Author Summary

Models are essential to make sense of scientific data, but they may also play a central role in how we assimilate sensory information. In this paper, we introduce a general model that generates or predicts diverse sorts of data. As such, it subsumes many common models used in data analysis and statistical testing. We show that this model can be fitted to data using a single and generic procedure, which means we can place a large array of data analysis procedures within the same unifying framework. Critically, we then show that the brain has, in principle, the machinery to implement this scheme. This suggests that the brain has the capacity to analyse sensory input using the most sophisticated algorithms currently employed by scientists and possibly models that are even more elaborate. The implications of this work are that we can understand the structure and function of the brain as an inference machine. Furthermore, we can ascribe various aspects of brain anatomy and physiology to specific computational quantities, which may help understand both normal brain function and how aberrant inferences result from pathological processes associated with psychiatric disorders. will be dealing with variables in generalised coordinates of motion, which will be denoted by a tilde; $\tilde{x} := \left[x, x', x'', \ldots\right]^T = \left[x^{[0]}, x^{[1]}, x^{[2]}, \ldots\right]^T$, where $x^{[i]}$ denotes i th order motion. A point in generalised coordinates can be regarded as encoding the instantaneous trajectory of a variable, in the sense it prescribes its location, velocity, acceleration etc.

Materials and Methods

Hierarchical Dynamic Models

In this section, we cover hierarchal models for dynamic systems. We start with the basic model and how generalised motion furnishes empirical priors on the dynamics of the model's hidden states. We then consider hierarchical forms and see how these induce empirical priors in a structural sense. We will try to relate these perspectives to established treatments of empirical priors in static and state-space models.

Hierarchical dynamic causal models. Dynamic causal models are probabilistic generative models $p(y, \vartheta)$ based on state-space models. As such, they entail the likelihood, $p(y \mid \vartheta)$ of getting some data, y , given some parameters $\vartheta = \{x, v, \theta, \lambda\}$ and priors on those parameters, $p(\vartheta)$. We will see that the parameters subsume different quantities, some of which change with time and some which do not. These models are causal in a control-theory sense because they are state-space models, formulated in continuous time.

State-pace models in generalised coordinates. A dynamic input-state-output model can be written as

$$y = g(x, v) + z \qquad (1)$$

$$\dot{x} = f(x, v) + w$$

The continuous nonlinear functions f and g of the states are parameterised by θ. The states $v(t)$ can be deterministic, stochastic, or both. They are variously referred to as inputs, sources or causes. The states $x(t)$ meditate the influence of the input on the output and endow the system with memory. They are often referred to as hidden states because they are seldom observed directly. We assume the stochastic terms (i.e., observation noise) $z(t)$ are analytic, such that the covariance of $\tilde{z} = [z, z', z'', \ldots]^T$ is well defined; similarly for the system or state noise, $w(t)$, which represents random fluctuations on the motion of the hidden states. Under local linearity assumptions (i.e., ignoring high-order derivatives of the generative model functions), the generalised output or response $\tilde{y} = [y, y', y'', \ldots]^T$ obtains from recursive differentiation with respect to time using the chain rule

$$y = g(x, v) + z \; \dot{x} = x' = f(x, v) + w$$

$$y' = g_x x' + g_v v' + z' \; \dot{x}' = x'' = f_x x' + f_v v' + w'$$

$$y'' = g_x x'' + g_v v'' + z'' \; \dot{x}'' = x''' = f_x x'' + f_v v'' + w'' \qquad (2)$$

$$\vdots$$

Note that the derivatives are evaluated at each point in time and the linear approximation is local to the current state. The first (observer) equation show that the generalised states $u = [\tilde{v}, \tilde{x},]^T$ are needed to generate a generalised response that encodes a path or trajectory. The second (state) equations enforce a coupling between neighbouring orders of motion of the hidden states and confer memory on the system.

