Hierarchical Dynamic Models

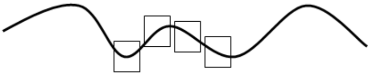
Will Penny

26th May 2011

AR Processes

Discrete time autoregressive processes of order p can be described

$$x_t = \sum_{k=1}^{p} a_k x_{t-k} + b_o z_t$$



They can also be written as

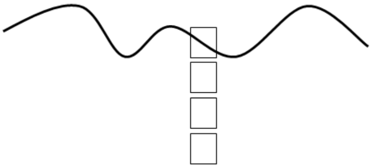
$$\sum_{k=0}^{p} a_k x_{t-k} = b_o z_t$$

OU Processes

One can define the analagous continuous time process as

$$\sum_{k=0}^{p} a_k x_t^k = b_o z_t$$

where x_t^k denotes the kth order derivative of x_t.



This is referred to as an $OU(p)$ process (Gardiner, 1983).

OU Processes

For an OU(1) process we have

$$a_0 x_t + a_1 \dot{x}_t = b_o z_t$$

We can rewrite this as

$$\begin{aligned} \dot{x}_t &= -\frac{a_0}{a_1} x_t + \frac{b_0}{a_1} z_t \\ &= -a x_t + b z_t \end{aligned}$$

In last weeks lecture we wrote this as

$$dx_t = -a x_t dt + b dW_t$$

where W_t is a Wiener process.

Embedding

A pth order differential equation describing y can be expressed as an ordinary differential equation comprising the p variables in u which is defined via an embedding. Consider, for example, an $OU(4)$ process.

$$a_0 x_t + a_1 \dot{x}_t + a_2 \ddot{x}_t + a_3 \dddot{x}_t + a_4 \ddddot{x}_t = b_0 z_t$$

We can divide through by a_4 and redefine coefficients giving

$$\begin{aligned} \dot{x}_t &= \dot{x}_t \\ \ddot{x}_t &= \ddot{x}_t \\ \dddot{x}_t &= \dddot{x}_t \\ \ddddot{x}_t &= -a_0 x_t - a_1 \dot{x}_t - a_2 \ddot{x}_t - a_3 \dddot{x}_t + b_0 z_t \end{aligned}$$

Embedding

If we define $u_t = [x_t, \dot{x}_t, \ddot{x}_t, \dddot{x}_t]^T$ we can express the process in vector form

$$\dot{u}_t = Au_t + Du_t + Bz_t$$

where

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -a_0 & -a_1 & -a_2 & -a_3 \end{bmatrix}$$

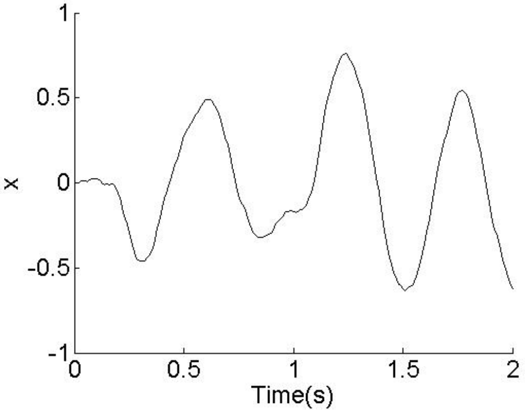
$$D = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & b_0 \end{bmatrix}$$

where D is a derivative operator.

OU(2) process

Single realisation of an OU(2) process with parameters $a_0 = 100, a_1 = 2, b_0 = 1$.



OU process

Defining $F = A + D$ we can write the OU process as

$$\dot{u}_t = F u_t + B z_t$$

This is equivalently written

$$du_t = F u_t dt + B dw_t$$

They have closed form solutions as follows

$$u_t = \exp(-Ft) u_0 + \int_0^t \exp(-F(t - t')) B dw'_t$$

This follows from the result in the last lecture (stochastic processes) - applying Ito's formula and integrating.

