Tina Memo No. 2001-005
Internal Report.

Model Selection and Convergence of the EM Algorithm.

N.A. Thacker, M. Pokrić and A.J.Lacey.

Last updated
30 / 10 / 2004

This document forms part of the **Statistics and Segmentation** Series (2008-001) available from www.tina-vision.net.

2007-008	Tutorial: Defining Probability for Science.
2001-007	Performance Characterisation in Computer Vision: The Role of Statistics in Testing and Design.
2002-007	The Effects of an Arcsin Square Root Transform on a Binomial Distributed Quantity.
2001-010	The Effects of a Square Root Transform on a Poisson Distributed Quantity.
2004-004	Shannon Entropy, Renyi Entropy, and Information.
2002-002	Validating MRI Field Homogeneity Correction Using Image Information Measures.
2004-001	Empirical Validation of Covariance Estimates for Mutual Information Coregistration.
2004-005	The Equal Variance Domain: Issues Surrounding the Use of Probability Densities in Algorithm Design.
2009-008	Avoiding Zero and Infinity in Sample Based Algorithms.
2001-008	Derivation of the Renormalisation Formula for the Product of Uniform Probability Distributions and Extension to Non-Integer Dimensionality.
2001-005	Model Selection and Convergence of the EM Algorithm.
2003-007	Noise Filtering and Testing for MR Using a Multi-Dimensional Partial Volume Model.
2002-004	A Novel Method for Non-Parametric Image Subtraction: Identification of Enhancing Lesions in Multiple Sclerosis from MR Images.
2001-014	Bayesian and Non-Bayesian Probabilistic Models for Image Analysis.
1997-001	The Bhattacharyya Metric as an Absolute Similarity Measure for Frequency Coded Data.
1999-001	The Bhattacharyya Measure requires no Bias Correction.
1999-004	B-Fitting: An Estimation Technique With Automatic Parameter Selection.
2005-008	Tutorial: Beyond Likelihood.

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

Model Selection and Convergence of the EM Algorithm

N. Thacker, M. Pokrić, A.J.Lacey,
Imaging Science and Biomedical Engineering, University of Manchester, UK
email: neil.thacker@man.ac.uk

Abstract

This document explains the proof of convergence for the EM algorithm. It presents a derivation based upon the 'classic' approach found in many texts [2, 1], formulating the problem as a likelihood maximisation. This is then evaluated to make explicit what is being optimised and the class of problems for which it is appropriate. We explain that strictly we should not expect this method to be capable of solving the model selection problem. We go on to show how the method can be applied to tissue labelling in medical image analysis tasks in the form of mixture modelling.

Introduction

The EM or Expectation-Maximisation algorithm is a maximum likelihood based function optimisation strategy applied to semi-parametric problems, in particular those methods based on *mixture models*. It is formulated as a procedure for updating the model parameters for each of the component models, in a direction which, it is claimed, is guaranteed to increase the log-likelihood of the data given the parameters.

Definitions

It is particularly important when working with probabilities to define clearly our notation, as often even relatively trivial algebraic manipulations lead to confusion due to subtleties in the definitions.

Expression	Meaning
$\mathbf{x} = \{x_1, x_2, ..., x_n\}$	a set of n observed (continuous) measurements.
$\mathbf{Z} = \{Z_1, Z_2, ..., Z_j\}$	a set of j unobserved discrete variables.
$\theta = \{\alpha_1, \alpha_2, ..., \alpha_m\}$	the set of m model parameters.
$P(\mathbf{x}) = p(\mathbf{x})\Delta\mathbf{x}$	the prior probability of the data being within an interval $\Delta\mathbf{x}$.
$P(\theta) = p(\theta)\Delta\theta$	the prior probability of the model parameters within an interval $\Delta\theta$.
$P(\mathbf{x}\|\theta) = p(\mathbf{x}\|\theta)\,\Delta\mathbf{x}$	the conditional probability of the data being within the interval $\Delta\mathbf{x}$.
$P(\mathbf{x}, \theta)$	the joint probability of the data and the model parameters.
$L(\theta) = ln(P(\mathbf{x}\|\theta))$	the log-likelihood of a particular parameter set.

So what should the EM algorithm buy me?

