Tina Memo No. 2004-006
Internal

Parameter Estimation for EM Mixture Modelling and its Relationship to Likelihood and EML.

N.A.Thacker.

Last updated
5 / 5 / 2008

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Parameter Estimation for EM Mixture Modelling and its Relationship to Likelihood and EML.

N.A.Thacker. 30/10/2004, updated 5/5/2008

Abstract

This document is intended for those who already know how to implement an Expectation Maximisation (EM) algorithm but might wish to understand better the relationship of the update process to probability theory. It aims to investigate the relationship between likelihood estimation of parameters describing a curve and parameter estimation methods used for histogram fitting and mixture modelling. This document lays out what I believe to be the true origin of these approaches from probability theory. It is shown that the standard solutions for density parameter terms can be derived from the definition of a Poisson sampling process, via the use of Extended Maximum Likelihood (EML)[1]*. This result is then compared to conventional function fitting of data with Gaussian errors, using the equal variance relationship. We conclude with some comments on the construction of histogram similarity measures and their extension for use with probability density distributions.*

Introduction

The standard proof that the average of a distribution of samples (x_j) is the maximum likelihood estimate of the mean parameter x_c can be derived directly on assumption of a Gaussian distribution. However, in order to apply the approach to a set of frequency measurements n_i describing a distribution over as space x_i, such as a histogram, we would be defining the likelihood in terms of the measurement variable rather than considering the stability of the sample estimates n_i. In addition, although we could perhaps estimate x_c this way, we would be unable to define a likelihood estimate for the normalisation of the distribution (A) using this approach. Under some circumstances, this could be considered as confusing measurement repeatability (error) with data distribution (signal), and therefore not necessarily a valid use of likelihood. Yet this process is exactly what is done in many situations when we wish to determine the parameters describing a distribution of data. The key question addressed here is; What is the relationship between; likelihood, estimation of the mean from a distribution and the Expectation Maximisation (EM) algorithm?

Conventional uses of the EM algorithm have the advantage that parameters from a distribution can be estimated via a very simple process (for example taking the weighted mean of a set of sample values). This avoids the need for downhill search. If we can convert a measurement and curve fit process (based upon an assumption of Gaussian residuals), such as a least-squares fit, into the equivalent of a sample of data we may be able to use these approaches to our advantage in the fitting of more general functions. At the same time it should give us some insight into the origins of parameter update schemes for mixture modelling.

The method to achieve this will be to make use of the square-root transform for Poisson sample data [6]. Generally, we use this relationship to transform a set of Poisson variables (eg: $n \pm \sqrt{n}$) to data with closely approximate equal variance ($d \pm 1/2$), in order to construct the likelihood for a set of data points with uniform variance (Figure 1);

$$-logP \;=\; 1/4 \sum_i (d_i \;-\; f(x_i))^2 \;\approx\; 1/4 \sum (\sqrt{n_i} \;-\; f(x_i))^2$$

This makes it possible to generate simplified solutions for systems of Poisson variables. Although an approximation, this approach is expected to be more accurate for determination of parameter estimates describing the density funciton $\lambda(x_i)$, from Poisson samples, than simply applying standard χ^2 measures, ie;

$$\chi^2 \;=\; \sum \frac{(n_i \;-\; \lambda(x_i))^2}{\lambda(x_i)}$$

due to the improved Gaussian approximation of Poisson data achieved following the square-root transform.

We wil show that for the EM algorithm, it is the Poisson nature of the data which makes simple closed form solution possible. We will therefore use the equal variance relationship in the other direction, to estimate an equivalent likelihood solution for Gaussian residual distributions from exact solutions for Poisson statistics. We will then compare the likelihood based and equivalent EM style parameter estimation.

[1] Previous drafts included a derivation which implied that sample moments were always Likelihood estimates, this contained an error and had been removed.

