Automatic Model Selection by Modelling the Distribution of Residuals

T.F. Cootes, N. Thacker, and C.J. Taylor

Department of Imaging Science and Biomedical Engineering,
University of Manchester,
Manchester M13 9PT, U.K.
t.cootes@man.ac.uk
http://www.isbe.man.ac.uk

Abstract. [1]
Many problems in computer vision involve a choice of the most suitable model for a set of data. Typically one wishes to choose a model which best represents the data in a way that generalises to unseen data without overfitting. We propose an algorithm in which the quality of a model match can be determined by calculating how well the distribution of model residuals matches a distribution estimated from the noise on the data. The distribution of residuals has two components - the measurement noise, and the noise caused by the uncertainty in the model parameters. If the model is too complex to be supported by the data, then there will be large uncertainty in the parameters. We demonstrate that the algorithm can be used to select appropriate model complexity in a variety of problems, including polynomial fitting, and selecting the number of modes to match a shape model to noisy data.

1 Introduction

Many problems in computer vision involve a choice of the most suitable model for a task. For instance, we may wish to choose which order of polynomial is best for some noisy data, how many nodes to use in a neural network, whether to fit a circle or an ellipse to some measured data or whether an affine or projective model is preferable for interpretting a 3D scene. Often one has a choice of increasingly complex models, and has to select the one which balances the desire to fit to the data accurately against the problem of overfitting.

This paper describes a novel, general approach for model selection. The key is to compare the distribution of residuals with that which one would expect for the correct model. The distribution has two components, one due to the measurement noise (assumed to be known), the other is due to the uncertainty on the estimate of the model parameters induced by the measurement noise (which can be estimated). The most suitable model is then the one for which the distribution of the residuals best matches their prediction. For an over-simple

[1] This paper appears in Proc.ECCV2002, Vol.4, pp.621-635

model, the residuals will be too large, for an over-complex model they will be too small.

In the following we explain the approach in more detail, both in the general case and in the common special case of linear models. We demonstrate that the approach can be used to select the order of polynomial to best fit data, and the complexity of a shape model to best represent noisy measurements of a shape.

2 Background

A common approach to deciding whether a model is a good fit to some data is to use the χ^2 method. Here one computes the weighted sum of squares of residuals $\chi^2 = \sum \frac{r_i^2}{\sigma_i^2}$ (where σ_i^2 is the variance on measurement i) and tests it against the expected χ^2-distribution [10]. However, such a statistic is not sufficient for choosing the better of two models, as the more general model will always give rise to a smaller error.

Akaike [2] developed a model selection criteria which chooses the model that minimizes the expected error of new observations with the same distribution as the available data. One should select the model which gives a minimum of *An Information Criterion*

$$AIC = -2 \log L + 2k \tag{1}$$

where $\log L$ is the log likelihood of the data and k is the number of degrees of freedom of the model. With a gaussian error model this becomes $AIC = \chi^2 + 2k$.

AIC takes the form of a log likelihood term and a second term which penalises the complexity of the model. There have been many alternatives proposed for the penalty term (a review is given by Torr [14]). A commonly used one is that due to Kanatani [7], who proposed a *Geometric Information Criterion* for comparing models which fit a d dimensional manifold to n data points,

$$GIC = -2 \log L + 2(nd + k) \tag{2}$$

Torr [14] generalised this using techniques from robust statistics in order to allow for outliers.

Many of the metrics are designed for large data sets, and do not work well on small samples. Various corrections have been proposed, one of the best of which is designed specifically for small data sets by Chappelle *et.al.*[9].

The approach examined below, rather than look at penalising a likelihood term, explicitly considers the distribution of residuals. This allows one to compare completely different types of model on the same data, automatically dealing with issues of differences in dimension of manifold fitted to the data.

There is a wide literature on model selection (overviews are included in [5, 14, 13]). Some of the most successful approaches are the metric-based methods of Schuurmans [4, 5]. These assume one is fitting a sequence of increasingly complex models h_i to the data, X, and can estimate a measure of distance between two models, $d(h_i, h_{i+1})$ and between each model and the data $d(h, X)$. The triangle inequality is invoked to suggest that $d(h_i, X) + d(h_{i+1}, X) \geq d(h_i, h_{i+1})$. The model to choose is then the most complex one for which this inequality holds with the previous model. This has been found to work effectively on a wide variety of problems [5]. However it does not take account of the uncertainty on the data (the measurement error), which is often known. The approach presented below assumes that the measurement error is known, and is used explicitly. Also the metric-based method is not designed to compare two completely different models matched to the data.

The approach taken below is a development of that taken by Thacker *et.al.*[12] who suggested using "Bhattacharyya fitting", in which the errors induced on the parameters are propagated back into the data space and convolved with the measurement noise. The overlap between this distribution and a distribution centred on the data was used to test the quality of fit. However, it can be shown that as the number of points increases the overlap drops, eventually falling to zero with infinite data. This is not a satisfactory result. The method described in this paper uses a similar initial approach, but compares the distribution of residuals with a predicted distribution. In the limit of infinite data the overlap tends to unity, which is more satisfactory.

