

3.10.4.5 Examples of Wigner Distribution: Application to Signals—Waviness* 

It was suggested earlier that the Wigner function might be useful in characterizing waviness. This is because the moments of the function clearly reveal the type of modulation that might be causing the waviness even if it is non-linear, where ordinary spectral decomposition breaks down.

According to the theory given earlier, the local moments in frequency are
\[
M_{f_0} = |z(x)|^2 = |F(\omega)|^2 = |a|^2 \quad \text{the instantaneous power}
\]
\[
M_{f_1} = \text{Im} \left[ \frac{z'(x)}{z(x)} \right] = \varphi'(x) \quad \text{the frequency of the signal}
\]
\[
M_{f_2} = -\frac{1}{2} \frac{d}{dx} \left( \frac{a'(x)}{a(x)} \right) \quad \text{the envelope of the signal} \tag{3.379}
\]
where the right-hand sides of the equations are for a signal \( z(x) = a(x) \exp(j\varphi(x)) \).

For a frequency-modulated type of surface or a non-linear chirp signal it seems evident that the first moment in frequency would reveal the most information because this moment is sensitive to the phase, whereas the second moment responds to the envelope. Hence the amplitude of the envelope is most likely to be seen here, so the figures in Figure 3.49 would be easily separated. As an example, consider a signal with a quadratic change in frequency with distance—commonly called a chirp signal, \( z(x) = a \exp(j(\alpha/2)x^2) \)
\[
\text{zero order } = a^2 \tag{3.380}
\]
\[
\text{first order } = \alpha x \tag{3.381}
\]
\[
\text{second order } = 0.
\]

For a frequency-modulated signal \( z(x) = a \exp[j(\omega_0 x + \phi_0 + b \sin(\omega_m x + \phi_m))] \) the moments are
\[
\text{zero order } = a^2 \tag{3.382}
\]
\[
\text{first order } = \omega_0 + b \omega_m \cos(\omega_m x + \phi_m) \tag{3.383}
\]
\[
\text{second order } = 0
\]

which are obviously very distinguishable from each other.

For a signal with amplitude modulation \( z(x) = a(x) \exp(j\phi(x)) \)
\[
\text{the zeroth moment } = a^2(x) \tag{3.384}
\]
\[
\text{the first moment } = 0 \tag{3.385}
\]
\[
\text{the second moment } = -\frac{1}{2} \frac{d}{dx} \left[ \frac{a'(x)}{a(x)} \right]. \tag{3.386}
\]

* See also Examples in Chapter 2.
Hence it can be seen that use of the local frequency moments can isolate the various forms that envelopes can take. This is most likely to be important in functional cases.

### 3.10.5 Comparison of the Fourier Transform, the Ambiguity Function, and the Wigner Distribution

A comparison of the properties of three transforms has been given in Table 3.6. All three are based on the Fourier kernel and so can be seen to have similar properties. However, whereas the Fourier transform is in either the frequency domain or the space domain the ambiguity and Wigner functions are in both.

It is useful to think of the space–frequency domain as a “music score”. The only difference is that the space–frequency “score” shows the intensity as well as the pitch and spacing of the notes.

The ambiguity function as seen in Figure 3.47 could be considered to be a way of scanning across the whole score for a spectral pattern or a time/space pattern within a given size space–frequency frame. The Wigner function on the other hand tends to give a space–frequency “feel” for the content of the score centered on a specific $x$ and $\omega$ as seen in Figure 3.48. The common feature of both is the “score” itself.

**FIGURE 3.49** The zeroth and first moments of the Wigner distribution for a chirp signal and a frequency-modulated signal.
The Fourier transform only gives the axes! The Wigner is easily represented as a real function because it can be decomposed into the sum of two integrals on the real axis (and so conserves energy as seen below) rather than be represented as a complex function as is the ambiguity function.

There are some interesting properties of these space frequency functions. For example in principle it should be possible to reconstitute the original profile signal from the ambiguity function.

Given

\[ A_f(\chi, \sigma) = \int f(x + \chi/2) f^*(x - \chi/2) \exp(-j\sigma x) \, dx \]

then

\[ \frac{1}{2\pi} \int A_f(\chi, \sigma) \exp(j\sigma x) \, d\sigma = f(x + \chi/2) f^*(x - \chi/2) \]

and letting \( x = \chi/2 \)

So,

\[ \frac{1}{2\pi} \int A_f(\chi, \sigma) \exp(j\sigma x) \, d\sigma = f(\chi) f^*(0) \]

\[ = \text{constant} \times f(\chi) \]  

(3.384)

which is the profile, apart from a constant which is not relevant for the roughness value. The question is how to measure the ambiguity function in the first place?

The ambiguity function is not very useful for differentiating between stationary profiles having high frequencies from those having low because the energy is always concentrated at the origin, but are effective in spotting instantaneous frequency changes but not where they occur, i.e.,

\[ |A_f(\chi, \sigma)| \leq \int |f(x)|^2 \, dx = A_f(0, 0) \]  

(3.385)

from the Schwarz inequality

\[ |A_f(\chi, \sigma)| \leq \left( \int |f(x + \chi/2)|^2 \, dx \right)^{1/2} \left( \int |f(x - \chi/2)|^2 \, dx \right)^{1/2} \]

\[ = \int |f(x)|^2 \, dx = A_f(0, 0). \]  

(3.386)

This is completely different from the energy picture given by the Wigner distribution which can be regarded as a measure of how the energy is distributed over the space–frequency domain, i.e.,

\[ \int W_f(x, \omega) \, dx = |F(\omega)|^2 \text{ and } \frac{1}{2\pi} \int W_f(x, \omega) \, d\omega = |f(x)|^2. \]

(3.387)

This indicates that the integral of the Wigner distribution over \( x \) at a given frequency gives the energy spectrum at that frequency and that the integral over frequency for a given position gives the energy at that position. Integrating over both gives the total energy i.e., the total value of the global power spectrum. From these comments it seems that the Wigner on balance may be the most useful function overall but the Fourier transform has the advantage that it can be measured optically by diffraction.

