# Lack-of-fit Detection using the Run-distribution Test

Andrew W. Fitzgibbon Robert B. Fisher
Department of Artificial Intelligence, Edinburgh University
5 Forrest Hill, Edinburgh EH1 2QL

**Abstract**

In this paper, we are concerned with the problem of deciding whether a fitted model accurately describes the data to which it has been fitted. We have developed an effective method of testing the lack-of-fit of a parametric model to data, with applications to the computer vision problems of robust estimation, model selection, and curve and surface segmentation.

The benefits of this technique are high sensitivity (large response to small outliers) and very low dependence on the noise distribution of the input data. Our test is new to the computer vision community in several ways:

- We look at the *distribution* of the residual errors, rather than basing statistics directly on their *values*.
- We assume a broad enough class of distributions as to be essentially distribution independent.
- The test requires *no knowledge of the sensor noise level*, and its response is essentially independent of that level.

We present results of experiments that compare the test with the standard $\chi^2$ statistic, and the median absolute deviation (MAD) measure used in robust estimation. The experiments are designed to represent typical vision tasks, namely feature tracking, robust fitting, and segmentation. We show that our test is comparable to the MAD and chi-square, but is cheaper than the MAD, and requires no knowledge of the noise level.

# Lack-of-fit Detection using the Run-distribution Test

**Abstract**

In this paper, we are concerned with the problem of deciding whether a fitted model accurately describes the data to which it has been fitted. We have developed an effective method of testing the lack-of-fit of a parametric model to data, with applications to the computer vision problems of robust estimation, model selection, and curve and surface segmentation.

The benefits of this technique are high sensitivity (large response to small outliers) and very low dependence on the noise distribution of the input data. Our test is new to the computer vision community in several ways:

- We look at the *distribution* of the residual errors, rather than basing statistics directly on their *values*.
- We assume a broad enough class of distributions as to be essentially distribution independent.
- The test requires *no knowledge of the sensor noise level*, and its response is essentially independent of that level.

We present results of experiments that compare the test with the standard $\chi^2$ statistic, and the median absolute deviation (MAD) measure used in robust estimation. The experiments are designed to represent typical vision tasks, namely feature tracking, robust fitting, and segmentation. We show that our test is comparable to the MAD and chi-square, but is cheaper than the MAD, and requires no knowledge of the noise level.

# 1 The Problem

It is very common in computer vision to wish to represent some large dataset in a concise way in order to extract geometric properties, attenuate noise, or simply to reduce the volume of data. In almost all cases, this is achieved by fitting an appropriate parametric model to the data set in the least squares sense. It is then vital to have some way of telling when the fit is wrong, and the model is not 'appropriate' to the data. Simple least squares techniques [7] assume the noise in the data to be strictly Gaussian of known variance, and then use the $\chi^2$ test to give an estimate of the probability that, under that assumption, the data fits the model. Robust estimators [5] approach the problem more directly, by effectively ignoring data points which do not fit the model. Robust models are, however, even more expensive to fit than unbiased nonlinear models, and do not help when the model is already fitted to the data, and simple verification is all that is needed. Our argument asserts that least squares is adequate for most purposes, *until its assumptions are violated.* Of course it is precisely these boundaries, at which the assumptions are violated, that are of most importance to the visual process. Hence, a quick and effective test which identifies such errors will allow a cheap estimator to be used on most of the signal, while the more expensive techniques are held in reserve until the cheaper methods fail.