A problem typical of the use of the EM algorithm is the labelling of image data (voxels or pixels), for medical data analysis these labels may be tissues. In order to label each of the pixels with a particular tissue we need to be able to calculate the probability that a particular grey-level is representative of a particular tissue [3]. That is, we want to know the probability that a particular model (tissue) θ_j is responsible for producing a particular measurement x_i (grey-level), $P(\theta_j|x_i)$. To be able to do this we need, for each of the tissues, a model of how likely each of the grey-levels would have been been *caused* by that tissue. These models will likely take the form of probability distributions such as probability density functions (pdf).

Although the basis form of these models can be chosen *a priori* (such as a Gaussian) the model parameters will depend on the data. The task is to modify the parameters of each model so that collectively they explain all of the data. It is necessary to use the data itself to parameterise the distributions, though it is not known a-priori how data is associated with each part of the model. This is the kind of problem the EM algorithm is able to tackle.

The question we need to ask ourselves is; What assumptions are made in order to achieve this? Knowledge of the answer to this question allows us to form conclusions regarding the kinds of problem which can be addressed using this technique. For example; Can EM solve the model selection problem? or; Is EM based upon Bayes theory?

Tissue labelling is an example of a problem where multiple distributions are used because it is known that multiple processes are involved; the three (or more) tissue types. The procedure is also valid in the case where a collection of models (distributions) are used to approximate another, perhaps more complex model. The classification labels can be, and often are, used as a form of missing data. The EM algorithm is specifically derived to deal with problems where there is *incomplete* data and the discussion at the end of this document attends to whether we believe the model selection and missing data problems can be reconciled in this way.

The 'Classic' Derivation

This section covers the derivation of the EM algorithm in a form which is related to that found in the majority of texts, though we have made an effort here to distinguish between probabilities 'P' (which obey the laws of probability) and probability densities 'p' (which do not). The intention here is to gain some insight into appropiate use of the EM algorithm, and in particular to understand the form of the statistical optimisation measure.

Finding the optimal decription of a set of data $\mathbf{x}$ using a model θ is the same as maximising the joint probability;

$$Maximise\ \left(P\left(\mathbf{x},\theta\right)\right)$$

In this case the goal of the EM algorithm is generally expressed as;

$$Maximise\ \left(L\left(\theta\right)\ =\ ln\ P\left(\mathbf{x}|\theta\right)\right)$$

where $P\left(\mathbf{x}|\theta\right)$ is directly computed from the density distribution which we intend to use to descirbe the data sample. Which (in common with all likelihood based methods) is equivalent to the joint probability on the assumption that the prior probabilities of each possible model $P(\theta)$ are all equal[1].

In order to optimise this expression in an iterative algorithm the aim is to find a model parameter set θ which has a greater likelihood than the current parameter set, θ'; a new estimate of the parameters which is more likely to have caused the data than the current set. Substituting in the definitions of likelihood from above;

$$L\left(\theta\right)-L\left(\theta'\right)\ =\ lnP\left(\mathbf{x}|\theta\right)-lnP\left(\mathbf{x}|\theta'\right)\ =\ ln\frac{p\left(\mathbf{x}|\theta\right)}{p\left(\mathbf{x}|\theta'\right)}\geq 0 \tag{1}$$

This states nothing more than the update of the parameters at each step should increase our likelihood function $L(\theta)$. In fact the correct statistical interpretation involves Extended Maximum Likelihood (EML) which can be explicitly derived for Poisson samples. This involves another term which cancels in the ratio of probability densities in this first step of the proof of convergence[2]. The full Likelihood is given later in the section describing application to mixture models.

The whole point of the EM approach is that if the density function can be constructed in a particular way, such that

$$p\left(\mathbf{x}|\theta\right)\ =\ \sum_{j}^{J}p\left(\mathbf{x}|Z_j,\theta\right)P\left(Z_j|\theta\right) \tag{2}$$

then the likelihood can be optimised using iterative use of the EM algorithm.

$$\theta\ =\arg\max_{\theta}\ \sum_{j}^{J}\left(P(Z_j|\mathbf{x},\theta')ln(p(\mathbf{x}|Z_j,\theta)P(Z_j|\theta))\right) \tag{3}$$

[1] It is debatable whether parameter estimation needs to be defined in the first place as a joint probability, as Likelihood has well defined quantitative statisticalproperties, independent of this assumed origin.

[2] see: Tina memo. 2004-006 .

