APPENDIX

Robust Statistics

A

At an early stage, science students learn that averaging is an effective way of eliminating noise and improving accuracy. However, Chapter 3 demonstrated unequivocally that median filtering of images is far better than mean filtering, both in retaining the form of the underlying signal and in suppressing impulse noise. Robust statistics is the subject of systematically eliminating outliers from visual or other data. This appendix aims to give useful insights into this important subject.

Look out for:

- the concepts "breakdown point" and "relative efficiency."
- M-, R-, and L-estimators.
- the idea of an influence function.
- the least median of squares (LMedS) approach.
- the RANSAC approach.
- the ways these methods can be applied in machine vision.

Although robust statistics is a relatively young discipline, dating largely from the 1980s, it has acquired a considerable following in machine vision, and is crucial, for example, in the development of robust 3-D vision algorithms. A basic problem to be tackled is the impossibility of knowing how much of the input data is in the form of outliers.

A.1 INTRODUCTION

We have found many times in this volume that noise can interfere with image signals and result in inaccurate measurements—e.g., of object shapes, sizes, and positions. Perhaps more important, however, is the fact that signals other than the particular one being focussed upon can lead to gross shape distortions and can thus prevent an object from being recognized or even being discerned at all. In many cases, this will render some obvious interpretation algorithm useless, although algorithms with intrinsic "intelligence" may be able to save the day. For this reason, the Hough transform has achieved some prominence: indeed, this

approach to image interpretation has frequently been described as "robust," although no rigorous definition of robustness has been ventured so far in this volume. This appendix aims to throw further light on the problem.

Research into robustness did not originate in machine vision but evolved as the specialist area of statistics now known as robust statistics. Perhaps the paradigm problem in this area is that of fitting a straight line to a set of points. In the physics laboratory, least-squares analysis is commonly used to tackle the task. Figure A.1(a) shows a straightforward situation, where all the data points can be fitted with a reasonably uniform degree of exactness, in the sense that the residual errors[1] approximate to the expected Gaussian distribution. Figure A.1(b) shows a less straightforward case, where a particular data point seems not to fall within a Gaussian distribution. Intuition indicates that this particular point represents data that has become corrupted in some way, for example, by misreading an instrument or through a transcription error. Although the wings of a Gaussian distribution stretch out to infinity, the probability that a point will be more than five standard deviations from the center of the distribution is very small, and indeed, $\pm 3\sigma$ limits are commonly taken as demarcating practical limits of correctness: it is taken as reasonable to disregard data points lying outside this range.

Unfortunately, the situation can be much worse than this simple example suggests. Suppose that there is a rogue data point that is a very long way off. In least squares analysis, it will have such a large leverage that the correct solution may not be found. And if the correct solution is not found, there will be no basis for excluding the rogue data point. This situation is illustrated in Fig. A.1(c), where the obviously correct solution has been ignored by the numerical analysis procedure.

A worse case of line fitting occurs when there are many rogue points, and it is not clear which points lie on the straight line and which do not (Fig. A.1(d)). In fact, it may not be known whether there are several lines to be fitted, or whether there are *any* lines to be fitted. Although this circumstance would appear not to occur while data points are being plotted in physics experiments, it can arise when high energy particles are being tracked; it also occurs frequently in images of indoor and outdoor scenes where a myriad of straight lines of various lengths can appear in a great many orientations and positions. Thus, it is a real problem for which answers are required. An attempt at a full statement for this type of problem might be: devise a means for finding all the straight lines—of whatever length—in a generalized[2] image so as to obtain the best overall fit to the dataset. Unfortunately, there are likely to be many solutions to any line fitting task, particularly if the data points are not especially accurate (if they are highly accurate, then the number of solutions will be small, and it should be easy to decide intuitively or automatically what the best solution is). In fact, a rigorous

[1]The residual errors or "residuals" are the deviations between the observed values and the theoretical predictions of the current model or current iteration of that model.

[2]That is, an image that might correspond to off-camera images, or to situations such as data points being plotted on a graph.

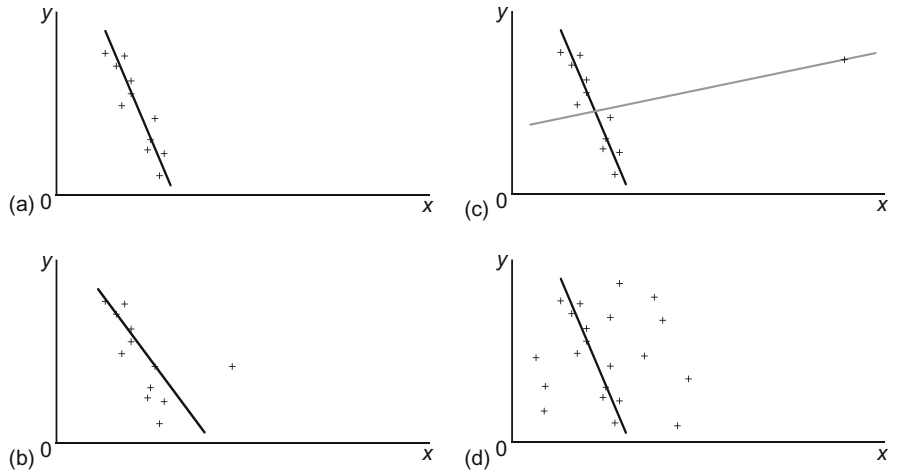

FIGURE A.1

Fitting of data points to straight lines. (a) Straightforward situation where all the data points can be fitted with reasonable precision; (b) a less straightforward case where a particular data point seems not to fall within a Gaussian distribution; (c) a situation where the correct solution has been ignored by the numerical analysis procedure; and (d) a situation where there are many rogue points, and it is not clear which points lie on the straight line and which do not: in such cases, it may not be known whether there are several, or any, lines to be fitted.

answer to the question of which solution provides the best fit requires the definition of a criterion function that in some way takes account of the number of lines and the *a priori* length distribution. We shall not pursue this line of attack here, as the purpose of this appendix is to give a basic account of the subject of robust statistics, not one that is tied to a particular task. Hence, we shall focus mainly on to the simpler case where there is only one line present in the generalized image, and there are a substantial number of rogue data points or "outliers" present.