# 2 Goodness-of-fit Testing

We denote the data points to which the model is to be fitted by $\{\mathbf{x}_i\}_{i=1}^n$ and the parameters of the model by $\{a_i\}_{i=1}^p$. We also assume that we have a distance metric $D(\mathbf{a}, \mathbf{x})$ which measures the signed distance between a particular data point and the fitted model. The model fitting process is assumed to have found the value of $\mathbf{a}$ for which $\epsilon = \sum_{i=1}^n \phi(D(\mathbf{a}, \mathbf{x}_i))$ is minimized. The function $\phi(x)$ is an influence function, which for classical least squares is $\phi(x) = x^2$. We do not need to know the form of $\phi$, simply that it must be symmetric or antisymmetric about $x = 0$. Having found the value $\mathbf{a}$, we can define the set of *residuals* $R = D(\mathbf{a}, \mathbf{x}_i)_{i=1}^n$. The task of goodness-of-fit testing is to determine, based on the values of the residuals, whether it is likely that the model describes the data. Lack-of-fit statistics say whether the model is unlikely to describe the data[1].

## 2.1 Chi-Square Test

Whaite [9] provides an accessible summary of the chi-square testing technique. The basic assumption is that each observed point $\hat{\mathbf{x}}_i$ is the exact point corrupted by an isotropic zero-mean Gaussian noise process of variance $\sigma^2$. If $\sigma^2$ is known, the chi-square statistic $\chi^2 = \sum_{i=1}^n (R_i/\sigma_i)^2$ has a known distribution. In fact the number $Q(\frac{n-p}{2}, \frac{\chi^2}{2})$ where $Q$ is the increasing incomplete gamma function gives a measure of how badly the model fits the data.

The disadvantages of the $\chi^2$ test are well known: the Gaussian noise model has repeatedly proved unrealistic in computer vision and the noise variance is often difficult to know in general. Additionally, the test, depending on a linearization of the residual equation, fails in the presence of high noise (see Figure 2c).

## 2.2 Median Absolute Deviation

The median absolute deviation (MAD) measure is not strictly a test, in the sense of providing a probability of error. However, because it is essentially the error metric used in robust estimators, it is interesting to see how its re-

[1] The distinction between lack of fit and goodness of fit is subtle and of great interest to statisticians, but we shall not make it here, treating the two terms as equivalent.

sponse compares with the RD test. The measure is simply the median of the absolute values of the residuals, and may be evaluated in about $O(n \log\log n)$ time. To use this measure as a test of goodness of fit, we need an estimate of the noise level. For Gaussian distributed residuals with a standard deviation $\sigma$, the median $M$ of the absolute values of the residuals satisfies

$$\frac{1}{\sigma\sqrt{2\pi}} \int_0^M e^{-\frac{t^2}{2\sigma^2}} dt = \frac{1}{4}$$

or $\mathrm{erf}(\frac{M}{\sigma\sqrt{2}}) = \frac{1}{2}$. From this, we can calculate the expected value of $M$ and threshold the MAD value accordingly.

**"RANSAC" Maximum Run Length Test:**

The "RANSAC" system of Fischler and Bolles [2] is the most similar test reported in the vision literature. Their system considers the *maximum* run length (see below) observed for a set of residuals. In our experiments, we have found this measure to be noise sensitive. In addition, we provide a possible extension to two dimensions.

# 3 Run-distribution Test

We now introduce our test, which we have called the run-distribution test. We describe the idea behind the test, the noise model which we assume, the actual test, and how it differs from similar tests in the literature.

The tests discussed above essentially extract one number from the set of residuals, and use that as a basis for discrimination. Instead we want to look at the set of residuals $R$, and decide whether that set is what we would expect, given data which is in concordance with both our parametric and noise models.

## 3.1 Noise model

We allow each point to be corrupted in each dimension by a scalar noise component sampled from a symmetric zero-median process plus an outlier process. Note that this is a very wide range of distributions, trivially including the normal distribution. Moreover, this particular type of distribution is common in computer vision. With such a distribution, the residuals after least-squares fitting will be similarly distributed. We can therefore detect outliers by quantifying the extent to which the *distribution* of the residuals matches our noise model.

