SLR Data Screening; location of peak of data distribution

A.T. Sinclair

Royal Greenwich Observatory, Madingley Road, Cambridge CB3 0EZ, England

1. Introduction

At the 5th Laser Ranging Instrumentation Workshop held at Herstmonceux in 1984 consideration was given to the formation of on-site normal points by laser stations, and an algorithm was formulated. The algorithm included a recommendation that an iterated 3.0 × rms rejection criterion should be used to screen the data, and that arithmetic means should be formed within the normal point bins of the retained data. From 1990 September onwards this algoritm and screening criterion have been brought into effect by various laser stations for forming on-site normal points, and small variants of the algorithm are used by most analysis centres for forming normal points from full-rate data, although the data screening criterion they use ranges from about 2.5 to 3.0 × rms. At the CSTG SLR Subcommission a working group was set up in 1991 March to review the recommended screening procedure. The working group consists of A.T. Sinclair (chairman), G.M. Appleby, R.J. Eanes, P.J. Dunn and T.K. Varghese. This paper has been influenced by the discussions of this working group, although the views expressed are primarily those of this author.

The main thrust of this paper is that, particularly for single photon systems, a more important issue than data screening is the determination of the peak of a data distribution, and hence the determination of the bias of the peak from the mean. Several methods of determining the peak are discussed.

2. The effect of skew data

The first stage of forming normal points (described by Appleby in these proceedings) is to fit a trend-function to the raw ranges or to their residuals from a predicted orbit so that all signature from the orbit is removed, and then the distribution of the trend-removed data can be examined. Some level of screening is needed in the iterative process of fitting a trend-function, but this is not critical; 3.0 × rms or perhaps even tighter should be fine, and finally a wider band of trend-removed data, say 5.0 × rms, should be retained for examination of the distribution and then final screening. If these data have a symmetrical distribution about the peak then the criterion used for final screening is not critical, and a 3.0 × rms screening and use of the arithmetic mean should be fine. However if the data have a skew distribution then the arithmetic mean will be biased away from the peak, and the amount of bias will probably be dependent on the level of final screening used.

Analysis of the raw data from numerous stations shows that stations operating at a multi-photon return level per shot tend to have a fairly symmetrical distribution of data, whereas those operating at a single photon return level frequently show a significant skewness in the distribution of the data, usually skewed towards long ranges. As is described in papers by

Appleby and by Kirchner in these proceedings, some satellites (particulary Ajisai) impose their own signature on the laser data and can cause skewness as has been detected by some single-photon systems. However it is probable that some of the skewness is caused by the laser systems themselves, and it is certainly not the intention to explain away features of the system as being due the satellite. However if the skewness is due to the satellite then a means must be found in software of handling it. If it is due to the system then one view is that it is an engineering problem and that the hardware should be adjusted, but another view is that hardware will never be perfectly adjusted, and that one should accept some level of mal-adjustment and calibrate the effect out with software. Also it is possible that some detectors such as avalanche photo-diodes have an inherent skewness which would be very difficult to remove by adjustment of the hardware.

Systems which operate at a multi-photon return level and use pulse-level detection are in effect using hardware to form a mean of the individual photon returns obtained each shot, and so will see little effect from the distribution structure of the individual returns. Systems which operate at multi-photon level and detect the first photon received will primarily see the leading part of the distribution and will see little effect of any skew tail. Hence these system use a hardware method to eliminate any effects of satellite signature, and must ensure that the hardware is set up so as not to cause any biases. The objective of this paper is to devise an equivalent software scheme for systems operating at single photon return levels.

3. Choice of reference point of data distribution

The current recommendation is that the reference point of a data distribution should be the arithmetic mean of the data retained after a 3.0 × rms screening. This may not be ideal for a skew distribution of data.