Covariance functions

Williams (2006) (see also Gardiner 1983, section 4.4.6) shows that, if $a_1^2 < 4a_0$, then the covariance function for an OU(2) process is given by

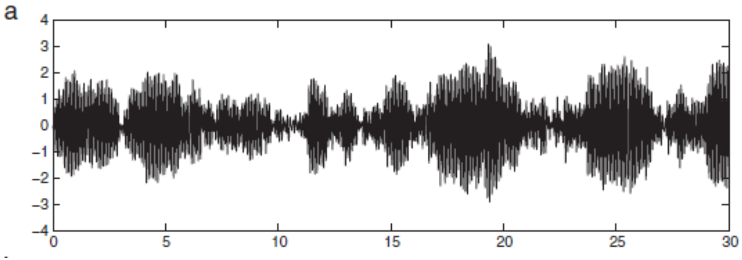
$$C(\tau) = \frac{b_0^2}{2a_0 a_1} \exp(-\alpha\tau)(\cos[\beta\tau] + \frac{\alpha}{\beta}\sin[\beta\tau])$$

where $\alpha = a_1/2$ and $\alpha^2 + \beta^2 = a_0$. For the parameters used to generate the sample path in the previous figure, this corresponds to a frequency of $f = \beta/(2\pi) = 1.6$Hz, which seems about right.

Similar results exist for OU(p) processes (Gardiner, 1983).

Resting state MEG

Hindriks et al. (2011) have fitted such an equation to resting state MEG data (showing alpha rhythm dynamics).



By matching the observed and model-based covariance functions they estimated the parameters to be $a_0 = 65.2^2, a_1 = 3.2, b_0 = 224.3$. This corresponds to a dominant frequency of $f = 10.4$Hz.

Dynamic models

We now add on an observation equation, implying that the dynamic processes of interest are not directly observed. Friston et al. (2008) developed a triple estimation DEM framework for estimation hidden states, parameters and hyperparameters. We first focus on filtering.

$$
\begin{aligned}
y &= g(x) + z \\
\dot{x} &= f(x) + w
\end{aligned}
$$

Here y is data to be modelled and x is a hidden state. The second equation here embodies a dynamical prior.

The noise terms z and w are stochastic innovations such that the covariance of the vector $[z, \dot{z}, \ddot{z}, ..]$ is well defined. Similarly for w.

In Friston et al's DEM formulation the noise processes are smooth Gaussian processes with non-zero correlation times. They are not Wiener processes.

Other Approaches

Before getting into the DEM formalism we note that the majority of work in filtering focuses on discrete time models

$$\begin{aligned} y_t &= g(x_t) + z_t \\ x_t &= f(x_{t-1}) + w_t \end{aligned}$$

Estimation of states then takes place using discrete time Bayesian filtering. Methods include (extended) Kalman filtering, particle filtering. There is a very large literature on this see eg. Bishop et al. (2006).

Daunizeau et al (2009) have developed a triple estimation discrete time framework using variational methods. For continuous time formulations (ie inference on SDEs) other than DEM see eg Archambeau (2011) and Calderhead (2009).

Generalised coordinates

The hallmark of the DEM approach is its use of generalised coordinates.

Under local linearity assumptions the state equation can be repeatedly differentiated with respect to time to give

$$
\begin{aligned}
\dot{x} &= f(x) + w \\
\ddot{x} &= f_x \dot{x} + \dot{w} \\
\dddot{x} &= f_x \ddot{x} + \ddot{w} \\
.. &= ..
\end{aligned}
$$

where f_x denotes the derivative df/dx.

Generalised coordinates

Instead however we write

$$
\begin{aligned}
x' &= f(x) + w \\
x'' &= f_x x' + \dot{w} \\
x''' &= f_x x'' + \ddot{w} \\
.. &= ..
\end{aligned}
$$

where $\tilde{x} = [x, x', x'', ..]$ are the 'Generalised Coordinates (GCs)' of x.

Filtering will take place in the space of GCs, and we will estimate $\tilde{x}$. Because we are estimating the states it is unlikely that $x' = \dot{x}$. However, this will be the case when tracking is good.

Generalised coordinates

The embedded state transition can be written as $\tilde{f} = [f, f'', f''', ...]$ where

$$\begin{aligned} f &= f(v, x, \theta) \\ f' &= f_x x' \\ f'' &= f_x x'' \\ .. &= .. \end{aligned}$$

and we write

$$\tilde{w} = [w, \dot{w}, \ddot{w}, ...]$$

Hence

$$D\tilde{x} = \tilde{f} + \tilde{w}$$

where D is the derivative operator.

Observation equation

Similarly the observation equation can be repeatedly differentiated to give

$$
\begin{aligned}
y &= g(x) + z \\
\dot{y} &= g_x x' + \dot{z} \\
\ddot{y} &= g_x x'' + \ddot{z} \\
.. &= ..
\end{aligned}
$$

where g_x denotes the derivative dg/dx.

