# B-Fitting: An Estimation Technique With Automatic Parameter Selection. *

N.A.Thacker, D.Prendergast, and P.I.Rockett.
Dept. of Electronic and Electrical Engineering
University of Sheffield
email: n.thacker@sheffield.ac.uk

**Abstract**

The problem of model selection is endemic in the machine vision literature, yet largely unsolved in the statistical field. Our recent work on a theoretical statistical evaluation of the Bhattacharyya similarity metric has led us to conclude that this measure can be used to provide a solution. Here we describe how the approach may be extended to solve the problem of model selection during the functional fitting process. This paper outlines the motivation for this work and a preliminary study of the use of the technique for function fitting. It is shown how the application of the method to polynomial interpolation provides some interesting insights into the behaviour of these statistical methods and suggestions are given for possible uses of this technique in vision.

## 1 System Identification

In some areas of machine vision the correct model required to describe a system can be derived without any real ambiguity. These situations are exemplified by the process of camera calibration where a good understanding of the physical processes involved allow us to define models which encompass the main optical and electronic effects. In other cases, however, selecting the correct model to describe a set of data is something that cannot be specified uniquely a-priori. Complicated models result in reduced ability to make predictions, but simple models may not adequately describe the data. This is the problem of model selection and unfortunately it is endemic in much machine vision research eg: object recognition, tracking, segmentation, 3D modelling and in fact most scene interpretation schemes! Needless to say, if the wrong function is selected to describe a particular data set then the associated machine vision algorithm will fail and provide data which is of little or no use for subsequent processes. The problem of automatic model selection should be regarded as an important research topic in machine vision.

The method of least-squares fitting will give us an estimate of the optimal set of parameters to describe a given data set with a particular model but unfortunately the Chi-squared measure is not directly suitable for model selection. The standard

*The work presented here is part funded by the EPSRC grant GR/K55288

BMVC 1996 doi:10.5244/C.10.58

method for optimal model selection is that suggested by Akaike. He showed that the Chi-squared test statistic is biased towards small values, due to the freedom that a model has to match the variation in the noise. An analysis for large data samples [1] shows that the bias could be estimated and compensated for using the test statistic:

$$\chi^2_C \;=\; \chi^2 \;+\; m/N$$

Where $N$ is the quantity of data and $m$ is the number of degrees of freedom for the parametric model. Under some limited circumstances this measure is sufficient to enable model selection but the method does have its limitations which are directly related to the definitions of the $N$ and $m$ terms and can best be understood by example. A 3x3 rotation matrix has 9 free parameters but only 3 degrees of freedom, which should we take as our value of $m$? The problems are not limited to the model. A well distributed data set can strongly constrain a set of model parameters but a tightly grouped set of data may not. Again the bias is data dependent in a way that is not taken into account.

Such problems lead to the conclusion that the number of model parameters is not necessarily the number we are currently using to define the model but needs to estimated in a different way, such as the number of linearly independent model parameters (which will be the same regardless of the specific choice of functional representation). However, if this is the case (and it is generally accepted that it is) we now have a problem because the definition of linear independence is data dependent so we would need a different value of $m$ for different data sets as well as for different model parameters.

Both of the above problems have arisen because the bias correction term is derived for a limiting case and does not take account of data dependent variations, particularly for small data sets. These problems will also occur in any technique where it is assumed that the effects of function complexity are something that can be computed a-priori and are data independent, such as [3]. Having said this, the Akaike measure can be successfully used for automatic model selection when "calibrated" on simulated data typical of the problem at hand by adjusting the process of estimation of $N$ and $m$ to give reasonable results. Over the years this approach has led to the generation of several "model selection" criteria for various applications. In this paper we try to get to grips with the problem more directly and suggest a new statistical technique which, although consistent with the standard methods, requires no problem dependent adjustment.

## 2 Suitability of the Bhattacharyya Measure.

The Bhattacharyya measure was originally defined on the basis of a geometric argument for the comparison of two probability distribution functions $P(a|x), P(b|x)$ [2] [1].

$$\int dx \sqrt{P(a|x)}\sqrt{P(b|x)}$$

[1] the definition given here excludes the prior probability P(x) as it is unnecessary for our purposes.