3 R-Fitting Algorithm

We consider the problem of estimating how well a particular model matches to some measured data, $\{\mathbf{y}_i\}$ for the measurement noise distribution is known.

In the following we will assume an explicit model of the form

$$\mathbf{y} = f(\mathbf{x} : \mathbf{a}) \tag{3}$$

where $\mathbf{y}$ is a vector of measurements made at points given by the elements of $\mathbf{x}$, and $\mathbf{a}$ is a vector of model parameters. We will use $\mathbf{Y}$ to represent the concatenation of all the N measurement vectors, $\mathbf{y}_i$, and $\mathbf{X}$ to represent the concatenation of all the vectors $\mathbf{x}_i$.

Note that the approach can be generalised to implicit models of the form $f(\mathbf{y} : \mathbf{a}) = 0$, such as the implicit equation of a circle $|\mathbf{y} - \mathbf{c}|^2 - r^2 = 0$.

The algorithm to estimate the quality of fit of the model to the data has the following steps

1. Match the model to the data and estimate the uncertainty in the parameters
2. Use leave-one-out fitting to obtain unbiased estimates of the residuals
3. Compute the theoretical distribution of the residuals
4. Compare the actual residuals with the theoretical distribution

We will first consider the general case, then show how efficient algorithms can be derived for certain common special cases.

3.1 Estimating the uncertainty on the parameters

Given a known distribution of noise on the measurements, we can estimate the implied uncertainty in the model parameters $\mathbf{a}$. This can usually be achieved analytically with error propogation. Suppose there is gaussian noise on the d dimensional measurements with covariance $\mathbf{S}_d$. Let $\mathbf{S}_n = \mathbf{I}_N \otimes \mathbf{S}_d$ (a block diagonal square matrix of dimensionality Nd x Nd). Then the covariance of the uncertainty on the parameters $\mathbf{a}$ can be shown to be

$$\mathbf{S}_{\mathbf{a}} \approx \frac{d\mathbf{a}}{d\mathbf{Y}}^T () \frac{d\mathbf{a}}{d\mathbf{Y}} \tag{4}$$

This is exact for linear models, but an approximation for non-linear models, though a good one in the case of small errors.

Where $\frac{d\mathbf{a}}{d\mathbf{Y}}$ is not easily computable, we can use Monte-Carlo techniques [10] to estimate the parameter distribution.

3.2 Un-biased estimates of residuals

The residual for the i^{th} data point, $\mathbf{r}_i$, is the difference between the measured point and the position suggested by the model. In the explicit case this is given by $\mathbf{r}_i = \mathbf{y}_i - f(\mathbf{x}_i : \mathbf{a})$. However, if the point $\mathbf{y}_i$ was used in the estimation of the model parameters, this will be a biased estimate of the residual, and will underestimate the errors one might get for unseen data. To deal with this we use a leave-one-out estimate, in which we fit the model to all the data except $\mathbf{y}_i$ to obtain the parameters $\mathbf{a}_i$. We then compute the residual difference from this model, $\mathbf{r}_i = \mathbf{y}_i - f(\mathbf{x}_i : \mathbf{a}_i)$.

3.3 Predicting the distribution of residuals

The distribution of the residuals is formed from convolving the distribution due to the uncertainty in the parameters with the measurement noise distribution.

The distribution of $\mathbf{y}$ due to the uncertainty in the parameters can be estimated using error propogation or Monte-Carlo methods. In the case of gaussian uncertainty on the parameters with covariance $\mathbf{S}_{\mathbf{a}}$,

$$\mathbf{S}_m \approx \frac{d\mathbf{Y}}{d\mathbf{a}}^T (\mathbf{S}_{\mathbf{a}}) \frac{d\mathbf{Y}}{d\mathbf{a}} \tag{5}$$

Again, this is exact for linear models, but an approximation for non-linear models.

This should be convolved with the distribution of measurement noise to obtain the predicted model distribution. Thus for gaussian noise we have covariance on the residuals of $\mathbf{S}_r = \mathbf{S}_m + \mathbf{S}_n$.

3.4 Correcting for Correllations

We wish to determine whether the measured residuals could have reasonably been generated by the predicted distribution. Unfortunately, the inclusion of the component due to the uncertainty in the model parameters introduces correllations in the predicted distributions - we cannot assume each measurement is independent.

Suppose the residuals are $\{\mathbf{r}_i\}$, and $\mathbf{r}^T = (\mathbf{r}_1^T \ldots \mathbf{r}_N^T)$ is a vector formed by concatinating the residual vectors together.

In the general case we have to rotate the space so that the projection of the distributions onto the axes are independent. Thus if the residuals are predicted to have covariance $\mathbf{S}_r$, then we must form

$$\mathbf{r}' = \mathbf{\Lambda}^{-0.5}\mathbf{\Phi}^T\mathbf{r} \tag{6}$$

where $\mathbf{\Phi}$ is the matrix of eigenvectors of $\mathbf{S}_r$ and $\mathbf{\Lambda}$ the diagonal matrix of eigenvalues.

If we split $\mathbf{r}'$ into N vectors, $\{\mathbf{r}'_i\}$, then these are linearly independent.

The difficulty is that in the general case we only have one sample from each individual distribution about each data point. One solution to this is given in Appendix A.