Generalised coordinates

The embedded predicted response can be written as $\tilde{g} = [g, g', g'', ...]$ where

$$
\begin{aligned}
g &= g(v) \\
g' &= g_x x' \\
g'' &= g_x x'' \\
.. &= ..
\end{aligned}
$$

and we also write

$$
\begin{aligned}
\tilde{y} &= [y, \dot{y}, \ddot{y}, ..] \\
\tilde{z} &= [z, \dot{z}, \ddot{z}, ..]
\end{aligned}
$$

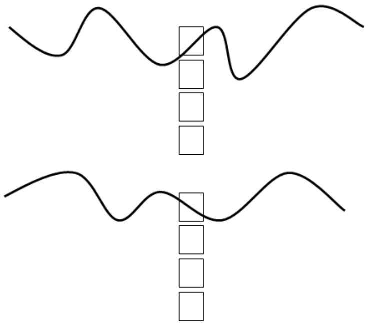
This means we can write the observation equation in generalised coordinates as

$$\tilde{y} = \tilde{g} + \tilde{z}$$

Compact expression

Hence, for locally linear systems

$$\begin{aligned}\tilde{y} &= \tilde{g} + \tilde{z} \\ D\tilde{x} &= \tilde{f} + \tilde{w}\end{aligned}$$



Joint Likelihood

The joint log likelihood of observatons and states is

$$L(\tilde{x}, t) = \log[p(\tilde{y}|\tilde{x})p(\tilde{x})]$$

Given Gaussian densities

$$\begin{aligned} p(\tilde{y}|\tilde{x}) &= \mathrm{N}(\tilde{y}; \tilde{g}, (\tilde{\Pi}^z)^{-1}) \\ p(\tilde{x}) &= \mathrm{N}(D\tilde{x}; \tilde{f}, (\tilde{\Pi}^w)^{-1}) \end{aligned}$$

where $\tilde{\Pi}^z$ and $\tilde{\Pi}^w$ are the precision matrices for the observations and state (more on this later). As a function of x we therefore have

$$L(\tilde{x}, t) = -\frac{1}{2}\tilde{e}_y^T \tilde{\Pi}^z \tilde{e}_y - \frac{1}{2}\tilde{e}_x^T \tilde{\Pi}^w \tilde{e}_x$$

where the observation and flow error terms are

$$\begin{aligned} \tilde{e}_y &= \tilde{y} - \tilde{g} \\ \tilde{e}_x &= D\tilde{x} - \tilde{f} \end{aligned}$$

Filtering

Filtering refers to the estimation of the hidden states $\tilde{x}$. Hidden states can be estimated by following the gradient of $L(\tilde{x}, t)$ with respect to $\tilde{x}$

$$
\begin{aligned}
j(\tilde{x}) &= \frac{dL(\tilde{x}, t)}{d\tilde{x}} \\
&= \frac{dL_1(\tilde{x}, t)}{d\tilde{e}_y} \frac{d\tilde{e}_y}{d\tilde{x}} + \frac{dL_2(\tilde{x}, t)}{d\tilde{e}_x} \frac{d\tilde{e}_x}{d\tilde{x}} \\
&= (I_p \otimes g_x) \tilde{\Pi}^z (\tilde{y} - \tilde{g}) + (I_p \otimes f_x - D) \tilde{\Pi}^w (D\tilde{x} - \tilde{f})
\end{aligned}
$$

where p is the embedding dimension. The hidden states could then be updated by following this gradient

$$
\dot{\tilde{x}} = j(\tilde{x})
$$

This optimisation would be sufficient for a stationary system where $L(\tilde{x}, t) = L(\tilde{x})$.

Mode Following

However, because the likelihood itself is changing we must add another term

$$\dot{\tilde{x}} = j(\tilde{x}) + D\tilde{x}$$

We can see that this makes sense because at the maximum likelihood value (or the 'mode') the gradient is zero, $j(\tilde{x}) = 0$. We then have

$$\dot{\tilde{x}} = D\tilde{x}$$

$$\begin{aligned} \tilde{x} &= [x, x', x'', ..] \\ D\tilde{x} &= [x', x'', x''', ..] \\ \dot{\tilde{x}} &= [\dot{x}, \dot{x}', \dot{x}'', ..] \end{aligned}$$

So at the mode, the generalised coordinates are equal to the time derivatives eg. $x' = \dot{x}$

Dynamic Model

We consider dynamic models of the form

$$\begin{aligned} \dot{x} &= f(x, v, \theta) + w \\ y &= g(x, v, \theta) + z \end{aligned}$$