### 3.10.6 Gabor Transform

The key to the Gabor transform is its discrete form. It is defined as

\[ z(i) = \sum_{m=0}^{\infty} \sum_{n=0}^{N-1} C_{m,n} h_{m,n}(i) \]

(3.388)

where

\[ C_{m,n} = \sum_{i=0}^{\infty} z(i) y_{m,n}^*(i) \]

(3.389)

where \( z(i) \) is the discrete space or surface signal, \( C_{m,n} \) is a Gabor coefficient, \( h_{m,n}(i) \) is the basis Gaussian function used in Gabor transformations, \( y_{m,n}(i) \) is called the bioradial auxiliary function, \( y^* \) is the conjugate and where

\[ h_{m,n}(i) = h(i - m\Delta M) W_{p|\Delta N} \]

\[ y_{m,n}(i) = \gamma(i - m\Delta M) W_{p|\Delta N} \]

(3.390)

Where \( \Delta M, \Delta N \) are sample intervals in the space and the frequency domain.

In general a Gabor transform is not unique. Depending on the selections of \( h(i) \) and \( \gamma(i) \) other transform pairs are possible.

The orthogonal-like Gabor transform maps a space (surface) signal which is composed of Gaussian functions separated in space into a joint space–frequency function with complex coefficients. These coefficients represent the amplitude of a space-shifted frequency-modulated Gaussian function centered at \( x_f, x_s f_s \) etc. for this space waveform. The Gabor coefficients can be used to reconstruct the original space series or to calculate power spectral density using the pseudo Wigner–Ville distribution. This is given by

\[ \text{Wigner–Ville}(i, k) \]

\[ = 2 \exp \left[ \left( \frac{i - m\Delta M}{\sigma^2} \right)^2 + \left( -\frac{2\pi\sigma}{L} \right)^2 (k - n\Delta N) \right] \]  

(3.391)

This equation can be precomputed and stored in a look-up table [54].
After the Gabor coefficients are obtained the Wigner–Ville distribution is applied to each of the time-shifted and frequency-modulated Gaussian functions.

One of the most promising applications of the Gabor transformation is to non-stationary surface signals, in much the same way as the Wigner function. However, it may be that it is better because it is virtually optimized in both space and frequency.

There is another possible benefit, which relates back to the discussions on waviness and filtering, and this concerns the Gaussian weighting function used to separate waviness. It may be that because the Gabor transform is also Gaussian, some practical benefit may be obtained.

So far the Gabor transform has not been used in surface analysis but it seems an obvious candidate for use.

3.10.7 Wavelets in Surface Metrology

The wavelet function is arbitrary but in order to be usable as a general operator the various scales of size have to be orthogonal—much as a sine wave and its harmonics are in the Fourier Series [58].

Wavelets can be very different in shape. Usually the actual shape used depends on the application. This causes two problems. One is that it is beneficial to have some knowledge of the basis of the data being operated on. Another problem is that there is usually no fall back shape to act as a contingency plan should the wavelet be unsuitable.

The general form is given by the wavelet transform.

\[ W(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} h^*(t-b/a) y(t) dt \]  

(3.392)

where \( h^* \) is the complex conjugate of the wavelet equation \( h(t) \).

This is in general a form for profile data. Notice that the form of \( h \) is arbitrary but the arguments \( a \) and \( b \) are not. The wavelet transform has a positional parameter \( b \) and a scale parameter \( a \). It does not contain frequency as such. The argument \( t-b/a \) or \( x-b/a \) could be more general still if raised to index \( n \), which more precisely indicates multiscale properties than just \( a \), especially in terms of fractal powers. This index is used for flexibility in other distributions such as the Weibull distribution although it has to be said that the kernel is exponential. However, \( h(t-b/a)^n \) has fractal and wavelet echoes.

The digital equivalent of Equation 3.392 is

\[ W(IT,a) = T \frac{1}{\sqrt{a}} \sum_{n} h^* \left[ \frac{(n-i)T}{a} \right] y(nT) \]  

(3.393)

where \( N \) is the number of samples and \( T_s \) is the sample interval. This representation is the discretized continuous wavelet transform (DCGT). This can be represented as a finite impulse response filter (see Figure 3.50).

Here the \( Z \)s refer to the \( Z \) transform and not height.

![FIGURE 3.50 Digitized continuous wavelet transform.](image_url)

The methods of calculation of these transforms can be graphical. (See Handbook of Surface Metrology, 1st Edition [55]).

3.11 Surface Generation

3.11.1 Profile Generation

The generation of surfaces in a computer for use in functional experiments requires a knowledge of digital characterization. This is essential in order to match the discrete behavior of surfaces as measured with those generated artificially. The third link is to relate the digital representation to the theory. How to do this has been described earlier in the surface characterization section (e.g., Ref. [99] in Chapter 2). Unless the digital generation is tied up properly, a comparison between simulation experiments will be meaningless. Consequently, a small review of the issues is needed to tie up the possible approaches to the generations that have been carried out in the past. To do this, consider first a profile. In its simplest form three running ordinates would be needed (to enable peaks as well as slopes to be described).

The joint probability density of three random variables is \( p(z_1, z_2, z_3) \). This can be rewritten in terms of conditional densities to give

\[ p(z_1, z_2, z_3) = p(z_1) p(z_2/z_1) p(z_3/z_2, z_1). \]  

(3.394)

The first approach to generation would be to select one ordinate from a random number generator. This first number could be generated from any random number generator. This usually means that there is a uniform chance of any height occurring. There are very many possibilities for this starting number and it is not within the remit of this book to consider every possibility. However, an example might be useful.