However, in the case in which the uncertainties are normally distributed, the elements of $\mathbf{r}'$ should come from a unit normal distribution. If we can obtain a measure of how well the distribution of measured residuals matches the predicted distribution, this tells us how well the model respresents the data.

3.5 Methods of Comparing Distributions

There are several methods of testing whether a set of 1D samples come from a particular distribution, the most well-known of which is the Kolmogorov-Smirnov Test (see [10] for details and a discussion of alternatives). This finds the magnitude of the largest difference between the cumulative distributions of the data and the model, which has a distribution which can be approximated easily.

An alternative approach is to fit a distribution, $p_d(x)$ to the samples and compute the Bhattacharyya overlap between this and the predicted distribution, $p_m(x)$,

$$B(p_d(x), p_m(x)) = \int \sqrt{p_d(x)p_m(x)}dx \tag{7}$$

This is zero if the distributions do not overlap at all, and 1 for identical distributions. It is symmetric in the choice of distribution and invariant to choice of co-ordinate frame. Unlike the KS-test it is simple to extend to multiple dimensions.

Though this measure is often described as an upper limit on the Bayes error for two distributions it can also be derived from a log-likelihood approach for Poisson distributed variables and so should be applicable to probability densities in the limit of large numbers [8].

To test the distribution of residuals we fit a distribution to them, which can either have a parametric form or be a kernel estimate [11]. Analytic forms exist for comparing two gaussians (Appendix B), but for more general distributions we can use a stochastic estimate (Appendix B.1).

In the following we use the term 'overlap' to mean the number in the range [0,1] indicating how well the distribution of residuals matches the predicted model distribution. Similar results are obtained if we choose to use the KS test to measure the difference between distributions.

3.6 Uncertainty in Overlap

We can estimate the uncertainty of the overlap using the *bootstrap method* [6, 10]. To do this we repeat the following n_r times

- draw a set of N values $\mathcal{R}_j$ from $\mathcal{R} = \{r'_i\}$ at random, with replacement, (an approximation to drawing from the underlying distribution of r'_i),
- compute the overlap, B_j, with the model distribution.

The scatter of these estimates, $\{B_j\}$, gives a measure of the uncertainty in the true estimate of the overlap.

3.7 Comparing Models

To select which model best fits some data, we perform the steps above for each model, j, to get a distribution of overlaps, $\{B_{ji}\}$. If we use the mean of each distribution we can make a choice of which is the best model. However, we can also use the scatter to decide whether one model is significantly better than another. A good policy is to choose the simplest model such that no more complex model is significantly better. This is demonstrated in the examples below.

3.8 Special Case: Gaussian Noise

Consider the special case of uniform independent gaussian noise on all the data points, with variance σ^2. (Note that any case with gaussian noise can be scaled and rotated so that the noise is independent and uniform.)

If our model is $\mathbf{y} = F(\mathbf{x} : \mathbf{a})$ then the covariance of the noise induced on the parameters, $\mathbf{a}$, is

$$\mathbf{S_a} = \frac{d\mathbf{a}}{d\mathbf{Y}}^T (\sigma^2 \mathbf{I}) \frac{d\mathbf{a}}{d\mathbf{Y}} = \sigma^2 \frac{d\mathbf{a}}{d\mathbf{Y}}^T \frac{d\mathbf{a}}{d\mathbf{Y}} \tag{8}$$

The error this causes back in the data space is

$$\mathbf{S_Y} = \sigma^2 \frac{d\mathbf{Y}}{d\mathbf{a}}^T \frac{d\mathbf{a}}{d\mathbf{Y}}^T \frac{d\mathbf{a}}{d\mathbf{Y}} \frac{d\mathbf{Y}}{d\mathbf{a}} \tag{9}$$

Since $\frac{d\mathbf{Y}}{d\mathbf{a}}$ is the pseudo-inverse of $\frac{d\mathbf{a}}{d\mathbf{Y}}$, this turns out to be simply σ^2 in the subspace spanned by the columns of $\frac{dF}{da}$ and zero in all directions orthogonal to this subspace.

Thus the distribution of the residuals (obtained by adding the measurement noise distribution, $\sigma^2\mathbf{I}$) is a gaussian with variance of $2\sigma^2$ in the subspace spanned by the columns of $\frac{dF}{da}$ and σ^2 in the orthogonal dimensions. If we estimate the residuals using a leave-one-out approach the variance due to uncertainty in parameters should be scaled by $\frac{N-1}{N}$, giving a variance in the subspace of $\frac{2N-1}{N}\sigma^2$.

If we scale the residuals in the subspace by a factor of $\alpha(N) = \sqrt{\frac{N}{2N-1}}$, then the resulting distribution is spherical with variance σ^2.

This can be achieved by the transformation

$$\mathbf{r}' = \mathbf{r} + (\alpha(N) - 1)\frac{d\mathbf{Y}}{d\mathbf{a}}\frac{d\mathbf{a}}{d\mathbf{Y}}\mathbf{r} \tag{10}$$

(which effectively projects $\mathbf{r}$ into the subspace, scales, then projects it back out again).

