Tina Memo No. 2011-004
Internal

Extended Maximum Likelihood vs Maximum Likelihood ; Error Estimates on Quantitative Bayesian Priors[1]

P. Tar

Last updated
20 / 9 / 2011

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

[1]Supervisor; N. Thacker

Extended Maximum Likelihood vs Maximum Likelihood Error Estimates on Bayesian Priors

Paul Tar

Abstract

This document updates the analysis of errors found in Tina Memo 2010-008. In particular we show that the errors presented there are exact for EML interpretations of parameter estimates, but only an approximation for conventional Likelihood. The work explains the error model used in Tina Memo 2000-007, for the analysis of distribution of cranial fluid as measured in a volumetric segmentation of MR images. This document provides useful insights into the practical distinctions which must be made between the two alternative definitions of Likelihood, with regard to parameter estimation and the use of algorithms such as EM.

1 Bayes Theorem and the EM Cost Function

Given a set of probability densities, $P(X|k)$, for data generation processes, $k \in K$, over some pattern space X, Bayes Theorem can be applied to determine the probability that a given observed pattern, X, was generated by any particular process:

$$P(k|X) = \frac{P(X|k)P(k)}{\sum_{l \in K} P(X|l)P(l)}$$

where $P(k)$ is proportional to the prior probability that class k was the source of the observation. If Bayes Theorem is to be applied quantitatively then prior terms must be made proportional to the actual frequencies with which the different classes of data generation process occur within a dataset. Given a dataset, $(X_1, X_2, \ldots, X_N)$, an appropriate prior could simple be the quantity of data points believed to be associated with each classification, $Q(k)$, such that $\sum_{k \in K} Q(k) = N$. This leads to the circular definition:

$$Q(k) = \sum_{d=1}^{N} \frac{P(X_d|k)Q(k)}{\sum_{l \in K} P(X_d|l)Q(l)}$$

The best fitting values of $Q(k)$ can be determined using the parameter update method of Expectation Maximisation, which minimises the cost function:

$$F = -\sum_{d=1}^{N} \ln \sum_{k \in K} P(X_d|k)Q(k)$$

This is monotonically related to the likelihood of observing all the data points given the chosen values of $Q(k)$.

1.1 Errors on priors

Assuming negligible noise on data points and probability densities, the largest source of uncertainty in the maximum likelihood estimates of $Q(k)$ will be due to the likelihood function itself. The Cramer Rao Lower Variance Bound can be applied to estimate the uncertainty caused by F. The CRB can give an inverse variance estimate using:

$$\frac{\partial^2 F}{\partial Q(k)^2} \geq \frac{1}{var(Q(k))}$$

However, there are two possible interpretations of the log-likelihood function, F, depending upon whether or not the total quantity of data, N, is known a-priori. If the total number of data points can fluctuate then the maximum likelihood estimates of $Q(k)$ must also estimate N. This estimate requires modification of the cost function to give the Extended Maximum Likelihood version, F_{eml}, given in the next section. On the other hand, if N is fixed there are fewer parameters to estimate as the sum over $Q(k)$ is constrained. This estimate requires a correctly normalised Maximum Likelihood version, F_{ml}. The rest of this document explores the variance estimate of both interpretations.

2 Unknown N: CRB with Extended Maximum Likelihood

If the total number of data points is a Poisson distributed quantity with a mean value of N then the Extended Maximum Likelihood values of $Q(k)$ are given by the modified cost function

$$F_{eml} = -\sum_d \ln \sum_{k \in K} P(X_d|k)Q(k) - \sum_{k \in K} Q(k)$$

Where $N = \sum_{k \in K} Q(k)$. It will be shown here that the inverse variance estimate of this interpretation of the cost function is identical to applying the CRB to the original unmodified function, F, with unconstrained N. Before continuing, to avoid potential ambiguities and to remove unnecessary summations, the cost function will be rewritten in terms of a single class k and all other classes $\bar{k}$, i.e. all classes which are not k:

$$F_{eml} = -\sum_d \ln[P(X_d|k)Q(k) + P(X_d|\bar{k})Q(\bar{k})] - [Q(k) + Q(\bar{k})]$$

Now, computing the first and second derivatives with respect to the quantity of class k gives:

$$\frac{\partial F_{eml}}{\partial Q(k)} = -\sum_d \left(\frac{P(X_d|k)}{P(X_d|k)Q(k) + P(X_d|\bar{k})Q(\bar{k})} \right) - 1$$