Later it was found to provide an upper bound on the Bayes classification error rate for a two class problem. In the meantime it (and the analogous Matusita measure) has been used as an empirical favourite for probability comparison in the field of statistical pattern recognition [5]. We have now shown that the Bhattacharyya measure is the correct (Maximum Likelihood) similarity metric for probability distributions. The measure can be shown to be both self consistent and unbiased [11]. In addition the Bhattacharyya (or Matusita) measure can be considered as a chi-squared test statistic with fixed bias. Thus in relation to the work of Akaike the measure requires no bias correction.

The relevance of the Bhattacharyya measure to system identification is due to the fact that both measurement and model prediction stability can be represented as a Pdf in the measurement domain. Optimal generalisation ability will be obtained when the prediction probability distribution most closely matches the observed data distributions (an analytic approximation to cross-validation).

# 3 Function Fitting

In order to construct the measure from measured data for model selection we execute the following steps;

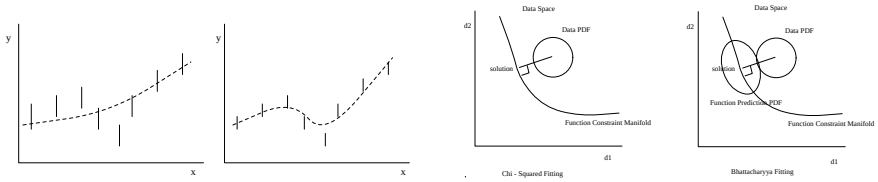
- compute the covariance of the function parameters.
- estimate the constraint provided by the function on the training data by error propagation.
- construct a probability distribution for the prediction of the output from the function.
- compare with the initial data set using the Bhattacharyya measure.

We have independently suggested the measure as the correct way to select models for a Kalman filter tracking system, and showed improved results over other approaches [6]. However, the benefits of the measure do not have to stop here. If we believe that the measure is a suitable indicator of prediction ability we would also be justified in constructing an optimisation procedure based on this measure. The reason for attempting this is as follows. If the Bhattacharyya measure is an unbiased estimate of prediction ability it should be possible to simultaneously minimise the output mapping accuracy and minimise the required internal degrees of freedom. All without the need for ad-hoc weighting of the significance of each of these to the combined cost function. We will call this process Bhattacharyya (or B) fitting. In the process we would hope to get improved fitting stability and prediction accuracy.

At this stage it would be right to ask the question; What extra information is being used in comparison to standard techniques that makes all of this possible? After all, the "bias/variance" dilemma (as it is referred to in the Neural Network literature [4]) is real so we must be providing new information if we are to solve it. The answer lies in the estimate of the errors on the output data. With normal least-squares no assumption is made about the absolute error on the output data.

All error scales will give the same least squares solution. With the Bhattacharyya measure the output is sensitive not only to the mean of the output distribution but also its variance. The more accurate the data the more accurately the function will be required to map it (figure a).

Interestingly, the requirement that the function 'concisely' fits the data emerges as a data driven Maximum Likelihood mechanism and does not require any assumption regarding prior probabilities of function suitability, unlike the Bayesian techniques described in [3].




(a) Effects of Data Accuracy

(b) Chi-square and B-fit optima

# 4 Properties of the Bhattacharyya Overlap

If we visualise the data as a hyper-sphere PDF in measurement space then the space of all allowable function parameter values defines a manifold though this space and the least squares solution is the point of closest approach of this manifold to the centre of the data PDF (Figure b), and the line from the centre of the data PDF to the manifold must be perpendicular to the manifold. The result of any fitting process must produce allowable parameter variations which are constrained to vary along the manifold. It will be seen below that the full covariance used for data prediction is a convolution of the data PDF with the function constraint PDF. This must be a hyper-ellipse with principle axes oriented along the function constraint manifold. For a fixed complexity function the optimal overlap with the data PDF must again be when the centre of the hyper-ellipse and the hyper-sphere are at their closest, ie at the same minimum as with least-squares. However, the solution can be expected to be different once we allow the functional complexity to vary. Thus the B fitting can be considered as an extension to least squares fitting and it is valid to use the standard techniques for estimating parameter covariance.