## 3.2 Motivation

We do this by creating the set $S = \mathrm{sign}(R - \mathrm{median}(R))$. By deleting the zeroes at the median from $S$, we now have a set whose elements may be represented as either **+** or **-**. Following von Mises [8, page 184] we define a *run* as a sequence of one or more symbols of the same sign. For example the set $S = \{$**+-+++--+**$\}$ contains runs of lengths 1,1,3,2,1 respectively. Intuitively, we would expect that if the model fits well, there will be a large number of short runs, with long runs of positive or negative residuals indicating that the model has been biased. This idea was used by Besl [1] to decide whether a model was of high enough order to describe the data. Besl also hints at the definition of an $n$ dimensional run: We assume that there is some topology defining adjacency between different data points – commonly the points are defined on a grid, implicitly providing such a topology. A run is then a connected set of points with the same label, the 'length' of the run becoming the volume of the connected set. Again, with gridded data, this value will be an integral multiple of some constant.

Measuring the likelihood of a particular distribution of runs is a problem that has been approached in the statistical literature [3, 4, 6]. In particular, having decided to measure the runs, the question arises as to how to quantify the deviation of a particular example from the general population. Kempthorne *et al* [4, page 234] calculate the expected value and variance of the total number of runs ($E[M] = n+1, E[M^2] = \frac{n(n-1)}{2n-1}$), and approximate the distribution by a Gaussian in order to calculate probabilities. This approach, taken also by Brownlee[3], von Mises [8] and Mood[6], simplifies the analysis, but reduces the sensitivity of the test. In this paper, we instead compare the "actual" distribution to the observed distributions using a modified Kolmogorov-Smirnoff test.

### 3.3 Comparing the distributions

If we make a histogram $H(j)$ where bin $j$ contains the number of runs of length $j$ in the residuals, then the sequence

$$C_k = \sum_{j=1}^{k} H(j), \qquad 1 \le k \le n$$

will approximate the cumulative distribution function. By comparing this function to the predicted cdf $P$ given by a zero-median process (see Figure 1), we can determine the extent to which the outlier process has corrupted the fit. Comparison of cdfs normally entails use of the Kolmogorov-Smirnoff test, where the likelihood is calculated from the known distribution of $D = \max \mid C_k - P(k) \mid$. However, this has the well-known disadvantage that the sample variance of $D$ varies with $k$. Our alternative, arrived at experimentally, was to calculate the weighted sum of distances

$$D = \frac{\sum_{k=1}^{n}(P(k) - C_k)w_k}{\sum w_k}$$

In the experiments described below, the weighting function used was a simple quadratic $w_k = k^2$ chosen to give more importance to longer runs.

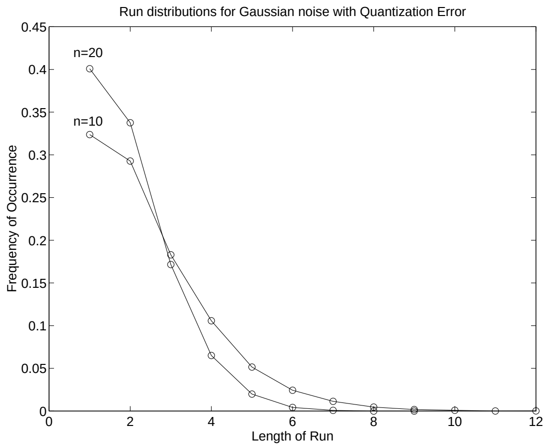


Figure 1: Empirically derived distributions of run frequencies for two values of $n$, the number of data points.

### 3.4 Determining the Actual Distribution

To enable use of the Kolmogorov-Smirnoff test, we must know the expected distribution of our measure. To this end we performed a Monte-Carlo simulation of the fitting process and recorded the results. We modelled the sensor noise process as a Gaussian plus quantization, which is an appropriate model for the laser range finder in use in our laboratory.

The distributions (graphed in Figure 1) were calculated as follows: For a given number of points $n$, the line $y = x+1, x = 1...n$ was corrupted by Gaussian noise of $\sigma = 5$, then quantized to the next lowest integer. The runs histogram was calculated using the residuals of a linear least-squares fit. Repeating this process 5000 times, and measuring the cumulative frequencies for each length of run gave the distributions shown. This technique was chosen because it was felt that the particular choice of this line would not alter the results. To test this conjecture, the line slope and noise were varied widely and the experiment repeated. Results were comparable to within about 0.2 percent. However, changing the model to a quadratic altered the frequencies by up to 10 percent, suggesting that in real applications, it is important to 'train' the test on the models expected.

