[32] M. Shadlen and W. Newsome. Noise, neural codes and cortical organization. *Current Opinion in Neurobiology*, 4:569–579, 1994.

[33] A. F. M. Smith. Bayesian computational methods. *Philosophical Transactions of the Royal Society of London A*, 337:369–386, 1991.

[34] W. R. Softky and C. Koch. The highly irregular firing of cortical-cells is inconsistent with temporal integration of random epsps. *Journal Of Neuroscience*, 13(1):334–350, 1993.

[35] Roger D. Traub, Robert K. S. Wong, Richard Miles, and Hillary Michelson. A model of a CA3 hippocampal pyramidal neuron incorporating voltage-clamp data on intrinsic conductances. *Journal of Neurophysiology*, 66(2):635–650, 1991.

[36] M. Vanier and J. Bower. A comparison of automated parameter-searching methods for neural models. In *Proceedings of the 1995 Computational Neuroscience Conference (CNS*95), Monterey (CA)*. 1995.

[17] E.T. Jaynes. Bayesian methods: General background. In J.H. Justice, editor, *Maximum Entropy and Bayesian Methods in Statistics*, pages 1–25. Cambridge University Press, Cambridge, 1986.

[18] W. H. Jeffreys and J. O. Berger. Ockham's razor and Bayesian analysis. *American Scientist*, 80:64–72, 1992.

[19] Christof Koch and Idan Segev. *Methods in Neuronal Modeling.* MIT Press, Cambridge, Massachusetts, 1989.

[20] M. S. Lewicki. Bayesian modeling and classification of neural signals. *Neural Computation*, 6(5):1005–1030, 1994.

[21] D.J.C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[22] D.J.C. MacKay. A practical Bayesian framework for back-propagation networks. *Neural Computation*, 4(3):448–472, 1992.

[23] Z. F. Mainen and T. J. Sejnowski. Reliability of spike timing in neocortical neurons. *Science*, 268:1503–1506, 1995.

[24] David A. McCormick, Barry W. Connors, James W. Lighthall, and David A. Prince. Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex. *Journal of Neurophysiology*, 54(4):782–806, 1985.

[25] R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report, Department of Computer Science, University of Toronto, Canada, 1993.

[26] R. M. Neal. Bayesian learning for neural networks. PhD Thesis, Department of Computer Science, University of Toronto, Canada, 1995.

[27] J. W. Pratt, H. Raiffa, and R. Schlaifer. *Introduction to Statistical Decision Theory.* MIT Press, Cambridge, MA, 1995.

[28] S. J. Press. *Bayesian statistics: principles, models, and applications.* John Wiley, New York, 1989.

[29] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C, 2nd. Ed.* Cambridge University Press, New York, 1992.

[30] E. De Schutter and J. M. Bower. An active membrane model of the cerebellar Purkinje cell i. simulation of current clamps in slice. *Journal of Neurophysiology*, 71:375–400, 1994.

[31] E. De Schutter and J. M. Bower. An active membrane model of the cerebellar Purkinje cell ii. simulation of synaptic responses. *Journal of Neurophysiology*, 71:401–419, 1994.

References

[1] Charles H. Anderson. Unifying perspectives on neuronal codes and processing. In *Proceedings of the XIX International Workshop on Condensed Matter Theories*. Caracas, Venezuela, 1995.

[2] P. Baldi and J. M. Bower. Multi-unit recording analysis by Bayesian modeling and learning. 1996. submitted.

[3] P. Baldi and F. Pineda. Contrastive learning and neural oscillations. *Neural Computation*, 3(4):526–545, 1991.

[4] V. Barnett. *Comparative Statistical Inference*. John Wiley, New York, 1982.

[5] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Verlag, New York, 1985.

[6] J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, 10(1):3–66, 1995.

[7] J. M. Bower and D. Beeman. *The Book of Genesis*. Springer Verlag, New York, 1995.

[8] P. Cheeseman. An inquiry into computer understanding. *Computational Intelligence*, 4:57–142, 1988. with discussion.

[9] R.T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14:1–13, 1964.

[10] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel. The Helmholtz machine. *Neural Computation*, 7(5):889–904, 1995.

[11] P. C. Gregory and T. J. Loredo. A new method for the detection of a periodic signal of unknown shape and period. *The Astrophysical Journal*, 398:146–168, 1992.

[12] S.F. Gull. Developments in maximum entropy data analysis. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods*, pages 53–71. Kluwer, Dordrecht, 1989.

[13] Bertil Hille. *Ionic Channels of Excitable Membranes, 2nd. Ed.* Sinauer Associates Inc., Sunderland, Massachusetts, 1992.

[14] G. E. Hinton, P. Dayan, B. J. frey, and R. M. Neal. The wake-sleep algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995.

[15] P. G. Hoel, S. C. Port, and C. J. Stone. *Introduction to Probability Theory*. Houghton Mifflin, Boston, 1971.

[16] E. T. Jaynes. *Probability Theory: The Logic of Science*. 1995. In preparation.