A.2 PRELIMINARY DEFINITIONS AND ANALYSIS

In the previous section, we saw that robustness is an important factor in deciding on a scheme for fitting experimental data to numerical models. It is clearly important to have an exact measure of robustness, and the concept of a "breakdown point" long ago emerged as such a measure. The breakdown point ε of a regression scheme is defined as the smallest proportion of outlier contamination, which may force the value of the estimate to exceed an arbitrary range. As we have seen, even a single outlier in a set of plots can cause least-squares regression to give completely erroneous results. However, a much simpler example is to hand,

Table A.1 Breakdown Points for Means and Medians

n	Mean	Median
1	1	1
3	1/3	2/3
5	1/5	3/5
11	1/11	6/11
∞	0	0.5

The table shows how the respective breakdown points for the mean and median approach 0 and 0.5 as n tends to infinity, in the case of 1-D data.

namely, a 1-D distribution for which the mean is computed: here again, a single outlier can cause the mean to exceed any stated bound. This means that the breakdown point for the mean must be zero. On the other hand, the median of a distribution is well known to be highly robust to outliers, and remains unchanged if nearly half the data is corrupted. Specifically, for a set of n data points, the median will remain unchanged if the lowest $\lfloor n/2 \rfloor$ points[3] are moved to arbitrary lower values, or the highest $\lfloor n/2 \rfloor$ points are moved to arbitrary higher values, but in either case the median value will be changed to an arbitrary value if $\lfloor n/2 \rfloor + 1$ points are so moved. By definition (see above), this means that the breakdown point of the median is $(\lfloor n/2 \rfloor + 1)/n$; this value should be compared with the value $1/n$ for the mean. In the case of the median, the breakdown point approaches 0.5 as n tends to infinity (Table A.1). Thus, the median attains the apparently maximum achievable breakdown point of 0.5, and is therefore optimal—at least in the 1-D case described in this paragraph.

In fact, the breakdown point is not the only relevant parameter for characterizing regression schemes. For example, the "relative efficiency" is also important, and is defined as the ratio between the lowest achievable variance and the actual variance achieved by the regression method. In fact, the relative efficiency depends on the particular noise distribution that the data is subject to. It can be shown that the mean is optimal for elimination of Gaussian noise, having a relative efficiency of unity, while the median has a relative efficiency of $2/\pi = 0.637$. However, when dealing with impulse noise, the median has a higher relative efficiency than the mean, the exact values depending on the nature of the noise. This point is discussed in more detail in the following paragraphs.

Time complexity is a further parameter that is needed for characterizing regression methods. We shall not pursue this aspect further here, beyond making the observation that the time complexity of the mean is $O(n)$, while that for the median

[3]The function $\lfloor \cdot \rfloor$ denotes the "floor" (rounding down) operation and indicates the largest integer less than or equal to the enclosed value. In the present case, we have $\lfloor n/2 \rfloor \leq n/2 \leq \lfloor n/2 \rfloor + 1$.

varies with the method of computation (e.g., $O(n)$ for the histogram approach of Section 3.3 and $O(n^2)$ when using a bubble sort): in any case, the absolute time for computing a median normally far exceeds than that for the mean.

Of the parameters referred to above, the breakdown point has been at the forefront of workers' minds when devising new regression schemes. While it might appear that the median already provides an optimal approach for robust regression, its breakdown value of 0.5 only applies to 1-D data. It is therefore worth considering what breakdown point could be achieved for tasks such as line fitting, bearing in mind the poor performance of least-squares regression. Let us take the method of Theil (1950) in which the slope of each pair of a set of n data points is computed, and the median of the resulting set of ${}^nC_2 = \frac{1}{2}n(n-1)$ values is taken as the final slope; in fact, the intercept can be determined more simply because the problem has at that stage been reduced to one dimension. As the median is used in this procedure, at least half the slopes have to be correct in order to obtain a correct estimate of the actual slope. If we assume that the proportion of outliers in the data is η, the proportion of inliers[4] will be $1-\eta$, and the proportion of correct slopes will be $(1-\eta)^2$, and this has to be at least 0.5. This means that η has to lie in the range:

$$\eta \leq 1 - \frac{1}{\sqrt{2}} = 1 - 0.707 = 0.293 \tag{A.1}$$

Thus, the breakdown point for this approach to linear regression is less than 0.3. In a 3-D data space where a best-fit plane has to be found, the best breakdown point will be even smaller, with a value $1-2^{-1/3} \approx 0.2$. The general formula for p dimensions is:

$$\eta_p \leq 1 - 2^{-1/p} \tag{A.2}$$

Clearly, there is a need for more robust regression schemes, which becomes more urgent for larger values of p.