We note that although the histogram should be calculated for all possible values of the number of data point $n$ (up to $10^6$ in a 2D system), there was no significant change in the frequencies after about $n = 100$, lightening the computational load significantly.

# 4 Experiments

A number of experiments were performed to assess the performance of the new test and compare it to existing test. The three tests were designed to be representative of 'everyday' vision tasks.

## 4.1 Tracking

Here we consider the problem of tracking a point through time or space while maintaining an estimate of its trajectory. The tracking can often be foiled when one point passes in front of another and the program begins to follow the second point. The error may be detected by examining the fit between the trajectory model and the data. In this experiment the track is represented by a line at 45 degrees which proceeds for 100 points (see Figure 2). The false trajectory is then represented by a second line of 50 points joining the first at an angle of 90 degrees.[2] The response is observed for two different noise levels.

### 4.1.1 Procedure

The following experiment was performed 1000 times for each noise level:

1. Gaussian noise was added to the trajectory described above.
2. For each $n$ between 3 and 150, a line was least-squares fitted to the noisy data points and the results of the three goodness-of-fit tests were recorded.

This generates 3 by 1000 traces of 147 response values.

### 4.1.2 Results

To combine these results, we consider the mean and 98th percentile responses for each $n$. The mean value gives a smoothed impression of the abilities of the tests to reject the incorrect model. These traces are shown on the left in Figure 2. The 98th percentile response indicates the potential for false negatives with each method. To ensure a false negative rate of less than 2%, it is necessary to threshold the test at a value above the highest 98th percentile response. These traces appear in the right hand column of Figure 2.

### 4.1.3 Discussion

The graphs of Figure 2 may be interpreted as follows. To the left of the dotted vertical line, false rejections will occur if the response is high. To the right, low values imply false acceptances. A perfect test will be a step function going from 0 on the left to 1 on the right. The *sensitivity* of a test may be thought of as the slope of the response curve at the breakpoint. The greater the slope, the more likely the test will correctly reject outliers.

The top left graph, for the low noise case, shows all three tests performing well, particularly for large $n$. The $\chi^2$, having been applied using the known noise variance shows the greatest sensitivity. Despite the tendency towards false rejections, as seen on the top right, a threshold of 0.95 will give excellent rejection.

[2] Although the choice of $90°$ may seem arbitrary, using smaller angles proved to be equivalent to increasing the noise level on the $90°$ case.