The interpretation of this formulation (equation 2) should be that we are attempting to account for the data in terms of a set of J mutually exclusive events Z_j where the expected number of such events is $P(Z_j|\theta)$. For example; in the case of tissue labelling these unseen variables could be regarded as a particular tissue type, i.e. grey matter, white matter or CSF. If properly constructed, running the EM algorithm would result in a functional description of the form of the distribution and also a tissue classification for each measurement. In system terms this process can be regarded as decoupling the driving function (the model) from the measurement process (the image) by inserting a classification layer.

We can now write equation 1 as;

$$L(\theta) - L(\theta') = ln\frac{\sum_j p(\mathbf{x}|Z_j,\theta)\,P(Z_j|\theta)}{p(\mathbf{x}|\theta')} \geq 0 \tag{4}$$

Proof that the EM algorithm satisfies this inequality can be achieved using *Jensen's inequality* which is expressed as;

$$ln\sum_j \lambda_j y_j \geq \sum_j \lambda_j ln y_j \qquad \text{such that } \lambda_j \geq 0 \text{ and } \sum_j \lambda_j = 1$$

In order to apply this relationship we need to organise equation 4 into the correct form, identifying suitable λ coefficients.

Note: In the classic derivation the posterior probability of Z_j, $P(Z_j|\mathbf{x},\theta')$ is used as λ. In the MR tissue example $P(Z_j|\mathbf{x},\theta')$ represents the probability that a pixel was located within a particular tissue. The tissue classes must be an all inclusive group (there is no other choice) and thus the probabilities for each class must sum to unity, i.e.;

$$\sum_j P(Z_j|\mathbf{x},\theta') = 1 \tag{5}$$

Also as a probability $P(Z_j|\mathbf{x},\theta') \geq 0$. Therefore, it does fulfil Jensen's criteria. However, Is $P(Z_j|\mathbf{x},\theta')$ the only choice here? What are the implications if an alternative substitution is made?

We will not make such a definite substitution for λ at this point, instead we will insert a λ ratio thus;

$$L(\theta) - L(\theta') = ln\sum_j \left(\frac{p(\mathbf{x}|Z_j,\theta)\,P(Z_j|\theta)}{p(\mathbf{x}|\theta')}\frac{\lambda_j}{\lambda_j}\right)$$

In this form we can now apply Jensen's inequality;

$$L(\theta) - L(\theta') \geq \sum_j \left(\lambda_j ln\frac{p(\mathbf{x}|Z_j,\theta)P(Z_j|\theta)}{p(\mathbf{x}|\theta')\lambda_j}\right)$$

Which can be re-written to allow a bound on $L(\theta)$ to be defined thus;;

$$L(\theta) \geq L(\theta') + \sum_j \left(\lambda_j ln\frac{p(\mathbf{x}|Z_j,\theta)P(Z_j|\theta)}{p(\mathbf{x}|\theta')\lambda_j}\right)$$

To simplify things we shall identify a Δ term thus;

$$L(\theta) \geq L(\theta') + \Delta(\theta,\theta') \tag{6}$$

$$\text{where } \Delta(\theta,\theta') = \sum_j \left(\lambda_j ln\frac{p(\mathbf{x}|Z_j,\theta)P(Z_j|\theta)}{p(\mathbf{x}|\theta')\lambda_j}\right) \tag{7}$$

Now in order to ensure that $L(\theta)$ is increased it is clear that the Δ term must be positive.

Expanding Δ gives;

$$\Delta(\theta,\theta') \;=\; \sum_j (\lambda_j ln(p(\mathbf{x}|Z_j,\theta)P(Z_j|\theta))) - \sum_j (\lambda_j ln(p(\mathbf{x}|\theta')\lambda_j))$$

If we assume λ_j is independent of θ then the second term in this equation is fixed relative to θ as θ' is our current, known estimate of the model parameters. It is also always either zero or negative in magnitude ($ln(x) \leq 0$ when $0 \geq x \leq 1$). For similar reasons the first term is always negative;

$$\Delta(\theta,\theta') \;=\; \underbrace{\sum_j (\lambda_j ln(p(\mathbf{x}|Z_j,\theta)P(Z_j|\theta)))}_{-ve} - \underbrace{\sum_j (\lambda_j ln(p(\mathbf{x}|\theta')\lambda_j))}_{-ve} \tag{8}$$