In this case the elements of $\mathbf{r}'$, $\{r'_i\}$, are distributed as zero mean gaussians with variance σ^2.

An elegant way of measuring the quality of match is as follows. If the models are approximately linear around the optimal parameters, then gaussian noise on the data creates gaussian noise on the residuals. We can match a gaussian to the residuals and compare this with the gaussian predicted by the model, $N(0, \sigma^2)$. An analytic form exists for the Bhattacharyya overlap between two gaussians (see Appendix B).

4 Example: Linear Models

Consider a linear model with parameter vector $\mathbf{b}$ of the form

$$\mathbf{y} = \mathbf{y}_0 + \mathbf{A}\mathbf{b} \tag{11}$$

Without loss of generality we can subtract $\mathbf{y}_0$ from both sides, and replace $\mathbf{A}$ with its singular value decomposition, $\mathbf{A} = \mathbf{U}\mathbf{W}\mathbf{V}^T$, in which the columns of $\mathbf{U}$ are orthogonal. We then obtain

$$\mathbf{y} = \mathbf{U}\mathbf{W}\mathbf{V}^T\mathbf{a} = \mathbf{U}(\mathbf{W}\mathbf{V}^T\mathbf{b}) = \mathbf{U}\mathbf{a} \tag{12}$$

where $\mathbf{a} = \mathbf{W}\mathbf{V}^T\mathbf{b}$. If we solve $\mathbf{y} = \mathbf{U}\mathbf{a}$ we can obtain the solution to our original problem using $\mathbf{b} = \mathbf{W}^{-1}\mathbf{V}\mathbf{a}$.

Thus we need only consider the linear models of the form

$$\mathbf{y} = \mathbf{U}\mathbf{a} \tag{13}$$

where $\mathbf{U}$ is a n x t matrix with orthogonal columns of unit length, and $\mathbf{a}$ is a t element vector of parameters. Note that this form of model is the typical output of eigen-decomposition of data (eg eigen-faces [15] or linear statistical shape models [3]).

Suppose we wish to fit such a model to a particular sample, $\mathbf{a}$, with gaussian noise of variance σ^2, and to choose the appropriate number of parameters, t, which best describe the data.

Consider calculating the residuals by missing out each element in turn. Let $\mathbf{W}_i$ be a diagonal n x n matrix with ones in the diagonal except at element (i, i), which is zero. $\mathbf{W}_i\mathbf{y}$ thus zeroes the i^{th} element of $\mathbf{y}$.

To compute the unbiased residual for the i^{th} element, we solve

$$\mathbf{W}_i\mathbf{y} = \mathbf{W}_i\mathbf{U}\mathbf{a}_i \tag{14}$$

to get the parameter estimates, $\mathbf{a}_i$, then compute

$$r_i = y_i - \mathbf{u}_i^T \mathbf{a}_i \tag{15}$$

where $\mathbf{u}_i^T$ is the i^{th} row of $\mathbf{U}$.
The solution for $\mathbf{a}_i$ is given by

$$\mathbf{a}_i = (\mathbf{U}^T \mathbf{W}_i \mathbf{U})^{-1} \mathbf{U}^T \mathbf{W}_i \mathbf{y} \tag{16}$$

It can be shown that

$$(\mathbf{U}^T \mathbf{W}_i \mathbf{U})^{-1} = (\mathbf{I} - \mathbf{u}_i \mathbf{u}_i^T)^{-1} = \mathbf{I} - \frac{1}{1 - |\mathbf{u}_i|^2} \mathbf{u}_i \mathbf{u}_i^T \tag{17}$$

and that $\mathbf{U}^T \mathbf{W}_i \mathbf{y} = \mathbf{a} - y_i \mathbf{u}_i$, where $\mathbf{a} = \mathbf{U}^T \mathbf{y}$.
Substituting into (15) and (16) gives

$$r_i = \frac{1}{1 - |\mathbf{u}_i|^2} (y_i - \mathbf{u}_i^T \mathbf{a}) \tag{18}$$

Note that $|\mathbf{u}_i|^2 = \sum_{j=1}^{t} u_{ij}^2$ where $\{u_{ij}\}$ are the elements of $\mathbf{U}$.

Thus the unbiased residuals found using a leave-one-out approach are simply a scaled version of the residuals computed with a leave-all-in method, and can be calculated swiftly.

Above we showed that in the gaussian noise case we can correct the residuals for correllations induced by the model by scaling in the subspace spanned by the gradient. In this case $\frac{d\mathbf{Y}}{d\mathbf{a}} \frac{d\mathbf{a}}{d\mathbf{Y}} = \mathbf{U}\mathbf{U}^T$, which we can substitute into Equation 10 to obtain

$$\mathbf{r}' = \mathbf{r} + (\alpha(N) - 1) \mathbf{U}\mathbf{U}^T \mathbf{r} \tag{19}$$

To test the model with t modes we compute the residuals as above, then test whether they appear to come from a gaussian with variance σ^2.