If the set of generated numbers is to have zero mean value and zero correlation between the ordinates, then methods such as congruent and mid-square techniques could be used. Both are based on elementary number theory.

As an example, if the set of numbers are denoted by \( [z_n], \) \( n = 0, 1, 2, 3, \ldots, \) the congruent method of generation is such that adjacent ordinates are related by the recursive relationship

\[ z_{n+1} = az_n + b \text{ modulo } T \]  

where \( b \) and \( T \) are prime numbers and \( T \) is matched to the word length of the computer, \( a \) and \( b \) can be chosen to give
certain statistical properties; for example by selecting $a$, $b$ and $T$ the correlation between ordinates can be made equal to $\rho_s$, where

$$\rho_s = \frac{1 - 6b_1(1 - b_1/T)}{a_s} + e$$  \hspace{1cm} (3.395)$$

and where $a_s = a'(\mod T)$, $b_1 = (1 + a + a^2 + \ldots + a^{|\frac{T}{2}})$ $b'(\mod T)$, and $|\frac{T}{2}| < a/T$ If $a - T^{1/2}$ then $\rho_s = T^{-1/2}$. Correlated data can also be generated for the linear correlation case by the recursive relationship

$$z_{n+1} = \rho z_n + \sqrt{12(1 - \rho^2)}(R_n - 0.5)$$  \hspace{1cm} (3.396)$$

where $R_n$ is obtained from a set of independent random numbers, for example obtained by squaring the two center digits of a six-figure number—not elegant, but effective.

In the method above one ordinate is correlated to the next. This is tantamount to a first-order Markov process. Most surfaces are more complicated and are produced by multiple processes. As a result the generator has to be at least second order, which means that an ordinate is related to the next ordinate and the one next to that. Furthermore the distribution of heights is not uniform: at best it is Gaussian, at worst it is not symmetrical and may well be skewed. These issues will be considered shortly.

The most important point in surface generation is to decide upon the correlation between ordinates. This is not arbitrary but can be at least estimated from the tribology prediction routines described in Ref. [56] and anticipated from the discrete analysis of random surfaces. For example, if it is required that a surface model needs to have curvature and slope variances of $\sigma_z^2$ and $\sigma_{z_m}^2$ to test a contact theory—then the required correlations between ordinates can be worked out from the previously determined formulæ:

$$\sigma_z^2 = (1 - \rho_z)/2h^2 \hspace{1cm} \sigma_{z_m}^2 = (6 - 8\rho + \rho_z)/h^4$$  \hspace{1cm} (3.397)$$

where $h$ is the spacing between ordinates taken here to be a profile; $\rho_z$ is the correlation between ordinates and $\rho_{z_m}$ is the correlation between alternate ordinates.

The described values of $\rho_1$ and $\rho_2$ from Equation 3.397 are given by

$$\rho_1 = 1 - h^2\sigma_m^2 - \frac{\sigma_z^2 h^4}{8} \hspace{1cm} \rho_2 = 1 - 2h^2\sigma_m^2.$$

(3.398)

Other parameters can be demonstrated in a similar way.

Given these values of $\rho_1$, $\rho_2$ and $h$, the generation can take place. Assume for example that the distribution is Gaussian. This can be arranged by shaping the uniform distribution of height frequency according to the Gaussian function. Thus the first ordinate can be chosen from the distribution

$$p(z_1) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{z_1^2}{2} \right).$$  \hspace{1cm} (3.399)$$

The second ($z_{2|z_1}$) represents the second ordinate given the first at height $z_1$;

$$p(z_2|z_1) = \frac{1}{\sqrt{2\pi(1 - \rho_z^2)}} \exp \left( -\frac{(z_2 - \rho z_1)^2}{2(1 - \rho_z^2)} \right)$$  \hspace{1cm} (3.400)$$

and the third is the probability of getting a height $z_3$, given that the second took a value $z_2$ and the first $z_1$, that is $p(z_3|z_2,z_1)$:

$$p(z_3|z_2,z_1) = \frac{1}{\sqrt{2\pi(1 - \rho_z^2)}} \left( 2(1 - \rho_z^2)(1 + \rho_z - 2\rho_z^2) \right)^{1/2} \exp \left[ \left( z_3 - \frac{\rho_1(1 - \rho_z)z_2 + (\rho_1^2 - \rho_z)z_1}{(1 - \rho_z^2)} \right)^2 \right].$$

(3.401)

All of these equations follow directly from the multinormal Equation 2.65.

The set of equations above read as follows. The value of $z_2$ given a previous $z_1$ has a standard deviation of $\sqrt{1 - \rho_z^2}$ and mean of $\rho z_1$ from Equation 3.392. The following $z_3$ distribution (Equation 3.401) has a mean value of

$$\frac{\rho_1(1 - \rho_z)}{(1 - \rho_z^2)} z_2 - \frac{\rho_1^2 - \rho_z}{(1 - \rho_z^2)} z_1,$$

and a standard deviation of

$$\frac{(1 - \rho_z)(1 + \rho_z - 2\rho_z^2)(1 + \rho_z)}{(1 - \rho_z^2)}.$$

The mean values and standard deviations are found simply by factorizing out first $z_1$, then $z_2$ and finally $z_3$ from the multinormal distribution. So, in sequence, $z_1$ is picked at random from a distribution which is shaped to be Gaussian of mean level zero and standard deviation unity (actually $R^2$). This value of $z_1$ determines the mean from which the distribution of $z$ is centered, $z_2$ is picked from such a Gaussian distribution at this mean having a standard deviation which is not unity but $\sqrt{1 - \rho_z^2}$; similarly for $z_3$. Having found these three ordinates, $z_3$ replaces $z_1$, $z_3$ replaces $z_2$, and a new $z_3$ is found. This repeated exercise generates the profile.