The development of robust multidimensional regression schemes took place relatively recently, in the 1970s. The basic estimators that were developed at that time, and classified by Huber in 1981, were the M-, R-, and L-estimators. The M-estimator is by far the most widely used, and appears in a variety of forms that encompass median and mean estimators and least-squares regression: we shall study this type of estimator in more detail below. The L-estimators employ linear combinations of order statistics, and include the alpha-trimmed mean, with the median and mean as special cases. However, it will be easier to consider the median and the mean under the heading of M-estimators, and in what follows we concentrate on this approach.

[4]Inliers are normal valid data points: the dataset is to be regarded as composed of inliers and outliers.

A.3 THE M-ESTIMATOR (INFLUENCE FUNCTION) APPROACH

M-estimators operate by minimizing the sum of a suitable function ρ of the residuals r_i. Normally, ρ is taken to be a positive definite function, and for least-squares (L_2) regression, it is the square of the residuals:

$$\rho(r_i) = r_i^2 \tag{A.3}$$

In general, it is necessary to perform the M-estimation minimization operation iteratively until a stable solution is obtained (at each iteration, the new set of offsets has to be added to the previous set of parameter values).

To improve upon the poor robustness of L_2 regression, reflected by its zero breakdown point, an improved function ρ must be obtained that is well adapted to the particular noise[5] and outlier content of the data. To understand this process, it is easiest to analyze the situation for 1-D datasets, and to consider the influence of each data point. We represent the influence of a data point by an influence function $\psi(r_i)$, where:

$$\psi(r_i) = \frac{\mathrm{d}\rho(r_i)}{\mathrm{d}(r_i)} \tag{A.4}$$

Note that minimizing $\sum_{i=0}^{n} \rho(r_i)$ is equivalent to reducing $\sum_{i=0}^{n} \psi(r_i)$ to zero, and in the case of L_2 regression:

$$\psi(r_i) = 2r_i \tag{A.5}$$

In one dimension, this equation has a simple interpretation—moving the origin of coordinates to a position where $\sum_{i=0}^{n} r_i = 0$, that is, to the position of the mean. Now that we have shown the equivalence of L_2 regression to simple averaging, the source of the lack of robustness becomes all too clear—however far away from the mean a data point is, it still retains a weight proportional to its residual value r_i. Accordingly, a wide range of possible alternative influence functions have been devised to limit the problem by cutting down the weights of distant points that are potential outliers.

An obvious approach is to limit the influence of a distant point to some maximum value: another is to eliminate its influence altogether once its residual error exceeds a certain limiting value (Fig. A.2). We could achieve this by a variety of schemes, either cutting off the influence suddenly at this limiting distance (as in the case of the $\pm 3\sigma$ points), or letting it approach zero according to a linear

[5]At this point, a certain ambiguity creeps into the discussion. "Noise" tends to originate from electronic processes in the image source, and typically leads to a Gaussian distribution in the pixel intensity values. By the time positions of objects are being measured, it is strictly speaking errors rather than noise that are being considered, and the error distribution is not necessarily identical to the noise distribution that gave rise to it. However, in the remaining sections of this appendix, we usually refer to noise and noise distributions: the term "noise" will be taken to refer either to the original noise source or to the derived errors, as appropriate to the discussion.

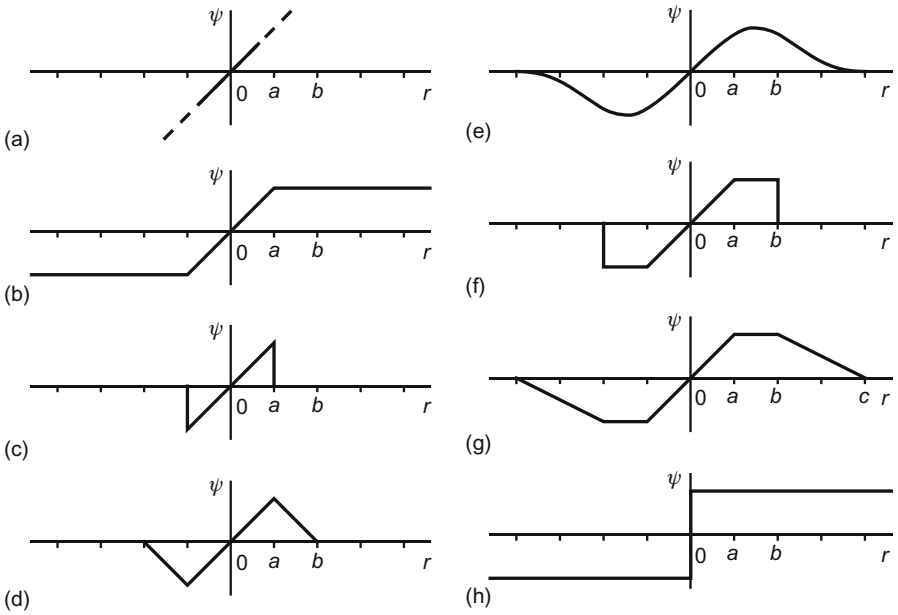

FIGURE A.2

Influence functions that limit the effects of outliers. (a) The case where no limit is placed on the influence of distant points; (b) how the influence is limited to some maximum value; (c) how the influence is eliminated altogether once the residual exceeds a certain maximum value; (d) a piecewise-linear profile that gives a less abrupt variation; (e) a mathematically more well-behaved influence function; (f) another possible piecewise-linear case; (g) a Hampel three-part redescending M-estimator that approximates the mathematically ideal case (e) with reasonable accuracy; and (h) the situation for a median estimator.

profile, or opting for a more mathematically ideal functional form with a smoother profile. In fact, there are other considerations, such as the amount of computation involved in dealing with large numbers of data points taken over a fair number of iterations. Thus, it is not surprising that a variety of piecewise linear profiles approximating to the smoother ideal profiles have been devised. In general, however, influence functions are linear near the origin, zero at large distances from the origin, and possess a region over which they give significant weight to the data points (Fig. A.2).