Note that the -1 would be omitted if the original cost function, F, was differentiated by this point, as it does not require the $\frac{\partial}{\partial Q(k)}[Q(k) + Q(\bar{k})]$ term. This term vanishes completely in the second derivative giving:

$$\frac{\partial^2 F_{eml}}{\partial Q(k)^2} = \frac{\partial^2 F}{\partial Q(k)^2} = \sum_d \frac{P(X_d|k)^2}{(P(X_d|k)Q(k) + P(X_d|\bar{k})Q(\bar{k}))^2}$$

Multiplying by a factor of $Q(k)^2/Q(k)^2$ allows Bayes Theorem to be applied, which lets the second derivative be written in terms of $P(k|X)$:

$$\frac{\partial^2 F_{eml}}{\partial Q(k)^2} = \frac{\partial^2 F}{\partial Q(k)^2} = \frac{\sum_d P(k|X_d)^2}{Q(k)^2}$$

Now, from the definition of the CRB an estimate of the variance of $Q(k)$ (our Baysian prior) can be given by:

$$var(Q(k)) \approx \frac{Q(k)^2}{\sum_d P(k|X_d)^2}$$

This estimate will, in general, give a larger variance than the constrained version examined next.

3 Known N: CRB with Maximum Likelihood

If the total number of data points is known then N constrains the quantities of all classes. In terms of class k, and all other classes $\bar{k}$, it must be true that:

$$Q(k) + Q(\bar{k}) = Q(k) + (N - Q(k)) = N$$

This constraint holds when performing parameter estimation using expectation maximisation (EM), and can therefore be seen as the basis of many image segmentation algorithms.

This leads to a strictly normalised Maximum Likelihood interpretation of the original cost function, F, which can be written

$$F_{ml} = -\sum_{d=1}^{N} \ln[P(X_d|k)Q(k) + P(X_d|\bar{k})(N - Q(k))]$$

Now, computing the first and second derivatives with respect to $Q(k)$ gives:

$$\frac{\partial F}{\partial Q(k)} = -\sum_{d=1}^{N} \frac{P(X_d|k) - P(X_d|\bar{k})}{P(X_d|k)Q(k) + P(X_d|\bar{k})(N - Q(k))}$$

$$\frac{\partial^2 F}{\partial Q(k)^2} = \sum_{d=1}^{N} \frac{[P(X_d|k) - P(X_d|\bar{k})]^2}{[P(X_d|k)Q(k) + P(X_d|\bar{k})(N - Q(k))]^2}$$

Letting $Q(X_d) = P(X_d|k)Q(k)+P(X_d|\bar{k})(N-Q(k))$ and inserting factors of $Q(k)/Q(k)$ and $(N-Q(k))/(N-Q(k))$ gives:

$$\frac{\partial^2 F}{\partial Q(k)^2} = \sum_{d=1}^{N} \left(\frac{P(X_d|k)Q(k)}{Q(X_d)Q(k)} - \frac{P(X_d|\bar{k})(N - Q(k))}{Q(X_d)(N - Q(k))} \right)^2$$

Which via Bayes Theorem becomes:

$$\frac{\partial^2 F}{\partial Q(k)^2} = \sum_{d=1}^{N} \left(\frac{P(k|X_d)}{Q(k)} - \frac{1 - P(k|X_d)}{N - Q(k)} \right)^2$$

Placing the bracketed terms over a common denominator then gives:

$$\frac{\partial^2 F}{\partial Q(k)^2} = \sum_{d=1}^{N} \left(\frac{P(k|X_d)(N - Q(k)) - Q(k)(1 - P(k|X_d))}{Q(k)(N - Q(k))} \right)^2$$

$$= \frac{\sum_{d=1}^{N} [NP(k|X_d) - Q(k)]^2}{[Q(k)(N - Q(k))]^2}$$

And finally, multiplying out the numerator and moving all terms which do not depend upon X_d out of the summation gives:

$$= \frac{\sum_{d=1}^{N} [N^2 P(k|X_d)^2 - 2NQ(k)P(k|X_d) + Q(k)^2]}{[Q(k)(N - Q(k))]^2}$$

$$= \frac{[NQ(k)^2] + [N^2 \sum_{d=1}^{N} P(k|X_d)^2] - [2NQ(k) \sum_{d=1}^{N} P(k|X_d)]}{[Q(k)(N - Q(k))]^2}$$

$$= \frac{[NQ(k)^2] + [N^2 \sum_{d=1}^{N} P(k|X_d)^2] - [2NQ(k)^2]}{[Q(k)(N - Q(k))]^2}$$

$$= \frac{[N^2 \sum_{d=1}^{N} P(k|X_d)^2] - [NQ(k)^2]}{[Q(k)(N - Q(k))]^2}$$

Now, from the definition of the CRB an estimate of the variance of $Q(k)$ can be given by:

$$var(Q(k)) \approx \frac{[Q(k)(N - Q(k))]^2}{[N^2 \sum_{d=1}^{N} P(k|X_d)^2] - [NQ(k)^2]}$$

This estimate will, in general, give a lower variance than the extended version derived previously. The differences between the two with varying quantities of classes is shown in the next section along with a discussion of appropriate usage.