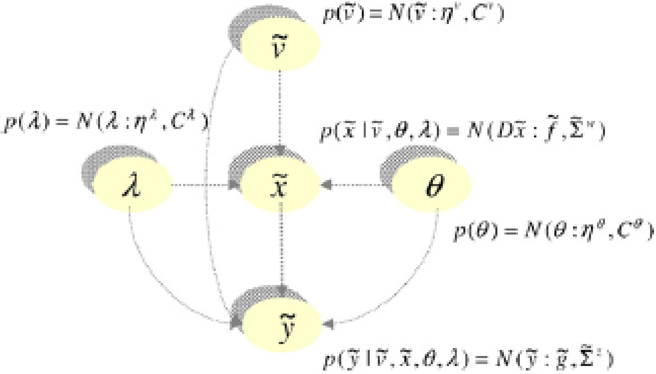
where x is a hidden state, v is a hidden cause, θ are model parameters, and y are observed time series.

We have now made explicit the dependence on parameters θ. The state and observation noise also depend on hyperparameters λ.

The DEM framework provides a method for estimating states, causes, parameters and hyperparameters.

Generative Model

$$
\begin{aligned}
y &= g(x,v) + z \\
\dot{x} &= f(x,v) + w
\end{aligned}
\quad \Rightarrow \quad
\begin{aligned}
\tilde{y} &= \tilde{g} + \tilde{z} \\
D\tilde{x} &= \tilde{f} + \tilde{w}
\end{aligned}
$$



Causes and Parameters

There is a Gaussian prior over the hidden causes

$$p(\tilde{v}) = \mathrm{N}(\tilde{v}; \eta^v, C^v)$$

where η^v is the mean and C^v is the covariance.

The prior over the parameters is

$$p(\theta) = \mathrm{N}(\theta; \eta^\theta, C^\theta)$$

where η^θ is the mean and C^θ is the covariance.

State Equation

The distribution governing the evolution of the states is

$$p(\tilde{x}|\tilde{v}, \theta, \lambda) = \mathrm{N}(D\tilde{x}; \tilde{f}, \tilde{\Sigma}^w)$$

The state noise covariance $\tilde{\Sigma}^w$ is parameterised by a hyperparameter, λ_w, and a parameter γ that governs the smoothness of the Gaussian innovations (Friston et al, 2008).

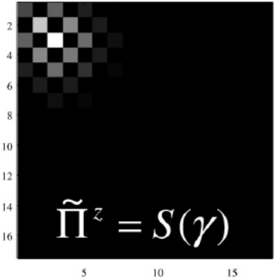
Previously we referred to the precision of the state noise covariance $(\tilde{\Pi}^w)^{-1} = \tilde{\Sigma}^w$.

Observation Equation

The distribution governing the evolution of the states is

$$p(\tilde{y}|\tilde{x}, \tilde{v}, \theta, \lambda) = \mathrm{N}(\tilde{y}; \tilde{g}, \tilde{\Sigma}^z)$$

The state noise covariance $\tilde{\Sigma}^z$ is the inverse of the state noise precision $\tilde{\Pi}^z$.



It is parameterised by a hyperparameter, λ_z, and a parameter γ that governs the smoothness of the Gaussian innovations.

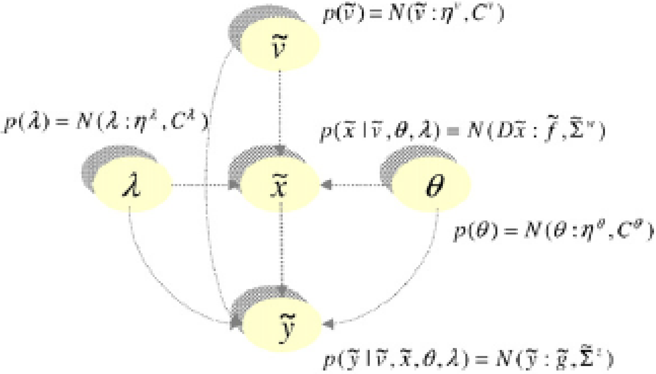
Hyperparameters

For the hyperparameters $\lambda = \{\lambda_w, \lambda_z\}$ we have a prior

$$p(\lambda) = \mathrm{N}(\lambda; \eta^{\lambda}, C^{\lambda})$$

Generative Model

$$
\begin{aligned} y &= g(x,v) + z \\ \dot{x} &= f(x,v) + w \end{aligned}
\quad \Rightarrow \quad
\begin{aligned} \tilde{y} &= \tilde{g} + \tilde{z} \\ D\tilde{x} &= \tilde{f} + \tilde{w} \end{aligned}
$$