Alternatively, but equally, a best-fit least-squares routine can be used from the linear relationship:

$$z_i + A z_{i+1} + B z_{i+2} + \ldots + e_1$$

(3.402)$$

where $e_i$ is the residual or error term. Minimizing $e^2_i$ and making $\sum e_i^2 = S$

$$\frac{\partial S}{\partial A} + \frac{\partial S}{\partial B} \ldots = 0.$$
gives the covariance matrix
\[
\begin{pmatrix}
1 & \cdots & \rho_{n-1} \\
\rho_1 & \ddots & \vdots \\
\rho_{n-1} & \cdots & 1
\end{pmatrix}
\begin{pmatrix}
A \\
\vdots \\
A
\end{pmatrix}
= 
\begin{pmatrix}
\rho_1 \\
\vdots \\
\rho_n
\end{pmatrix}
(3.404)
\]
from which it emerges that
\[A = \frac{\rho_1(1 - \rho_2)}{1 - \rho_1^2}\]
and
\[B = \frac{\rho_1^2 - \rho_2}{1 - \rho_1^2}.
\]

The standard deviation is obtained simply from the determinant of the matrix and is of course equal to that given in Equation 3.402.

If a non-Gaussian surface profile is to be generated, then this is straightforward. All that has to be done is to pick the \( z_i \) from a non-Gaussian distribution. Obviously if the distribution of the residuals from which \( z_i \) is picked is skewed, then the final surface will be skewed; similarly for kurtosis. There are relationships between the moments of the generated profile and the moments of the distribution from which the ordinate is picked.

Take for the simplest case the simplest surface. This is one which has an exponential correlation function. It corresponds to a first-order Markov process in the sense that there are only two terms in the generator \( z_{i-1} \). Thus if \( \sigma_a, S_{i}, K_p \) are the moments of the profile and \( \sigma_r, S_r, K_r \) are those of the distribution from which the surface is being generated [57],

\[
\begin{align*}
\sigma_r^2 &= \frac{1}{\rho_1^2} \sigma_a^2 \\
S_{i} &= \frac{S_{a_i}^2}{1 - \rho_1^2} \\
K_r &= K_a (1 - \rho_1^2) + 6 \rho_1^2
\end{align*}
(3.405)
\]

where \( \sigma_j^2 \) is obtained from the relationship

\[
M^j_\rho = \frac{1}{N} \sum_{n=1}^{N} (z - \overline{z})^j p(z)
\]
and so on for \( n = 2, 3, \) and \( 4 \), from which \( S_{ip} \) can be evaluated in terms of \( S_{ip} \) and

\[
K_a = \frac{M_i^4}{(\sigma_i^2)^2} \quad K_r = \frac{M_i^4}{(\sigma_r^2)^2}
\]

where \( M_i^4 \) is obtained from the relationship

\[
M_i^4, (A - BC) = M_i^4 D + \sigma_r^2 \sigma_p^2 (E + BF)
\]
and where

\[
\begin{align*}
A &= (1 - \varphi_1^4 - \varphi_2^4 - 4 \varphi_1^2 \varphi_2^2)(1 - \varphi_1^2) - 6 \varphi_1^2 \varphi_2^2 \\
B &= \frac{4 \varphi_1^2 \varphi_2^2 (1 + 2 \varphi_1^2 + \varphi_2^2)(1 - \varphi_1^2)}{1 - \varphi_1^2 - \varphi_2^2 + \varphi_1^2 + 3 \varphi_1^2 \varphi_2^2 (1 + \varphi_2^2)} \\
C &= \varphi_1 (\varphi_1^2 + \varphi_2^2 - \varphi_1^2 - 2 \varphi_1 \varphi_2) \\
D &= 1 - \varphi_2^2 \\
E &= 6(\varphi_1^2 + \varphi_2^2 - \varphi_1^2 - 2 \varphi_1 \varphi_2 (1 + \varphi_2^2)) \\
F &= 3 \varphi_1 (1 + \varphi_2^2 + \varphi_2^2)
\end{align*}
(3.407)
\]

where, for both Equations 3.406 and 3.407, the \( \varphi \) can be evaluated in terms of the correlation coefficients between ordinates \( \rho_1 \) and \( \rho_2 \). Thus

\[
\varphi_1 = \frac{\rho_1 (1 - \rho_2)}{1 - \rho_1^2} \quad \text{and} \quad \varphi_2 = \left( \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \right).
(3.408)
\]

General expressions for more complicated surfaces than second order are possible but decidedly complex. However, in practice, generations to more than second order are rarely needed. Other techniques are given in Ref. [57].

### 3.11.2 Areal Surface Generation

As stated earlier, the real requirement for functional studies is in areal rather than profile generation. However, as would be expected a number of extra problems arise. The first problem is to decide on the numerical model. It could be argued that an ARMA model would be as appropriate to the areal as to the surface profile. This is difficult to justify in practice because of the more difficult numerical stability problems associated with areal analysis.

The possible spatial configurations for areal generation are considerable even in its simplest form. For the reasons given above, only second-order surfaces will be considered—and even this produces problems.

Possible models are shown in Figure 3.51a through c. The correlation model is shown in Figure 3.51c.
There has to be a compromise between the amount of coverage needed to achieve a certain accuracy and parsimony of the model to minimize computing time.

Figure 3.51 shows three configurations using the common grid sampling. There are others which have advantages in some respects but they will not be considered for generation here [20]. The model most often used is that shown in Figure 3.51c. Notice that even for this simple arrangement eight correlation coefficients are needed to cover the cross-correlation adequately between the x and y directions (here y is taken as the other horizontal axis and z the vertical axis).