Prominent among these possibilities are the Hampel three-part redescending M-estimator, whose influence function is composed simply of convenient linear components, and the Tukey biweight estimator (Beaton and Tukey, 1974) that takes a form similar to that shown in Fig. A.2(e):

$$\begin{aligned} \psi(r_i) &= r_i(\gamma^2 - r_i^2)^2 \quad |r_i| \le \gamma \\ &= 0 \qquad\qquad\quad |r_i| > \gamma \end{aligned} \tag{A.6}$$

It was remarked above that the median operation is a special case of the M-estimator: here all data points on one side of the origin have a unit positive weight, and all data points on the other side of the origin have unit negative weight:

$$\psi(r_i) = \text{sign}(r_i) \tag{A.7}$$

Thus, if more data points are on one side than the other, the solution will be pulled in that direction, iteration proceeding until the median is at the origin.

It is important to appreciate that while the median has exceptionally useful outlier suppression characteristics, it actually gives outliers significant weight: in fact, the median clearly ignores how far away an outlier is, but it still counts up how many outliers there are on either side of the current origin. As a result, the median is liable to produce a biased estimate. This is a good reason for considering other types of influence function for analyzing data. Finally, note that the median influence function leads to the value of ρ for L_1 regression:

$$\rho(r_i) = |r_i| \tag{A.8}$$

When selecting an influence function, it is important not only that the function must be appropriate but also that its scale must match that of the data. If the width of the influence function is too great, too few outliers will be rejected; if the width is too small, the estimator may be surrounded by a rather homogeneous sea of data points with no guarantee that it will do more than find a locally optimal fit to the data. These factors mean that preliminary measurements must be made to determine the optimal form of the influence function for any application.

It is now clear that we need a more scientific approach, which would permit the influence function to be calculated from the noise characteristics. Hence, if the expected noise distribution is given by $f(r_i)$, the optimal form of the influence function (Huber, 1964) has to be:

$$\psi(r_i) = -\frac{f'(r_i)}{f(r_i)} = -\frac{\mathrm{d}}{\mathrm{d}r_i}\ln[f(r_i)] \tag{A.9}$$

The logarithmic form of this solution is interesting and helpful, as it simplifies the situation for exponential-based noise distributions such as the Gaussian and double exponential functions. For the former, $\exp(-r_i^2/2\sigma^2)$, we find:

$$\psi(r_i) = \frac{r_i}{\sigma^2} \tag{A.10}$$

and for the latter, $\exp(-|r_i|/s)$:

$$\psi(r_i) = \frac{\text{sign}(r_i)}{s} \tag{A.11}$$

Since the constant multipliers may be ignored, we conclude that the mean and median are optimal estimators for signals in Gaussian and double exponential noise, respectively.

Gaussian noise may be expected to arise in many situations (most particularly because of the effects of the central limit theorem), demonstrating the intrinsic value of employing the mean or L_2 regression. On the other hand, the double exponential distribution has no obvious justification in practical situations. However, it represents situations where the wings of the noise distribution stretch out rather widely, and it is good to see under what conditions the widely used median would be optimal. Nevertheless, our purpose in wanting an explicit mathematical form for the influence function was to optimize the detection of signals in arbitrary noise conditions and specifically those where outliers may be present.

Let us suppose that the noise is basically Gaussian, but that outliers may also be present and that these would be drawn approximately from a uniform distribution: there might, for example, be a uniform (but low-level) distribution of outlier values over a limited range. An overall distribution of this type is shown in Fig. A.3. Near $r_i = 0$, the uniform distribution of outliers will have relatively little effect and $\psi(r_i)$ will approximate to r_i. For large $|r_i|$, the value of f' will be due mainly to the Gaussian noise contribution, whereas the value of f will arise mainly from the uniform distribution f_u, and the result will be:

$$\psi(r_i) \approx \frac{r_i}{s^2 f_u} \exp\left(-\frac{r_i^2}{2\sigma^2}\right) \tag{A.12}$$

a function that peaks at an intermediate value of r_i. This essentially proves that the form shown in Fig. A.2(e) is reasonable. However, there is a severe problem in that outliers are by definition unusual and rare, so it is almost impossible in most cases to be able to produce on optimum form of $\psi(r_i)$ as suggested above. Unfortunately, the situation is even worse than this discussion might indicate. Redescending M-estimators are even more limited in that they are sensitive to local densities of data points, and are therefore prone to finding false solutions—unique solutions are *not* guaranteed. Non-redescending M-estimators are guaranteed to arrive at unique solutions, although the accuracy of the latter depends on the accuracy of the preliminary scale estimate. In addition, the quality of the initial approximation tends to be of very great importance for M-estimators, particularly for redescending M-estimators.

Finally, we should point out that the above analysis has concentrated on optimization of accuracy and is ultimately based on maximum likelihood strategies

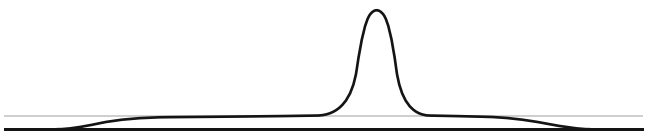

FIGURE A.3

Distribution resulting from Gaussian noise and outliers. Here, the usual Gaussian noise contribution is augmented by a distribution of outliers, which is nearly uniform over a limited range.