If a skew data distribution is entirely caused by the laser ranging system, and if the ranges to the terrestrial calibration target have the same skew distribution as the satellite ranges, then it probably does not matter what screening criterion is used, provided the same is used for both calibration and satellite ranging, and it will probably be satisfactory to take the arithmetic mean of the data, even though this does not give the peak of a skew data distribution. However if the satellite is adding a significant contribution to the distribution, or if for some other reason the distributions of calibration and satellite data are different, then it is probably best that some means of processing the data should be devised such that the peak can be located. This is because, in the complicated convolution of the signatures of the system and the satellite, the range represented by the peak corresponds to the distance travelled by a photon from the peak of the laser pulse to the peak reflection point of the satellite (and so the centre of mass correction for the satellite should be that corresponding to peak reflection as determined in pre-launch testing).

Note that for a satellite pass it is not the peak of the trend-removed data for the whole pass that is required, but the peak of the distribution of the data within each normal point bin. This is a problem as there may not be sufficient points within a bin to give a reasonable indication of the distribution. The solution we recommend is that both the peak and arithmetic mean of the whole pass should be determined, and the difference, or bias, of the mean from the peak should thus be determined for the whole pass. Then within each normal point bin

just the arithmetic mean should be determined, but it should be corrected by applying to it the bias of the whole pass.

4. Methods of determining the peak of the data distribution

The usual method to determine the distribution of data is to plot a histogram, but this by itself does not give a good indication of the precise location of the peak. There is some arbitrariness introduced by the choice of bin width, and this is likely to be much coarser than the resolution required for the peak. An improvement can be made by fitting a curve such as a Gaussian profile to the histogram, but this total process is rather complicated, and as a Gaussian profile is symmetrical it will to some extent be influenced by a skew data set and be pulled away from the peak. In this paper we propose and examine three simpler techniques, and compare their performance on a variety of passes tracked by RGO Herstmonceux.

4.1 Data smoothing

After fitting and removing of a trend function, a plot of the data against time should be just a scatter plot about the mean, exhibiting no trend, but possibly not uniformly distributed about the mean. In order to examine this distribution we no longer consider the data as a time series, but just as points as lying along an x-axis, and our requirement is to plot in the y-direction some function describing the distribution of the points. The usual procedure is to plot a histogram, but we consider an alternative, in which each plotted residual is regarded as the most probable location of the measurement, and so we spread (or smooth) the effect of the residual each side of it using a Gaussian probability distribution. The result is that at any given location on the x-axis there will be contributions from all of the residuals, which can be summed and plotted on the y-axis. The peak of this plotted curve will give the most probable mean value of all of the residuals. The mathematical description of the method is very simple. Let $x_i, (i = 1, n)$ be the residuals of the range values from the trend function. Then for a range of values of x at, say, 10 ps intervals, evaluate and plot the quantity y, given by:

$$y = k \sum_{i=1}^{n} \exp[-\frac{1}{2}(x - x_i)^2/\sigma^2]$$

where σ is the somewhat arbitrary standard deviation of the smoothing function, although it would be reasonable to choose a value close to the single shot precision of the system. We regard the scale of y as arbitrary, and k is an arbitrary factor chosen to give some convenient maximum value of y.

Figure 1 shows a series of plots of this distribution function for a pass of Ajisai for a range of values of σ, with the conventional histogram plotted also. Apart from very small values of σ the peak is well-defined, and can be determined precisely. These plots are centred on the arithmetic mean of the distribution, so it is seen that the peak differs from the mean by about 2 cm, showing the large effect of the skewness (which is primarily caused by Ajisai - see paper by Appleby in this proceedings). A problem with the method is that the location of the peak depends on σ. As σ is increased the skewness has an increasing effect on the location of the peak, and in the plots the peak moves to the right by 2.3 mm as σ varies from 40 to 80 ps. A further problem is that for very sparse passes, possibly affected by a

significant amount of noise, the method fails to give a single main peak, or requires a large value of σ in order to do so.