For the preferred model it is assumed that the $z_{ij}$ value is approximated by the recursive relationship

$$z_{ij} = A_0 z_{i-1,j} + B z_{i-2,j} + C z_{i-2,j-1} + D z_{i,j-2} + E z_{i-1,j-1} + F z_{i+j-2} - G z_{i-2,j-1} + H z_{i-1,j-2} + e_{ij} \quad (3.409)$$

where $e_{ij}$ is the random distribution of residuals. Taking expected values of the difference between the true and modeled values of $z_{ij}$ and using the least-squares linear technique yields the following unwieldly covariance matrix for solution:

$$\begin{pmatrix} 1 & \rho_1 & \rho_5 & \rho_8 & \rho_6 & \rho_4 & \rho_3 & \rho_5 \\ \rho_1 & 1 & \rho_3 & \rho_5 & \rho_4 & \rho_8 & \rho_6 & \rho_3 \\ \rho_5 & \rho_3 & 1 & \rho_4 & \rho_7 & \rho_6 & \rho_5 & \rho_7 \\ \rho_8 & \rho_6 & \rho_3 & 1 & \rho_1 & \rho_2 & \rho_6 & \rho_2 \\ \rho_6 & \rho_5 & \rho_7 & \rho_1 & 1 & \rho_7 & \rho_3 & \rho_7 \\ \rho_4 & \rho_6 & \rho_3 & \rho_5 & \rho_1 & 1 & \rho_4 & \rho_5 \\ \rho_3 & \rho_5 & \rho_6 & \rho_4 & \rho_2 & \rho_1 & 1 & \rho_2 \\ \rho_3 & \rho_5 & \rho_6 & \rho_4 & \rho_2 & \rho_1 & \rho_2 & 1 \end{pmatrix} A = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \rho_4 \\ \rho_5 \\ \rho_6 \\ \rho_7 \\ \rho_8 \end{pmatrix} \quad (3.410)$$

The significance of the correlation values are shown in Figure 3.51. Obviously because of symmetry, the expected value of the product of $z_{i-1,j-1}$ is the same, for example, as $z_{i-2,j-1}$ (which in this case is $\rho_5$).

There is no restriction in this equation on the distribution of the residual (i.e., the distribution from which the current point is selected). Non-Gaussian statistics can be generated in exactly the same way as for the profile, but the relationship between the skew and kurtosis of the noise distribution (residual) and the values for the generated surface are much more complicated and not presented here. By the same token, as before, the effect of non-Gaussian skew and kurtosis values of the residual distribution is considerably attenuated in the same parameter of the surface. Whatever the non-normality of the parameters the desired correlation coefficients are maintained.

Equation 3.410 has to be inverted to give the values of $A-H$ for any given set of values of $\rho_{ij}$. However, it has to be realized that because of the cross-terms, the values of $\rho$ have to be consistent with each other, so

$$|A| = |\text{cov}|^{1/2} |p| \quad (3.411)$$

where $|\text{cov}|$ is the covariance matrix which by definition is symmetrical about its main diagonal.

In order that a solution is possible the covariance matrix has to be positive definite. Furthermore, the matrix values of $\rho$ have to be self-consistent as mentioned above.

Thus, if it is assumed that the surface structure comprises two independent lay patterns, for example $z(x,y) = s(x)s(y)$, then the autocorrelation-coefficient corresponding to a diagonal trace $A(\alpha,\beta$) is given by

$$A(\alpha,\beta) = \frac{1}{E} \int_0^E \int_0^E s(x) s(x+\alpha) s(y) s(y+\beta) \, dx \, dy \quad (3.412)$$

which corresponds in the figure to the conclusion that, for example, $\rho_2 = \rho_{24}$, $\rho_4 = \rho_{32}$, etc. If $\rho_2$, $\rho_4$, $\rho_{32}$, etc., are not made the matrix in Equation 3.410 will be ill-conditioned.

As a further check on the conditioning, the alphabetic constants $A-H$ are related in the same way that is $E = -AC$, $F = -BD$ and so on.

The variance of the distribution for the surface given by the model shown in Figure 3.51c is

$$\text{var} = 1 + A^2 + B^2 + C^2 + D^2 + E^2 + F^2 + G^2 + H^2$$

$$-2(ABho_2 + ACho_1 + ADho_5 + AE\rho_4 + AF\rho_3 + AG\rho_8 + AH\rho_6)$$

$$+ 2(AB\rho_3 + AC\rho_1 + AD\rho_8 + AE\rho_4 + AF\rho_3 + AG\rho_8 + AH\rho_6)$$

$$+ BC\rho_7 + BD\rho_5 + BE\rho_3 + BF\rho_1 + BG\rho_8 + BH\rho_6$$

$$+ CD\rho_3 + CE\rho_1 + CF\rho_3 + CG\rho_8 + CH\rho_6$$

$$+ DE\rho_5 + DF\rho_3 + DG\rho_8 + DH\rho_6 + \rho_6$$

$$+ EF\rho_5 + EG\rho_3 + EH\rho_8$$

$$+ FG\rho_3 + FH\rho_1 + GH\rho_8 \quad (3.413)$$
From this it is obvious that the variance has to be less than unity and greater than zero, which provides some constraint on the chosen values of \( \rho \). Unfortunately, owing to the size of the array, it is difficult to identify troublesome computations. Note that the direct terms \( A \rho \), etc. tend to increase the general correlation of the model, whereas the cross-terms \( AB \rho \) tend to reduce it.

A further consideration is concerned with the physical sense of the correlation coefficients being modeled. The correlation coefficient values have to be realistic in the second-order system sense, otherwise the generation suffers. One consideration is that \( 1 + \rho_2 - 2 \rho_1 \) has to be positive. This has to be true for \( \rho_1 \) and \( \rho_2 \) and for any other possible linear track across the grid.

