A P P E N D I X

3

Expectation maximization

K. Friston

INTRODUCTION

This appendix describes expectation maximization (EM) for linear models using statistical mechanics (Neal and Hinton, 1998). We connect this formulation with classical methods and show the variational free energy is the same as the objective function maximized in restricted maximum likelihood (ReML). In Appendix 4, we show that EM itself is a special case of variational Bayes (Chapter 24).

The EM algorithm is ubiquitous in the sense that many estimation procedures can be formulated as such, from mixture models through to factor analysis. Its objective is to maximize the likelihood of observed data $p(y|\lambda)$, conditional on some hyperparameters, in the presence of unobserved variables or parameters θ. This is equivalent to maximizing the log-likelihood:

$$\begin{aligned} \ln p(y|\lambda) &= \ln \int p(\theta,|\lambda) d\theta \geq \\ F(q,\lambda) &= \int q(\theta) \ln p(\theta, y|\lambda) d\theta - \int q(\theta) \ln q(\theta) d\theta \end{aligned} \qquad \textbf{A3.1}$$

where $q(\theta)$ is *any* density on the model parameters (Neal and Hinton, 1998). Eqn. **A3.1** rests on Jensen's inequality that follows from the concavity of the log function, which renders the log of an integral greater than the integral of the log. F corresponds to the [negative] free energy in statistical thermodynamics and comprises two terms: the energy and entropy. The EM algorithm alternates between maximizing F and, implicitly, the likelihood of the data, with respect to the distribution $q(\theta)$ and the hyperparameters λ, holding the other fixed:

$$\textbf{E}\text{-step: } q(\theta) \leftarrow \max_q F(q(\theta), \lambda)$$

$$\textbf{M}\text{-step: } \lambda \leftarrow \max_\lambda F(q(\theta), \lambda)$$

This iterative alternation performs a coordinate ascent on F. It is easy to show that the maximum in the **E**-step obtains when $q(\theta) = p(\theta|y, \lambda)$, at which point Eqn. **A3.1** becomes an equality. The **M**-step finds the ML estimate of the hyperparameters, i.e. the values of λ that maximize $p(y|\lambda)$ by integrating $\ln p(\theta, y|\lambda) = \ln p(y|\theta, \lambda) + \ln p(\theta|\lambda)$ over the parameters, using the current estimate of their conditional distribution. In short, the **E**-step computes sufficient statistics (in our case the conditional mean and covariance) of the unobserved parameters to enable the **M**-step to optimize the hyperparameters, in a maximum likelihood sense. These new hyperparameters re-enter into the estimation of the conditional density and so on until convergence.

The E-step

For linear models, under Gaussian (i.e. parametric) assumptions, the **E**-step is trivial and corresponds to evaluating the conditional mean and covariance as described in Chapter 22:

$$\begin{aligned} y &= X\theta + \varepsilon \\ \overline{y} &= \begin{bmatrix} y - X\theta \\ \eta_\theta \end{bmatrix} \overline{X} = \begin{bmatrix} X \\ I \end{bmatrix} \overline{C}_\varepsilon = \begin{bmatrix} \sum \lambda_i Q_i & 0 \\ 0 & C_\theta \end{bmatrix} \\ \eta_{\theta|y} &= C_{\theta|y} \overline{X}^T \overline{C}_\varepsilon^{-1} \overline{y} \\ C_{\theta|y} &= (\overline{X}^T \overline{C}_\varepsilon^{-1} \overline{X})^{-1} \end{aligned} \qquad \textbf{A3.2}$$

where the prior and conditional densities are $p(\theta) = N(\eta_\theta, C_\theta)$ and $q(\theta) = N(\eta_{\theta|y}, C_{\theta|y})$. This compact form is a result of absorbing the priors into the errors by augmenting the linear system. As described in Chapter 22, the same augmentation is used to reduce hierarchal models with empirical priors to their non-hierarchical form. Under local linearity assumptions, non-linear models can be reduced to a linear form as described in Chapter 34. The resulting conditional density is used to estimate the hyperparameters of the covariance components in the **M**-step.

