Tina Memo No. 2007-010
Internal.

Parameter Estimation for EM in the Presence of Noise.

N. A. Thacker and P.A.Bromiley.

Last updated
21 / 8 / 2007

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

Parameter Estimation for EM in the Presence of Measurement Noise.

N.A.Thacker, P.A.Bromiley. 1/12/2006

Abstract

This document provides a short analysis of the extensions required in order to take account of measurement noise in an EM density analysis for Gaussian mixtures. This is done as an exercise which demonstrates the more general issue of the assumption that measurement noise can be neglected in the construction of pattern recognition systems. These observations have particular relevance for recent techniques which attempt non-linear characterisation of data.

Introduction

The problem of constructing artificial intelligent systems is characterised by the need to build systems which learn. This is conventionally done by using sample data sets to parameterise the dependencies or correlations present in multi-dimensional data in the form of probability densities, and then applying probability theory in order to define and evaluate specific hypotheses. These techniques can be used for a variety of tasks, including noise filtering, prediction and classification (interpretation).

With the exception of factor analysis, the conventional treatment of distribution fitting for pattern recognition assumes that it is unnecessary to have to consider the underlying noise process which generated the data set. Instead, the combined effects of noise and signal variation are treated as inseparable sources of data variation, with the combined variation giving rise to the observed distribution. This makes it possible to consider analysis of data sets without regard to the underlying measurement processes which generated them. This is often justified on the basis that noise is negligible in comparison to the variation induced in the data by the signal process which we are trying to model.

There are several established techniques which allow us to efficiently estimate the parameters of distributions confined to linear subspaces, such as PCA. These methods can be derived from conventional statistical methods on the basis of assumptions regarding the intrinsic homogeneity of noise in multi-dimensional data. However, for many years the nonlinear character of data distributions from the real world have led researchers to seek methods which can determine the parameters of complex distributions, without, it must be said, very much success. Recently, there have been several approaches (Kernal PCA, Latent variable models, support vector machines) which have at their heart a non-linear transformation which converts non-linear distributions into linear ones so that the well known solutions can be exploited. Unfortunately, the process which linearises the distributions also modifies the local noise properties.

Some might argue that this problem simply amounts to assuming that the noise is homogenous on the new transformed variables, rather than the original data. However, it would be fortunate in the extreme to discover that the non-linear mapping (either explicit or implicit as in kernal methods) which transforms our initial data space to a linear-subspace also transforms the initial measurement errors so they are homogenous. As a consequence, methods based upon an assumption of homogenous errors will no longer be valid tools for analysis. It is for this reason that the successful application of these techniques cannot be predicted in advance. Every new dataset requires experimental evaluation and comparisons between methods. We can have no confidence that performance is limited only by intrinsic information content of the data. In order to grasp this problem more fully it is necessary to understand the origins of the "negligible noise" assumption. This can be done by deriving conventional methods to see at which point and in what ways the effects of noise are assumed unimportant.

The standard processes for estimating the parameters of a mixture model using EM (Expectation Maximisation) do not take account of the noise associated with the data. In the sections which follow we attempt to consider the modifications needed to take explicit account of the noise process so that we can stabilise density estimation in the presence of large variable errors. We find that update equations can be derived for this case, but must be iterated within the Maximisation step in order to obtain the relevant Likelihood solution. Also, we show that the standard update equations are better considered as consistent with an assumption, not of negligible measurement noise, but the more general requirement of uniform measurement noise.

EM Based Likelihood

We will start from the standard result that the convergence of the EM algorithm to the minimum of the combined likelihood;

$$L \;=\; -\sum_{i}^{N} log[\sum_{m}^{M} P(m)\lambda_m(x_i)]$$

is guaranteed for a mixture ($P(m)$ weighted sum) of density distributions $\lambda_m(x) \;=\; p(x|\theta_m)$ provided that the Maximisation step minimises a probability weighted Likelihood for each mixture component [1]

$$L_m \;=\; -\sum_{i}^{N} P(m|x_i)\; log(\lambda_m(x_i))$$

where is $P(m|x_i)$ is the probability that data i was generated by the mth component and takes a fixed value in each Maximisation step. The presence of this (constant) probability in the sum has the result that any parameter estimate, derived by differentiating the Likelihood and setting to zero, will produce a solution in which any sums over the data (i) will always include this weighting factor.

For a 1D Gaussian distribution we can include the effects of a noisy measurement process as follows;

$$\lambda_m(x_i) \;=\; \frac{1}{\sqrt{2\pi(v_m \;+\; \eta_i)}}\; exp - (x_i \;-\; \mu_m)^2/2(v_m \;+\; \eta_i)$$

where v is the intrinsic variance of the distribution and η_i is the variance (measurement error) associated with observation x_i.

The normalisation factors $P(m)$ are computed at the Maximisation step via (see below);

$$P(m) \;=\; \sum_{i}^{N} P(m|x_i)$$

Mean Estimation

The update function for μ_m can be derived from the Likelihood, but can be constructed more quickly using the argument in the previous section by inserting the probability weighting term into a variance weighted estimate of the mean.

$$\mu_m \;=\; \sum_{i}^{N} \frac{P(m|x_i)x_i}{v_m \;+\; \eta_i} / \sum_{i}^{N} \frac{P(m|x_i)}{v_m \;+\; \eta_i}$$

For the case of uniform measurement noise the variance terms cancel leaving the conventional update equation. Notice however, in the absence of this, estimation of μ_m requires knowledge of not only η_i but also v_m.