In order to try out the basic statistical ideas behind calculation of the Bhattacharyya measure and B fitting we have first confined ourselves to the relatively simple problem of polynomial function fitting with a fixed number of parameters. This will be done to establish that the fitting metric has a minimum at the expected number of parameters before allowing the number of free parameters to be modified to automatically locate this minimum.

# 5 Fixed Order Polynomial Fitting.

Assuming a locally quadratic least-squares minimum (which is always true at a sufficiently small error scale) the inverse covariance matrix for a set of parameters $a$ of a function $f(x_i, a)$ can be estimated from a set of data $y_i$ from

$$ {}^{t}C_a^{-1} \;=\; \sum_{i}^{N} \frac{1}{\sigma_i^2} (\frac{\partial f_i}{\partial a_n}) \otimes (\frac{\partial f_i}{\partial a_m}) \;=\; \sum_{i \neq j}^{N} \frac{1}{\sigma_i^2} J_i^T \otimes J_i $$

where $\sigma_i^2$ is an estimate of the data measurement accuracy $var(Y_i)$ and $J_i$ is the Jacobian. For the particular case of polynomials, the second derivatives computed this way can be shown to be independent of the data measurements $Y_i$ and only dependent on the ordinal location of the data $x_i$. This means that the expected parameter accuracy can be estimated once the measurement ordinates are specified. Similarly, therefore the parameter covariance is completely independent of the estimated parameter values. This represents a considerable simplification in comparison to more complicated models such as those used for camera calibration or neural networks, where the covariance on the data would be expected to be parameter and measurement dependent, (As parameter correlations can remove degrees of freedom from the 'effective' network function.).

To calculate the overlap measure we first need an estimate of the functions ability to estimate the data. In order to generate an unbiased estimate of the Bhattacharyya measure the estimated covariance on the parameters must be constructed in a manner that excludes the data value used in the Bhattacharyya overlap. This can be achieved if the B fitting process is visualised as a 'leave one out' chi-squared fitting process of $m$ fits, each excluding one data point. Such a fitting process, when optimally combined, would still give the same result as a chi-squared fit on the entire data set (exactly as predicted above) but gives us some insight on how to compute the unbiased covariance.

We can define the inverse covariance for the excluded data point $j$ as

$$ {}^{j}C_a^{-1} \;=\; \sum_{i \neq j}^{N} \frac{1}{\sigma_i^2} J_i^T \otimes J_i $$

The covariance of the optimally combined set of fits would then be given by the average of the parameter covariance estimated from

$$ C_a \;=\; 1/N \sum_{i}^{N} {}^{j}C_a $$

This estimation process can be considered as an extension to the normal technique for unbiased variance estimation on $n$ data points when weighting with a factor of $1/(n-1)$. It would be wrong to compute the full covariance matrix from these fits from the summed inverse covariance, in the normal fashion for optimal combination, as in this case the data sets are almost perfectly correlated. The adequacy of this multiple fit interpretation of B-fitting can be validated by comparing the data prediction capability with that predicted by $C_a$. This particularly

cumbersome calculation can be rationalised by making use of the matrix inversion lemma (Appendix A). The errors on the predicted data points will be correlated and the constraint on the data given the model must be propagated fully back to the measurement domain.

$$C_y = \nabla_a f C_a \nabla_a f^T$$

where $\nabla_a f$ is the matrix of function derivatives for the entire data set. Computation of the ability of the fitted function to predict the data is now given by

$$C_f = C_y + \sigma^2$$

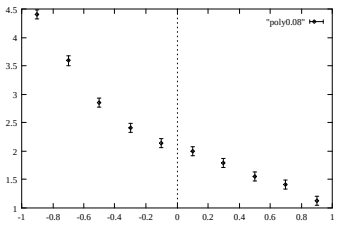
where $\sigma^2$ is the diagonal matrix of independent measurement errors. The overlap between this probability distribution and the data distribution can not be computed directly and the data needs to be rotated into a space where the principle axes are orthogonal to the ordinates. This is achieved using Singular Value Decomposition [9]. The full Bhattacharyya measure is then estimated along the axes defined by the eigen vectors of this matrix by multiplying each independent contribution from the 1 D form of the Gaussian overlap (Appendix B). As explained above, when fitting with this technique with a fixed number of parameters, we would expect to find the same function minima as located by least squares.