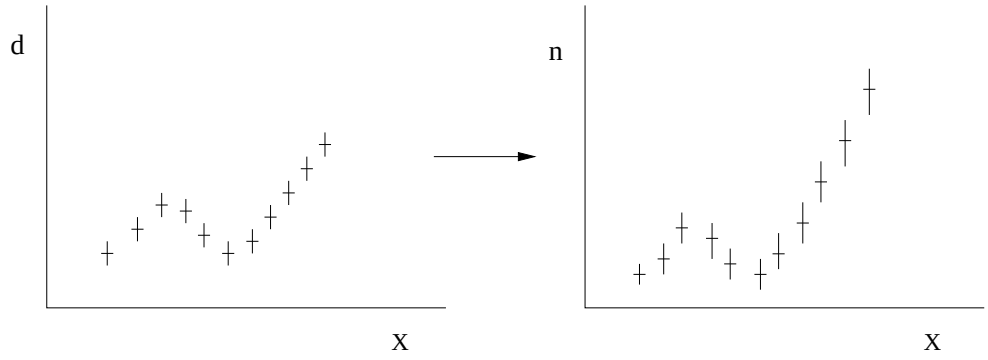

Figure 1: *Transformation of a uniform variance data set to a statistically equivalent Poisson sample.*

Fitting a Data Curve

If we wish to fit the curve $y_i = Af(x_i, x_c)$ to a set of data d_i with independent uniform random Gaussian noise, we should minimise;

$$\sum_i (d_i - Af(x_i, x_c))^2 = \sum_i d_i^2 + \sum_i A^2 f^2(x_i, x_c) - 2\sum_i d_i A f(x_i, x_c) \qquad -(1)$$

differentiation with respect to A and setting to zero we get the exact likelihood estimate for A;

$$A = \frac{\sum_i d_i f(x_i, x_c)}{\sum_i f^2(x_i, x_c)} = \mathbf{d}.\mathbf{f}/|\mathbf{f}|^2$$

Which, by setting $|\mathbf{f}| = 1$ can be written as

$$A = |\mathbf{d}| cos(\theta)$$

where the $cos(\theta)$ is the angle between the data and model vectors in the data space. This term is a measure of the detailed match between the 'shape' of the data and the function, such that if the model is not located over the data the normalisation term will be penalised. From the point of view of curve fitting this is a perfectly reasonable behaviour.

Now by differentiating (1) with respect to x_c and setting to zero we can determine a constraint for the likelihood estimate of x_c

$$A^2 \frac{\partial |\mathbf{f}|^2}{\partial x_c} + 2A \sum_i d_i \frac{\partial f(x_i, x_c)}{\partial x_c} = 0$$

Once again defining $|\mathbf{f}| = 1$ and for $A \neq 0$ gives

$$\sum_i d_i \frac{\partial f(x_i, x_c)}{\partial x_c} = 0 \qquad -(2)$$

Solution of which should give the likelihood estimate of x_c.

A True Likelihood for EM

We will now consider the process of estimating the parameters for a distribution which models a sample of data measured as a set of n_i frequencies. The aim is to define the probability of observing a set of sample data $\mathbf{x}_i$ in a multi-dimensional (continuum) space based upon a specific probability density function $\lambda(\mathbf{x}_i)$. To do this we must define a finite set of sample cells ($\mathbf{X}_i$) and compute the probability of observing data within each and take the continuum limit. The correct statistical model for this is the Poisson distribution, so that the probability of observing a given number of samples within each discrete region within the non-zero portion of the probability density is;

$$P(n_i, \lambda) = \frac{exp(-\lambda)\lambda_i^n}{n!}$$

We can therefore write down the log probability for the model in terms of the data as the sum of independent terms for each quantity of observation $n_i = 0$,$n_i = 1$,$n_i = 2$ ...;

$$log(P(\lambda)) = \sum_i^{N_0} log\ P(0, \lambda(\mathbf{X_i})) + \sum_i^{N_1} log\ P(1, \lambda(\mathbf{X_i})) + \sum_i^{N_2} log\ P(2, \lambda(\mathbf{X_i})) + ...$$

