# PARAMETER ESTIMATION USING VOLTERRA SERIES

*Mark C.M. Hsieh and Peter J.W. Rayner*

Signal Processing and Communications Laboratory
Department of Engineering, University of Cambridge
Trumpington Street, Cambridge, CB2 1PZ
England
mcmh1@eng.cam.ac.uk

**ABSTRACT**

A polynomial approximation to the likelihood function allows for marginalised estimates of model parameters to be obtained in the form of a Volterra series. The series can be applied directly to the observed data vector in an iterative fashion, to converge upon a set of parameter MAP estimates with low computational cost. A sample application towards OCR is used as an illustration.

## 1. INTRODUCTION

Bayesian analysis is based upon the supposition of a collection of alternative hypotheses, that are responsible for generating observable data. In the light of such data, the plausibility (quantified by probability) is evaluated for each hypothesis. This process is termed Bayesian inference.

Assume a set of models $H_1..H_S$, that are believed to have an initial or prior probability of $P(H_1)..P(H_S)$. When data $\mathbf{d}$ is observed, Bayes rule provides a formalism for evaluating the probability of model $H_i$ given the data

$$P(H_i|\mathbf{d}) = \frac{P(H_i)p(\mathbf{d}|H_i)}{p(\mathbf{d})} \tag{1}$$

Models typically have a set of parameters expressed as a vector $\boldsymbol{\theta}$ (noise parameters included). Bayes rule may be applied again to find the posterior probability of $\boldsymbol{\theta}$ from its prior probability

$$p(\boldsymbol{\theta}|\mathbf{d}, H_i) = \frac{p(\mathbf{d}|\boldsymbol{\theta}, H_i)p(\boldsymbol{\theta}|H_i)}{p(\mathbf{d}|H_i)} \tag{2}$$

or

$$posterior = \frac{likelihood \times prior}{evidence}$$

Thus alongside deducing the plausibility of each model, this equation represents the probability density of the parameters given the observed data and the model.

These two levels of inference, that of evaluating the plausibility of model $H_i$ and the plausibility of its parameters $\boldsymbol{\theta}$, are linked by the evidence

$$p(\mathbf{d}|H_i) = \int p(\mathbf{d}|\boldsymbol{\theta}, H_i)p(\boldsymbol{\theta}|H_i)\, d\boldsymbol{\theta} \tag{3}$$

Rewriting equation 1

$$P(H_i|\mathbf{d}) = \frac{P(H_i)}{p(\mathbf{d})} \int p(\mathbf{d}|\boldsymbol{\theta}, H_i)p(\boldsymbol{\theta}|H_i)\, d\boldsymbol{\theta} \tag{4}$$

$p(\mathbf{d})$ is a normalising agent, as it is independent of $H_i$, and is generally not computed. Thus equation 4 provides the framework for model selection or classification.

## 2. DEVELOPMENT

If additive Gaussian white noise is assumed, the likelihood is formulated from the error between the observed data $d_n$ and the model prediction $y_n^{(i)}(\mathbf{w})$. Thus the error $e_n^{(i)}(\mathbf{w}) = d_n - y_n^{(i)}(\mathbf{w})$ is a function of the model $H_i$ and its parameters $\mathbf{w}$ (noise parameters excluded).

For convenience of notation, we will write $e_n(\mathbf{w}) \equiv e_n^{(i)}(\mathbf{w})$ and $y_n(\mathbf{w}) \equiv y_n^{(i)}(\mathbf{w})$.

The likelihood is then the Gaussian distribution of the error, with standard deviation $\sigma$

$$\begin{aligned} p(\mathbf{d}|\mathbf{w}, \sigma, H_i) &= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left[-\frac{\Phi(\mathbf{w})}{2\sigma^2}\right] \\ \Phi(\mathbf{w}) &= \sum_n e_n^2(\mathbf{w}) \end{aligned} \tag{5}$$

Introducing an inverse chi prior for $\sigma$

$$p(\sigma|H_i) = K\sigma^{-C_0} \exp\left[-\frac{C_1}{2\sigma^2}\right] \tag{6}$$

of which Jeffrey's [2] non-informative scale prior $p(\sigma|H_i) = \frac{K}{\sigma}$ is a particular case, we can integrate $\sigma$ out of the likelihood

$$\begin{aligned} p(\mathbf{d}|\mathbf{w}, H_i) &= \int p(\mathbf{d}|\mathbf{w}, \sigma, H_i)p(\sigma|H_i)\, d\sigma \\ &= K \int_0^\infty (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left[-\frac{\Phi(\mathbf{w})}{2\sigma^2}\right] \sigma^{-C_0} \exp\left[-\frac{C_1}{2\sigma^2}\right] d\sigma \\ &= K \frac{\pi^{-\frac{N}{2}} \Gamma\left(\frac{N+C_0-1}{2}\right)}{2^{\frac{3-C_0}{2}}} \left[C_1 + \Phi(\mathbf{w})\right]^{\frac{1-N-C_0}{2}} \end{aligned} \tag{7}$$