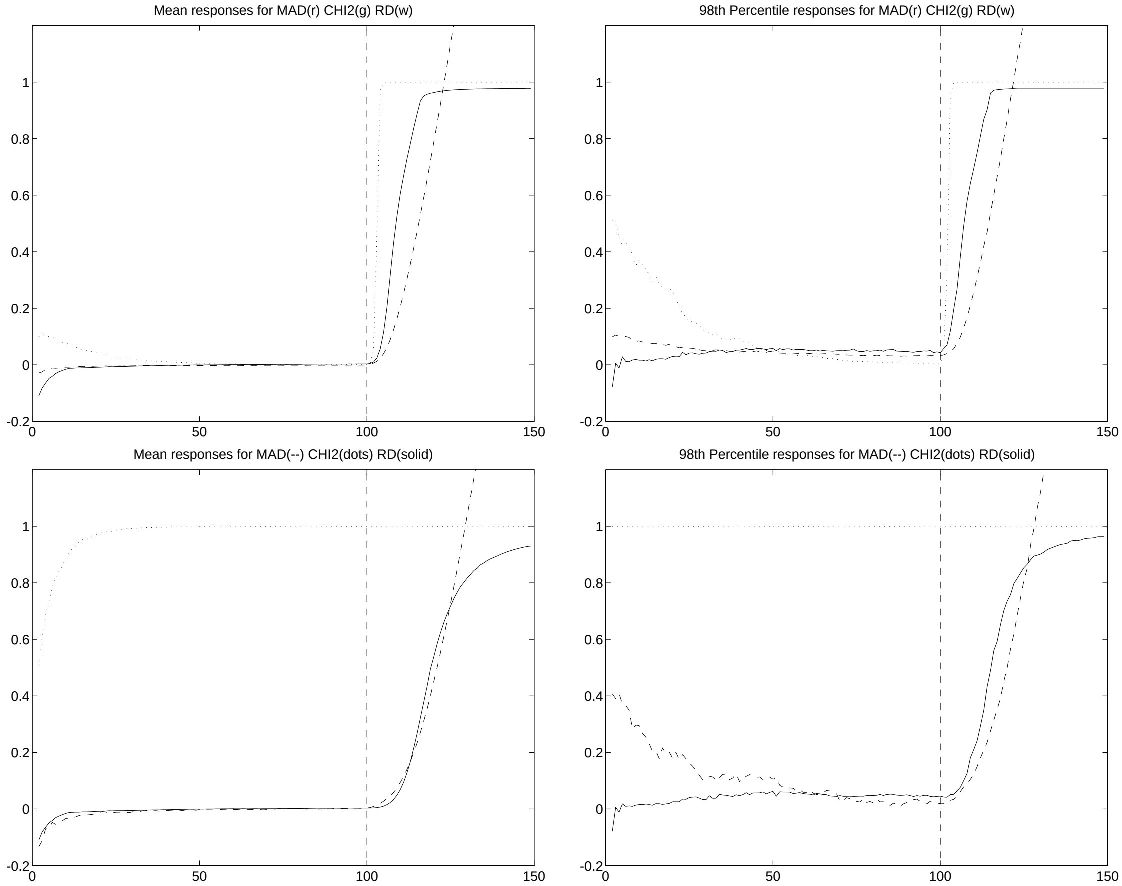


Figure 2: Performance of the $\chi^2$, MAD and RD tests on the tracking task. See section 4.1 for details.

With the RD test, the low false rejection rate means that a much lower threshold will give similar results.

The real advantage of the RD test becomes apparent as noise is increased. The $\chi^2$ test, with a slightly incorrect *a priori* noise model ($\sigma = 4$ rather than $\sigma = 5$) fails drastically, rejecting almost every point.

## 4.2 Segmentation

The test was applied to the problem of conic curve segmentation, with results as shown in Figure 3. This experiment indicates the ability of the test to identify subtle changes in model, at the $C_2$ discontinuity between line and circle for example. Curves were fitted to the 2D boundary of a 3D plane using Taubin's generalized eigenvector fit and the RD test used to identify outliers. This model was chosen to be similar to that used by Whaite [9], but the results are not comparable without knowing the use to which the segmentation is intended to be put.