the first being N_0 empty cells, the second being N_1 cells containing one sample and N_2 cells containing two samples, etc. Equally we can write this as;

$$log(P(\lambda)) = \sum_i^{N_0+N_1+N_2+...} log\ P(0, \lambda(\mathbf{X_i})) +$$

$$\sum_i^{N_1} log\ P(1, \lambda(\mathbf{X_i})) - log\ P(0, \lambda(\mathbf{X_i})) + \sum_i^{N_2} log\ P(2, \lambda(\mathbf{X_i})) - log\ P(0, \lambda(\mathbf{X_i})) + ...$$

Substituting the appropriate Poisson terms gives;

$$log(P(\lambda)) = \sum_i^{N_1} log(\lambda(\mathbf{X}_i)\ exp(-\lambda(\mathbf{X}_i))/1!) + (\lambda(\mathbf{X}_i)) + \sum_i^{N_2} log(\lambda^2(\mathbf{X})\ exp(-\lambda(\mathbf{X}_i))/2!) + (\lambda(\mathbf{X}_i)) + ...$$

$$- \sum_i^{N_0+N_1+N_2+...} \lambda(\mathbf{X}_i)$$

Which simplifies to;

$$= \sum_i^{N_1+N_2+...} n_i log(\lambda(\mathbf{X}_i)) - \sum_i^{N_0+N_1+N_2+...} \lambda(\mathbf{X}_i) + k_1$$

where k_1 is a constant[2] defined by the sample of data and n_i is the quantity of data at $\mathbf{X}_i$. In this form we can see that the second term (which corresponds to the probability of a zero entry in every location) is simply the integrated probability density which we can make constant as part of the model definition. Therefore, only the first term is data dependant. This has the obvious benefit that we do not need to attempt to compute probability terms for the empty cells. In addition, this expression supports the calculation of the normalisation parameter B, for **arbitrary data distributions**. Writing $\lambda(\mathbf{X}_i) = Bf^2(\mathbf{X_i})$, differentiating with respect to B and setting equal to zero in the usual way gives;

$$\frac{N_1 + 2N_2 + 3N_3...}{B} - \sum_i^{N_0+N_1+N_2+...} f^2(\mathbf{X}_i) = 0$$

Solving with $\sum_i^N f^2(\mathbf{X}_i) = 1$ gives the standard mixture update estimate of $B = N_1 + 2N_2 + 3N_3 + ...$ That is the **total number of observations** M.

We can write;

$$log(P(\lambda)) = \sum_i^N n_i log(\lambda(\mathbf{X}_i)) - v + k_1$$

where v is the integral of the density function and N is the number of non zero cells. Summing now over the M individual data entries j , the log probability is given by;

$$log(P(\lambda)) = \sum_j^M log(\lambda(\mathbf{X}_j)) - v + k_1$$

This approach is also directly applicable to the continuum ($\mathbf{x}_j$).

$$log\ L = \sum_j^M log(\lambda(\mathbf{x}_j)) - k_2 \int \lambda(\mathbf{x})\ d\mathbf{x} \qquad -(3)$$

[2]Provided we do not respecify the sampling process, such as happens in some bootstrap likelihood techniques [8].

Where k_2 is an unknown constant which relates probability density $\lambda(\mathbf{x}_j)$ to discrete probability $\lambda(\mathbf{X}_j)$ [3].

As a consequence the data dependant term of the log probability **looks like** a likelihood formulated using the probability density $\lambda(\mathbf{x}_j)$. Perhaps because of this, the measure (first described by Fermi) is referred to as Extended Maximum Likelihood (EML) [1]. It is extended in the sense that, rather than simply computing the likelihood for a fixed density distribution it allows parameters describing the density distribution to also be determined. A-priori knowledge of the density distribution allows the regeneration of the conventional likelihood [4] (first term in equation (3)). This result also tells us that this first term is sufficient for joint determination of the parameters of the density model provided that we fix the normalisation [5].