To find $P(H_i|\mathbf{d})$ in equation 4 we must further integrate equation 7 and the prior $p(\mathbf{w}|H_i)$ *w.r.t.* $\mathbf{w}$. This analysis is essentially intractable, with the exception of the General Linear Model [1] [5].

Numerical integration methods (Monte Carlo Markov chain) may be applied directly, but for many problems, it is common for $p(\mathbf{d}|\mathbf{w}, H_i)p(\mathbf{w}|H_i)$ to have a strong peak around its mode

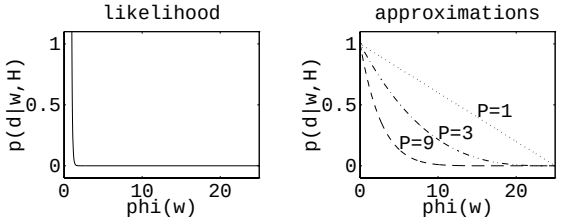


Figure 1: Approximations to likelihood

$\mathbf{w}_{MAP}$. Therefore if we are able to estimate $\mathbf{w}_{MAP}$, we can improve the efficiency and accuracy of sampled estimates, or alternatively enable us to apply approximation methods such as Gaussian approximations to the integrand, as employed by MacKay [3].

## 3. FORMATION OF THE VOLTERRA PARAMETER ESTIMATOR

The posterior density for $\mathbf{w}$ is given by

$$p(\mathbf{w}|\mathbf{d}, H_i) = \frac{p(\mathbf{d}|\mathbf{w}, H_i)p(\mathbf{w}|H_i)}{p(\mathbf{d}|H_i)} \tag{8}$$

from which we can find the expected value of the parameters

$$\begin{aligned} E[\mathbf{w}|\mathbf{d}, H_i] &= \int_{\Re^M} \mathbf{w}.p(\mathbf{w}|\mathbf{d}, H_i)\, d\mathbf{w} \\ &= \frac{1}{p(\mathbf{d}|H_i)} \int_{\Re^M} \mathbf{w}.p(\mathbf{d}|\mathbf{w}, H_i)p(\mathbf{w}|H_i)\, d\mathbf{w} \end{aligned} \tag{9}$$

We intend $E[\mathbf{w}|\mathbf{d}, H_i]$ to be used as an estimate for $\mathbf{w}_{MAP}$. The accuracy of this estimate is dependent upon the general shape of the posterior $p(\mathbf{w}|\mathbf{d}, H_i)$, but the prevailing factor is the relative difference in magnitude between the global maximum at $\mathbf{w}_{MAP}$ and other local maxima. Essentially, the greater the difference, the more accurate the estimate, with equality attained in the limit $p(\mathbf{w}|\mathbf{d}, H_i) \to \delta(\mathbf{w}_{MAP})$.

As mentioned in the previous section, the posterior derived from the likelihood of the form $[C_1 + \Phi(\mathbf{w})]^{\frac{1-N-C_0}{2}}$ in equation 7 would be well suited as a strongly peaked function to provide a good estimate, but is not readily integrable *w.r.t.* $\mathbf{w}$. Numerical solutions would require inefficient recomputation of the expectation for every data observation $\mathbf{d}$. A significantly more desirable formulation would be a functional, rather than point estimate of $E[\mathbf{w}|\mathbf{d}, H_i]$.

We therefore consider replacing the likelihood with a similar function that will allow our analysis to continue, and meet our criteria of estimating $\mathbf{w}_{MAP}$ accurately with a functional estimate of $E[\mathbf{w}|\mathbf{d}, H_i]$. To this end, we require a simple monotonically decreasing function of $\Phi(\mathbf{w})$ that resembles the sharp peakedness of the likelihood function illustrated in figure 1.

A possible family of functions (figure 1) are $\left[1 - \frac{1}{M}\Phi(\mathbf{w})\right]^P$ where $P$ is the order of the function, and $M$ is a practical limit for $\Phi(\mathbf{w})$.