# 6 Predicting Model Order.

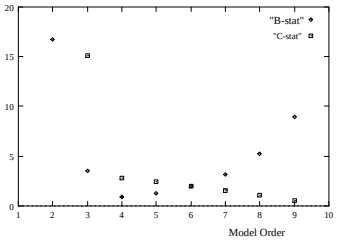
The Bhattacharyya calculation described above was tested on a simple polynomial expression

$$y = 2 - x + x^2 - x^3$$

for ten data points generated in the range -1 to 1. Data generated from this model with a uniform random error of 0.08 (Figure c) was fitted using least squares and the prediction accuracy of the model for unseen data was compared with the estimated value of $log(B_{tot})$ as a function of the number of fit parameters. The data shown in Figure d below demonstrates the observed behaviours for the average of the resulting test statistic from 100 fits ("B-stat") in comparison to the chi-squared ("C-stat").




(c) Typical Polynomial data




(d) Behaviour of B and Chi Statistics

The Bhattacharyya measure predicts correctly the model order of the data and correlates directly with the prediction ability estimate of the obtained fit. The least-squares fit measure continues to reduce with increasing numbers of model parameters as expected while its ability to predict unseen data worsens with decreasing model stability (Figure e "C"). This result is entirely consistent with the results obtained for model selection in [6]. Identical fitting results are obtained with both the Bhattacharyya and Chi-squared measures at this stage.

# 7 Gating Function Parameters.

The results from the previous sections confirm that the Bhattacharyya metric is a reasonable statistic for estimating the correct model order as a statistic directly related to generalisation ability. So far we have described the behaviour of the B fitting process on a model of fixed complexity, but the whole motivation for using this measure is to allow function complexity to be modified during the fitting process. This requires a mechanism to eliminate unwanted degrees of freedom in the fitted function.

As SVD is a standard technique for matrix inversion which identifies linearly independent degrees of freedom we can use a modified SVD algorithm to identify unwanted parameters. The standard techniques for matrix inversion eliminate the infinite contributions to the parameter covariance during it's calculation from the inverse according to the 'condition' of each eigen value almost exactly as required. The only change needed is that variances must be eliminated when correlations are considered sufficiently strong to remove a particular linear combination of parameters from the model rather than due to numerical stability ie: when a linearly independent parameter is consistent with zero. We call this process function gating and this procedure is entirely in keeping with our discussion on the effective number of linearly independent parameters given above. It is done by modifying the estimate of the inverse singular values $s_i$ from the eigen values $e_i$ (returned from SVD) as follows:

$$s_i = 1/e_i \quad if \quad e_i a_i'^2 > c$$

$$s_i = e_i a_i'^4/c^2 \quad else$$

Where $c$ is our confidence limit and $a_i'$ is a linearly independent parameter obtained by rotating the initial parameter set using the $i$ th eigen vector $v_i$.

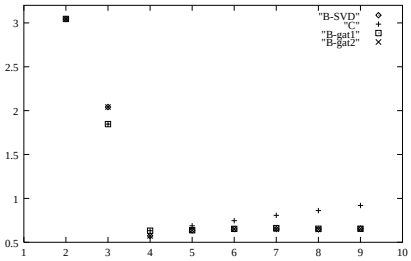
$$a_i' = a.v_i$$

The covariance for the initial parameter set is then computed in the usual manner from the outer products of the eigen vectors.

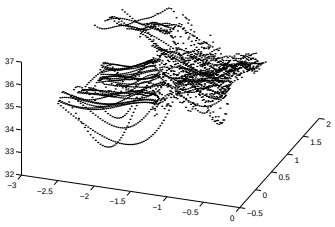
$$C_a = \sum_i s_i v_i \otimes v_i$$

This combined process can be thought of as a probabilistic weighting of contributions to the matrix inverse from the significant linearly independent parameters in the model fit.

