Computing the objective function in DCM

Klaas Enno Stephan, Karl J. Friston & Will D. Penny

The Wellcome Dept. of Imaging Neuroscience, 12 Queen Square, WC1N 3BG, London

Technical Report
05 May 2005
(minor corrections on 18 March 2006, 21 January 2007 and 14 June 2009)

The goal of this technical report is to demonstrate how the EM objective function in DCM can be derived mathematically, under Gaussian assumptions about the densities involved. For this purpose, we derive the equation for the objective function as it is used by the SPM routine `spm_nlsi_GN` from first principles. We do not deal with the actual principles of parameter estimation in the EM algorithm; these are described in Friston (2002) and Friston et al. (2003) and simply represent a maximisation of this objective function using conventional ascent schemes.

Motivation of the objective function

As described in Friston et al. (2003), Eq. A.5, the objective function for EM is defined as

$$F(q,\lambda) = \int q(\theta) \cdot \log \frac{p(y,\theta;\lambda,u)}{q(\theta)} d\theta \qquad (1)$$

where θ is a vector comprising the parameters of the model, λ contains the hyperparameters, $q(\theta)$ is an approximation to the true posterior $p(\theta \mid y;\lambda)$ and u represents the external inputs to the system.

What is the motivation underlying Eq. 1? What we would like to maximise is the model evidence (= marginal likelihood) where the parameters θ are integrated out. Because of the strict monotonicity of the log function, this is equivalent to maximising the marginal log likelihood:

$$\begin{aligned} L(\lambda) &= \log p(y|\lambda;u) \\ &= \log \int p(y \mid \theta,\lambda;u) p(\theta \mid \lambda;u)\, d\theta \\ &= \log \int p(y,\theta \mid \lambda;u)\, d\theta \end{aligned} \qquad (2)$$

In other words, we want to find hyperparameters such that the marginal log likelihood is maximised. Maximising Eq. 2 directly is usually difficult, but we can define $F(q,\lambda)$ as a lower bound on $L(\lambda)$, using Jensen's inequality:

$$\begin{aligned} L(\lambda) &= \log p(y|\lambda;u) = \log \int q(\theta)\frac{p(y,\theta|\lambda;u)}{q(\theta)}d\theta \\ &\geq \int q(\theta)\log\frac{p(y,\theta|\lambda;u)}{q(\theta)}d\theta = F(q,\lambda) \end{aligned} \tag{3}$$

Therefore, instead of maximising $L(\lambda)$, we can maximise its lower bound $F(q,\lambda)$. This can be accomplished by means of the EM algorithm (see Dayan & Abott 2001 and Ghahramani 2002 for introduction to the principles of EM). In the context of DCM, the EM algorithm updates the parameters in the E-step and the hyperparameters in the M-step at each iteration such that the following inequality holds (the superscript indexes the iteration step):

$$L(\lambda^{(k-1)}) \overset{E-step}{=} F(q(\theta^{(k)}),\lambda^{(k-1)}) \overset{M-step}{\leq} F(q(\theta^{(k)}),\lambda^{(k)}) \overset{Jensen}{\leq} L(\lambda^{(k)}) \tag{4}$$

The equality $L(\lambda^{(k-1)}) = F(q(\theta^{(k)}),\lambda^{(k-1)})$ achieved by the E-step is valid if the posterior can be matched precisely by $q(\theta)$. This is the case for the Laplace approximation which assumes that the posterior can be approximated by a Gaussian density whose mean is centered on the posterior maximum (see below).

The principles of parameter and hyperparameter estimation in DCM are described in detail by Friston (2002) and Friston et al. (2003). Here, we describe how the value of F can be computed at each iteration step in DCM under Gaussian assumptions. There are several ways of decomposing F into computationally more tractable terms than Eq. 1; we derive two of these approaches in this tutorial.