Statistical Parametric Mapping, by Karl Friston et al.
ISBN–13: 978-0-12-372560-8 ISBN–10: 0-12-372560-7

The M-step

Given that we can reduce the problem to estimating the error covariances of the augmented system in Eqn. **A3.2**, we only need to estimate the hyperparameters of the error covariances (which contain the prior covariances). Specifically, we require the hyperparameters that maximize the first term of the free energy (i.e. the energy) because the entropy does not depend on the hyperparameters. For linear systems, the free energy is given by (ignoring constants):

$$\log p(\theta, y\,|\lambda) = -\frac{1}{2}\ln|C_\varepsilon| - \frac{1}{2}(\overline{y}-\overline{X}\theta)^T C_\varepsilon^{-1}(\overline{y}-\overline{X}\theta).$$

$$\int q(\theta)\ln p(\theta, y\,|\lambda)\,d\theta = -\frac{1}{2}\ln|C_\varepsilon| - \frac{1}{2}r^T C_\varepsilon^{-1} r - \frac{1}{2}tr\{C_{\theta|y}\overline{X}^T C_\varepsilon^{-1}\overline{X}\}$$

$$\int q(\theta)\log q(\theta) = -\frac{1}{2}\ln|C_{\theta|y}| \qquad \textbf{A3.3}$$

$$F = \frac{1}{2}\ln|C_\varepsilon^{-1}| - \frac{1}{2}r^T C_\varepsilon^{-1} r - \frac{1}{2}tr\{C_{\theta|y}\overline{X}^T C_\varepsilon^{-1}\overline{X}\} + \frac{1}{2}\ln|C_{\theta|y}|$$

where the residuals $r = \overline{y} - \overline{X}\eta_{\theta|y}$. By taking derivatives with respect to the error covariance we get:

$$\frac{\partial F}{\partial C_\varepsilon^{-1}} = \frac{1}{2}C_\varepsilon - \frac{1}{2}rr^T - \frac{1}{2}\overline{X}C_{\theta|y}\overline{X}^T \qquad \textbf{A3.4}$$

When the hyperparameters maximize the free energy this gradient is zero and:

$$C(\lambda)_\varepsilon = rr^T + \overline{X}C_{\theta|y}\overline{X}^T \qquad \textbf{A3.5}$$

(cf. Dempster *et al.*, 1981: 350). This means that the ReML error covariance estimate has two components: that due to differences between the data and its conditional prediction; and another due to the variation of the parameters about their conditional mean, i.e. their conditional uncertainty. This is not a closed form expression for the unknown covariance because the conditional covariance is a function of the hyperparameters. To find the ReML hyperparameters, one usually adopts a Fisher scoring scheme, using the first and expected second partial derivatives of the free energy:

$$\Delta\lambda = -E\left(\frac{\partial^2 F}{\partial\lambda_{ij}^2}\right)^{-1}\frac{\partial F}{\partial\lambda_i}$$

$$\frac{\partial F}{\partial\lambda_i} = tr\left(\frac{\partial F}{\partial C_\varepsilon^{-1}}C_\varepsilon^{-1}Q_i C_\varepsilon^{-1}\right) = -\frac{1}{2}tr\{PQ_i\} + \frac{1}{2}\overline{y}^T P^T Q_i P\overline{y}$$

$$\frac{\partial^2 F}{\partial\lambda_{ij}^2} = \frac{1}{2}tr\{PQ_iPQ_j\} - \overline{y}^T PQ_iPQ_jP\overline{y} \qquad \textbf{A3.6}$$

$$E\left(\frac{\partial^2 F}{\partial\lambda_{ij}^2}\right) = -\frac{1}{2}tr\{PQ_iPQ_j\}$$

$$P = C_\varepsilon^{-1} - C_\varepsilon^{-1}\overline{X}C_{\theta|y}\overline{X}^T C_\varepsilon^{-1}$$

Fisher scoring corresponds to augmenting a Gauss-Newton scheme by replacing the second derivative or curvature with its expectation. The curvature or Hessian is referred to as Fisher's information matrix[1] and encodes the conditional prediction of the hyperparameters. In this sense, the information matrix has a close connection to the degrees of freedom in classical statistics. The gradient can be computed efficiently by capitalizing on any sparsity structure in the constraints and by bracketing the multiplications appropriately. This scheme is general in that it accommodates almost any form for the covariance through a Taylor expansion of $C\{\lambda\}_\varepsilon$.

Once the hyperparameters have been updated they enter the **E**-step as a new error covariance estimate to give new conditional moments which, in turn, enter the **M**-step and so on until convergence. A pseudo-code illustration of the complete algorithm is presented in Figure 22.4 of Chapter 22. Note that in this implementation one is effectively performing a single Fisher scoring iteration for each **M**-step. One could postpone each **E**-step until this search converged, but a single step is sufficient to perform a coordinate ascent on F. Technically, this renders the scheme a generalized EM or GEM algorithm.