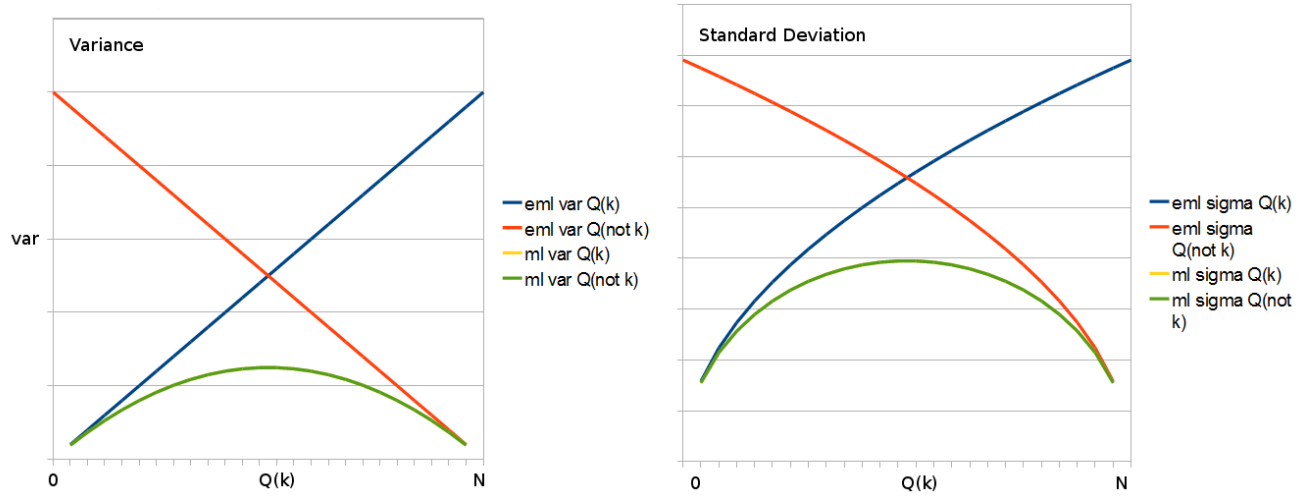

Figure 1: Left: Unambiguous cases for ML and EML variance of Q(k) as Q(k) increases from 0 to N. Right: Standard deviation.

4 Discussion

Both the Extended Maximum Likelihood and Maximum Likelihood variance estimates contain a $\sum_d P(k|X_d)^2$ term. If the density estimate for class k is unambiguous, i.e. there is no overlap between that and other classes, then for any sample, X, it must be the case that $P(k|X)$ is equal to exactly 1 or 0. As the sum is over all data, in unambiguous cases $\sum_d P(k|X_d)^2 = Q(k)$. In such cases the variance estimates become:

$$var_{eml}(Q(k)) \approx \frac{Q(k)^2}{Q(k)}$$

$$var_{ml}(Q(k)) \approx \frac{[Q(k)(N - Q(k))]^2}{[N^2 Q(k)] - [NQ(k)^2]}$$

Curves for both of these are shown in Figure ¡FIG¿ for classes k and $\bar{k}$, along with the corresponding standard deviations which would be the basis of any confidence intervals. It can be seen there is an exact correlation between the variances of k and $\bar{k}$, with EML giving an inverse relationship and ML giving a direct relationship. Intuitively, the shapes of these curves follow the assumptions of a fixed or Poisson distributed total quantity of data.