Now

$$\begin{aligned} \Phi(\mathbf{w}) &= \sum_n \left(d_n - y_n(\mathbf{w})\right)^2 \\ &= D + Y(\mathbf{w}) - 2\sum_n d_n y_n(\mathbf{w}) \end{aligned} \tag{10}$$

where

$$\begin{aligned} D &= \sum_n d_n^2 \\ Y(\mathbf{w}) &= \sum_n y_n^2(\mathbf{w}) \end{aligned}$$

The function $\left[1 - \frac{1}{M}\Phi(\mathbf{w})\right]^P$ when expanded generates a discrete Volterra series of order $P$ in $d_n$, $y_n(\mathbf{w})$, $D$, and $Y(\mathbf{w})$.

$$\begin{aligned} p_v^{P=1}(\mathbf{d}|\mathbf{w}, H_i) &= 1 - \frac{1}{M}D - \frac{1}{M}Y(\mathbf{w}) + \frac{2}{M}\sum_n d_n y_n(\mathbf{w}) \\ p_v^{P=2}(\mathbf{d}|\mathbf{w}, H_i) &= 1 - \frac{2}{M}D - \frac{2}{M}Y(\mathbf{w}) + \frac{4}{M}\sum_n d_n y_n(\mathbf{w}) \\ &\quad + \frac{2}{M^2}DY(\mathbf{w}) - \frac{4}{M^2}Y(\mathbf{w})\sum_n d_n y_n(\mathbf{w}) \\ &\quad - \frac{4}{M^2}D\sum_n d_n y_n(\mathbf{w}) + \frac{1}{M^2}D^2 + \frac{1}{M^2}Y^2(\mathbf{w}) \\ &\quad + \frac{4}{M^2}\sum_m\sum_n d_m d_n y_m(\mathbf{w}) y_n(\mathbf{w}) \\ p_v^{P=3}(\mathbf{d}|\mathbf{w}, H_i) &= \ldots \end{aligned} \tag{11}$$

The integration in equation 9 over the parameters $\mathbf{w}$ can now be performed numerically to obtain an approximation function for the parameter expectation

$$\begin{aligned} E_v[\mathbf{w}|\mathbf{d}, H_i] &= \frac{1}{p_v(\mathbf{d}|H_i)} \int_{\Re^M} \mathbf{w}.p_v(\mathbf{d}|\mathbf{w}, H_i)p(\mathbf{w}|H_i)\, d\mathbf{w} \\ E_v^{P=1}[\mathbf{w}|\mathbf{d}, H_i] &= \frac{1}{p_v(\mathbf{d}|H_i)}\Big[\mathbf{U}^0 - \frac{1}{M}\mathbf{U}^Y - \frac{1}{M}\mathbf{U}^0 D \\ &\quad + \frac{2}{M}\sum_n \mathbf{U}_n^y d_n\Big] \\ E_v^{P=2}[\mathbf{w}|\mathbf{d}, H_i] &= \frac{1}{p_v(\mathbf{d}|H_i)}\Big[\mathbf{U}^0 - \frac{2}{M}\mathbf{U}^Y + \frac{1}{M^2}\mathbf{U}^{YY} \\ &\quad - \frac{2}{M^2}(N\mathbf{U}^0 - \mathbf{U}^Y)D + \frac{1}{M^2}\mathbf{U}^0 D^2 \\ &\quad + \frac{4}{M^2}\sum_n (N\mathbf{U}_n^y - \mathbf{U}_n^{Yy})d_n - \frac{4}{M^2}D\sum_n \mathbf{U}_n^y d_n \\ &\quad + \frac{4}{M^2}\sum_m\sum_n \mathbf{U}_{mn}^{yy} d_m d_n\Big] \\ E_v^{P=3}[\mathbf{w}|\mathbf{d}, H_i] &= \ldots \end{aligned} \tag{12}$$

where the coefficients $\mathbf{U}$ are given by

$$\mathbf{U}^0 = \int_{\Re^M} \mathbf{w}.p(\mathbf{w}|H_i)\, d\mathbf{w} \;,\; \mathbf{U}^Y = \int_{\Re^M} \mathbf{w}.Y(\mathbf{w})p(\mathbf{w}|H_i)\, d\mathbf{w}$$

$$\begin{aligned} \mathbf{U}_n^y &= \int_{\Re^M} \mathbf{w}.y_n(\mathbf{w})p(\mathbf{w}|H_i)\, d\mathbf{w} \\ \mathbf{U}^{YY} &= \int_{\Re^M} \mathbf{w}.Y^2(\mathbf{w})p(\mathbf{w}|H_i)\, d\mathbf{w} \\ \mathbf{U}_n^{Yy} &= \int_{\Re^M} \mathbf{w}.Y(\mathbf{w})y_n(\mathbf{w})p(\mathbf{w}|H_i)\, d\mathbf{w} \\ \mathbf{U}_{mn}^{yy} &= \int_{\Re^M} \mathbf{w}.y_m(\mathbf{w})y_n(\mathbf{w})p(\mathbf{w}|H_i)\, d\mathbf{w} \end{aligned}$$