As an example of this type of generation, correlation values can be selected and the surface generated in two dimensions. Obviously in order to start the generation off two initial random tracks have to be generated in the \( x \) and \( y \) direction. These four tracks should be discarded at the end of the run. A typical 2D surface would have 100 \( \times \) 100 tracks. One example of such a generation is shown in Figure 3.52.

The points to note are that in quadrants \( (\rho_1, \rho_2), (-\rho_1, \rho_2), (-\rho_1, -\rho_2) \) a good agreement between the arbitrary \( \rho \) values is possible. In the quadrant \( (+\rho_1, +\rho_2) \) the second-order constraint becomes obvious as it is to some extent in the \( (-\rho_1, \rho_2) \) quadrant, simply because of the difficulty of making a 2D second-order fit.

Despite the problems of consistency in the modeling of the 2D case, which will depend on its severity on the nature of the degree of isotropy, this method provides very convincing surfaces for the computer simulation of tribological functions. No doubt there are other methods available, but perhaps these do not incorporate the basic surface metrology needed in such investigations.

The foregoing 2D generation has been concerned with random surfaces. In most tribological situations a geometrical form has to be superimposed. This might be the shape of a cam, a circle or an involute. On top of this there might be a sinusoidal variation due to chatter; the possibilities are endless. Nevertheless it is the surface texture which is the essential and difficult component to which this type of generation is addressed. The other is to a greater extent deterministic.

### 3.12 Atomistic Considerations and Simulations

#### 3.12.1 General

The word atomistic implies that atoms are involved. This does not just mean that the scale of size is atomic it also has another meaning which is associated with the method of processing in which individual atoms are involved. In the mid-1980s atomic simulations for the analysis and description of the “collective behavior” of the liquid–solid state and the gas–solid state began to be introduced. Such behavior included the investigation of equilibrium and non-equilibrium systems where the evolving surface or interface may be subject to redistribution or the flow of particles such as in growth mechanisms or etching. The behavior can be broadly be broken down into two:

- **3.12.1.1 Microscopic Mechanics**
  - This includes the formation of defects and other surface features where a wide range of microscopic mechanisms might be involved and which need to be taken into account.

- **3.12.1.2 Macroscopic Propagating Surfaces**
  - The surface or interface between surfaces may exhibit a complex evolution e.g., anisotropic etching which can lead to complex multi valued surface fronts propagating with different rates and displaying different morphologies depending on the exposed surface orientation, particularly relevant for constructions used in MEMS.

There are three main approaches, Molecular Dynamics (MD), Monte Carlo (MC), and Cellular Automata (CA). Whereas the MD method tends to deal with individual atoms as they are, the MC and CA methods simplify the nature of the interactions between the atoms: simplifying them by simple rules which may be tailored to retain the essential behavior of the system. This approach enables the simulation of large system activity involving tens of millions of atoms or cells. Consideration of the differences between MC and CA has led to the adoption of the “octree” (octal tree) in both of the simulation methods because it offers substantial reductions in memory as well as allowing fast addressing and searching procedures. To illustrate the methods an example in anisotropic etching is used.

The main difference between the MC and CA methods is that the MC as its name suggests deals with a random approach whereas the CA is deterministic. When simulating using the MC technique the propagation i.e., the development of the process is stochastic. In a variant KMC (kinetic Monte Carlo) the times during the simulation process are faithfully measured and monitored so that if total evolved time or some other time factor is needed it is available.
facility could well be useful if times taken to converge to the steady state, from the transient state, are needed. In CA, however, the system evolves in a deterministic way and time is an integral feature. Just because the method is deterministic does not necessarily mean the sometimes random-like patterns do not emerge.

In addition to the random/deterministic difference there is another fundamental difference which is that MC is sequential in terms of dealing with atoms but the CA is completely parallel and deals in averages. In both methods the atoms of the surface or interface are visited one by one and their neighborhoods are inspected in order to determine the removal rate (or probability), deciding whether the atom is to be removed or to remain attached. The essential difference is that in the KMC method the atom is removed as soon as it has been decided whereas in the CA method it first decides which atoms are to be removed then it removes them altogether. So, the CA takes two loops to do a time step and the KMC takes one. This and other processing reasons make KMC more suited to the microscopic collective processes and CA better suited for surface propagation.

The MCA method has been used by Popov and colleagues, an example of which is given below in the friction problems associated between wheel and rail.

**3.12.2 Mobile Cellular Automata MCA**

The behavior of the friction mechanism between the wheel and the track of rolling stock is still not properly understood. For this reason there has been some attempts at developing new theories about the friction mechanisms involved. For example see Bucher et al. [63].

They develop a model based on “Movable Cellular Automata” (MCA). The influence of pressure and sliding velocity can be investigated using this technique which apparently has been devised in Russia just for complicated processes like this [64]. The reader is advised to consult this paper by Popov and Psakhie.

In the movable cellular automata method a medium is represented as an ensemble of discrete elements such as the center of mass position, the value of the plastic deformation, the rotation vector and discrete variables characterizing “connectivity” between the neighboring automata. The evolution of the modeled ensemble of automata was defined on the basis of the numerical solution of Newton–Euler equations of motion by the so called “integration scheme of the second order of accuracy”. This probably means that the model is a second order of approximation to the true situation. To give an example friction to a first order is simply a number i.e., the coefficient of friction whereas the relationship to the sliding velocity would have to be included to make it a second order approximation to the actual, very complex frictional behavior [65]. The MCA model seems to be a good tool to investigate the formation of a “quasi-fluid layer which is supposed to be generated during the rubbing process between the two solids. In it physical as well as chemical processes take place such as fracture, deformation and welding all at this very small scale of size. Figure 3.53a shows the structure at $F_y = 255$ Mpa and

![Figure 3.53](image-url)
Each movable automaton is further characterized by a set of mechanical parameters corresponding to mechanical properties of the simulated material. Automata can have a size chosen for a specific application but in the case of the rubbing of solid surfaces it is usually taken as the size of a typical asperity of the surfaces in contact. In the case of this investigation the modeled object consisted of four layers: an upper layer of automata moved horizontally at velocities between 1 and 10 m/s in different numerical simulations. Two intermediate levels with initial roughness in the “nm” range representing the upper and lower surface regions making contact. A lower layer arranged to represent an immovable object.