If N is a Poisson distributed quantity, as is assumed in EML, then $var(N) = N$, as is the case for all Poisson distributions. It follows that when the data is randomly divided into classes the quantities of each class should also be Poisson distributed with $var(Q(k)) = Q(k)$, which is exactly what is found in unambiguous cases:

$$\frac{Q(k)^2}{Q(k)} = Q(k)$$

On the other hand, if N is fixed as is the case with ML, when the data is divided into classes the quantities of each class should be binomially distributed with $var(Q) = NP(1 - P)$ where P is the probability of class k, i.e. $P = Q(k)/N$. This can be shown to be the case by first formulating the standard result for the variance of a binomial in terms of $Q(k)$:

$$NP(1 - P) = N\frac{Q(k)}{N}(1 - \frac{Q(k)}{N}) = \frac{Q(k)(N - Q(k))}{N}$$

Which is exactly the same as the unambiguous variance estimate:

$$\frac{[Q(k)(N - Q(k))]^2}{[N^2 Q(k)] - [NQ(k)^2]} = \frac{Q(k)^2(N - Q(k))^2}{NQ(k)(N - Q(k))} = \frac{Q(k)(N - Q(k))}{N}$$

In real data ambiguity is to be expected. As $P(k|X)$ can not be larger than 1 it follows that in real data $\sum_d P(k|X_d)^2 \leq Q(k)$. The effect of this is to increase the variance, making the Poisson and Binomial variances the

best-case accuracies for Bayesian priors determined via Expectation Maximisation given the assumptions made by the EM cost function interpretations presented here.

A subtle issue surrounds the additional variance resulting from ambiguous data. To appreciate the subtlety it is necessary to fully understand the true meaning of the Poisson or Binomial best-cases. The best-case variances assume the events leading to a data point being generated by a particular class of data generation process is fundamentally driven by either a Poisson or Binomial process. If they are not, e.g. if the data is generated by a deterministic process with no variance, then these contributions are invalid. The additional variance from ambiguous data is something different - it is the ability to count what is there. The additional variance can be viewed as the ability to separate out what is actually in data, whereas the Poisson or Binomial components relate to what could have been in the data. Such considerations may be important when creating Monte-Carlo simulations, as the Poisson or Binomial contributions may have to be artificially added to ground truths on top of the randomness of simulated testing data.

All of the above raises the question: what is the 'correct' interpretation for EM? In an image analysis context the number of data points will typically be fixed by the number of pixels within a region under analysis, and so N may be known a-priori. In this case it is more appropriate to use a correctly normalised ML interpretation for the purposes of variance estimation. However, there may be instances when data points are not immediately related to the number of pixels. For instance, if data points are taken only at strong edges then there may be no prior knowledge of N making an EML interpretation more appropriate. Even if edges are determined ahead of time, counted, then held fixed during the application of an EM algorithm, the true number of edges may not be known due to noise in the original edge detection stage. If in doubt, the EML interpretation will give overestimated variances, making it perhaps a safer option.

In previous work (Tina memo 2000-007) we estimated the proportion of cranial fluid in a number of regions across the skull using an EM based segmentation algorithm. These measurements were found empirically to behave as Poisson samples. This can be explained here as due to inherent binomial statistical behaviour approximating the Poisson case for for an un-ambiguous class and low class proportions ($\leq 10\%$). In this work it was therefore legitimate to re-represent these measurements via an equal variance (square-root) transform, prior to construction of a nearest neighbour classifier, which employed a Euclidean distance metric. This earlier work therefore provides a practical use of the analysis presented in this document.

5 Conclusions

A prior term in Bayes Theorem can simply be the raw frequency with which a particular class of data point occurs in a data set under analysis. The maximum likelihood estimates of these priors can be determined by minimising the Expectation Maximisation cost function. If the total number of data points in a data set is not known a-priori, being Poisson distributed, then the EM cost function should be interpreted as an Extended Maximum Likelihood problem, whereas if the number of data points is fixed then it should be interpreted as a properly normalised Maximum Likelihood problem. The variance on the prior terms in either case can be estimated using the Cramer Rao Lower Variance Bound, which gives a larger variance estimate for EML. In the best case when data points are unambiguous an EML interpretation gives an accuracy with variance equal to a Poisson distribution. In the best case an ML interpretation gives an accuracy with variance equal to a Binomial distribution for a two class problem (k and $\bar{k}$) which should expand to a Multinomial distribution for multiple classes. The choice of interpretation will be down to an application's needs. For example, if the features being counted are from a Poisson source, perhaps counting impact craters on a planetary surface, then EML may be appropriate. On the other hand, if there is a fixed number of pixels in a binary image segmentation problem then ML may be better. The best case variances should be seen as Bayes Theorem's ability to count what could have been in a dataset, whereas the additional variation introduced due to ambiguous data should be seen as the ability to count what actually is found in a dataset. If a data generation process does not conform to either a Poisson or Binomial distribution then careful consideration must be given to the meaning of the variance estimations. In the case of deterministic, i.e. no variance, data generation it may be appropriate to either artificially insert additional uncertainty, or to subtract off the Poisson / Binomial contributions.