Finally

$$
\begin{aligned}
p_v(\mathbf{d}|H_i) &= \int_{\Re^M} p_v(\mathbf{d}|\mathbf{w}, H_i) p(\mathbf{w}|H_i)\, d\mathbf{w} \\
p_v^{P=1}(\mathbf{d}|H_i) &= V^0 - \frac{1}{M} V^Y - \frac{1}{M} V^0 D + \frac{2}{M} \sum_n V_n^y d_n \\
p_v^{P=2}(\mathbf{d}|H_i) &= \ldots
\end{aligned} \tag{13}
$$

where the coefficients $V$ are defined similarly to $\mathbf{U}$

$$
\begin{aligned}
V^0 &= \int_{\Re^M} p(\mathbf{w}|H_i)\, d\mathbf{w} \\
V^Y &= \int_{\Re^M} Y(\mathbf{w}) p(\mathbf{w}|H_i)\, d\mathbf{w} \\
V_n^y &= \ldots
\end{aligned}
$$

Thus the expectation functions $E_v[\mathbf{w}|\mathbf{d}, H_i]$ and model evidence $p_v(\mathbf{d}|H_i)$ are described by a set of discrete Volterra series of order $P$ in $d_n$ and $D = \sum d_n^2$ that can be directly applied to the data.

If the parameters are chosen carefully, they can be largely independent, and a quasi-random Sobol [4] sequence may be used for efficient Monte Carlo integration to calculate the coefficients $\mathbf{U}$ and $V$. The speed and accuracy of using Monte Carlo at this preprocessing stage needs to be reasonable but is not critical, and can be improved upon if there is prior knowledge of the parameter distribution $p(\mathbf{w}|H_i)$, otherwise a practical non-informative distribution is used.

Using equation 1, the Volterra series for the model evidence $p_v(\mathbf{d}|H_i)$ can be used to estimate $P(H_i|\mathbf{d})$ for model selection or classification. Accuracy is good enough for first stage reduction of the number of candidate hypotheses.

### 3.1. An iterative method for finding $\mathbf{w}_{MAP}$

The graphs of the likelihood approximations in figure 1 show that higher orders of P increase the relative magnitude of the global maximum and therefore the accuracy of the Volterra estimates, but at a cost of handling a high order Volterra series.

With a low order approximation $E_v[\mathbf{w}|\mathbf{d}, H_i]$ to $E[\mathbf{w}|\mathbf{d}, H_i]$, however, we are able to iterate towards the mode of the posterior by using successive parameter estimates to inverse transform the data. For example, if one of the parameters $t_d$ represents time or phase shift of a time series, and the first estimate for $E[t_d|\mathbf{d}, H_i] = \hat{t}_d^0$, then we apply an inverse time shift *i.e.* $-\hat{t}_d^0$ to the observed data $\mathbf{d}$ to obtain $\mathbf{d}^1$. Reapplying the Volterra estimator to $\mathbf{d}^1$, we obtain $\hat{t}_d^1$, and again inverse time shift the original data $\mathbf{d}$ by $-(\hat{t}_d^0 + \hat{t}_d^1)$ to obtain $\mathbf{d}^2$. This continues until $\mathbf{d}^n$ converges to our model prototype $\mathbf{y}^{(i)}(0)$, $E[t_d|\mathbf{d}^n, H_i] \to 0$ and $\sum_{k<n} \hat{t}_d^k \to E[t_d|\mathbf{d}, H_i]$.

### 3.2. Incomplete Data

If only a subset of the observed data $\mathbf{d} = \{\mathbf{d}_v, \mathbf{d}_o\}$ is valid, *e.g.* due to occlusion, the Volterra estimators can still be applied to the visible subset $\mathbf{d}_v$ by eliminating the occluded $d_n$ terms in equations 12 and 13, and the corresponding coefficient terms $\mathbf{U}_n$ and $V_n$ from the full Volterra series. Therefore, providing the precalculated Volterra coefficients are stored in an accessible form, the preliminary numerical integration used to calculate the coefficients need not be reapplied in order to form the reduced Volterra series, although the overhead associated with the coefficient elimination process increases with $O(N^P)$.