(Huber, 1964). It is really concerned with maximizing relative efficiency on the assumption that the underlying distribution is known. Robustness measured according to the breakdown point criterion is not optimized, and this factor will be of vital importance in any situation where the outliers form part of a totally unexpected distribution, or do not form part of a predictable distribution.[6] Clearly, methods that are intrinsically highly robust must be engineered according to the breakdown point criterion. This is what motivated the development of the least median of squares (LMedS) approach to regression during the 1980s.

A.4 THE LEAST MEDIAN OF SQUARES APPROACH TO REGRESSION

In Sections A.2 and A.3, we have seen that a variety of estimators exist, which can be used to suppress noise from numerical data, and to optimize the robustness and accuracy of the final result. The M-estimator (or influence function) approach is extremely widely used and is successful in eliminating the main problems associated with the use of least-squares regression (including, in 1-D, use of the mean). However, it does not in general achieve the ideal breakdown value of 0.5 and requires careful setting up to give optimal matching to the scale of the variation in the data. Accordingly, much attention has been devoted to a newer approach—LMedS regression.

The aim of LMedS regression is to capitalize on the known robustness of the median in a totally different way—by replacing the mean of the least (mean) squares averaging technique by the far more robust median. The effect of this is to ignore errors from the distant parts of the distribution and also from the central parts where the peak is often noisy and ill-defined, and to focus on the parts about halfway up and on either side of the distribution. Minimization then balances the contributions from the two sides of the distribution, thereby sensitively estimating the mode position, although clearly this is achieved rather indirectly. Perhaps the simplest view of the technique is that it determines the location of the narrowest width region that includes half the population of the distribution. In a 2-D straight-line location application, this interpretation amounts to locating the narrowest parallel-sided strip that includes half the population of the distribution (Fig. A.4). In principle, in such cases, the method operates just as effectively if the distribution is sparsely populated—as happens where the best-fit straight line for a set of experimental plots has to be determined.

The LMedS technique involves minimizing the median of the squares of the residuals r_j for all possible positions in the distribution that are potentially mode positions, that is, it is the position x_i that minimizes $M = \text{med}_j(r_j^2)$. While it might be thought that M is also equal to $M = \text{med}_j(|r_j|)$, this is not so if there are two

[6]It is perhaps a philosophical question whether an outlier distribution does not exist, cannot exist, or cannot be determined by any known experimental means, for example, because of rarity.

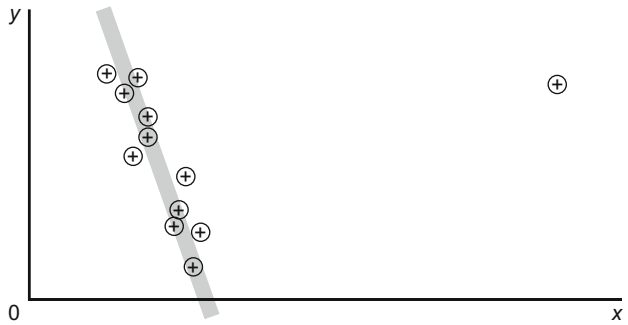

FIGURE A.4

Application of the least median of squares technique. Here, the narrowest parallel-sided strip is found that includes half the population of the distribution, in an attempt to determine the best-fit line. Note the effortless superiority in performance when compared with the situation in Fig. A.1(c).

adjacent central positions giving equal responses (as in Fig. A.5(a−c)); however, the form of M guarantees that a position midway between these two will give an appropriate minimum. For clarity, we shall temporarily ignore this technicality and concentrate on M: the reason for doing this is to take advantage of piecewise linear responses that considerably simplify theoretical analysis.

Figure A.5(a) shows the response M when the original distribution is approximately Gaussian. There is a clear minimum of M at the mode position, and the method works perfectly. Figure A.5(b) shows a case where there is a very untidy distribution, and there is a minimum of M at an appropriate position. Figure A.5(c) shows a more extreme situation in which there are two peaks, and again the response M is appropriate, except that it is now clear that the technique can only focus on one peak at a time. Nevertheless, it gets an appropriate and robust answer for the case it is focussing on. If the two peaks are identical, the method will still work, but will clearly not give a unique solution.

The LMedS approach to regression (Rousseeuw, 1984) has acquired considerable support, since it has the maximum possible breakdown point of 0.5. In particular, it has been used for pattern recognition and image analysis applications (see, e.g., Kim et al., 1989). In these areas, the method is useful for (a) location of straight lines in digital images, (b) location of Hough transform peaks in parameter space, and (c) location of clusters of points in feature space.

Unfortunately, the LMedS approach is liable to give a biased estimate of the modes if two distributions overlap, and in any case focusses on the main mode of a multimodal distribution. Thus, the LMedS technique has to be applied several times, alternating with necessary truncation processes, to find all the cluster centers, while weighted least-squares fitting is required to optimize accuracy. The result is a procedure of some complexity and considerable computational load. Indeed, the load is in general so large that it is normally approximated by taking