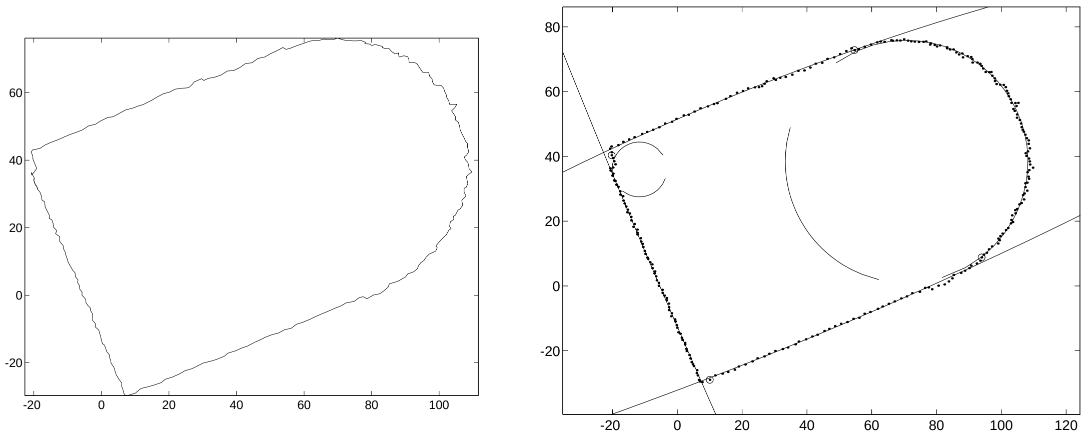


Figure 3: Segmentation results. The tracked edge data on the left has been segmented into the lines and circles shown on the right. The RD test is used to identify the breakpoints (shown as dots on the right).

# 5 Conclusions

We have introduced a new method of testing the hypothesis that some unknown data set is a noisy instance of a parametric model. Our method is superior to existing methods that make unrealistic assumptions about the noise characteristics of the input data. The method is fast, and can in most cases be made to have $O(n)$ time and space complexity. Sensitivity to small deviations in the model is high, while the false rejection rate is extremely low, even when the data are heavily corrupted by noise. The major advantage of our test however is that there is no need to know the input noise level.

A problem with the system is that in situations where quantization error grossly exceeds sensor error, the noise model is violated and the false rejection rate increases sharply. This can be avoided by adding a little Gaussian noise to the data, but this is obviously not an ideal solution.

# 6 Current Work

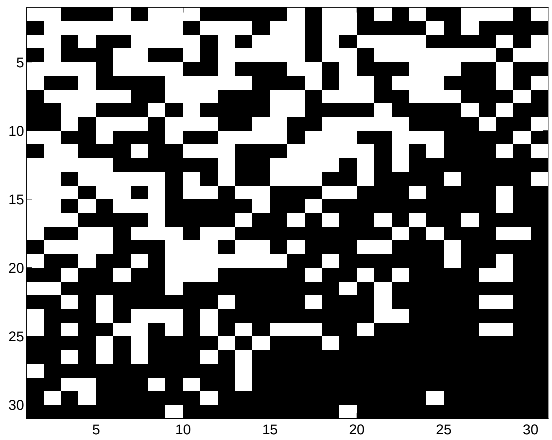


Figure 4: Example residuals sign map for a plane fit corrupted by several $10\sigma$ outliers clustered in the lower right corner.

The 2D version of the test is still under development (see Figure 4), but preliminary tests indicate similar performance to the 1D test. Using area as the equivalent to 'length' of a run may need to be changed to a fractal measure of slightly lower dimension. This is currently implemented by using morphological operators to approximate the dimensionality reduction, and then measuring areas.

## References


[1] P. J. Besl and R. C. Jain. Segmentation through variable-order surface fitting. *IEEE T-PAMI*, 10(2):167–192, March 1988.

[2] R. C. Bolles and M. A. Fischler. A RANSAC-based approach to model fitting and its application to finding cylinders in range data. In *Proceedings, IJCAI*, pages 637–643, 1981.

[3] K.A. Brownlee. *Statistical Theory and Methodology in Science and Engineering.* Wiley, 1960.

[4] O. Kempthorne and L. Folks. *Probability, Statistics, and Data Analysis.* Iowa State University Press, Ames, Iowa 50010, 1971.

[5] P. Meer, D. Mintz, A. Rosenfeld, and D.Y. Kim. Robust regression methods for computer vision: A review. *International Journal of Computer Vision*, 6(1):59–70, 1991.

[6] A.M. Mood. The distribution theory of runs. *Ann. Math. Stat.*, 14:217–226, 1940.

[7] W. H. Press et al. *Numerical Recipes.* Cambridge University Press, 2nd edition, 1992.

[8] Richard von Mises. *Mathematical Theory of Probability and Statistics.* Academic Press, 1964.

[9] P. Whaite and F. Ferrie. Active exploration: Knowing where we're wrong. In *Proceedings, ICCV*, 1993.