Equation (3) is the statistical basis for the EM algorithm. In the context of the use of an EM optimisation we can therefore estimate a new set of parameters by minimising the first term subject to the constraint of fixed normalisation of the density function, (so that the arbitrary constant k_2 plays no role in the solution).

The presence of an unknown constant term (k_2) prevents determination of an absolute normalisation when applied to a density function over a continuous space. This is an expected consequence of transforming to continuous probability densities. Notice also that the definition of a probability density in this context is as the large sample limit of a Poisson sampling process.

As menitioned above, it is well known that the likelihood estimate of the mean of a Gaussian distribution is given by the mean of the sample. However, more general solutions can be derived from the above likelihood. Differentiating $log\ L$ with respect to x_c (the translation parameter [6]) we get the constraint;

$$\sum_j \frac{\partial \lambda(\mathbf{x}_j)}{\partial \mathbf{x}_c} / \lambda(\mathbf{x}_j) \ = \ 0 \qquad -(4)$$

Comparison of Gaussian Likelihood and EML

We now compare the above results with parameter estimation for Gaussian based likelihood. The method for estimation of the normalisation constant from a sample of n_i data follows from EML and can be written as;

$$B \ = \ \sum_i n_i$$

This corresponds to the simplest case for the EM algorithm, where all data has been identified with one component of the EM model. The more general case is obtained by introducing the appropriate probability weighting terms (which can be justified on the basis of the standard proof for convergence of the EM algorithm [2]).

Applying the variance normalising transform to the previous result $n_i \rightarrow d_i^2$ this would be equivalent to

$$B \approx A^2$$

The approximation here is due to the approximation of the Poisson distribution to a Gaussian via a square-root transform. Which can strictly only be expected to be exact for large samples.

Comparing these two approaches we have;

$$\sum_i n_i \approx \sum_i d_i^2 \ \rightarrow \ \sum_i d_i^2 cos^2(\theta)$$

The implication seems to be that the normalisation parameter in the EM algorithm ignores the shape dependency of the agreement between the data and fitted curve (ie $cos(\theta) \ = \ 1$).

Considering now the localisation parameter x_c. Taking the same approach as previously , the single mixture component EM algorithm is of course;

$$x_c \ = \ \frac{\sum_i n_i x_i}{\sum_i n_i} \ \approx \ \frac{\sum_i d_i^2 x_i}{\sum_i d_i^2}$$

[3] Strictly k_2 is the interval on $\mathbf{x}$ which specifies an independant sample and in order for it to take the form of a simple constant (and not be a function of $\mathbf{x}$, $k_2(\mathbf{x}_i)$), we must assume that this interval is the same everywhere within the measurement space; ie: we are in an equal variance domain [9].

[4] A tribute to the self consistency of probablity theory.

[5] The probability densitities must also be specified in the equal variance domain.

[6] Notice that the derivative of the integral of the density function with respect to pure translation is zero.

For the specific case of a Gaussian curve fit

$$f(x_i, x_c) \;=\; G(x_i, x_c, \sigma) \;=\; k\; exp - (x_i \;-\; x_c)^2/2\sigma^2)$$

constraint equation (2), for the fitting of data with Gaussian residuals, becomes;

$$\sum_i d_i(x_i \;-\; x_c)G(x_i, x_c, \sqrt{2}\sigma) \;=\; 0$$

which we can write as

$$\sum_i d_i^2(x_i \;-\; x_c) \;+\; \sum_i d_i(x_i \;-\; x_c)(G(x_i, x_c, \sqrt{2}\sigma) \;-\; d_i) \;=\; 0$$

In this form, we can see we must assume that the second term is on average zero for the correct model (which is true once again when $cos(\theta) \;=\; 1$), then

$$\sum_i d_i^2(x_i \;-\; x_c) \;=\; 0$$

such that the standard solution for x_c is generated.