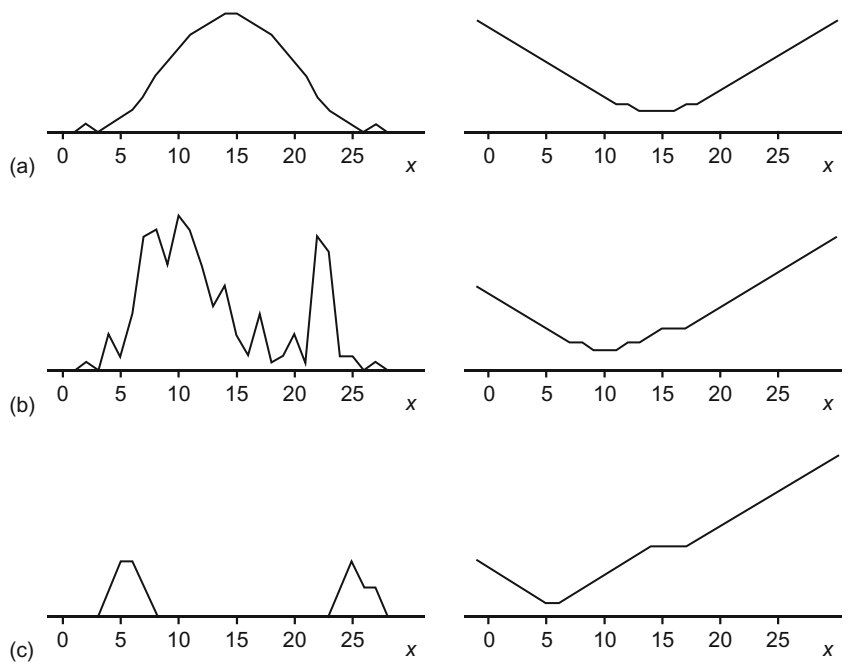

FIGURE A.5

Minimizing *M* for various distributions. The figure shows (left) the original distributions and (right) the resulting response functions *M*, in the following cases: (a) an approximately Gaussian distribution, (b) an "untidy" distribution, and (c) a distribution with two peaks.

subsets of the data points, although this aspect cannot be examined in detail here (see, e.g., Kim et al., 1989). Once this has been carried out, the method can give quite impressive results.

Ultimately, the value of the LMedS approach lies in its increased breakdown point in situations of multidimensional data. If we have n data points in p dimensions, the LMedS breakdown point is:

$$\varepsilon_{\text{LMedS}} = \frac{(\lfloor n/2 \rfloor - p + 2)}{n} \tag{A.13}$$

which tends to 0.5 as n approaches infinity (Rousseeuw, 1984). This value must be compared with a maximum of

$$\varepsilon = \frac{1}{(p+1)} \tag{A.14}$$

for standard methods of robust regression such as the M-, R-, and L-estimators discussed earlier (Kim et al., 1989). (Eq. (A.2) represents the suboptimal solution achieved by the Theil approach to line estimation.) Thus, in these latter cases,

0.33 is the best breakdown point that can be achieved for $p = 2$, while the LMedS approach offers 0.5. However, the relative efficiency of LMedS is relatively low (ultimately because it is a median-based estimator); as stated above, this means that it has to be used with the weighted least-squares technique. We should also point out that the LMedS technique is intrinsically 1-D, so it has to be used in a "projection pursuit" manner (Huber, 1985), concentrating on one dimension at a time. For implementation details, the reader is referred to the literature (see Section A.8).

A.5 OVERVIEW OF THE ROBUSTNESS PROBLEM

For greatest success in solving the robustness and accuracy problems—represented, respectively, by the breakdown point and relative efficiency criteria—it has been found in the foregoing sections that the LMedS technique should be used for finding signals (whether peaks, clusters, lines, or hyperplanes, etc.), and weighted least squares regression should be used for refining accuracy, the whole process being iterated until satisfactory results are achieved. This is a complex and computationally intensive process, but reflects an overall strategy that has been outlined several times in earlier chapters (particularly Chapters 11–14)—namely, search for an approximate solution, and then refinement to optimize location accuracy. The major question to be considered at this stage is: what is the best method for performing an efficient and effective initial search? In fact, there is a further question that is of especial relevance: is there any means of achieving a breakdown point of greater than 0.5?

We now consider the extent to which the Hough transform tackles and solves these problems. First, it is a highly effective search procedure, although in some contexts its computational efficiency has been called into question (however, in the present context, it must be remembered that the LMedS technique is especially computationally intensive). Second, it seems able to yield breakdown points far higher than 0.5 and even approaching unity. Consider a parameter space where there are many peaks and also a considerable number of randomly placed votes. Then any individual peak includes perhaps only a small fraction of the votes, and the peak location proceeds without difficulty in spite of the presence of 90–99% contamination by outliers (the latter arising from noise and clutter). Thus, the strategy of searching for peaks appears to offer significant success at avoiding outliers. Yet this does not mean that the LMedS technique is valueless, since subsequent application of LMedS is essentially able to *verify* the identification of a peak, to locate it more accurately via its greater relative efficiency, and thus to feed reliable information to a subsequent least-squares regression stage. Overall, we can see that a staged progression is taking place from a high breakdown point, low relative efficiency procedure, to a procedure of intermediate breakdown point and moderate relative efficiency, and finally to a procedure of low breakdown point and high relative efficiency. We summarize the progression by giving possible figures for the relevant quantities in Table A.2.

Table A.2 Breakdown Points and Efficiency Values for Peak Finding

	HT	LMedS	LS	Overall
ε	0.98	0.50	0.2	0.98
η	0.2	0.4	0.95	0.95

The table gives possible breakdown points ε and relative efficiency values η for peak finding. A Hough transform is used to perform an initial search for peaks; then the LMedS technique is employed for validating the peaks and eliminating outliers; finally, least-squares regression is used to optimize location accuracy. The result is far higher overall effectiveness than that obtainable by any of the techniques applied alone; however, computational load is not taken into account, and is likely to be a major consideration.

A.6 THE RANSAC APPROACH

Over a good many years, RANSAC has become one of the most widely used outlier rejection and data fitting tools: it has achieved particular value in 3-D vision. RANSAC is an acronym for *random sample consensus*, and involves repeatedly trying to obtain a consensus (set of inliers) from the data until the degree of fit exceeds a given criterion.