Approach I: Direct decomposition of F

We can rewrite $F(q,\lambda)$ from Eq. 1 as

$$\begin{aligned} F(q,\lambda) &= \int q(\theta)\log\frac{p(y,\theta;\lambda,u)}{q(\theta)}d\theta \\ &= \int q(\theta)\log\frac{p(y\mid\theta;\lambda,u)\,p(\theta;\lambda)}{q(\theta)}d\theta \\ &= \int q(\theta)\log p(y|\theta,\lambda;u)d\theta - \int q(\theta)\log\frac{q(\theta)}{p(\theta)}d\theta \\ &= \left\langle \log p(y|\theta,\lambda;u)\right\rangle_q - KL(q(\theta),p(\theta)) \end{aligned} \tag{5}$$

Eq. 5 shows that F is comprised of an accuracy term (the expected log likelihood) and the Kullback-Leibler (KL) divergence between the approximate posterior $q(\theta)$ and the prior $p(\theta)$.

First, we will rewrite the expected log likelihood. Using the general definition of a multivariate Gaussian distribution for an *n*-dimensional vector *x* with mean μ and covariance Σ

$$N(x;\mu,\Sigma)=(2\pi)^{-n/2}\left|\Sigma\right|^{-1/2}\exp(-\frac{1}{2}(x-\mu)^{T}\Sigma^{-1}(x-\mu)) \tag{6}$$

the likelihood is

$$p(y\left|\theta,\lambda;u\right)=N(y;h(\theta),C_{e}) \tag{7}$$

where $y=h(\theta)+e$. The expected log likelihood can then be written as

$$\begin{aligned}\left\langle \log p(y\left|\theta,\lambda;u\right)\right\rangle_{q} &=\left\langle -\frac{d_{y}}{2}\log 2\pi-\tfrac{1}{2}\log\left|C_{e}\right|-\tfrac{1}{2}(y-h(\theta))^{T}C_{e}^{-1}(y-h(\theta))\right\rangle_{q}\\ &=\left\langle -\frac{d_{y}}{2}\log 2\pi-\tfrac{1}{2}\log\left|C_{e}\right|-\tfrac{1}{2}r_{1}(\theta)\right\rangle_{q}\end{aligned} \tag{8}$$

where d_{y} is the dimensionality of the data.

The challenge now is to rewrite $r_{1}(\theta)$ from Eq. 8 such that the expectation with regard to $q(\theta)$ can be computed. Before we can do that, we need to introduce and define a few things. First of all, we use a Gaussian as the prior density:

$$p(\theta)=N(\theta;\theta_{p},C_{p}) \tag{9}$$

Furthermore, we make use of the Laplace approximation, i.e. as the approximate posterior $q(\theta)$, we use a Gaussian density whose mean is centered on the maximum of the true posterior. Under Gaussian assumptions about the posterior, this approximation becomes an identity after each E-step in the EM algorithm (see Eq. 4):

$$\begin{aligned}&q(\theta)=N(\theta;\theta_{MP},C_{MP})\\ &=p(\theta\,|\,y)\end{aligned} \tag{10}$$

(Note: this nomenclature corresponds to that of Friston et al. 2003 in the following way: $\theta_{MP}=\eta_{\theta|y},C_{MP}=C_{\theta|y},C_{p}=C_{\theta}$)

The Laplace approximation allows us to express $h(\theta)$ as a Taylor expansion around θ_{MP}:

$$\begin{aligned} &h(\theta) = h(\theta_{MP}) + J(\theta - \theta_{MP}) \\ &J = \frac{dh(\theta)}{d\theta} \end{aligned} \tag{11}$$