Thus if we are to ensure that the Δ term is positive, which we must if we are to increase the likelihood, the magnitude of the first term must be made as small as possible. Given that this term is always negative this is the same as finding the maximum value. This implies that we choose θ on the basis of;

$$\theta \;= \arg\max_\theta \sum_j (\lambda_j ln(p(\mathbf{x}|Z_j,\theta)P(Z_j|\theta))) \tag{9}$$

It is not yet obvious that by maximising equation 9 it can be guaranteed that the likelihood $L(\theta)$ is also maximised. In particular, how can it be claimed that finding the maximum of equation 9 *ensures* that $L(\theta) \;\geq\; L(\theta')$ when the maximum of equation 9 is independent of the second term in equation 8? As we are maximising Δ with respect to θ one way to ensure this would be if $\Delta(\theta',\theta') \;=\; 0$. From this we can identify a value for λ.

$$\Delta(\theta',\theta') \;=\; \sum_j \left(\lambda_j ln \frac{p(\mathbf{x}|Z_j,\theta')P(Z_j|\theta')}{p(\mathbf{x}|\theta')\lambda_j}\right)$$

if $\Delta(\theta',\theta') \;=\; 0$ then;

$$p(\mathbf{x}|Z_j,\theta')P(Z_j|\theta') \;=\; p(\mathbf{x}|\theta')\lambda_j$$

which may be rearranged as;

$$\frac{p(\mathbf{x},Z_j|\theta')}{p(\mathbf{x}|\theta')} = \;\lambda_j$$

$$\lambda_j \;=\; P(Z_j|\mathbf{x},\theta')$$

Thus if $\theta = \theta'$ it follows that $\Delta(\theta',\theta') = 0$. In an iterative algorithm θ' is available as a potential solution to θ. Therefore, $\Delta(\theta,\theta')$ should at least be zero and with this specific choice for λ_j $L(\theta) = L(\theta')$ becomes a lower bound on $L(\theta)$.

Thus we have;

$$\theta \;= \arg\max_\theta \sum_j (P(Z_j|\mathbf{x},\theta')ln(p(\mathbf{x}|Z_j,\theta)P(Z_j|\theta)))$$

which is the same as equation 3 (QED).

Equation 3 may be simplified further using $p(\mathbf{x}|Z_j,\theta)P(Z_j|\theta) \;=\; p(\mathbf{x},Z_j|\theta)$ and equating over the expected value of $\mathbf{Z}$ thus;

$$\theta \;= \arg\max_\theta \; E_{\mathbf{Z}|\mathbf{x},\theta'} p(\mathbf{x},\mathbf{Z}|\theta) \tag{10}$$

Where $E_{\mathbf{Z}|\mathbf{x},\theta}$ represents the expectation of $\mathbf{Z}$ given $\mathbf{x}$ and θ'.

Mixture Modelling

For a Poisson sample of independant data points x_i we can write the extended maximum likelihood (EML) as;

$$L(\theta) \;=\; \sum_i^n ln \sum_j p\left(x_i|Z_j,\theta\right) P\left(Z_j|\theta\right) \;-\; k \int \sum_j p\left(x_i|Z_j,\theta\right) P\left(Z_j|\theta\right)\; dx \tag{11}$$

where k is an unknown arbitrary constant which relates probability densities to probabilities and the last term is simply the integrated density of the model.

$P(Z_j|\theta)$ (the density normalisation) is effectively a prior term for each model component which can be optimised independently (due to the EML form of the likelihood) giving the likelihood estimate up to the unknown (constant) scale factor.

$$P(Z_j|\theta) \;=\; \frac{1}{k}\sum_i^n P(Z_j|x_i,\theta') \qquad s.t. \qquad \sum_j P(Z_j|\theta) \;=\; n/k$$

Updating the model parameters so as to fix the total density normalisation at a value equal to the sample size n (the EML likelihood estimate) allows us to otherwise ignore the last term in equation (11) during optimisation.

Based upon 3, all other terms can therefore be computed as though we were optimising;

$$\theta \;= \arg\max_\theta \sum_i^n \sum_j P(Z_j|x_i,\theta') ln(p(x_i|Z_j,\theta)P(Z_j|\theta)) \tag{12}$$

where $(p(x_i|Z_j,\theta)$ is the probability of obtaining x_i from the Z_j component of the underlying data generation process.

To use this in an iterative scheme requires two steps; the Expectation of $P(Z_j|\mathbf{x},\theta')$ followed by a Maximisation of $p(\mathbf{x}|Z_j,\theta)P(Z_j|\theta)$ over θ and Z.