Joint Likelihood and Posterior

The joint log-likelihood of the data, hidden states, causes, parameters and hyperparameters is

$$L(\tilde{x}, \tilde{v}, t, \theta, \lambda) = \log \left[p(\tilde{y}|\tilde{x}) p(\tilde{x}|\tilde{v}) p(\tilde{v}) p(\theta) p(\lambda) \right]$$

We assume an approximate posterior that factorises over states, parameters and hyperparameters

$$q(\tilde{x}, \tilde{v}, t, \theta, \lambda) = q(\tilde{x}, \tilde{v}, t) q(\theta) q(\lambda)$$

Energies and Actions

The variational energies

$$
\begin{aligned}
I(\tilde{x}, \tilde{v}, t) &= \int\int L(\tilde{x}, \tilde{v}, t, \theta, \lambda) q(\theta) q(\lambda) d\theta d\lambda \\
I(\theta) &= \int\int\int\int L(\tilde{x}, \tilde{v}, t, \theta, \lambda) q(\tilde{x}, \tilde{v}, t) q(\lambda) d\tilde{x} d\tilde{v} d\lambda dt \\
I(\lambda) &= \int\int\int\int L(\tilde{x}, \tilde{v}, t, \theta, \lambda) q(\tilde{x}, \tilde{v}, t) q(\theta) d\tilde{x} d\tilde{v} d\theta dt
\end{aligned}
$$

The latter two are known as 'actions' as the integral is also over time.

Dynamic Expectation Maximisation

The Dynamic Expectation Maximisation (DEM) algorithm implements a triple estimation of states (D-step), parameters (E-step) and hyperparameters (M-step).

- The D-step updates $q(\tilde{x}, \tilde{v}, t)$ so as to minimise the variational energy $I(\tilde{x}, \tilde{v}, t)$.
- The E-step updates $q(\theta)$ so as to minimise the variational action $I(\theta)$.
- The M-step updates $q(\lambda)$ so as to minimise the variational action $I(\lambda)$

D-Step

The state estimates are updated (approximately) using

$$\dot{\tilde{x}} = j(\tilde{x}) + D\tilde{x}$$

Previously (when just filtering) the gradient was of the joint log likelihood

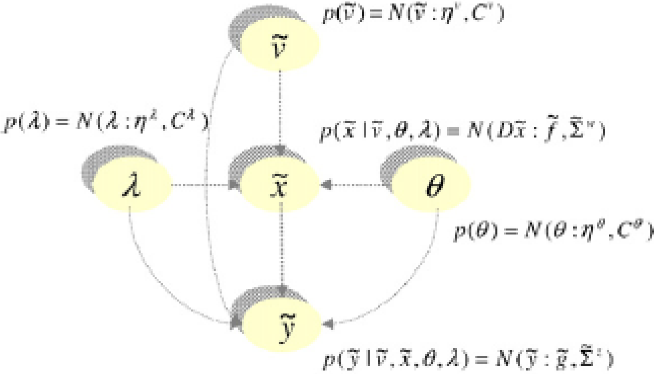
$$j(\tilde{x}) = \frac{dL(\tilde{x}, t)}{d\tilde{x}}$$

But for the D-step we use the gradient of the variational energy

$$j(\tilde{x}) = \frac{dI(\tilde{x}, t)}{d\tilde{x}}$$

Generative Model

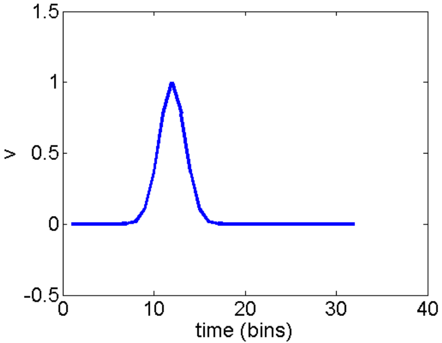
$$\begin{aligned} y &= g(x,v) + z \\ \dot{x} &= f(x,v) + w \end{aligned} \quad \Rightarrow \quad \begin{aligned} \tilde{y} &= \tilde{g} + \tilde{z} \\ D\tilde{x} &= \tilde{f} + \tilde{w} \end{aligned}$$



Linear Convolution model

We generate data from a single input multiple output linear dynamical model with input