Finally, we need the following expression for the posterior precision C_{MP}^{-1} which, given the linear approximation in Eq. 11 and Gaussian assumptions about the posterior, can be derived from the definition of the posterior according to Bayes theorem (see Eqs. 10-13 in Friston 2002):

$$C_{MP}^{-1} = J^T C_e^{-1} J + C_p^{-1} \tag{12}$$

Now, using the identity

$$\begin{aligned} &h(\theta) = h(\theta) - h(\theta_{MP}) + h(\theta_{MP}) \\ &y - h(\theta) = \left(y - h(\theta_{MP})\right) + \left(h(\theta_{MP}) - h(\theta)\right) \end{aligned} \tag{13}$$

we can write $r_1(\theta)$ from Eq. 8 as (14)

$$\begin{aligned} r_1(\theta) &= (y - h(\theta_{MP}))^T C_e^{-1}(y - h(\theta_{MP})) + (h(\theta_{MP}) - h(\theta))^T C_e^{-1}(h(\theta_{MP}) - h(\theta)) \\ &= (y - h(\theta_{MP}))^T C_e^{-1}(y - h(\theta_{MP})) + (h(\theta) - h(\theta_{MP}))^T C_e^{-1}(h(\theta) - h(\theta_{MP})) \\ &= (y - h(\theta_{MP}))^T C_e^{-1}(y - h(\theta_{MP})) + (J(\theta - \theta_{MP}))^T C_e^{-1}(J(\theta - \theta_{MP})) \\ &= (y - h(\theta_{MP}))^T C_e^{-1}(y - h(\theta_{MP})) + r_2(\theta) \end{aligned}$$

Using Eqs. 11 and 12, we can write $r_2(\theta)$ as

$$\begin{aligned} &r_2(\theta) = \left(J(\theta - \theta_{MP})\right)^T C_e^{-1}\left(J(\theta - \theta_{MP})\right) \\ &= (\theta - \theta_{MP})^T J^T C_e^{-1} J(\theta - \theta_{MP}) \\ &= (\theta - \theta_{MP})^T (C_{MP}^{-1} - C_p^{-1})(\theta - \theta_{MP}) \end{aligned} \tag{15}$$

Given these expressions for $r_1(\theta)$ and $r_2(\theta)$, we can write Eq. 8 as

$$\begin{aligned} &\left\langle \log p(y|\theta, \lambda; u) \right\rangle_q = \\ &\left\langle -\frac{d_y}{2}\log 2\pi - \tfrac{1}{2}\log|C_e| - \tfrac{1}{2}\left((y - h(\theta_{MP}))^T C_e^{-1}(y - h(\theta_{MP})) + r_2(\theta)\right)\right\rangle_q \\ &= -\frac{d_y}{2}\log 2\pi - \tfrac{1}{2}\log|C_e| - \tfrac{1}{2}(y - h(\theta_{MP}))^T C_e^{-1}(y - h(\theta_{MP})) - \tfrac{1}{2}\int q(\theta) r_2(\theta)\, d\theta \\ &= -\frac{d_y}{2}\log 2\pi - \tfrac{1}{2}\log|C_e| - \tfrac{1}{2}(y - h(\theta_{MP}))^T C_e^{-1}(y - h(\theta_{MP})) \\ &- \tfrac{1}{2}\int q(\theta)(\theta - \theta_{MP})^T (C_{MP}^{-1} - C_p^{-1})(\theta - \theta_{MP})\, d\theta \end{aligned} \tag{16}$$

We now make use of the following lemma that holds for any Gaussian density *x* with $p(x) = N(x;\mu,\Sigma)$:

$$\left\langle x^T A x \right\rangle = \mu^T \Sigma \mu + Tr(A\Sigma) \qquad (17)$$

Note that under the assumption of the Laplace approximation

$$(\theta - \theta_{MP}) \sim N(0, C_{MP}) \qquad (18)$$

Therefore, the above lemma (Eq. 17) can be used to rewrite the integral from Eq. 16 as

$$\begin{aligned} &-\tfrac{1}{2}\int q(\theta)(\theta-\theta_{MP})^T (C_{MP}^{-1} - C_p^{-1})(\theta - \theta_{MP})\, d\theta \\ &= -\tfrac{1}{2} Tr\Big((C_{MP}^{-1} - C_p^{-1})C_{MP}\Big) \\ &= -\tfrac{1}{2} Tr(C_{MP}^{-1} C_{MP}) + \tfrac{1}{2} Tr(C_p^{-1} C_{MP}) \\ &= -\tfrac{1}{2} d_\theta + \tfrac{1}{2} Tr(C_p^{-1} C_{MP}) \end{aligned} \qquad (19$$

where d_θ is the dimensionality of the parameter vector θ.