At this point, readers familiar with standard state-space models may be wondering where all the extra equations in Equation 2 come from and, in particular, what the generalised motions; $w', w'', \ldots$ represent. These terms always exist but are ignored in standard treatments based on the theory of Markovian processes [6]. This is because standard Markovian (c.f., Wiener) processes have generalised motion that has infinite variance and are infinitely 'jagged' or rough. This means $w', w'', \ldots$ and $x'', x''', \ldots$ have no precision (inverse variance) and can be ignored with impunity. It is important to realise that this approximation is not appropriate for real or actual fluctuations, as noted at the inception of the standard theory; "a certain care must be taken in replacing an actual process by Markov process, since Markov processes have many special features, and, in particular, differ from the processes encountered in radio engineering by their lack of smoothness... any random process actually encountered in radio engineering is analytic, and all its derivative are finite with probability one" ([6], pp 122-124). So why have standard state-space models, and

their attending inversion schemes like Kalman filtering, dominated the literature over the past half-century? Partly because it is convenient to ignore generalised motion and partly because they furnish reasonable approximations to fluctuations over time-scales that exceed the correlation time of the random processes: "Thus the results obtained by applying the techniques of Markov process theory are valuable only to the extent to which they characterise just these 'large-scale' fluctuations" ([6], p 123). However, standard models fail at short time-scales. This is especially relevant in this paper because the brain has to model continuous sensory signals on a fast time-scale.

Having said this, it is possible to convert the generalised state-space model in Equation 2 into a standard form by expressing the components of generalised motion in terms of a standard [uncorrelated] Markovian process, $\varsigma(t)$:

$$w + \gamma_1 w' + \gamma_2 w'' + \ldots + \gamma_n w^{[n]} = \varsigma \Longrightarrow$$

$$\dot{x} = x'$$

$$\dot{x}' = x''$$

(3)

$$\vdots$$

$$\gamma_n \dot{x}^{[n]} = f(x,v) + \sum_{i=1}^{n} \left(\left(\gamma_i f_x - \gamma_{i-1} \right) x^{[i]} + \gamma_i f_v v^{[i]} \right) + \varsigma$$

$$\ln p(\theta) = \frac{1}{2} \ln \left| \Pi^{\theta} \right| - \frac{1}{2} \varepsilon^{\theta T} \Pi^{\theta} \varepsilon^{\theta}$$

$$\ln p(\lambda) = \frac{1}{2} \ln \left| \Pi^{\lambda} \right| - \frac{1}{2} \varepsilon^{\lambda T} \Pi^{\lambda} \varepsilon^{\lambda} \qquad (14)$$

$$\varepsilon^{\theta} = \theta - \eta^{\theta}$$

$$\varepsilon^{\lambda} = \lambda - \eta^{\lambda}$$

$$g_v = \begin{bmatrix} g_v^{(1)} & ¿ \\ 0 & ¿ \ddots ¿ g_v^{(m)} ¿ 0 ¿ \end{bmatrix} g_x = \begin{bmatrix} g_x^{(1)} & ¿ \\ 0 & ¿ \ddots ¿ g_x^{(m)} ¿ 0 ¿ \end{bmatrix} \qquad (13)$$

$$f_v = \begin{bmatrix} f_v^{(1)} & ¿ \\ \ddots & ¿ f_v^{(m)} \end{bmatrix} f_x = \begin{bmatrix} f_x^{(1)} & ¿ \\ \ddots & ¿ f_x^{(m)} \end{bmatrix}$$

Note that the partial derivatives of $g(x,v)$ have an extra row to accommodate the top level. To complete model specification we need priors on the parameters and hyperparameters. We will assume these are Gaussian, where (ignoring constants)

Summary. In this section, we have introduced hierarchical dynamic models in generalised coordinates of motion. These models are about as complicated as one could imagine; they comprise causes and hidden states, whose dynamics can be coupled with arbitrary (analytic) nonlinear functions. Furthermore, these states can have random fluctuations with unknown amplitude and arbitrary (analytic) autocorrelation functions. A key aspect of the model is its hierarchical form, which induces empirical priors on the causes. These recapitulate the constraints on hidden states, furnished by the hierarchy implicit in generalised motion. We now consider how these models are inverted.