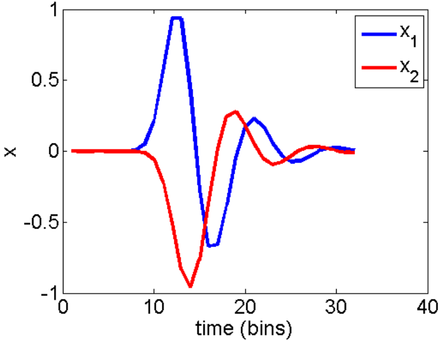
$$v = \exp(\frac{1}{4}[t - 12]^2)$$



Linear Convolution model

State variables were then generated according to the equations

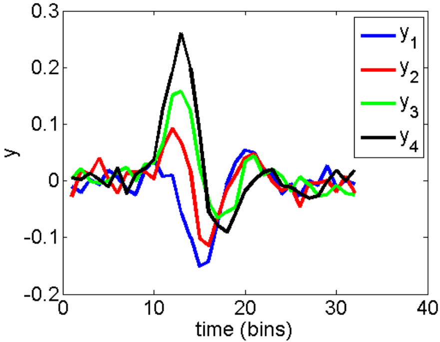
$$
\begin{aligned}
\dot{x}_1 &= -0.25x_1 + x_2 + v + w_1 \\
\dot{x}_2 &= -0.50x_1 - 0.25x_2 + w_2
\end{aligned}
$$



Generated Data

Observations were then created from the state variables

$$
\begin{aligned}
y_1 &= 0.125x_1 + 0.1633x_2 + z_1 \\
y_2 &= 0.125x_1 + 0.0676x_2 + z_2 \\
y_3 &= 0.125x_1 - 0.0676x_2 + z_3 \\
y_4 &= 0.125x_1 - 0.1633x_2 + z_4
\end{aligned}
$$



Linear Convolution model

This conforms to a dynamical model

$$\begin{aligned} \dot{x} &= f(x, v, \theta) + w \\ y &= g(x, v, \theta) + z \end{aligned}$$

where

$$\begin{aligned} f(x, v) &= Fx + hv \\ g(x, v) &= Gx \end{aligned}$$

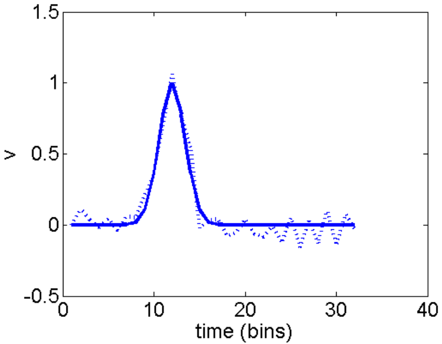
Linear Convolution model

The parameters $\theta = \{F, h, G\}$ are given by

$$
\begin{aligned}
F &= \begin{bmatrix} -0.25 & 1 \\ -0.50 & -0.25 \end{bmatrix} \\
h &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\
G &= \begin{bmatrix} 0.125 & 0.1633 \\ 0.125 & 0.0676 \\ 0.125 & -0.0676 \\ 0.125 & -0.1633 \end{bmatrix}
\end{aligned}
$$

Filtering

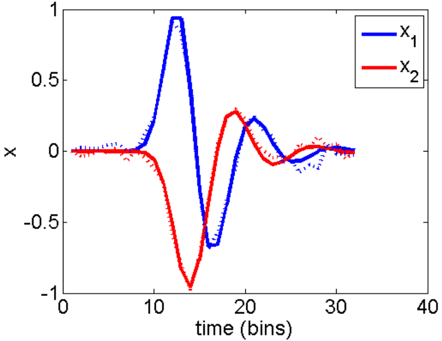
We now run the DEM algorithm with (infinitely) tight priors on the parameters and hyperparameters. Effectively, DEM just implements the D-step ie filtering.



The figure shows the true (solid) and estimated (dotted) causes, v. Implementation in 'DEMdemoconvolution.m' from the DEM toolbox of SPM.

Linear Convolution model

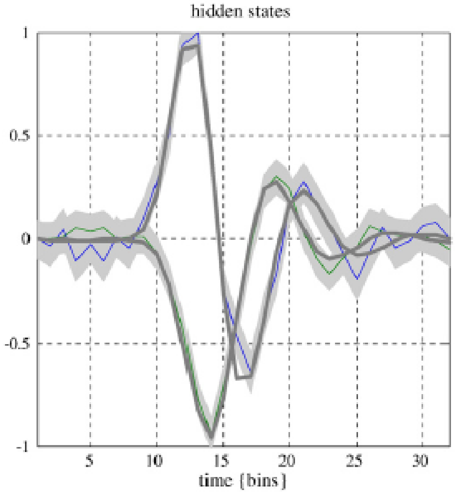
The figure shows the true (solid) and estimated (dotted) hidden variables, x.