Substituting Eq. 19 into Eq. 16, we now have a complete expression for the expected log likelihood from Eq. 5:

$$\begin{aligned} &\left\langle \log p(y|\theta,\lambda;u) \right\rangle_q \\ &= -\frac{d_y}{2}\log 2\pi - \tfrac{1}{2}\log|C_e| - \tfrac{1}{2}(y - h(\theta_{MP}))^T C_e^{-1}(y - h(\theta_{MP})) - \tfrac{1}{2} d_\theta + \tfrac{1}{2} Tr(C_p^{-1} C_{MP}) \end{aligned} \qquad (20)$$

It now remains to transform the *KL* term from Eq. 5.
The general definition of the *KL* divergence between two densities *v* and *w* is

$$KL(v,w) = \int v(\theta) \log \frac{v(\theta)}{w(\theta)} d\theta \qquad (21)$$

If *v* and *w* are *d*-dimensional normal densities, $KL(v,w)$ can be written as

$$KL(v,w) = -\frac{1}{2}\log|\Sigma_v| + \frac{1}{2}\log|\Sigma_w| + \frac{1}{2} Tr(\Sigma_w^{-1}\Sigma_v) + \frac{1}{2}(\mu_v - \mu_w)^T \Sigma_w^{-1}(\mu_v - \mu_w) - \frac{d}{2} \qquad (22)$$

(this can be derived by substituting the definitions of *v* and *w* as multivariate Gaussians [see Eq. 6] into Eq. 21; see also Penny 2001).

Applying this to the *KL* divergence between the approximate recognition density $q(\theta)$ and the prior density $p(\theta)$ in Eq. 5 gives

$$KL\big(q(\theta), p(\theta)\big)$$
$$= -\frac{1}{2}\log\left|C_{MP}\right| + \frac{1}{2}\log\left|C_p\right| + \frac{1}{2}Tr(C_p^{-1}C_{MP}) + \frac{1}{2}(\theta_{MP} - \theta_p)^T C_p^{-1}(\theta_{MP} - \theta_p) - \frac{d_\theta}{2} \qquad (23)$$

Substituting the results from Eqs. 20 and 23 into Eq. 5 we obtain the following expression for $F(q,\lambda)$:

$$F(q,\lambda) = \left\langle \log p(y|\theta,\lambda;u)\right\rangle_q - KL(q(\theta), p(\theta))$$

$$= -\frac{d_y}{2}\log 2\pi - \tfrac{1}{2}\log\left|C_e\right| - \tfrac{1}{2}(y - h(\theta_{MP}))^T C_e^{-1}(y - h(\theta_{MP})) - \frac{d_\theta}{2} + \frac{1}{2}Tr(C_p^{-1}C_{MP})$$
$$+ \frac{1}{2}\log\left|C_{MP}\right| - \frac{1}{2}\log\left|C_p\right| - \frac{1}{2}Tr(C_p^{-1}C_{MP}) - \frac{1}{2}(\theta_{MP} - \theta_p)^T C_p^{-1}(\theta_{MP} - \theta_p) + \frac{d_\theta}{2} \qquad (24)$$

$$= -\frac{d_y}{2}\log 2\pi - \tfrac{1}{2}\log\left|C_e\right| + \frac{1}{2}\log\left|C_{MP}\right| - \frac{1}{2}\log\left|C_p\right|$$
$$- \tfrac{1}{2}(y - h(\theta_{MP}))^T C_e^{-1}(y - h(\theta_{MP})) - \frac{1}{2}(\theta_{MP} - \theta_p)^T C_p^{-1}(\theta_{MP} - \theta_p)$$

This is exactly the expression used in `spm_nlsi_GN` to compute the value of *F* at each iteration step.