For the general case of centroid estimation, the two constraint equations ((2)[7] and (4)) can be related directly;

$$\sum_i n_i \frac{\partial \lambda(x_i)}{\partial x_c}/\lambda(x_i) \approx 2 \sum_i d_i^2 \frac{\partial f(x_i, x_c)}{\partial x_c}/f(x_i, x_c) \rightarrow \sum_i d_i \frac{\partial f(x_i, x_c)}{\partial x_c}$$

Once again this association can be resolved by assuming that the data and the model match exactly at the solution.

Conclusions

This document has described the likelihood estimation of parameters from data with Gaussian errors and Poisson samples. It has then compared the form of these solutions via the use of the equal-variance (variance normalising) transform.

We are now in a position to understand how the Maximisation step in the EM algorithm manages to directly estimate the key parameters of Gaussian mixture distributions. Transformation of the result for arbitrary distributions of data illustrates that the main difference between the methods is due to the inability to form a solution, from a Gaussian derived likelihood, without assuming that the data and model will be identical at the solution. This would only appear to be true in the limit of infinite sample sizes, where both solutions coverge [8].

Because of the relationship between least-squares fitting and Poisson statistics via the equal-variance transform, we would appear to have a choice for ways of estimating parameters A and x_c for a curve fit of data d_i to the function $Af(x_i, x_c)$.

The standard derivation from likelihood (normalising the function with $|\mathbf{f}| = 1$) would give;

$$A \;=\; \sum_i d_i f(x_i, x_c)$$

Which can be usefully applied in the context of an iterative optimisation scheme such as the 'Simplex' algorithm of Nealder and Mead. This algorithm iteratively adjusts a set of parameters and evaluates the success of this manipulation by estimating the cost function. Results such as the one above can be used to eliminate one free parameter prior to each function evaluation, within the function evaluation subroutine. In particular this has been used in the TINA software to eliminate the scale parameter during fitting of contrast time course curves for analysis of MRI perfusion [4]. In fact this idea is just a simple example of the more general method of partitioning the free parameters in terms of linear and non-linear terms. The linear terms can be estimated by direct solution (such as this above) before calculation of the cost function using the remaining non-linear terms. In many cases the effective reduction in dimensionality of the search space can have real benefits in execution speed at no cost to theoretical rigour. An example of the use of this technique in medical image analysis can be found in [10].

[7] Notice we are summing over locations x_i not samples j.

[8] Like conventional likelihood it strictly requires that we work in a measurement domain of equal variance.

The equivalent form based upon the mixture model update approach would be;

$$A \approx \sqrt{\sum_i d_i^2}$$

The advantage of this formula over the previous one being that it is independent of the agreement between location and shape of the fitted function.

The analysis presented in this document also makes it possible to identify three different approaches to the fitting of histogrammes of discrete bins X_i with frequency values n_i to a density function $\lambda(X_i)$. In order of accuracy of approximation to the Poisson distribution, we can minimise;

the conventional χ^2

$$\chi^2 = \sum_i \frac{(n_i - \lambda(X_i))^2}{\lambda(X_i)} \qquad -(5)$$

the equal variance approximation (expected to be twice as accurate);

$$log(P) \approx \frac{1}{4}\sum_i (\sqrt{n_i} - \sqrt{\lambda(X_i)})^2 \qquad -(6)$$

or the Extended Maximum Likelihood

$$log(L) = \sum_i n_i \; log(\lambda(X_i)) - v \qquad -(7)$$

which will give an **exact** likelihood estimate for Poisson distributed data.

However, we should remember that the probability calculations (for the probability of the data given the model), from which these expressions are derived, strictly **only deliver valid probability estimates for the true value of the parameters** $\lambda^*(x_i)$. This is a subtle but important point, as $\lambda(x_i)$ moves away from the true value the computed statistics no-longer have construct validity. In fact the exact likelihood basis for (7) may be of less value than we might think, as the likelihood estimate for $\lambda(x_i)$ will always be slightly different to $\lambda^*(x_i)$. Though *parameter estimates* are possibly unbiased (as defined over a series of possible estimations), this difference accounts for the well known bias of likelihood based *statistics*.