Importantly, DEM also provides an approximate posterior density over the hidden states (not just the posterior mean - dotted line).

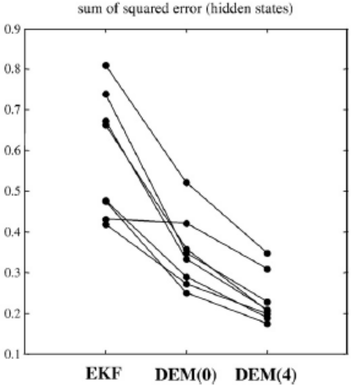
Linear Convolution model

Importantly, DEM also provides an approximate posterior density over the hidden states, $q(\tilde{x}, t)$.



Extended Kalman Filtering

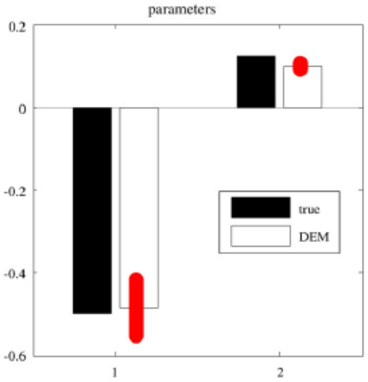
State estimation is more accurate using DEM as the innovations can be smooth Gaussian processes. With the (Extended) Kalman Filter (EKF) the Gaussian innovations are assumed IID (ie rough) (Bishop, 2006).



Additionally, DEM can invert models with hidden causes and hierarchical structure (see later).

Triple Estimation

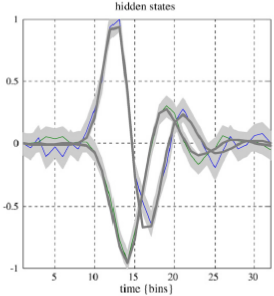
The DEM algorithm was then run with uninformative priors over two of the parameters.

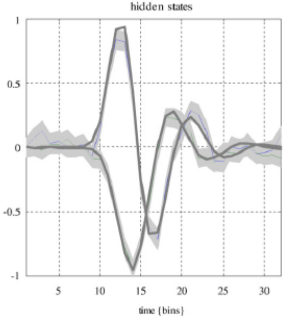


These were (1) the parameter coupling the first hidden state to the second (true value -0.5) and (2) the parameter coupling the first hidden state to the output (true value 0.125).

State Posterior

Posterior state distribution, $q(\tilde{x}, t)$, from DEM filtering (left) - where the parameters are assumed known - versus DEM triple estimation (right) - where the parameters and hyperparameters are estimated.

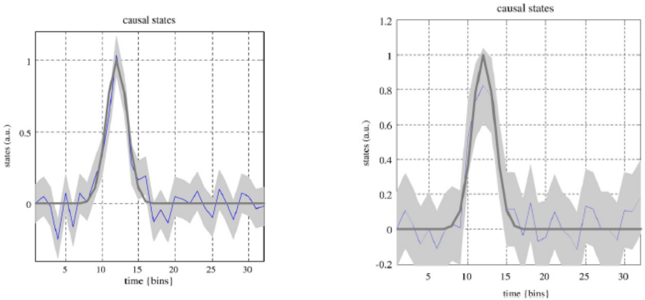




The gray tubes mark 90 percent confidence intervals.

Cause Posterior

Posterior cause distribution, $q(\tilde{v}, t)$, from DEM filtering (left) - where the parameters are assumed known - versus DEM triple estimation (right) - where the parameters and hyperparameters are estimated.



The gray tubes mark 90 percent confidence intervals.

Hierarchical Dynamic Models

We now make the causes dependent on higher order dynamics and causes. This continues in hierarchical fashion up to the *n*th level cause

$$
\begin{aligned}
y &= g(x_1, v_1, \theta_1) + z_1 \\
v_1 &= g(x_2, v_2, \theta_2) + z_2 \\
.. &= .. \\
v_{n-1} &= g(x_{n-1}, v_n, \theta_{n-1}) + e_{n-1}
\end{aligned}
$$

These hierarchical relations embody what might be termed structural priors.