To understand the process, let us first return to the LMedS approach, which is useful both in providing a graphic presentation of what it achieves and in requiring no parameters to be set in order to make it work. In fact, the latter feature is in many ways its undoing, because if the proportion of outliers in the data exceeds 50%, the resulting fit is liable to be heavily biassed. A simple modification of the method is to require a smaller number of inliers—indeed, whatever proportion would be expected in the incoming data. Thus, we may go for 20% inliers, 80% outliers if this seems appropriate. This naturally leads to problems, as ideally we will have to estimate the proportion of inliers in advance, or as part of the fitting process, and then apply the resulting value as part of the technique.

Once the "cleanness" of the LMedS method is lost, a variety of alternative solutions become possible. In fact, the RANSAC method involves not taking the proportion of inliers as fixed and finding how the residual distance (e.g., from a best-fit straight line) varies, but rather specifying a threshold residual distance t and finding how the proportion of inliers varies. Here, the word "inlier" is not a good term to use; as it implies, we already know that these data are acceptable points: instead they should be called consensus points—at least until the end of the process. In summary, we set a threshold residual distance t and ask how much consensus this gives. Note that in principle at least, t has to be iterated as part of the whole process of finding the best fit. However, it is possible to work on the basis that the experimental uncertainty is known in advance, and if, for example, t is made equal to three standard deviations, this should not lead to too much error in the final fit obtained.

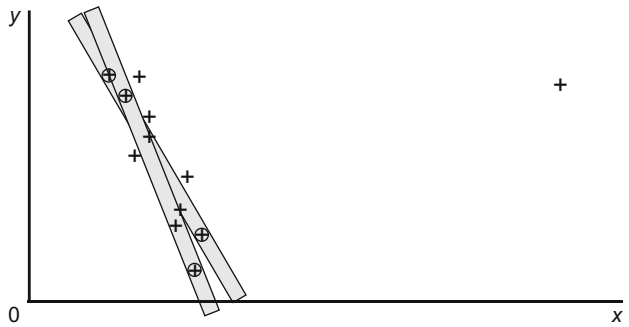

FIGURE A.6

The RANSAC technique. Here, the + signs indicate data points to be fitted, and two instances of pairs of data points (indicated by ⊕ signs) leading to hypothesized lines are also shown. Each hypothesized line has a region of influence of tolerance $\pm t$ within which the support of maximal numbers of data points is sought. The line with the most support indicates the best fit (although weighted least-squares analysis may subsequently be applied to improve it further).

Another aspect of RANSAC that must be brought out is the random extraction of n data points to specify each initial potential fit, following which the hypothesized solution is tested to find how much consensus there is; then out of k trials, the best solution is the one with the greatest consensus, and at this final stage, we can interpret the consensus as the set of inliers.

Finally, the number of data points n needed to specify a potential fit is made equal to the number of degrees of freedom of the data—two for a straight line in a plane, three for a circle, four for a sphere, and so on (Fig. A.6). All that remains to be specified is the number of iterations k of sets of n data points in order to reach the final best-fit solution. One way of estimating k is to calculate the risk that all the k sets of n data points chosen will contain only outliers so that no good data will be examined. Clearly, k must be sufficiently large to reduce the risk of this eventuality to a low enough level. Formulas to estimate k on this basis appear in several sources, for example, Hartley and Zisserman (2000).

Finally, note that, as happens with many other outlier identification processes, improved fits can be obtained by a final stage in which normal or weighted least-squares analysis is applied to the remaining (inlier) data.

A.7 CONCLUDING REMARKS

This appendix has aimed to place the discussion of robustness on a sounder basis than might have been thought possible in the earlier chapters of the book (particularly Chapters 11–14), where a more intuitive approach was presented. It has been necessary to delve quite deeply into the maturing and highly mathematical subject

of robust statistics, and there are certain important lessons to be learnt. In particular, three relevant parameters have been found to form the basis for study in this area. The first is the breakdown point of an estimator, which shows the latter's resistance to outliers and provides the core meaning of robustness. The second is the relative efficiency of an estimator, which provides a measure of how efficiently it will use the inlier data at its disposal to arrive at accurate estimates. The third is the time complexity of the estimator when it is implemented as a computer algorithm. While the last parameter is a vital consideration in practical situations, available space has not permitted it to be covered in any depth here, although it is clear that the most robust techniques (especially LMedS) tend to be highly computationally intensive. It is also found that there is a definite tradeoff between the other two parameters—techniques that have high breakdown points have low relative efficiencies and vice versa.[7] These factors make it reasonable, and desirable, to use several techniques in sequence, or iteratively in cycle, in order to obtain the best overall performance. Thus, LMedS is frequently used in conjunction with least-squares regression (see, e.g., Kim et al., 1989).

Finally, it is worth pointing out that the basis of robust statistics is that of statistical analysis of the available data: there is thus a tendency to presume that outliers are rare events due typically to erroneous readings or transcriptions. Yet, in computer vision, the most difficult problems tend to arise from the clutter of irrelevant objects in the background, and only a tiny fraction of the incoming data may constitute the relevant inlier portion. This makes the problem of robustness all the more serious, and in principle *could* mean that until a whole image has been interpreted satisfactorily, no single object can finally be identified and its position and orientation measured accurately. However, it is rare that we need to take such an extreme view in practical applications of vision.