4.1 Polynomial Fitting

Suppose we wish to fit a polynomial model of degree $(t-1)$ to some data,

$$y = F(x : \mathbf{a}) = \sum_{i=0}^{t-1} a_i x^i = (1, x, x^2, \ldots, x^{t-1})^T \mathbf{a} \tag{20}$$

For a set of measurement positions, $\{x_i\}$, we obtain

$$\mathbf{Y} = F(\mathbf{X} : \mathbf{a}) = \mathbf{D}\mathbf{a} \tag{21}$$

where $\mathbf{D}_t$ is the N x t *design matrix* (see [10]), the i^{th} row of which is $(1, x_i, x_i^2, \ldots, x_i^{t-1})$.

This is then a linear model. The columns of $\mathbf{D}_t$ are not orthogonal so we decompose it using SVD and solve as described above. By examining the overlap between residuals and prediction we can determine how well a polynomial of degree $(t-1)$ matches the data, and can automatically determine which degree gives the most suitable compromise between fitting the given data and generalising to new data.

5 Experiments with Polynomial Fitting

We performed the following experiment. [2] We evaluated the quartic $f(x) = 3x^4 - 3x^3 - x^2 + 2x$ at 10 points in the range $[0, 1]$ and added noise to the measurements. We then matched polynomials of degrees 0-8 and evaluated the distributions of the corrected residuals using the algorithms described above. We measured the Bhattacharyya overlap between a generalised nearest-neighbour gaussian kernel estimate of the data [11] with the predicted distribution. We used bootstrap resampling of the residuals to estimate the uncertainty in the overlap estimate. Figure 1 shows the underlying polynomial and the noisy measurements of it, with noise $\sigma_n = 0.1$. Figure 2 shows the scatter of the overlaps against the degree. To select the most appropriate polynomial we choose the simplest one for which all higher degree polynomials lead to distributions of B which are not significantly higher. We say one distribution is significantly better than another if on average the 60% of the samples in the second are lower than each one in the first. Using this approach, with $\sigma_n = 0.1$ we see that having anything more than a linear model is not justified.

Figures 3 and 4 show that with less noise ($\sigma_n = 0.05$) we can justify a cubic approximation.

Figures 5 and 6 show that with noise of $\sigma_n = 0.01$ we can justify a full quartic fit.

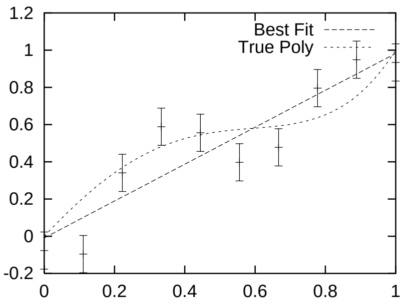

Fig. 1. Best polygon fit with $\sigma_n = 0.1$ (Linear model)

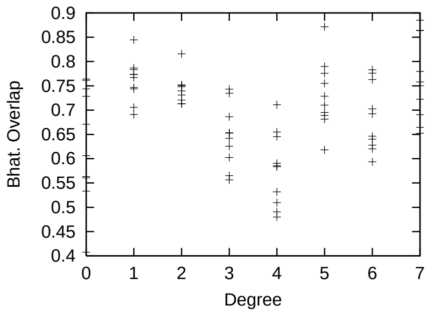

Fig. 2. Distribution of B values for $\sigma_n = 0.1$. Degree 1 best.

Figure 7 shows the distribution of the chosen degree obtained by fitting curves to noisy data 1000 times. Three methods are compared; the R-fitting approach, the Akaike orrection to the χ^2 measure (Equation 1) and metric-based approach of chuurmans *et.al.*[4, 5]. This shows hat both the R-fitting approach and that based on AIC gives answers much closer to the true degree (4) than the metric

[2] The algorithms have been implemented in the VXL C++ vision library (www.sourceforge.net/projects/vxl). The classes to do the experiments are in the mul package.

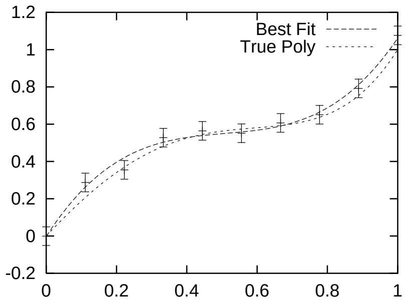

Fig. 3. Best polygon fit with $\sigma_n = 0.05$ (Cubic model)

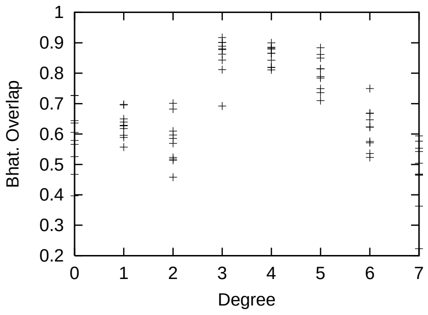

Fig. 4. Distribution of B values for $\sigma_n = 0.05$. Degree 3 best.