Variance Estimation

The update function for the variance v_m is not quite so obvious. Substituting into L and differentiating with respect to v_m we get

$$\frac{\partial L_m}{\partial v_m} \;=\; \frac{1}{2}\sum_{i}^{N} P(m|x_i)\frac{1}{(v_m \;+\; \eta_i)} \;-\; \frac{1}{2}\sum_{i}^{N} P(m|x_i)\frac{(x_i \;-\; \mu_m)^2}{(v_m \;+\; \eta_i)^2}$$

so that the minimum of the Likelihood is obtained when

$$\sum_{i}^{N} P(m|x_i)\frac{1}{(v_m \;+\; \eta_i)} \;=\; \sum_{i}^{N} P(m|x_i)\frac{(x_i \;-\; \mu_m)^2}{(v_m \;+\; \eta_i)^2} \qquad (1)$$

[1] This is consistent with the Extended Maximum Likelihood derivation for Likelihood, provided we enforce an integral normalisation of λ.

If η_i is constant and equal to η for all data i then

$$(v_m \ + \ \eta) \ = \ \frac{\sum_i^N P(m|x_i)(x_i \ - \ \mu_m)^2}{\sum_i^N P(m|x_i)}$$

which is the conventional EM mixture model update equation for the variance.

For variable η_i, we cannot re-write (1) to solve directly for v_m. However, we can use (1) to generate an iterative update formula for v at time $t+1$ from an estimate at time t

$$v_{m(t+1)} \ = \ \sum_i^N \frac{P(m|x_i)[(x_i \ - \ \mu_m)^2 \ - \ \eta_i]}{(v_{mt} \ + \ \eta_i)^2} / \sum_i^N \frac{P(m|x_i)}{(v_{mt} \ + \ \eta_i)^2}$$

One way to justify the form of this expression is by noting that the error (SD) on an estimated variance is proportional to that variance. This update expression, in analogy to the update function for the mean, can therefore be considered as a probability and variance ($(v_{mt} \ + \ \eta_i)^2$)) weighted estimate.

Notice, the estimation of v_m requires knowledge of μ_m. The update process for these parameters must therefore be iterated for fixed $P(m|x_i)$ in order to converge on the value which minimises the probability weighted Likelihood.

Other Considerations

This document has considered only the effects of measurement noise on parameter estimation within an EM mixture model. Clearly however, once the data variability has been analysed in terms of separate components there is no reason why this logic should not extend to the Expectation stage of EM. In particular, the probability $P(m|x_i)$ used in these calculations, is measurement error dependant due to λ.

$$P(m|x_i) \ = \ \frac{P(m)\lambda_m(x_i)}{\sum_l^M P(l)\lambda_l(x_i)}$$

So that poorly localised data will be ambiguously attributed between particular model components in comparison to a well localised data. This property is clearly missing from standard algorithms, though it is exactly what we might expect to require of any valid solution. The measurement based approach would therefore seem to open up the possibility of building models for, and interpreting, input data with varying levels of information. It is not difficult to imagine situations in which this would be useful. For example, such data would arise in computer vision tasks due to poor illumination, (for simple grey level based representations), or uninformative orientations, (for geometrical based representations). Such examples can never be disregarded as they can happen at any time due to quite reasonable and expected imaging circumstances.

Conclusions

Although researchers often apply pattern recognition systems in a way which makes the implicit assumption that we do not need to know the measurement noise, this stance is not consistent with the requirement that any effects of noise are negligible in comparison to other sources of variability. How can we know this is true unless we take the trouble to find out?

The above analysis illustrates that the conventional assumption of negligible noise, illustrated here in the context of fitting density distributions using EM algorithms, is probably better considered as an assumption of uniform noise. It is this property which allows exact cancellation of the terms and not numerically or proportionally small values. However, when applying non-linear functions to data spaces, initially homogenous noise characteristics (as often found in real data), will take on spatially varying characteristics including elongation and associated orientation. If there are no constraints on the specific forms of non-linearity applied, it would be quite easy to construct data vectors for which the expected noise process is neither negligible nor homogenous. Indeed, the tasks we undertake in many pattern recognition problems will often guarantee this to be the case. For example; a key feature of efficient learning systems is the need for invariant representations, which eliminate the variations produced by physical changes in the world leaving only that information pertinent to the required analysis. Yet transformations which aim to eliminate variation by computing invariant quantiles must be expected, if successful, to generate data for which the noise is ultimately the dominant source of remaining variation. Moreover, as invariant transformations are always a many to one mapping, the amount of noise observed at any location in the transformed space will vary

across the data set, invalidating even the more general assumutmption of local consistency, unless the measurement process (which generates the data) is designed to avoid it.

The extension of our analysis to multi-dimensions is non-trivial, a summary of the problems involved may be added to later versions of this document. For uncorrelated data components however, the above is sufficient to illustrate the consequence of varying noise characteristics due to applying a non-linear mapping. This would change the relative significance of each sample of data in the estimation of the key distribution parameters. In addition, subsequent classification processes (estimates of $P(m|x_i)$) will not be computed correctly.

In the light of this observation, the current set of non-linear analysis tools available to those wishing to perform pattern recognition must, despite their mathematical complexity, be considered theoretically lacking. While some may wish to brush such observations aside with the claim that "these methods seem to work", we can no-longer guarantee valid solutions nor should anyone claim theoretical legitimacy for these techniques. The consequence, as we have already said, is that there is then a certain amount of luck involved in finding problems for which the output results are useful. This must be considered as a poor starting point for any theory of learning. I would have included some references if I had ever seen a publication which corroborates the conclusions of this document.