The problem is that although surface roughness affects performance in a dynamic system just those physical phenomena which cause the effect on performance also can be instrumental in changing the surface itself.

A constant normal pressure was applied to all the automata of the upper layer, the size of the automata was kept fixed at 2.5 μm in all the numerical simulations. It was assumed that the surfaces obeyed fractal characteristics, which as usual, caused problems in the matter of determining the real area of contact and the pressures at the very small sizes. The authors proceeded to get around the fractal problems by resorting to material properties such as the hardness of the asperities being finite.

The MCA technique allows various activities between bonds and couplings to be performed which enable a simulation to take place of the destruction and regeneration of the surface topographies of both the contacting surfaces. Figure 3.54 shows a path which has been drawn connecting all bonds of one of the contacting surfaces, in this case the lower surface. Having done this it is then possible to analyze this simulated profile. Figure 3.55 shows the power spectrum.

This simulation method imposes a digital increment corresponding to half the size of the automaton in the formulae below designated by δx. The profile length is l. The Fourier coefficients (a) and (b) are given by the standard formulae.

$$a(K_N) = 2 \frac{1}{l} \int_0^l f(x) \sin(K_N x) dx = 2 \frac{1}{l} \sum_{n=1}^M f_m \int_{(n-1)\delta x}^{n\delta x} \sin(K_N x) dx$$

$$b(K_N) = 2 \frac{1}{l} \int_0^l f(x) \cos(K_N x) dx = 2 \frac{1}{l} \sum_{n=1}^M f_m \int_{(n-1)\delta x}^{n\delta x} \cos(K_N x) dx$$

Here K is the wave number corresponding to the frequency. It has a decaying spectrum which the authors claim is typical of a fractal surface. It is statements like this which reduce
the credibility of the conclusions reached in this and similar work. A considerable amount of thought has gone into the MCA method of simulation and how it can describe and explain the behavior of mass and energy in the quasi-fluid zone. In fact in other work Popov and Schargott [66] liken the process of material movement and deposition to a diffusion process which seems to be a plausible mechanism.

Their argument goes as follows, let the height of the lower surface be \( h(x) \) piled up (in this intermediate region) made up of particles momenta-like. The density of such particles \( j(x, t) \) is given by,

\[
j(x, t) = -D \frac{\partial}{\partial x} h(x, t) \quad (3.416)
\]

and the equation describing the transportation of these particles as a function of time and position is

\[
\dot{h}(x, t) + \frac{\partial}{\partial x} j(x, t) = 0 \quad (3.417)
\]

The combination of the two equations describe a diffusion equation, namely

\[
\dot{h}(x, t) = D \frac{\partial^2}{\partial x^2} h(x, t) \quad (3.418)
\]

Equation 3.418 is of average behavior in a stochastic sense. At this point they introduce a stationary Gaussian distributed noise in space and time which is where the problem arises because the probability of distribution of impacts, deformation, subsequently wear in space surely has to be uniform. This automatically imposes an autocorrelation envelope which is exponential due to the Poissonian and hence Markov statistics. The power spectrum of the surface must have, therefore, a decay at large wave numbers of \( K^{-2} \) (i.e., \( 1/\omega^2 \)) which is not dependent on fractal mechanisms and so this conclusion should not be reached unless there is evidence that this MCA technique produces any decay which is significantly different from the value of \(-2\) quoted. The other non fractal indication is that all the spectra shown have a flattening at low values of \( K! \)

Unfortunately what has been described is only a taste of the very interesting technique proposed by Popov and others. It is not possible to follow the arguments properly without a knowledge of the MCA method they propose in more detail than that given here but it seems that the pressures they evaluate in the simulations are very large at the outset with the result that practically all the main assumptions are concerned with plastic flow effects almost at the start of contact thereby removing the initial surface finish aspects: the original surfaces appear to have little impact on the surfaces subsequently generated. Also, the assertion that the surfaces produced are fractal is not sustainable without more evidence. Their main conclusion is that the dynamic processes of plastic deformation and fracture at the nano-level is likely to be of most importance.

3.12.3 Simulation Considerations

To simulate means to pretend. In science this is not taken to be an attempt to fraud or any other such activity it is taken to mean an attempt to get information about an event or procedure or experiment which cannot take place for reasons of practicality, safety, cost or whatever. It is often useful to carry out an experiment, in a computer, which needs to be done but which cannot by normal means. A model of the experiment is set up in the computer having the same constituents as the actual experiment: the components, the system, the inputs and the same laws of behavior as in practice. Also the same limits on the range of operation have to be set. In all this it is assumed that the knowledge and experience of the investigator is sufficient to ensure that the model and its operation are valid i.e., that all the required constraints have been included and that the laws and rules of behavior adequately cover the practical case. Once this is assured simulation is a valuable tool which can give relevant information quickly, in a conducive environment, to allow judgments and hopefully predictions to be made which enable the science to move forward. The danger is that the simulation reflects the knowledge of the investigator and does not allow for the unknown to creep in as often happens in the practical case.

It is becoming increasingly obvious that many properties of surfaces cannot be fully explained just by using macro arguments. Examples are in friction and wear and even machining. More and more use is being made of atomistic simulations which look at the detailed interaction and motion between atoms and molecules to explain some aspects at least of complex macro properties. Such simulations use what is often referred to as “molecular dynamics.”