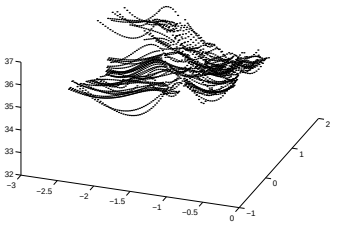
We repeated the multiple fit tests performed in the previous sections but now with gated parameters. As expected for the B fitting process, the unnecessary parameters were eliminated and their contribution to the parameter covariance was reduced to zero. More importantly the generalisation ability of the fitted model achieved optimal, least-squares fit, performance for the correct model order, regardless of the number of parameters in the fitted model (Figure e "B-SVD"), eliminating entirely the problem of over-fitting.




(e) Residuals from the true function



(f) 4th order Chi



(g) 4th order B fit

Finally we turn our attention to parameter correlation. We wish to establish that the B fitting technique is independent of the particular choice of equivalent function parameterisation. To do this the parameters must be allowed to correlate in a manner dependent on the parameter values.

The effects of parameter correlation are best described by giving examples. In the case of a simple polynomial function, as we have already said, the covariance of the parameters is a constant, fixed once the ordinates of the measurements are specified. It is independent of both the actual data measurements $Y_i$ and therefore also the parameters describing the data. However, if we construct a function of the form

$$y \;=\; (a_0(1 + a_1x(1 + a_2x(1 + a_3x(...)))))$$

we see that if at any point an $a_n$ becomes zero all following terms effectively get switched off thus giving a limited form of gating and improved generalisation characteristics. The simplest function which identifies individual terms for elimination

is

$$y = a_0|a_0| + a_1|a_1|x + a_2|a_2|x^2 + a_3|a_3|x^3 + ...$$

Under normal circumstances these functions would behave inconsistently both in terms of performance of the fitted function and estimation of the least-squares and Akaike measures. When B fitting however, both functional forms give equivalent and optimal (least squares generalisation for the true number of parameters) performance regardless of the number of fit parameters once the data complexity is matched (Figure e "B-gat1","B-gat2").

# 8 Conclusions.

In this work the Bhattacharyya statistic, a maximum likelihood estimator, has been described and compared to standard techniques for model selection. A method has been described for computing this measure from expected data variance estimates for the purposes of function fitting. The method appears to be able to correctly identify the model of appropriate order with which to describe the data. Fitting results in a model with a generalisation ability that is consistent with an optimally selected least squares solution. We have introduced the concept of 'gating' and showed how parameters can be eliminated from the Bhattacharyya evaluation function during the minimisation process. This process results in optimal least-squares performance for the fitting process regardless of the number of fit parameters.

B fitting can be regarded as $n$ simultaneous least-squares fits on $n$ data points, where each fit excludes one of the data. The net effect produces a parameter covariance which can be estimated as the average from all fits. This interpretation is also in agreement with a minimum which is co-incident with the standard LSF on the full data set. Thus B fitting can be regarded as an analytic approximation to the jack-knife cross validation procedure. However, in this case we can now fit with all of the data and the resulting fit value gives a direct estimation of generalisation capabilities of an individual solution rather than a set of solutions.

The ability to compute a statistic which directly estimates a model's capabilities for data prediction gives us a valuable tool for modelling and interpretation. B-fitting will eliminate the ad-hoc selection of an appropriate fitting model (or equivalently function priors), replacing this instead with a functional mapping with the minimum number of linearly independent parameters required to accurately account for the data from the class of functions selected. This is an important result even if the 'true' underlying model of the data is not representable. We are currently applying the technique developed here for stereo mapping of image data between known edge correspondences, this application is described in a companion paper at this conference [7]. The value of this technique is shown in figures f and g where the technique has been used to constrain the mapping solution to the one which describes the data with the minimum functional complexity [2].

The statistic is also being used to develop a neural network training algorithm which will literally optimise the network generalisation ability. Such an approach is not supported by standard statistical measures due to limitations such as those

[2] Thanks are due to Tony Lacey for helping us to generate these figures

described in this paper. This work is to be evaluated on the analysis of SPOT images for land use management. The measure is very similar to one suggested previously in this area of network selection [13]. Previously the same statistical measure has been used in the development of object recognition and neural network auto generation algorithms [10, 12]. The measure and the fitting technique described here would also clearly have use in the problem of image segmentation and geometric approaches to model based vision as an alternative to approaches such as those described in [3]. In fact the formulation of this technique for ellipse fitting is found to have many similarities to the empirically motivated "bias corrected Kalman Filter" [8].