Robust statistics is at the core of any practical vision system. This appendix has aimed to cover the intricacies of the subject in an accessible way, dealing with important concepts such as "breakdown point" and measurement "efficiency." What is really in question is *how* robust statistics will be incorporated into any practical vision system, not whether it needs to be.

A.8 BIBLIOGRAPHICAL AND HISTORICAL NOTES

This appendix has given a basic introduction to the rapidly maturing subject of robust statistics that has made a substantial impact on machine vision over the past 25 years or so. The most popular and successful approach to robust statistics must still be seen as the M-estimator (influence function) approach (which is

[7]The reason for this may be summarized as the aim of achieving high robustness requiring considerable potentially outlier data to be discarded, even when this could be accurate data that would contribute to the overall accuracy of the estimate.

broad enough to include least-squares regression and median filtering), although in high-dimensional spaces its robustness is called into question, and it is here that the newer LMedS approach has gathered a firm following. More recently, the value of using a sequence of estimators that can optimize the overall breakdown point and relative efficiency has been pointed out (Kim et al., 1989): in particular, the right combination of Hough transform (or other relevant technique), LMedS, and weighted least-squares regression would seem especially powerful.

Robust statistics has been applied in a number of areas of machine vision, including robust window operators (Besl et al., 1989), pose estimation (Haralick and Joo, 1988), motion studies (Bober and Kittler, 1993), camera location and calibration (Kumar and Hanson, 1989), and surface defect inspection (Koivo and Kim, 1989), to name but a few.

The original papers by Huber (1964) and Rousseeuw (1984) are still worth reading, and the books by Huber (1981), Hampel et al. (1986), and Rousseeuw and Leroy (1987) are valuable references, containing much insight and useful material. On the application of the LMedS technique, and for more reviews of robust regression in machine vision, see Meer et al. (1990, 1991).

Note that the RANSAC technique (Fischler and Bolles, 1981) was introduced before LMedS and presaged its possibilities: thus RANSAC was of great historical importance. The work of Siegel (1982) was also important historically in providing the background from which LMedS could take off, while the work of Steele and Steiger (1986) showed how LMedS might be implemented with attainable levels of computation.

While much of the work on robust statistics dates from the 1980s, one has only to look at the book by Hartley and Zisserman (2003) to see how deeply embedded it is in the current methodology and thinking on machine vision. An example of its application to 3-D correspondence matching is provided by Hasler et al. (2003): they consider exactly where the outlier data originates and model the whole process. Unexpected motion, occlusion of points in some views, and also viewing of convex boundaries from different positions all lead to mismatches and outliers; they arrive at a new way of calculating outliers in image pairs, which helps to put the subject area on a more secure footing.

A.8.1 More Recent Developments

In many applications, RANSAC requires an overly large number of hypotheses to be made before converging to an acceptable solution: this applies especially when searching in high-dimensional spaces. Many attempts have been made to overcome this problem. Myatt et al. (2002) tackled it by noting that in general inliers tend to be closer to one another than to outliers. Their algorithm, called NAPSAC, samples sets of adjacent points in a hypersphere: thereby the probability of selecting an inlying set is significantly increased—as demonstrated using wide baseline stereo matching data. Torr and Davidson (2003) also produced an improved version of RANSAC, which they called *importance sampling consensus*

(IMPSAC). It works in a hierarchical manner and is initialized at the coarsest level by RANSAC, but then goes on to sample at a finer level to refine relevant *a posteriori* estimates. While IMPSAC has been applied to 3-D matching tasks, it embodies statistical techniques that can be applied to a wide variety of statistical problems to eliminate outlier corrupted data.

Chum and Matas (2005) developed another idea for improving RANSAC. Instead of using randomly chosen hypotheses, they start by testing the most promising hypotheses and gradually revert to uniform sampling as diminishing returns set in. Their method, called PROSAC, achieves large computational savings and can be as much as 100 times faster than RANSAC, for example, with wide baseline stereo data. Effectively, PROSAC gains by ordering the hypotheses in an appropriate way. The worst-case performance essentially equals that of RANSAC, although no proof exists of this. Ni et al. (2009) developed another variant of RANSAC called GroupSAC. This relies on the assumption that there exists a grouping of the data in which some of the groups have a high inlier ratio while the others contain mostly outliers. When tested on wide baseline stereo data, GroupSAC was found to be much faster than PROSAC "most of the time," and RANSAC was always slower than either. Méler et al. (2010) devised yet another variant of RANSAC called BetaSAC. This was formulated as a general framework for including any relevant information for improving performance. BetaSAC offers a conditional sampling that is able to generate more suitable samples than pure random during the initial iterations. The only hypothesis required is that suitable samples can be built by successive data point selections. In the case of random ranking of samples, the method reverts to the same performance as RANSAC. When used for homography estimation, the method is always faster than RANSAC and typically 10−40 times faster than PROSAC.

A.9 PROBLEM

1. a. What is meant by the *breakdown point* of a data analysis method? Show how it is related to the concept of robustness. Consider also how accuracy of measurement is affected by the proportion of data points that are fully utilized by the data analysis method. Discuss the situation in relation to (i) the mean, (ii) the median, and (iii) the result of applying a Hampel three-part redescending M-estimator.

b. A method for locating straight lines in digital images involves taking every pair of edge points and finding where a line through both points of a pair intercepts the x- and y-axes. Then medians for all such intercepts are found and the positions of any straight lines are deduced. Show that the effect of taking pairs is to reduce the breakdown point from 50% to around 30%, and give an exact answer for the breakdown point. (*Hint:* start by assuming that the fraction of outliers in the original set of edge points is ε and work out the probability of half the intercept values being correct.)