It should be noted that the search for the maximum of F does not have to employ Fisher scoring or indeed the parameterization of C_ϵ used above. Other search procedures, such as quasi-Newton searches, are commonly employed (Fahrmeir and Tutz, 1994). Harville (1977) originally considered Newton-Raphson and scoring algorithms, and Laird and Ware (1982) recommend several versions of EM. One limitation of the linear

[1] The derivation of the expression for the information matrix uses standard results from linear algebra and is most easily seen by differentiating the gradient, noting:

$$\frac{\partial P}{\partial\lambda_j} = -PQ_jP$$

and taking the expectation, using

$$E(tr(PQ_iP\overline{y}\overline{y}^T PQ_j)) = tr\{PQ_iPC_\varepsilon PQ_j\} = tr\{PQ_iPQ_j\}$$

hyperparameterization described above is that it does not guarantee that C_ϵ is positive definite. This is because the hyperparameters can take negative values with extreme degrees of non-sphericity. The EM algorithm employed by *multistat* (Worsley *et al.*, 2002) for variance component estimation in multisubject fMRI studies, uses a slower but more stable algorithm that ensures positive definite covariance estimates.

In Appendix 4, we will revisit this issue and look at linear hyperparameterizations of the precision. The common aspect of all these algorithms is that they (explicitly or implicitly) optimize free energy. As shown next, this is equivalent to restricted maximum likelihood.

RELATIONSHIP TO REML

ReML or *restricted maximum likelihood* was introduced by Patterson and Thompson in 1971, for estimating variance components in a way that accounts for the loss in degrees of freedom that result from estimating fixed effects (Harville, 1977), i.e. that accounts for conditional uncertainty about the parameters. It is commonly employed in standard statistical packages (e.g. SPSS). Under the present model assumptions, ReML is formally identical to EM. One can regard ReML as embedding the **E**-step into the **M**-step to provide a single log-likelihood objective function: substituting $C_{\theta|y} = \left(\overline{X}^T C_\varepsilon^{-1} \overline{X}\right)^{-1}$ into the expression for the free energy gives:

$$F = -\frac{1}{2}\ln|C_\varepsilon| - \frac{1}{2} r^T C_\varepsilon^{-1} r - \frac{1}{2}\ln|\overline{X}^T C_\varepsilon^{-1}\overline{X}| \qquad \textbf{A3.7}$$

This is the ReML objective function (see Harville, 1977: 325). Critically, its derivatives with respect to the hyperparameters are exactly the same as those in the **M**-step.[2] Operationally, the **M**-step can be re-formulated to give a ReML scheme by removing any explicit reference to the conditional covariance using:

$$P = C_\varepsilon^{-1} - C_\varepsilon^{-1}\overline{X}(\overline{X}^T C_\varepsilon^{-1}\overline{X})^{-1}\overline{X}^T C_\varepsilon^{-1} \qquad \textbf{A3.8}$$

The resulting scheme is formally identical to that described in Section 5 of Harville (1977). Because one can eliminate the conditional density, one could think of ReML as estimating the hyperparameters in a subspace that is *restricted* in the sense that the estimates are conditionally independent of the parameters. See Harville (1977) for a discussion of expressions, comparable to the terms in Eqn. **A3.7** that are easier to compute, for particular hyperparameterizations of the variance components.

Having established ReML is a special case of **EM**, in Appendix 4, we take an even broader perspective and look at EM as a special case of variational Bayes.

REFERENCES

Dempster AP, Rubin DB, Tsutakawa RK (1981) Estimation in covariance component models. *J Am Stat Assoc* **76**: 341–53

Fahrmeir L, Tutz G (1994) *Multivariate statistical modelling based on generalised linear models*. Springer-Verlag Inc., New York, pp 355–56

Harville DA (1977) Maximum likelihood approaches to variance component estimation and to related problems. *J Am Stat Assoc* **72**: 320–38

Laird NM, Ware JH (1982) Random effects models for longitudinal data. *Biometrics* **38**: 963–74

Neal RM, Hinton GE (1998) A view of the EM algorithm that justifies incremental, sparse and other variants. In *Learning in graphical models*, Jordan MI (ed.). Kluwer Academic Press, Dordrecht, pp 355–68

Worsley KJ, Liao CH, Aston J *et al.* (2002) A general statistical analysis for fMRI data. *NeuroImage* **15**: 1–15

[2] Note

$$\frac{\partial \ln|\overline{X}^T C_\varepsilon^{-1}\overline{X}|}{\partial \lambda_i} = tr\left((\overline{X}^T C_\varepsilon^{-1}\overline{X})^{-1}\frac{\partial \overline{X}^T C_\varepsilon^{-1}\overline{X}}{\partial \lambda_i}\right)$$
$$= -tr\{C_{\theta|y} X^T C_\varepsilon^{-1} Q_i C_\varepsilon^{-1}\overline{X}\}$$