4.2 Tight rejection criterion

The arithmetic mean of a set of data will be biased away from the peak due to any skewness of the data, but the amount of bias will be reduced if a tighter rejection level is used in forming the mean. We look at the effect of using various rejection levels, expressed as multiples of the root mean square difference from the mean (rms). However the rms of the retained data varies and usually gets smaller as the rejection limit is reduced, so for clarity we first determine the mean and rms using a 3.0 × rms iterated rejection level. Then subsequent rejection levels are expressed as multiples of this fixed rms. It also aids convergence with a tight rejection level if the rejection level itself does not vary as the iterations proceed. The table below gives the results of using various rejection levels on the Ajisai pass shown in Figure 1, with the various determinations of the mean given relative to the mean obtained using 3.0 × rms rejection.

Rej.	Mean(cm)	No.Pts.
3.0 × rms	0.00	1104
2.5 × rms	0.19	1081
2.0 × rms	0.58	1030
1.5 × rms	1.18	936
1.0 × rms	1.72	799
0.5 × rms	2.13	501

The peak of the distribution, as given by the smoothing method, is about 2 cm from the initial mean, and it is seen that the successive estimates of the mean move closer to the peak as the rejection level is reduced. For the rejection level of 0.5 × rms a large number of points have been rejected, and also in tests on various passes some difficulty was experienced in obtaining convergence. So for subsequent tests we have adopted a level of 1.0 × rms.

Objections that are frequently raised to using a tight rejection level are that too much data is being discarded, and that the data are being made to look better than they really are. However what we are proposing is that this tight rejection level is used only for the purpose of obtaining an estimate of the peak of the pass distribution, so that the bias of the peak from the 3.0 × rms mean can be determined. The means in the normal point bins and the value of the rms of the whole pass will be calculated from the data that remain after making a 3.0 × rms rejection.

4.3 Pearson curves

A distribution of n points x_i with mean $\bar{x}$ is characterised to a large extent by its moments μ_2, μ_3, μ_4 where

$$\mu_j = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^j$$

The second moment is the square of the standard deviation. The following quantities are also defined:

Skewness = μ_3^2/μ_2^3 indicates deviation from symmetry, = 0 for symmetry about $\bar{x}$

Kurtosis = μ_4/μ_2^2 indicates degree of peakiness, = 3 for Gaussian distribution.

These quantities are dimensionless and of restrained magnitude, and so are in some applications more convenient than the 3rd and 4th moments. This conventional definition of skewness has the disadvantage that the sign of the 3rd moment is lost, and it is this which describes the direction of the skewness. (It may be better to define skewness as $\mu_3/\mu_2^{3/2}$.) In the plots in this paper we attach the sign of the 3rd moment to the skewness.

There is a method in statistics of deriving a distribution function from values of these three moments obtained from a set of data. These are the Pearson distributions (see description by M.G. Kendall, The Advanced Theory of Statistics, Vol 1, 1947). The distribution function f of the quantity x is defined by a differential equation

$$\frac{df}{dx} = \frac{(x-a)f}{b_0 + b_1 x + b_2 x^2}$$

where

$$a = -\mu_3(\mu_4 + 3\mu_2^2)/A$$

$$b_0 = -\mu_2(4\mu_2\mu_4 - 3\mu_3^2)/A$$

$$b_1 = -\mu_3(\mu_4 + 3\mu_2^2)/A$$

$$b_2 = -(2\mu_2\mu_4 - 3\mu_3^2 - 6\mu_2^3)/A$$

$$A = 10\mu_4\mu_2 - 18\mu_2^3 - 12\mu_3^2$$

and where the origin of x is now at its mean value. The peak of the distribution curve is at $x = a$.