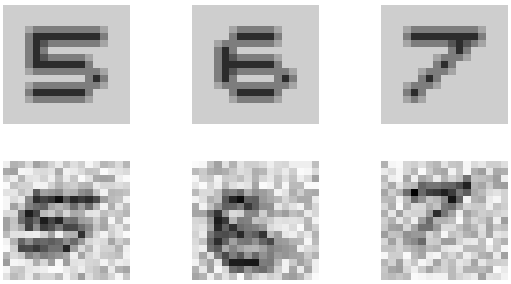

Figure 2: Sample model prototypes and test data

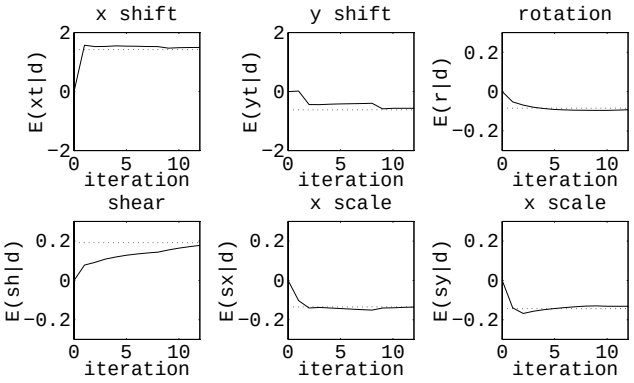


Figure 3: Progress of parameter estimates for a test sample

## 4. EXAMPLE IMPLEMENTATION OF A BASIC CHARACTER RECOGNISER

A set of ten synthetic greyscale numerals centred on a $17 \times 17$ grid were created to represent the basis of ten class models, which have parameters representing the affine transformations linear shear $sh$, rotation $r$, scaling $s_x$ and $s_y$, translation $x_t$ and $y_t$ and additive white noise standard deviation $\sigma$.

Figure 2 shows sample model prototypes alongside random test observations simulated from the model class prototypes with uniformly distributed random transformations and additive noise described by the following table

| | |
|---|---|
| typical SNR | 5 dB |
| $x_t$, $y_t$ range | $\pm 1.6$ pixels |
| $s_x$, $s_y$ range | $\pm 0.2 \frac{\triangle x}{x}$, $\frac{\triangle y}{y}$ |
| $r$ range | $\pm 0.2$ radians |
| $sh$ range | $\pm 0.2 \frac{\triangle x}{y}$ |

A set of first order Volterra series ($P = 1$) for the estimates of the parameters and class probabilities were generated, based upon the full $17 \times 17$ data set, and a subsampled $9 \times 9$ data set. This hierarchical arrangement allows for coarse parameter estimation over a greater range, followed by application of a finer set of parameter estimators.

Figure 3 shows the progress of the estimates for each of the parameters for a sample two stage process. The first eight iterations are performed on the $9 \times 9$ interpolated data, and the last four on the $17 \times 17$ grid.

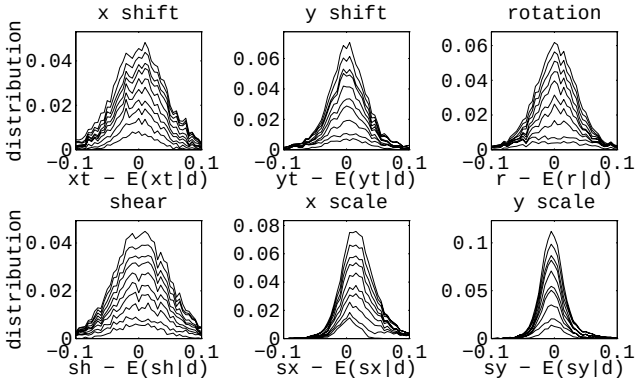


Figure 4: Histograms for errors in parameter estimates with 5dB additive white noise in data