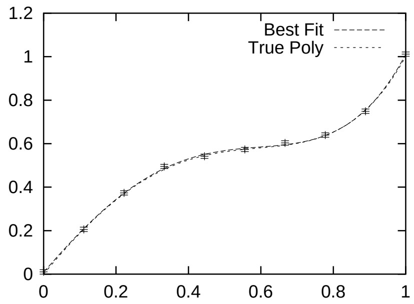

Fig. 5. Best polygon fit with $\sigma_n = 0.01$ (Quartic model)

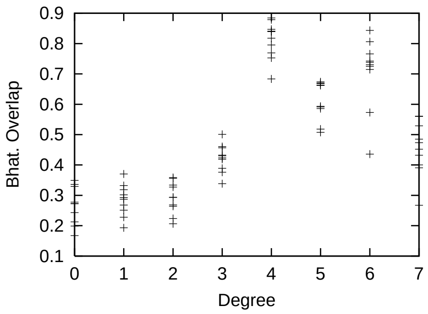

Fig. 6. Distribution of B values for $\sigma_n = 0.01$. Degree 4 best.

approach. It should be noted that the metric-based approach makes no explicit use of the (known) noise, so its relatively poor performance is not too surprising.

Figure 7 repeats this, with noise s.d. of $\sigma_n = 0.1$. The R-fitting and AIC tend to select lower order polynomials (usually linear or cubic approximations), as there is insufficient information to justify higher orders. The R-fitting method is more likely to select a linear model than using Akaike.

6 Shape Fitting

We can use the approach to automatically select the number of components to using in a shape model given noisy data. As an example, consider using a fourier model for closed shapes,

$$x(\theta) = \sum_{i=0}^{t} a_{2i} sin(i\theta) \qquad y(\theta) = \sum_{i=0}^{t} a_{2i+1} cos(i\theta) \tag{22}$$

Suppose we have sampled points from a closed shape with some noise σ_n, at approximately steps along the shape. We can estimate the parameters a_i of

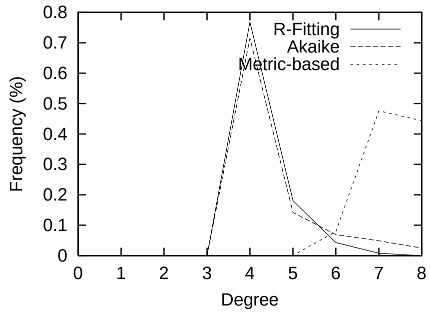

Fig. 7. Distributions of polynomial degree selections, $\sigma_n = 0.01$

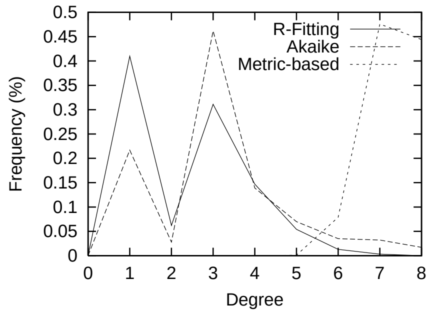

Fig. 8. Distributions of polynomial degree selections, $\sigma_n = 0.10$

the shape model which best match to the data using generalised least squares [10]. Like the polynomial fitting, this is a linear problem and we can use the same approach to matching and estimating the quality of match. We can thus choose the complexity of the model (the degree t) most suitable for the data. If we assume that we can miss out one ordinate at a time, we can use the fast methods described above (strictly we should miss out one point at a time).

Figure 9 shows the best fitting model to a fourier shape generated with $t = 5$ and noise of $\sigma_n = 0.01$. 20 points were sampled. The best fitting model has $t = 4$ in this case. Figure 10 shows the corresponding distribution of Bhattacharyya overlaps for different model complexities.

Figures 11-14 show the effect of increasing the noise. The complexity of the preferred model decreases as the noise increases.

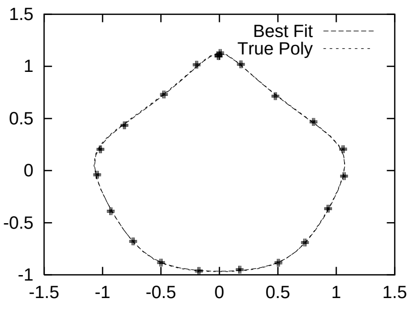

Fig. 9. Best fourier shape fit with $\sigma_n = 0.01$ (t=4)

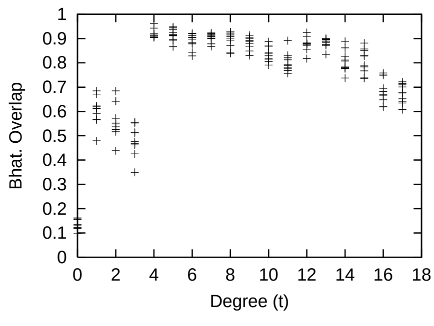

Fig. 10. Distribution of B values for $\sigma_n = 0.01$. $t = 4$ selected

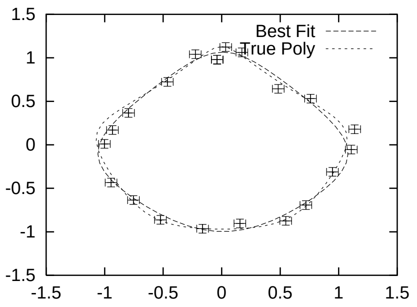

Fig. 11. Best fourier shape fit with $\sigma_n = 0.05$ (t=3)