This differential equation has several forms of analytical solution depending on the values of the moments. For values likely to be met in practice its solution is of the form

$$f = k(1 + \frac{x}{a_1})^{m_1}(1 - \frac{x}{a_2})^{m_2}$$

but for the precise values of the moments corresponding to a Gaussian distribution this solution becomes singular, and it has an alternative solution which is in fact the Gaussian distribution. So unfortunately the analytical solution is close to a singularity in the region likely to be met in practice, and so is not a very useful way of deriving the shape of the curve. However it is easy to solve the differential equation by numerical integration starting from the peak, although there can be problems as the singularity on the x axis is approached

which require a little fudging. Some solution curves are shown in Figure 2 for a range of values of skewness and kurtosis. No attempt has been made to normalise the area under the curves; they are all plotted with the same height at the peak. It is seen that the shapes of the curves are close to a Gaussian curve until fairly extreme values of the parameters are reached, eg., skewness of 0.6 or kurtosis of 2.2, and then the shapes are not particulary typical of what is seen in some SLR data, and so it can be expected that this method will give a good estimate of the peak for distributions close to Gaussian, but not be so useful in more extreme cases.

This discussion of how to plot the Pearson curves is given here for completeness, but it is not proposed that this should be a normal procedure for SLR data handling. However the method provides a simple estimator of the location of the peak of a distribution of data. Expressed in terms of the standard deviation σ, the skewness s, and the kurtosis k, with consideration given to the sign of the 3rd moment, the displacement of the peak from the mean is

$$a = \frac{-s^{1/2}\sigma(k+3) \times \text{sign}(\mu_3)}{10k - 18 - 12s}$$

5. Comparison of methods of peak determination

Figure 3 shows the results of applying these various methods of peak determination to a number of passes tracked by RGO Herstmonceux. The passes were selected to provide a good test of the methods, and are not necessary typical passes from the station. The figures give the following information:

- the conventional histogram, using a bin of 40 ps (= 6 mm)
- the smoothing-method distribution function plotted as a solid curve
- the Pearson distribution function plotted as a dashed curve
- the 3 × rms-rejection mean shown as a solid vertical line from the top
- the 1 × rms-rejection mean shown as a dashed vertical line from the top.

The two distribution curves are plotted with slightly different peak heights for clarity, and the peak height of the histogram is limited if necesssary to be slightly below these two curves.

The information given in the captions includes:

- Smoothing parameter σ in ps (multiply by 0.15 to get mm)
- Bias(1): difference of smoothing-method peak from 3 × rms mean
- Bias(2): difference of 1 × rms mean from 3 × rms mean
- Bias(3): difference of Pearson peak from 3 × rms mean.

Figures 3(a) and 3(b) show passes with insignificant skewness, in which all methods of determining the peak and mean agree well.

Figures 3(c), 3(d) and 3(e) show passes with significant skewness, in which smoothing, Pearson and 1 × rms agree well, but are significantly different from the 3 × rms mean.

Figures 3(f) and 3(g) show passes in which smoothing and 1 × rms agree well, but the Pearson peak stands off, and all differ from the 3 × rms mean.

Figures 3(h) and 3(i) show passes in which the Pearson peak and 1 × rms agree fairly well, but the smoothing peak stands off.

From these and tests on numerous other passes we conclude that:

- for a single-photon station there is often a significant difference of the peak from the 3 × rms-rejection mean
- the 1 × rms-rejection mean usually agrees with one or other of the smoothing peak and Pearson peak, and often with both.

6. Recommendations

In conclusion we recommend the following:

a) the ranges to a calibration target or the trend-removed data from a whole satellite pass should be screened at an iterated 3 × rms level, and in the process determine *rms* and *mean* of the retained data

b) the skewness and kurtosis of the retained data should be determined

c) using this fixed value of *rms* a second determination of the mean should be made using an iterated 1 × *rms* rejection. This provides an estimate of *peak*. Then the bias of the calibration or pass is $bias = peak - mean$

d) for a calibration run, use the value of *peak* as the calibration value

e) for a satellite pass, form normal points from the screened data within each bin in the usual way, but add the correction *bias* to the normal point.