## Appendix A. Recursive estimation of Covariance

For the particular case of estimating ${}^jC_a$ from

$$ {}^jC_a^{-1} \;=\; {}^tC_a^{-1} \;-\; J_i^T \otimes J_i^T/\sigma_i^2 $$

we can use the matrix inversion lemma to get the equivalent form for the inverse, where K is often called the Kalman Gain;

$$ {}^jC_a \;=\; (I \;-\; K \otimes J_j)\,{}^tC_a \qquad K \;=\; {}^tC_a J_j^T ({J_j}^t C_a J_j^T \;-\sigma_i^2)^{-1} $$

## Appendix B. Analytical 1D Bhattacharyya

With Gaussian distributions the Bhattacharyya integral becomes

$$ B_l \;=\; -ln\frac{1}{\sqrt{2\pi\sigma_a\sigma_b}} \int_{\infty}^{\infty} exp - \frac{1}{4}((x-\mu_a)^2/\sigma_a^2 + (x-\mu_b)^2/\sigma_b^2)dx $$

$$ = \; -ln\frac{exp\frac{(\mu_a-\mu_b)^2}{4(\sigma_a^2+\sigma_b^2)}}{\sqrt{2\pi\sigma_a\sigma_b}} \int_{\infty}^{\infty} exp - \left( \frac{\sigma_a^2+\sigma_b^2}{4\sigma_a^2\sigma_b^2} \left( x - \frac{\sigma_b^2\mu_a + \sigma_a^2\mu_b}{\sigma_a^2+\sigma_b^2} \right)^2 \right) dx $$

$$ = \; -ln(\frac{\sqrt{2\sigma_a\sigma_b}}{\sqrt{\sigma_a^2+\sigma_b^2}}) \;+\; \frac{(\mu_a-\mu_b)^2}{4(\sigma_b^2+\sigma_a^2)} $$

# References


[1] H.Akaike, 'A new Look at Statistical Model Identification', IEEE Trans. on Automatic Control, **19, 716,** (1974).

[2] A.Bhattacharyya. 'On a Measure of Divergence Between Two Statistical Populations Defined by their Probability Distributions',Bull, Calcutta Math Soc., 35, 99 (1943).

[3] T.Darrell and A.Pentland. Coorperative Robust Estimation Using Layers of Support. IEEE Trans, PAMI, 17,5,1995.

[4] S.Geman, E.BienenStock and R.Doursat, 'Neural Networks and the Bias/Variance Dilemma', Neural Computation 4(1),1 (1992).

[5] K.Fukenaga, 'Introduction to Statistical Pattern Recognition', 2ed, Academic Press, San Diego (1990).

[6] A.J.Lacey, N.A.Thacker and N.L.Seed, 'Feature Tracking and Motion Classification Using a Switchable Model Kalman Filter.' Proc. BMVC, York, Sept. 1994.

[7] A.J.Lacey, N.A.Thacker and R.B.Yates, 'Surface Approximation from Industrial SEM Images.', submitted to BMVC, 1996.

[8] J.Porrill, 'Fitting Ellipses and Predicting Confidence Envelopes using a Bias Corrected Kalman Filter.' Proc. 5th. Alvey Vision Conference, 175-185, Sept. 1989.

[9] W.H.Press B.P.Flannery S.A.Teukolsky W.T.Vetterling, Numerical Recipes in C, Cambridge University Press 1988.

[10] N.A.Thacker and J.E.W.Mayhew, 'Designing a Network for Context Sensitive Pattern Classification.' Neural Networks 3,3, 291-299, 1990.

[11] N.A.Thacker, F.J.Aherne and P.I.Rockett, 'The Bhattacharyya Metric as an Absolute Similarity Measure for Frequency Coded Data', Submitted to Pattern Recognition (1994).

[12] N.A.Thacker, P.A.Riocreux, and R.B.Yates, 'Assessing the Completeness Properties of Pairwise Geometric Histograms", Image and Vision Computing, 13, 5, 423-429, 1995.

[13] D.J.C.MacKay, 'Bayesian Modelling and Neural Networks', Research Fellow Dissertation, Trinity College, Cambridge (1991).