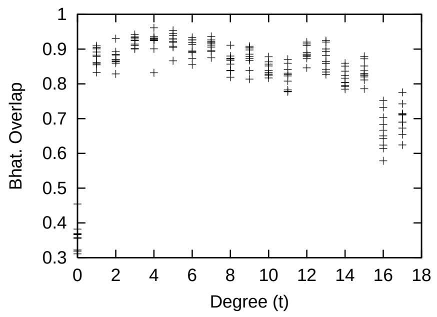

Fig. 12. Distribution of B values for $\sigma_n = 0.05$. $t = 3$ selected

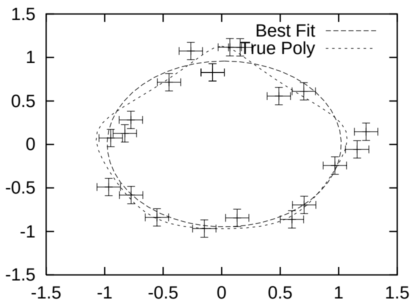

Fig. 13. Best fourier shape fit with $\sigma_n = 0.10$ (t=1 - an ellipse)

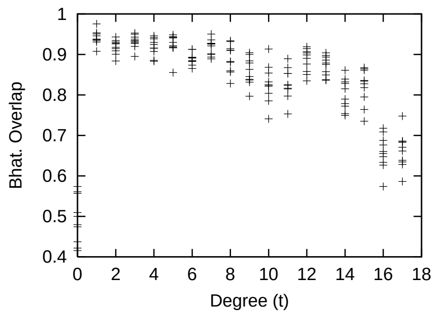

Fig. 14. Distribution of B values for $\sigma_n = 0.10$. $t = 1$ selected

7 Discussion

Model selection is an important subject in computer vision, and a key problem when designing vision systems which learn. They need a way of selecting the most suitable model to represent the training data and to avoid the overfitting problem. We have presented an approach to selecting models by examining how well the distribution of the residuals matches the distribution predicted by applying error propogation to the model matching. Not only do we obtain an estimate of the quality of fit of the model which can be compared with that of any other model matched to the same data, but we also estimate the uncertainty in that measure. This enables us to select the most appropriate model by choosing the simplest model such that no more complex model gives a significantly better match.

We have demonstrated that for linear models we can compute the quantities required using efficient algorithms. It should be noted that any function of the form $y(x) = \sum a_i f_i(x)$ where $f_i(x)$ are arbitrary (but known) functions and a_i

are the parameters, leads to a general linear system which can be solved using a design matrix [10] as described above.

Though only linear systems have been demonstrated, the approach extends to non-linear systems, and allows us to compare the match of two or more completely different models to a given set of data. We can thus try a variety of different model forms and choose the most appropriate.

We have used the Bhattacharyya overlap as a method of comparing distributions, but any other method (eg the Kolmogorov-Smirnov test) could also be used. Similar results to those described above have been obtained using the KS-statistic. The Bhattacharyya overlap has the advantage that it is simple to extend into two or more dimensions - useful when matching to sets of higher dimensional points.

There is a fundamental difference between the approach taken here and that taken in the vast majority of model selection literature. Here we start explicitly from the definition of the problem as optimisation of generalisation. Techniques such as Akaike and its variations start with log-likelihood and then attempt to correct bias in generalisation. These approaches must therefore operate within the assumptions already imposed by log-likelihood, such as the true model being from the assumed model of generating functions. The inability to deal with this issue is perhaps one of the factors which has led to the plethora of modifications to the basic approach which have been suggested for each family of data. For the particular problem presented here (polynomial fitting) the new approach is relatively complicated and appears to perform no better than Akaike. However, we believe that these results add credence to our claim that direct optimisation of generalisation defined as an overlap between probability distributions is the correct way to formulate the model selection problem [1]. This formulation is not restricted by the assumptions of the log-likelihood formulation in the same way. In principle absolute measures of generalisation capability can be compared across families of generation models without empirical tweaking. It is hope therefore that this insight will be of general value in the search for solutions to the model selection problem in both the machine vision and artificial neural network areas.

Acknowledgements

The authors would like to thank their colleagues R.H.Davies and A.L.Lacey for their helpful comments on this work.

A Comparing samples from many distributions

Suppose we wish to determine if a set of samples, $\{x_i\}$, are likely to have come from a corresponding set of distributions, $p_i(x)$. The problem is that we only have one sample from each distribution. If we can transform the data so the transformed data all come from the same distribution, we can use standard tests to see in the data is plausible. Here we consider a way of doing just that.

Suppose we have a p.d.f., $p(x)$. Consider the following function,

$$f_p(x) = 1 - \int_{p(x')<p(x)} p(x')dx' \tag{23}$$

This is the integral of $p(x')$ in all those regions for which $p(x') < p(x)$. Effectively it is the probability that a random sample from the distribution, x' has a lower value of the density than that at x, $f_p(x) = P(p(x') < p(x))$.

If $p(x)$ is symmetric about 0 and monatonically decreasing away from 0, then $f_p(x)$ is the area in the tails of the distribution beyond x;

$$f_p(x) = 2\int_{x'=|x|}^{\infty} p(x')dx' \tag{24}$$

In the case of univariate gaussians with zero mean and variance σ^2, this is given by $f_g(x) = erf\left(x/\sqrt{2}\sigma\right)$.