The first term in equation 12 can be computed from Bayes theory for the correct model as the Expectation step;

$$P(Z_j|x_i,\theta') \;=\; \frac{p(x_i|Z_j,\theta')P(Z_j|\theta')}{\sum_k^J \left(p(x_i|Z_k,\theta')P(Z_k|\theta')\right)} \tag{13}$$

prior to estimation of the paramters.

The above form of the optimisation function makes Maximisation possible by optimising in turn each mixture component of the model θ_j.

$$\theta_j \;= \arg\max_{\theta_j} \sum_i^n P(Z_j|x_i,\theta') ln(p(x_i|Z_j,\theta_j)P(Z_j|\theta_j)) \tag{14}$$

Pure scaling on the probability density of this part of the model by $P(Z_j|\theta_j)$ has no effect on the estimates of other mixture component parameters θ_j. So this is now equivalent to;

$$\theta_j \;= \arg\max_{\theta_j} \sum_i^n (P(Z_j|x_i,\theta') ln(p(x_i|Z_j,\theta_j)) \tag{15}$$

Which simply results in a probability weighted version of the standard parameter estimates for each term in the model.

For example the mean value of a density component x_m

$$x_m \;=\; \frac{\sum_i^n P(Z_j|x_i,\theta')\; x_i}{\sum_i^n P(Z_j|x_i,\theta')}$$

All other parameters are computed in a similar fashion, for example a sample variance C_m around x_m is given by;

$$C_m \;=\; \frac{\sum_i^n P(Z_j|x_i,\theta')\; (x_i \;-\; x_m)^2}{\sum_i^n P(Z_j|x_i,\theta')}$$

Discussion

The classic derivation for proof of convergence of the EM algorithm begins by defining the goal of the EM algorithm maximising equation 4 which is based on likelihoods (specifically EML). Suggestions that the EM algorithm actually optimise other quantities are spurious [1].

In answer to the questions in the introduction, though aspects of its implimentation require Bayes formula, the definition of the data analysis process could not strictly be described as 'Bayesian'. There is no role for prior probabilties, independant of normalisation terms which are determined using Likelihood.

The use of data density terms in EML derives directly from the conventional definition of likelihood and should not be interpretted as any form of bias correction (in the Akaike sense), which may have made model selection a possibility. A likelihood, by definition, quantifies the expectation of measuring that data given the model. The model is therefore already specified and we would not expect EM to be capable of performing model selection (ie: correctly identifying the required number of parameters in a mixture model). The likelihood function being optimised will simply continue to reduce with the addition of extra model components until there are sufficient parameters in the model to describe the entire distribution and measurement noise. However, as EM is a likelihood based approach, it should be possible to compute covariances on estimated parameters and so perform a Bhattacharyya overlap with the original data PDF in order to confirm the requirement for each model component [4, 5]. Otherwise the method can only be applied in problems where the number of mixtures can be specified a-priori.

Application of the the method to mixture modelling involves splitting the optimisation process into a set of separate stages, one for each part of the model θ_j. The fact that the algorithm will converge when the $P(Z_j|x_i, \theta')$ term is fixed by the previous estimate of parameters θ' is crucial to allowing this process.

The proof of convergence shows that in any problem where the underlying density distribution model can be written as a sum of terms (mixtures) can be optimised using an EM approach. This freedom has been exploited in the probability estimation techniques used on the IERAPSI project in a way which allows partial volume terms to be included in the tissue labelling model [6].

References

1. B.D.Ripley, Appendix A in Pattern Recognition and Neural Networks, Cambridge University Press, 1996.
2. C.M.Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 66 ff, 1995.
3. Laidlaw DH, Fleischer KW, Barr AH, Partial-volume bayesian classification of material mixtures in MR volume data using voxel histograms, IEEE Trans. Med. Imag., vol. 17, no. 1, 74-86, Feb. 1998.
4. 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.
5. N.A.Thacker, D.Prendergast and P.I.Rockett,'B-Fitting: A Statistical Estimation Technique with Automatic Parameter Selection.',Proc, BMVC, Edinburgh, 1996.
6. M. Pokric, N.A. Thacker, M.L.J.Scott and A.Jackson, Multi-Dimesional Medical Image Segmentation with Partial Voluming, MIUA, Birmingham, 77-80, 2001.