Model Inversion

This section considers variational inversion of models under mean-field and Laplace approximations, with a special focus on HDMs. This treatment provides a heuristic summary of the material in [2]. Variational Bayes is a generic approach to model inversion that approximates the conditional density $p(\vartheta \mid y, m)$ on some model parameters, ϑ, given a model m and data y. This is achieved by optimising the sufficient statistics (e.g., mean and variance) of an approximate conditional density $q(\vartheta)$ with respect to a lower bound on the evidence (marginal or integrated likelihood) $p(y \mid m)$ of the model itself. These two quantities are used for inference on the parameters of any given model and on the model per se. [11-15]. The log-evidence for any parametric model can be expressed in terms of a free-energy $F(\tilde{y}, q)$ and a divergence term, for any density $q(\vartheta)$ on the unknown quantities

$$\ln p(\tilde{y} \mid m) = F + D\left(q(\vartheta) \,/\!/\, p(\vartheta \mid \tilde{y}, m)\right) \Longrightarrow \qquad (15)$$

$$F = \langle \ln p(\tilde{y}, \vartheta) \rangle_q - \langle \ln q(\vartheta) \rangle_q$$

The free-energy comprises the internal energy, $U(y, \vartheta) = \ln p(y, \vartheta)$ expected under $q(\vartheta)$ and an entropy term, which is a measure of its uncertainty. In this paper, energies are the negative of the corresponding quantities in physics; this ensures the free-energy increases with log-evidence. Equation 15 indicates that $F(\tilde{y}, q)$ is a lower-bound on the log-evidence because the cross-entropy or divergence term is always positive.

The objective is to optimise $q(\vartheta)$ by maximising the free-energy and then use $F \approx \ln p(\tilde{y} \mid m)$ as a lower-bound approximation to the log-evidence for model comparison or averaging. Maximising the free-energy minimises the divergence, rendering $q(\vartheta) \approx p(\vartheta \mid y, m)$ an approximate posterior, which is exact for simple (e.g., linear) systems. This can then be used for inference on the parameters of the model selected.

Invoking an arbitrary density, $q(\vartheta)$ converts a difficult integration problem (inherent in computing the evidence; see discussion) into an easier optimisation problem. This rests on inducing a bound that can be optimised with respect to $q(\vartheta)$. To

finesse optimisation, one usually assumes $q(\vartheta)$ factorises over a partition of the parameters

$$q(\vartheta) = \prod_i q(\vartheta^i) \qquad (16)$$

$$\vartheta = \{u, \theta, \lambda\}$$

In statistical physics this is called a mean-field approximation. This factorisation means that one assumes the dependencies between different sorts of parameters can be ignored. It is a ubiquitous assumption in statistics and machine learning. Perhaps the most common example is a partition into parameters coupling causes to responses and hyperparameters controlling the amplitude or variance of random effects. This partition greatly simplifies the calculation of things like t -tests and implies that, having seen some data, knowing their variance does not tell you anything more about their mean. Under our hierarchical dynamic model we will appeal to separation of temporal scales and assume, $q(\vartheta) = q(u(t))q(\theta)q(\lambda)$, where $u = [\tilde{v}, \tilde{x},]^T$ are generalised states. This means that, in addition to the partition into parameters and hyperparameters, we assume conditional independence between quantities that change (states) and quantities that do not (parameters and hyperparameters).

In this dynamic setting $q(u(t))$ and the free-energy become functionals of time. By analogy with Lagrangian mechanics, this calls on the notion of action. Action is the anti-derivative or path-integral of energy. We will denote the action associated with the free energy by $\acute{F}$, such that $\partial_t \acute{F} = F$. We now seek $q(\vartheta^i)$ that maximise the action. It is fairly easy to show [2] that the solution for the states is a function of their instantaneous energy, $U(t) := U(u/\theta, \lambda) = \ln p(\tilde{y}, u/\theta, \lambda)$

$$q(u(t)) \propto \exp(V(t)) \qquad (17)$$

$$V(t) = \langle U(t) \rangle_{q(\theta)q(\lambda)}$$

where $V(t) = \partial_t \acute{V}^u$ is their variational energy. The variational energy of the states is simply their instantaneous energy averaged over their Markov blanket (i.e., averaged over the conditional density of the parameters and hyperparameters). Because the states are time-varying quantities, their conditional density is a function of time-dependent energy. In contrast, the conditional density of the parameters and hyperparameters are functions of their variational action, which are fixed for a given period of observation.