It can be shown that if x is drawn from $p(x)$ then the distribution of $f_p(x)$ is flat, and in particular is unity in the range [0,1] and zero elsewhere (this follows from the interpretation of $f_p(x) = P(p(x') < p(x))$). Thus if we draw many samples from a distribution and compute $f_p(x)$ for each, we will get a flat distribution of f_p.

It therefore follows that if we have N distributions, $p_i(x)$, and draw one sample from each and evaluate f_{p_i} at each, we will obtain a flat distribution.

Thus if we test whether the distribution of values $\{f_{p_i}(x_i)\}$ is flat, we get a measure of how reasonable our original assumption was.

B The Bhattacharya overlap

The Bhattacharya metric has many desirable statistical properties [12] that make it a suitable measure of the divergence of two pdfs. The 1D analytical Bhattacharya measure for a Gaussian distribution is:

$$B = \frac{\sqrt{2\sigma_d\sigma_m}}{\sqrt{\sigma_d^2+\sigma_m^2}} \exp\left(-\frac{(\mu_m-\mu_d)^2}{4(\sigma_d^2+\sigma_m^2)}\right) \tag{25}$$

where $\mu_m - \mu_d$ are the means, σ_d is the (error) variance of the data pdf and σ_m is the probable variance of the model pdf.

For multivariate gaussians with means differing by $d\bar{\mathbf{x}}$ and covariances $\mathbf{S}_1$ and $\mathbf{S}_2$, the overlap is

$$B = \frac{|0.5(\mathbf{S}_1^{-1}+\mathbf{S}_2^{-1})|^{-\frac{1}{2}}}{|\mathbf{S}_1|^{\frac{1}{4}}|\mathbf{S}_2|^{\frac{1}{4}}} \exp\left[-\frac{1}{4}d\bar{\mathbf{x}}^T\left(\mathbf{S}_1^{-1} - \mathbf{S}_1^{-1}(\mathbf{S}_1^{-1}+\mathbf{S}_2^{-1})^{-1}\mathbf{S}_1^{-1}\right)d\bar{\mathbf{x}}\right] \tag{26}$$

B.1 Stochastic Estimation of Bhattacharyya Overlap

$$B = \int \sqrt{p_1(\mathbf{x})p_2(\mathbf{x})}d\mathbf{x} = \int \left(\sqrt{\frac{p_1(\mathbf{x})}{p_2(\mathbf{x})}}\right) p_2(\mathbf{x})d\mathbf{x} \tag{27}$$

Thus to estimate B we can draw a large number, N, of samples, $\mathbf{x}_i$ from distribution $p_2(\mathbf{x})$. The estimate is then given by

$$B = \int \sqrt{p_1(\mathbf{x})p_2(\mathbf{x})}d\mathbf{x} \approx \frac{1}{N}\sum_{i=1}^{N} \sqrt{p_1(\mathbf{x}_i)/p_2(\mathbf{x}_i)} \tag{28}$$

References

1. N. A.J.Lacey and N.L.Seed. Feature tracking and motion classification using a switchable model kalman filter. In E. Hancock, editor, 5^{th} *British Machine Vison Conference*, pages 599–608. BMVA Press, Sept. 1994.
2. H. Akaike. A new look at statistical model identification. *IEEE Trans. on Automatic Control*, 19:716–723, 1974.
3. T. F. Cootes, C. J. Taylor, D. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, Jan. 1995.
4. D.Schuurmans. A new metric-based approach to model selection. In *National Conf. on Artificial Intelligence (AAAI97)*, 1997.
5. D.Schuurmans and F.Southey. Metric-based methods for adaptive model selection and regularization. *Machine Learning*, page To Appear, 2001.
6. B. Efron. *The Jackknife, the Bootstrap, and other Resampling Plans.* S.I.A.M, Philadelphia, 1982.
7. K.Kanatani. *Statistical Optimization for Geometric Computation: Theory and Practise.* Elsevier Science, Amsterdam, 1996.
8. N.A.Thacker, F.Ahearne, and P.I.Rockett. The bhattacharryya metric as an absolute similarity measure for frequency coded data. *Kybernetika*, 34(4):363–368, 1997.
9. O.Chapelle, V.Vapnik, and Y.Bengio. Model selection for small sample regression. In *NIPS2000 Workshop: Cross-Validation, Bootstrap and Model Selection*, 2000.
10. W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C (2nd Edition).* Cambridge University Press, 1992.
11. B. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.
12. N. Thacker, D. Prendergast, and P.I.Rockett. B-fitting: An estimation technique with automatic parameter selection. In 7^{th} *British Machine Vison Conference*, pages 283–292, Edinburgh, UK, 1996.
13. P. Torr. Model selection for two view geometry. Technical report, http://research.microsoft.com/-philtorr, 1998.
14. P. H. S. Torr. Geometric motion segmentation and model selection. In J. Lasenby, A. Zisserman, R. Cipolla, and H. Longuet-Higgins, editors, *Philosophical Transactions of the Royal Society A*, pages 1321–1340. Roy Soc, 1998.
15. M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.