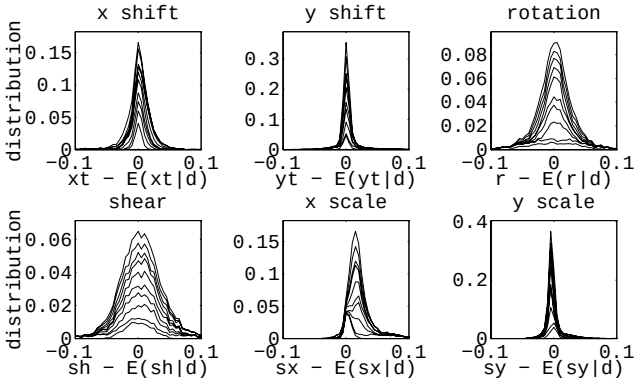


Figure 5: Histograms for errors in parameter estimates with no additive noise in data

The Volterra estimates of the class probabilities (equation 13) $p_v(H_i|\mathbf{d}) \propto p_v(\mathbf{d}|H_i)$ are compared for classification after every iteration, and numerical integration only performed to evaluate the true model evidence if the two top candidates are very close after the final iteration. Data is never rejected at any stage in order that the basic substitution error rate could be measured.

### 4.1. Results

| Stage of hierarchy | Errors of 100000 | % error |
|---|---|---|
| Initial iteration $9 \times 9$ | 35871 | 35.9 |
| Final $(8^{th})$ iteration $9 \times 9$ | 7007 | 7.0 |
| Initial iteration $17 \times 17$ | 452 | 0.45 |
| Final $(4^{th})$ iteration $17 \times 17$ | 74 | 0.07 |
| Final classification $17 \times 17$ | 0 | 0.00 |

Despite the low SNR, the relatively large data dimensionality of $17 \times 17$ ensures a very small theoretical probability of error, providing the mode of the posterior $\mathbf{w}_{MAP}$ is estimated correctly. The final classification error rate realises this theoretical probability and the decreasing error rate at succesive stages validates the convergence of the Volterra estimator upon the mode.

Histograms of the error in the parameter estimates for 10000 test data samples with additive white noise and no noise are displayed in figures 4 and 5 respectively. The histograms are each initially computed by class, and then displayed cumulatively over the classes. Thus the lowest line represents the error histogram for class 0, the next for classes 0 and 1, and so forth, until the top line represents the histogram over all classes.

As expected, the variance of the error is reduced when the additive noise is removed, though markedly more so for the translation parameters $x_t$ and $y_t$. One may suggest that the latter disparity is related to the correlation between the parameters and the data. At one end of the scale, translation affects all pixels equally, whereas scaling, rotation and shear have a significant effect only at increasing distances from their neutral points. Estimates for the scaling parameters $s_x$ and $s_y$ are, however, generally improved because of their influence on the $D = \sum d_n^2$ power term.

Further examination of the means of the error histogram distributions reveals that the parameter $s_x$ is systematically underestimated and $s_y$ systematically overestimated by about $1\%$ for all character classes except $0$ and $1$. It is believed that this is due to interpolation errors arising whilst performing the affine tranformations over the discrete space. This is also considered to be the cause of the local peaks and troughs in the histograms by way of creating local maxima close to the global mode of the posterior.

## 5. CONCLUSION

Bayesian analysis provides a framework for model selection or classification. In general, the analysis is intractable, and requires numerical or approximation techniques to obtain estimates for the probability of each model. The estimates can be enhanced if an estimate of the mode of the posterior parameter distribution is available. By substituting a polynomial function for the likelihood, we can generate an estimator in the form of a Volterra series that can be applied directly to the data. Accuracy improves with higher polynomial orders, at the expense of handling an exponentially increasing number of terms. Lower orders, however, can be employed iteratively, and by example of a basic character recogniser, are shown to be both effective and efficient multi-dimensional parameter estimators.

The Volterra parameter estimator has been further applied successfully to occluded object recognition, using its ability to generate estimates from incomplete data, while applicability to other fields including motion estimation, image registration, range estimation, and time series classification are to be investigated.

## 6. REFERENCES

[1] M.Hsieh and P.Rayner, *Extension of the General Linear Model to include prior parameter information*, IEEE International Conference on Acoustics, Speech, and Signal Processing pages 3569-3572, 1997.

[2] H.Jeffreys, *Theory of Probability*, Oxford University Press, 1939.

[3] D.Mackay, *Bayesian Methods for Adaptive Models*, PhD Thesis submitted to California Institute of Technology, 1991

[4] W.Press, S.Teukolsky, W.Vetterling, and B.Flannery, *Numerical Recipes in C*, Cambridge University Press, 2nd edition, 1992.

[5] J.O Ruanaidh and W.Fitzgerald, *Numerical Bayesian Methods applied to Signal Processing*, Cambridge University Engineering Department, 1995.