(Note that earlier versions of `spm_nlsi_GN` actually omitted the constants terms $-\frac{d_y}{2}\log 2\pi$ and $-\tfrac{1}{2}\log\left|C_p\right|$.)

Approach II: Determining F via computing the log evidence under the assumptions of the Laplace approximation

In this second approach, we show that the expression for the log evidence under the Laplace approximation is identical to the definition of *F* above.

Starting with the definition of the densities under Gaussian assumptions, we have

$$\begin{array}{ll} \text{prior:} & p(\theta) = N(\theta;\theta_p, C_p) \\ \text{likelihood:} & p(y\,|\,\theta) = N(y; h(\theta), C_e) \\ \text{posterior:} & p(\theta\,|\,y) = N(\theta;\theta_{MP}, C_{MP}) \end{array} \qquad (25)$$

The evidence can be computed by integrating out the parameters (marginalising):

$$
\begin{aligned}
&p(y) = \int p(y\,|\,\theta)\,p(\theta)\,d\theta \\
&= (2\pi)^{(-d_y-d_\theta)/2}\left|C_e\right|^{-\frac{1}{2}}\left|C_p\right|^{-\frac{1}{2}}\int \exp\left[-\frac{1}{2}\big((y-h(\theta))^T C_e^{-1}\big((y-h(\theta)\big)-\frac{1}{2}\Big(\big(\theta-\theta_p\big)^T C_p^{-1}\big(\theta-\theta_p\big)\Big)\right]d\theta \\
&= (2\pi)^{(-d_y-d_\theta)/2}\left|C_e\right|^{-\frac{1}{2}}\left|C_p\right|^{-\frac{1}{2}}\Lambda
\end{aligned}
\qquad (26)
$$

Here, d_θ is the dimensionality of the parameter vector θ and d_y is the dimensionality of the data.

Our goal is to find an expression that allows to compute the log evidence at every iteration step in a convenient fashion. Initially, we use the following trick to re-write the two terms from Λ in Eq. 24:

$$
\begin{aligned}
&\theta = \theta + \theta_{MP} - \theta_{MP} \\
&\rightarrow \theta - \theta_p = (\theta - \theta_{MP}) + (\theta_{MP} - \theta_p) \\
&\rightarrow (\theta-\theta_p)^T C_p^{-1}(\theta-\theta_p) = (\theta-\theta_{MP})^T C_p^{-1}(\theta-\theta_{MP}) + (\theta_{MP}-\theta_p)^T C_p^{-1}(\theta_{MP}-\theta_p)
\end{aligned}
\qquad (27
$$

$$
\begin{aligned}
&h(\theta) = h(\theta) + h(\theta_{MP}) - h(\theta_{MP}) \\
&\rightarrow y - h(\theta) = \big(y - h(\theta_{MP})\big) + \big(h(\theta_{MP}) - h(\theta)\big) \\
&\rightarrow \big(y-h(\theta)\big)^T C_e^{-1}\big(y-h(\theta)\big) = \big(y-h(\theta_{MP})\big)^T C_e^{-1}\big(y-h(\theta_{MP})\big) + \big(h(\theta_{MP})-h(\theta)\big)^T C_e^{-1}\big(h(\theta_{MP})-h(\theta)\big)
\end{aligned}
$$

Now we have decomposed the exponent from Eq. 26 into 4 terms. Using Eqs. 11 and 12, we can show that two of them are identical to the exponent of the posterior (note that this is only valid under the assumptions of the Laplace approximation, see above):