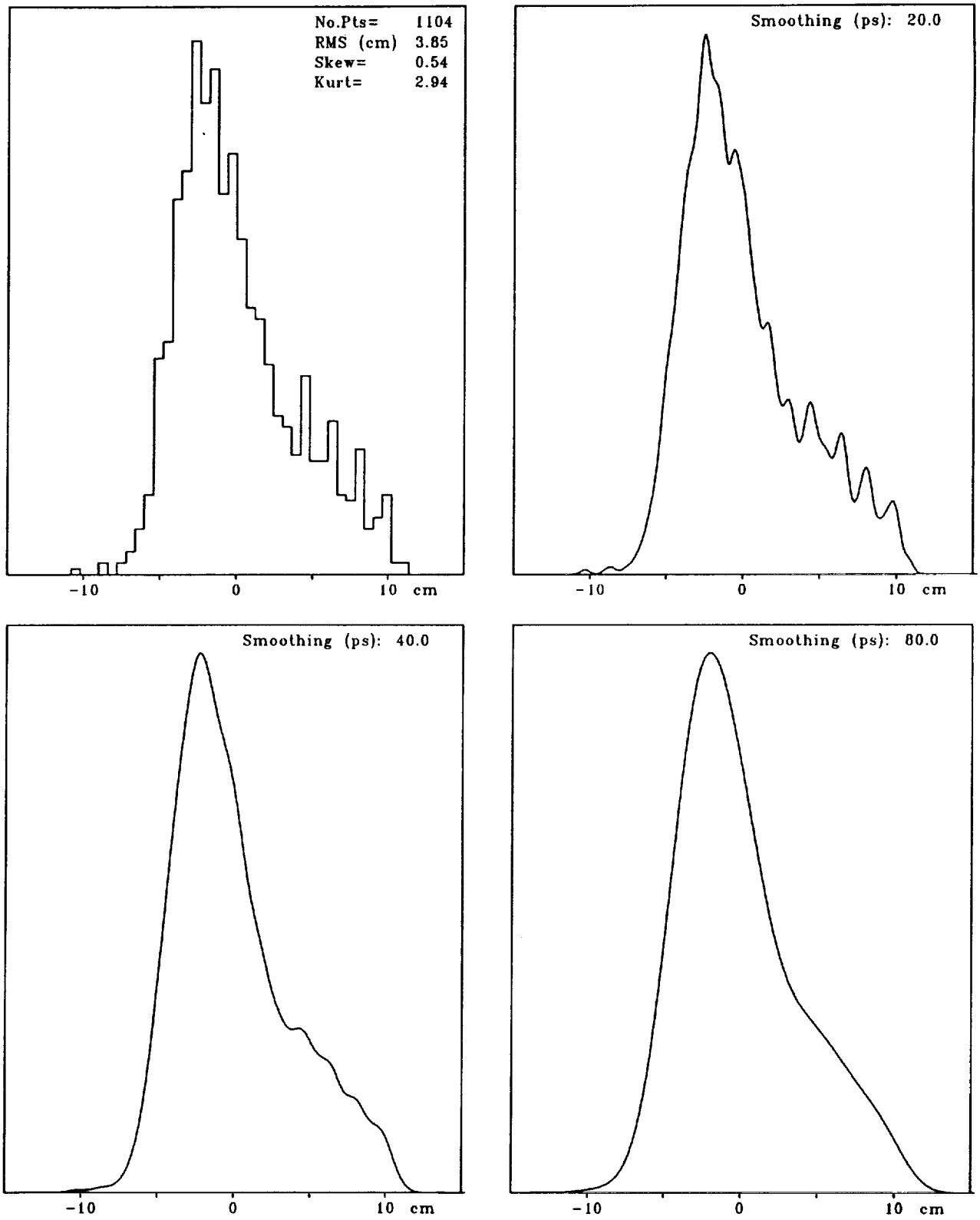

Figure 1. Use of the smoothing method to determine the distribution of an Ajisai pass.