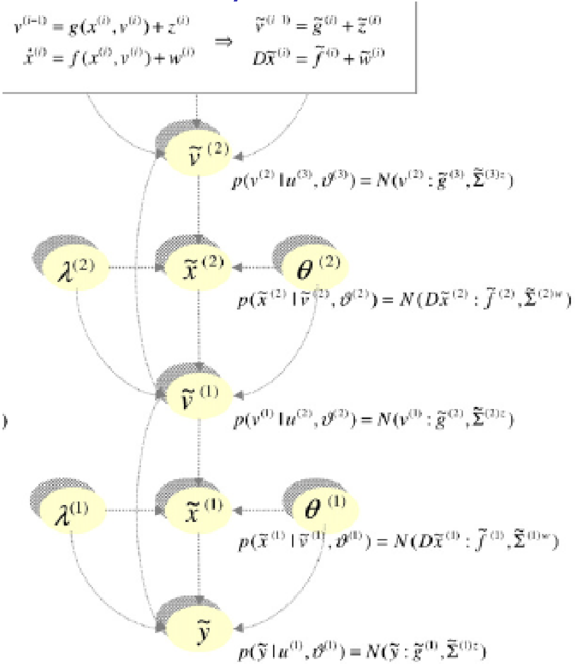
The generative models are identical to those in the hierarchy lecture (number 4) with the addition of hidden variables x_i that allow the model to remember the causes v_i of previous data y.

Hierarchical Dynamic Models

The hidden variables evolve dynamically and are also constrained by higher level causes (and parameters at the same level).

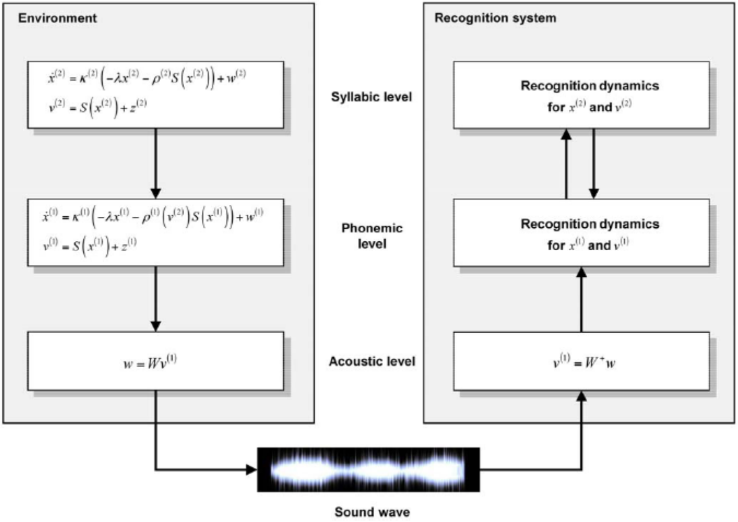
$$
\begin{aligned}
\dot{x}_1 &= f(x_1, v_1, \theta_1) \\
\dot{x}_2 &= f(x_2, v_2, \theta_2) \\
.. &= .. \\
\dot{x}_{n-1} &= f(x_{n-1}, v_n, \theta_{n-1})
\end{aligned}
$$

Hierarchical Dynamic Models



Speech Perception

Kiebel et al (2009) proposed that speech perception might correspond to Bayesian inference in dynamical systems comprised of a hierarchy of SHCs.



References

C. Archambeau (2011) Approximate Inference for continuous-time Markov processes. In D. Barber, A. T. Cemgil, and S. Chiappa, Inference and Learning in Dynamic Models. Cambridge University Press.

C. Bishop (2006) Pattern Recognition and Machine Learning, Springer.

B. Calderhead et al (2009) Accelerating Bayesian Inference over Nonlinear Differential Equations with Gaussian Processes Advances in Neural Information Processing Systems, Vol. 21, 217-224, MIT Press.

J. Daunizeau et al. (2009) Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models. Physica D: nonlinear phenomena, 238: 2089-2118.

K. Friston et al. (2008) DEM: A variational treatment of dynamic systems. Neuroimage 41, 849-885.

K. Friston (2008) Variational filtering in generalised coordinates of motion. Conference on Approximate Inference in Stochastic Processes and Dynamic Systems. http://videolectures.net.

C. Gardiner (1983) Handbook of Stochastic Methods. Springer-Verlag, 1983.

R. Hindriks et al (2011) Dynamics underlying spontaneous human alpha oscillations: a data-driven approach. Neuroimage, in press.

S. Kiebel et al. (2009) Recognizing sequences of sequences. PLoS Computational Biology, e1000464.

C. Williams (2006) A tutorial introduction to stochastic differential equations. Invited talk in NIPS

workshop on Dynamical Systems, Stochastic Processes and Bayesian Inference Whistler, BC, Canada.

http://videolectures.net.