Equation (7) also has the advantage that we can extend this to continuous probability densities $\lambda(x_i)$ (equation (3)). It is this last property which makes this measure the statistical basis for parameter estimation within the EM algorithm. It follows from the standard proof of convergence that any valid likelihood estimate for mixture model terms can be used within the EM framework.

Additional Comments

Given the material I have written regarding this subject previously [9], it would be remiss of me not to comment on the relationship of these conclusions to pattern recogntiion. We can extend the above analysis further and for the limit of infinite sample sizes we can derive expressions for the comparison of probabilities distributions from histogram similarity measures. At first sight this might seem to suggest that equation (7) will generate the appropriate way to compare probability densities.

$$L_{KL} = \int p(x) log(\lambda(x)) \; dx$$

This would lead to the Kullback-Liebler divergence. This obervation may be one of the reasons for the widspead acceptance of this measure as the 'correct' metric. However, simple inspection of this result tells us that something has gone wrong. The measure is not symmetric under interchange of $p(x)$ and $\lambda(x)$. The reason for this is that although we have been careful to approach $p(x)$ from the definition as a Poisson limit, we were not so careful regarding $\lambda(x)$. As a consequence, the Kullback-Liebler measure gives the correct similarity measure only for the trivial case of two identical probability densities, away from this point it reverts to an approximation.

The correct way to compare probability distributions across a space must formulate a similarity measure with more care. Specifically, for a Euclidean form this must be done in an equal variance domain (as explained in [5]) in order to be consistent with the concept of Fisher information. Following this approach we see that it is the second

of these alternatives (equation (6)) which (being exact in the probability limit ($n/M \rightarrow p$ and also $l/M \rightarrow \lambda$)) leads to the correct way to compare probabilities using the Matusita measure.

$$L_M = \int (\sqrt{p(x)} - \sqrt{\lambda(x)})^2 \, dx$$

Similarly, we would probably prefer to use equation (6) rather than (7) if we wish to define a distance between dissimilar histogrammes rather than trying to establish equivalence.

References

[1] R.J. Barlow. Statistics: A Guide to the use of Statistical Methods in the Physical Sciences. John Wiley and Sons, U.K., 1989.

[2] A.J.Lacey, N.A.Thacker and M.Pokric, Model election and the EM Algorithm, Tina memo, 2001-005, 2005.

[3] M. Pokric, N.A. Thacker, M.L.J. Scott and A. Jackson, Multi-dimensional Medical Image Segmentation with Partial Voluming. Tina memo, 2001-009, 2001.

[4] N.A.Thacker, A new Approach for the Estimation fo MTT in Bolus Passage Perfusion Techniques. Tina memo. 2001-003, 2001.

[5] 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.

[6] N.A. Thacker and P.A. Bromiley, The Effects of a Square Root Transform on a Poisson Distributed Quantity. Tina memo, 2001-010, 2001.

[7] N.A.Thacker, M. Pokric, Noise Filtering and Testing for MR Using a Multi-Dimensional Partial Volume Model, Tina memo, 2003-007, 2003.

[8] N.A. Thacker, P.A. Bromiley and M. Pokric, Computing Covariances for Mutual Information Co-registration. Tina memo, 2001-013, 2001.

[9] N.A.Thacker, P.Bromiley, The Equal Variance Domain: Issues Surrounding the use of Probabulity Densities for Algorithm Construction. Tina memo, 2004-005, 2004.

[10] D.C.Williamson and N.A.Thacker, Assessment of the Padé Approximant as a Method for Quantifying ^{1}H Magnetic Resonance Spectroscopic Data. Tina memo, 2003-008, 2003.