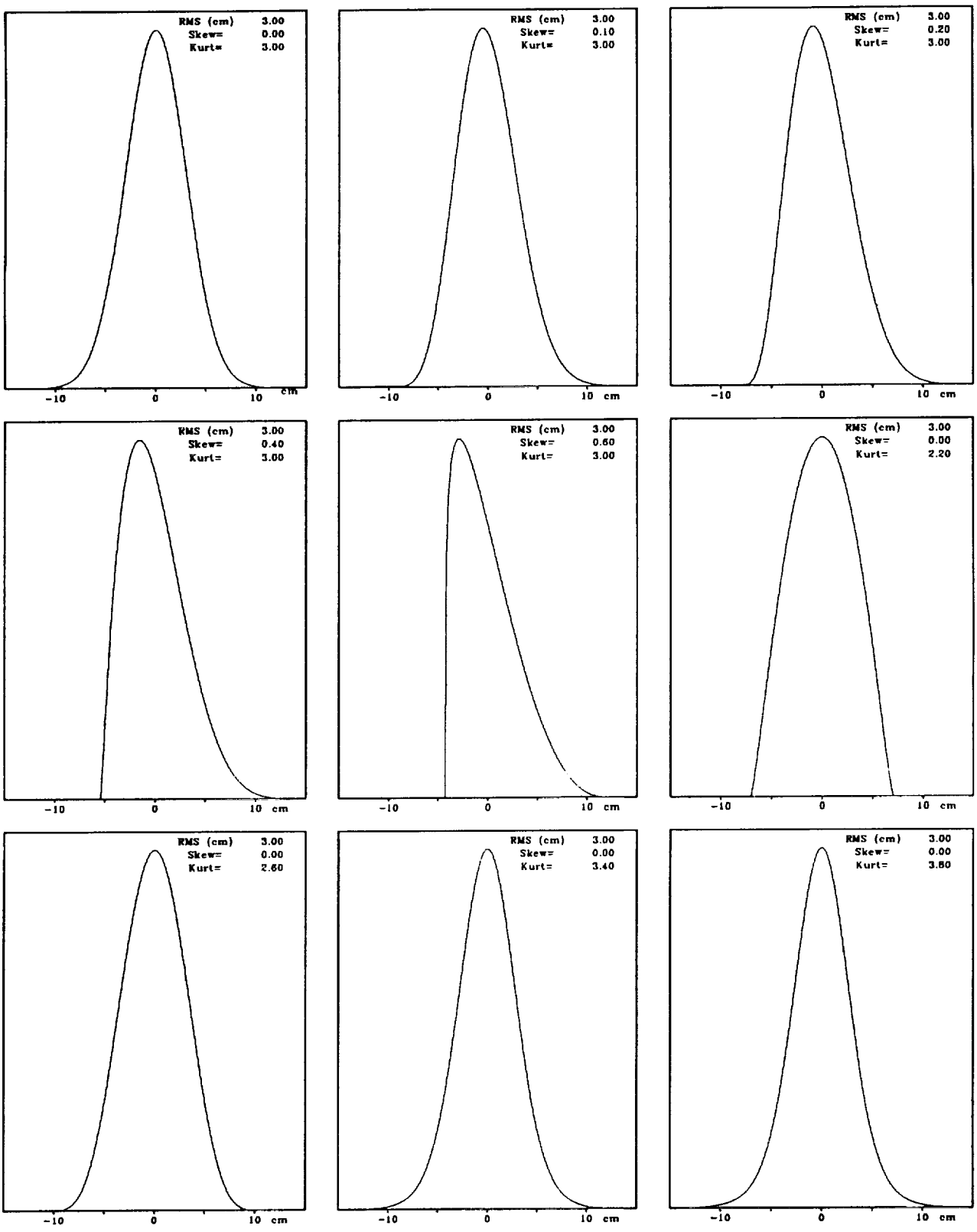

Figure 2. A range of Pearson distribution curves.