$$
\begin{aligned}
&\big(h(\theta_{MP})-h(\theta)\big)^T C_e^{-1}\big(h(\theta_{MP})-h(\theta)\big) + (\theta-\theta_{MP})^T C_p^{-1}(\theta-\theta_{MP}) \\
&= \big(h(\theta)-h(\theta_{MP})\big)^T C_e^{-1}\big(h(\theta)-h(\theta_{MP})\big) + (\theta-\theta_{MP})^T C_p^{-1}(\theta-\theta_{MP}) \\
&= \big(J(\theta-\theta_{MP})\big)^T C_e^{-1} J(\theta-\theta_{MP}) + (\theta-\theta_{MP})^T C_p^{-1}(\theta-\theta_{MP}) \\
&= (\theta-\theta_{MP})^T J^T C_e^{-1} J(\theta-\theta_{MP}) + (\theta-\theta_{MP})^T C_p^{-1}(\theta-\theta_{MP}) \\
&= (\theta-\theta_{MP})^T (J^T C_e^{-1} J + C_p^{-1})(\theta-\theta_{MP}) \\
&= (\theta-\theta_{MP})^T C_{MP}^{-1}(\theta-\theta_{MP})
\end{aligned}
\qquad (28)
$$

Because $\int p(\theta\,|\,y)\,d\theta = 1$ and because, according to Eq. 25,

$$
p(\theta\,|\,y) = (2\pi)^{-d_\theta/2}\left|C_{MP}\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\theta-\theta_{MP})^T C_{MP}^{-1}(\theta-\theta_{MP})\right] \qquad (29)
$$

it follows that

$$
\begin{aligned}
&\int (2\pi)^{-d_\theta/2}|C_{MP}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\theta-\theta_{MP})^T C_{MP}^{-1}(\theta-\theta_{MP})\right]d\theta = 1 \\
&(2\pi)^{-d_\theta/2}|C_{MP}|^{-\frac{1}{2}}\int \exp\left[-\frac{1}{2}(\theta-\theta_{MP})^T C_{MP}^{-1}(\theta-\theta_{MP})\right]d\theta = 1 \qquad (30) \\
&\int \exp\left[-\frac{1}{2}(\theta-\theta_{MP})^T C_{MP}^{-1}(\theta-\theta_{MP})\right]d\theta = (2\pi)^{d_\theta/2}|C_{MP}|^{\frac{1}{2}}
\end{aligned}
$$

Substituting the results from Eqs. 27 and 30 into Λ, this allows us to write

$$
\begin{aligned}
&\Lambda = \\
&(2\pi)^{d_\theta/2}|C_{MP}|^{\frac{1}{2}}\int \exp\left[-\frac{1}{2}((y-h(\theta_{MP}))^T C_e^{-1}((y-h(\theta_{MP}))-\frac{1}{2}((\theta_{MP}-\theta_p)^T C_p^{-1}(\theta_{MP}-\theta_p))\right]d\theta \qquad (31)
\end{aligned}
$$

Substituting this into Eq. 26 and taking logs gives

$$
\begin{aligned}
&\log p(y) = \\
&-\frac{d_y}{2}\log 2\pi - \frac{1}{2}\log|C_e| - \frac{1}{2}\log|C_p| + \frac{1}{2}\log|C_{MP}| \qquad (32) \\
&-\frac{1}{2}((y-h(\theta_{MP}))^T C_e^{-1}((y-h(\theta_{MP}))-\frac{1}{2}((\theta_{MP}-\theta_p)^T C_p^{-1}(\theta_{MP}-\theta_p))
\end{aligned}
$$

This expression for the log evidence is identical to the expression for *F* derived by the first approach above.

References

Dayan P & Abott LF (2001) Theoretical neuroscience. Cambridge, MA: MIT Press.

Friston KJ (2002) Bayesian estimation of dynamical systems: an application to fMRI. *NeuroImage* 16: 513-530.

Friston KJ, Harrison LM, Penny WD (2003) Dynamic causal modelling. *NeuroImage* 19: 1273-1302.

Ghahramani (2002) Graphical Models: Parameter Learning. In: Arbib (ed): The Handbook of Brain Theory and Neural Networks (2nd edition).

Penny WD (2001) KL-divergences of Normal, Gamma, Dirichlet and Wishart densities. Technical report, Wellcome Dept. of Imaging Neuroscience.