3.12.4 Molecular Dynamics

It is beyond the scope of this book to go into detail on this subject but some words of introduction describing the basic concepts, based loosely on the work of Harrison and Brenner [67], will be given here.

They start off by making the point that at the time their article was published there were two basic approaches which illustrate the way in which atomic activity can be tackled. In the first it is assumed that the potential energy of the system is determined by the relative positions of the atoms and can be represented mathematically in a form which allows the presence of a number of free parameters. These parameters are then chosen to describe some set of physical properties of the system of interest. Forces necessary are obtained by taking the gradient of the potential energy. The problem is that unless this potential energy function is chosen correctly the parameters extracted are determined from a limited set of properties and therefore are not entirely credible.

An alternative approach is to calculate the forces from first principles using the electronic structure. This approach limits the number of unknown parameters because of the quantum constraints which therefore allows the forces to be predicted more readily.
According to Harrison and Brenner, however, the simplest approach to developing a continuous potential energy function is to assume that the binding energy can be written as a "pair sum" between atoms i.e., the atoms are taken in pairs, indicated by the index i for one atom and j for the other.

\[ E_b = \sum_i \sum_{j \neq i} V_{\text{pair}}(r_{ij}) \]  

(3.419)

where \( E_b \) is the binding energy, \( r_{ij} \) is the distance between the atoms and \( V_{\text{pair}}(r_{ij}) \) is the assumed functional form for the energy, thus,

\[ V_{\text{pair}}(r_{ij}) = DX(X-1) \]  

(3.420)

\( D \) is a parameter which determines the minimum energy for the pair of atoms. The two common forms are Morse where

\[ X = \exp(-\beta r_{ij}) \]

and the Lennard-Jones form \( X = \left( \sigma/r_{ij} \right)^6 \)

(3.421)

in the form of "12–6" as described in Chapter 8.

These forms are quite good at describing the force behavior because the Morse exponent form models the repulsive forces between atoms and the Lennard-Jones describes the leading term in the long range dispersive forces. A compromise between the two is sometimes used called the Buckingham potential. This uses the exponential form for repulsion and a 1/r^6 form for the attraction. It is sometimes referred to as the 'exponential-6'. Although these methods are somewhat crude they have proved to be effective in giving some idea of the behavior.

An extension is to assume that the boundary energy can be written as a many body expression rather than the simple pairing in the above system. In principle this many body system should include all of the possible surrounding atoms, i.e.,

\[ E_b = \frac{1}{2} \sum_i \sum_{j \neq i} V_{2-\text{body}} + \frac{1}{3} \sum_i \sum_{j \neq i} \sum_{k \neq j} V_{3-\text{body}} + \ldots \]  

(3.422)

In practice terms above these are ignored because the influence is considered to be very small. Also the computing required is prohibitive although it is conceivable that using the parallel computing advances with DNA many more terms could be included.

This formula does not say anything about the actual nature of the interaction between the atoms; there are many possibilities outside the scope of this book but a common form would be that of a valence force field in which the atomic interactions are modeled with a Taylor series expansion incorporating bond stretches angle bends etc. For an excellent survey of the possibilities consult the work of Harrison e.g., Ref. [67].

One point which emerges when dealing with the generation of surfaces by using simulations is that it is not just a case of predicting atom and molecular trajectories under a system of forces: some very important factors have also to be taken into account. Typical of these in the metal removal process is the inclusion of a mechanism for generating and distributing the heat which is produced in a practical situation of turning for example. Under these conditions it may be necessary to introduce non-equilibrium equations of motion to maintain a constant temperature. This method has the advantage that it introduces an element of friction into the atomic forces [68] Even then factors such as the lack of tool rigidity are often omitted – the tool is considered to be infinitely stiff! A classic example often used to investigate the cutting process is a diamond tool cutting copper [69] which might be relevant for the diamond turning of copper cylinders or mirrors but is not typical of general turning.

In the next chapter, specific ways of measuring surfaces will be considered ranging from the well-known stylus methods, both traditional and modern, to optical and other methods. Chapter 4 should therefore enable people to make a better choice of instrument to use in a particular situation.

### 3.13 SUMMARY

This chapter has concentrated on ways of processing the many surface parameters which are in use. Considerable attention has been paid to filtering because of its importance in separating different geometrical components and also because of its use in providing reference lines. Also, sections have been devoted to evaluating references using minimax methods. This is because they are now being used more often, especially as dimensional and surface metrology become more integrated. Some of the ways of using new transforms have also been given. Obviously, much work has been carried out in all these fields elsewhere, but little has been targeted at surface metrology, which has its own set of completely different problems. It is hoped that this section shows how to address some of these problems computationally.

This chapter is necessarily written at different levels of difficulty because of the wide range of complexity of the metrological parameters and references which have to be evaluated. It should, however, be most useful to researchers embarking on a surface research program and possibly people who are building or modifying instruments. It is not likely to be so useful to people operating instruments, but it may give an insight into how the parameters are evaluated. There has throughout the chapter been some emphasis on mathematic techniques. This has been intentional. It is hoped that by exploring some of these techniques, the subject itself can be advanced. The purely digital methods described here have usually been discrete ways of obtaining parameters described mathematically in Chapter 2, although in some instances more than one method has been included. The digital/ discrete emphasis is necessary because this is how surfaces are now measured. Nanometrology does not need the same bandwidth as surface metrology and, from the processing point
of view, is simpler. Also, nanometrology is generally planar and so does not often involve complex 3D shape, however some information on simulation has been included because nanotechnology and hence the metrology is becoming heavily dependent on atomic simulation methods.

REFERENCES

34. BS 7172. British Standard Institute