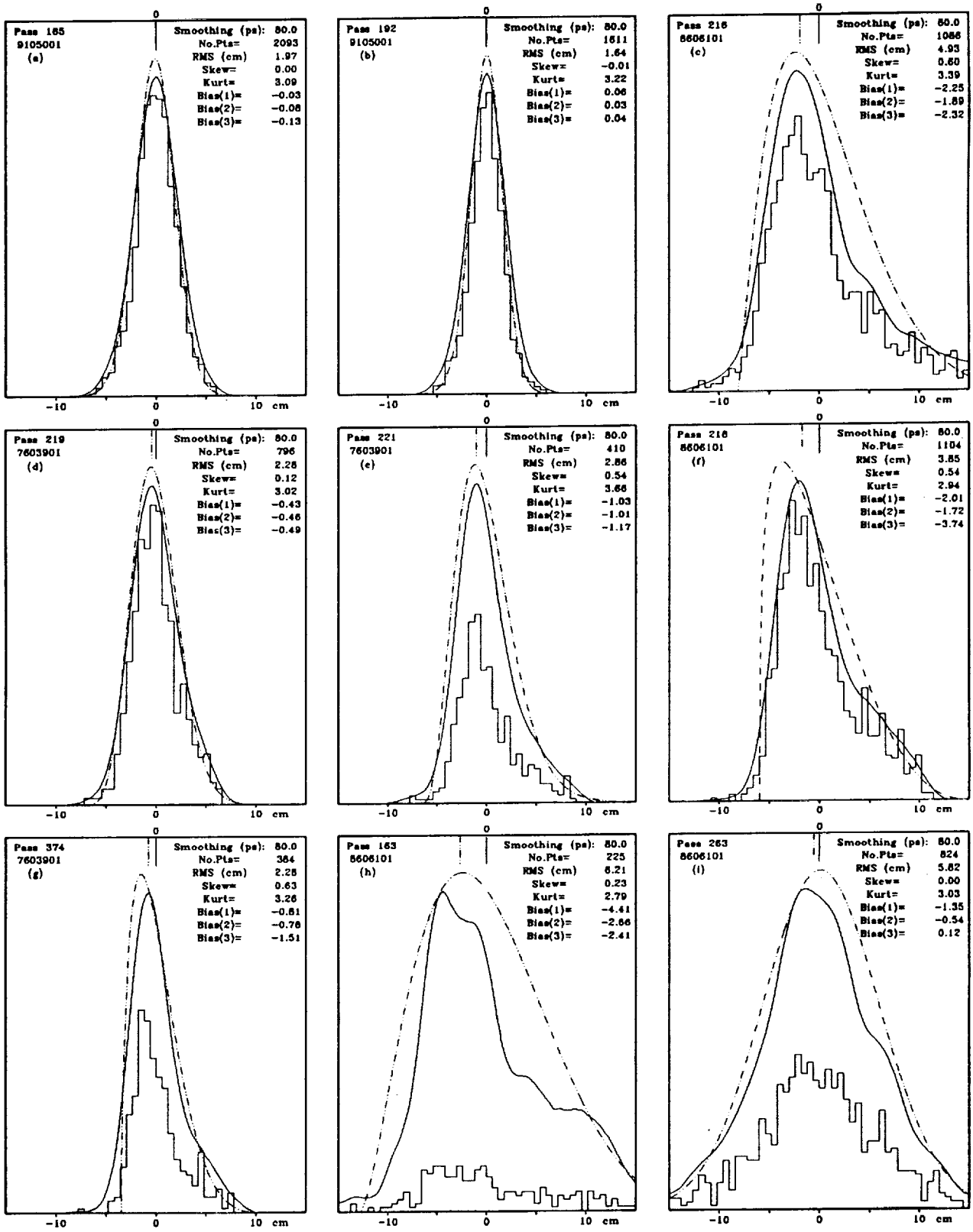

Figure 3. Comparison of methods of peak determination.