Method:	Within-class M_4^{opt1}	Within-class M_4^{opt2}	Between-class M_4^{opt3}	Between-class M_3^{opt}	Whole-class C_4	Whole-class C_3
Variable-rate Poisson						
CMs run:	1000x with noise		1000x with noise		1x without noise	
Parameters:						
μ:	estimated from outputs of simple noise model runs		estimated from outputs of simple noise model runs		estimated as time of spikes in deterministic component	
σ:	estimated from outputs of simple noise model runs		estimated from outputs of simple noise model runs		2 $msec$	
C:	$1/(\sqrt{2\pi}\sigma)$		$1/(\sqrt{2\pi}\sigma)$		$1/(\sqrt{2\pi}\sigma)$	
bin size:	1 $msec$		1 $msec$		1 $msec$	
Gaussian ISI						
CMs run:	1000x with noise		1000x with noise		1x without noise	
Parameters:						
μ:	estimated from outputs of simple noise model runs		estimated from outputs of simple noise model runs		estimated as time of spikes in deterministic component	
σ:	estimated from outputs of simple noise model runs		estimated from outputs of simple noise model runs		1 $msec$	
x:	estimated separately for each current from outputs of simple noise model runs		estimated separately for each current from outputs of simple noise model runs		0.968	0.921
y:	not needed		not needed for M_4; see note for M_3		$(1-x)$ for $M=0$; else $(1-x)/2$	
T:	400 $msec$		400 $msec$		400 $msec$	

Table 6

conductance models with different parameters and with 3-conductance models with random parameter values from within the allowed range (not shown).

Fig. 4: The histogram of ISI standard deviation values for a 4-conductance model with noise. The parameters of this model were chosen randomly from within the parameter ranges. Note that the majority of the values cluster in the range (0,1) msec. This was analysis was repeated for 5 randomly chosen models in each of the two model classes (model 1 and model 2) and similar results were obtained in all cases. Absolute spike time standard deviation values were somewhat larger (not shown) but were mainly concentrated in the (0,2) msec range.

Fig. 5: (a) A raster plot of target model without noise vs. best matching model for two 4-compartment models without noise. These models were used in the "within-class" model comparisons. The parameters for each model were determined using the parameter-search methods described above. Each raster represents one spike. The response for different levels of current injection are offset on the y-axis, with the lowest level displayed as current 1 and the highest as current 6. The blue rasters represent the target model, the green rasters represent the first four-parameter model and the red rasters represent the second four-parameter model. The time axis begins at the beginning of the current injection *i.e.* at 50 *msec* after the beginning of the simulation. (b) Raster plots for the "between-class" model comparison, also without added noise. Here the blue rasters represent the target model, the green rasters represent the best matching four-parameter model and the red rasters represent the best matching three-parameter model. Note the extremely good match between both of the models and the target data, except for current 1, where the target and the four-parameter model have no spikes but the three-parameter model has a spike early in the trace.

Table 6: Parameter estimation methods used for the variable-rate Poisson model and for the Gaussian ISI model in all the comparisons.

Acknowledgements

The work of PB is supported by a grant from the ONR.

Figure Captions

Fig. 1: (a): The output of the target model without noise. Input currents are 0.1, 0.2, 0.4, 0.6, 0.8 and 1.0 nA, from bottom trace to top respectively. Traces represent membrane potential of the cell (V_m) and are offset on the y-axis for clarity. The current injection begins at 50 $msec$ and continues to the end of the trace (at 250 $msec$). Note the initial high rate of spiking in the top four traces, followed by pronounced spike frequency adaptation. (b): The output of the same model with noise, for five different current injections of 0.8 nA. To facilitate comparison, traces were overlaid. Note that the initial few spikes (four spikes in an initial burst) are nearly coincident for all traces, while later ones have larger standard deviations.

Fig. 2: A group of ISI distribution histograms from a four-conductance model with noise and a range of input currents. The parameter settings were chosen randomly from within the allowed range. Each distribution represents a particular spike which occurred in most or all of 1000 runs of the model; 9 spikes were chosen at random for display in this fashion. The x-axis is in units of seconds while the y-axis represents number per bin. The vertical line represents the position of the ISI for the model without noise. Note that the means of the distributions closely approximate the positions of the deterministic ISIs.

Fig. 3: The normalized covariance matrix of a sample 4-conductance model. The parameter settings were chosen randomly from within the allowed range. The model was run 1000 times with noise, and the standard deviations of the resulting ISI distributions were calculated for each spike. The x and y axes represent the spike number; the higher the spike number the later the spike in the trace. Spikes from different currents are given different spike numbers; thus spikes 1-2 are from current 1, spikes 3-7 are from current 2, etc. The bar scale on the left indicates the color map, with light colors representing positive covariance and darker colors representing negative covariance. Since the covariance matrix has been normalized all diagonal entries are equal to 1. Off-diagonal entries are usually close to 0, implying that spike times are largely independent of one another. Some anticorrelation exists between adjacent spikes in many cases. Additionally, the last current injection (spikes 65-106) shows a positive correlation between ISI times for the earlier spikes; in this case the amplitude of the input current was sufficiently large relative to the amount of current noise that the cell's firing pattern before spike frequency adaptation set in is quite regular, leading to high covariances. Similar results were obtained for other 4-

$$\alpha_n = \frac{-398.4 - 16000\ V_m}{-1 + exp(-(V_m + 0.0249)/0.005)} \tag{53}$$

$$\beta_n = \frac{250}{exp((V_m + 0.040)/0.040} \tag{54}$$

K_A channel:

$$\alpha_x = \frac{-938 - 20000\ V_m}{-1 + exp(-(V_m + 0.0469)/0.01} \tag{55}$$

$$\beta_x = \frac{348.25 + 17500\ V_m}{-1 + exp((V_m + 0.0199)/0.01)} \tag{56}$$

$$\alpha_y = \frac{1.6}{exp((V_m + 0.073)/0.018)} \tag{57}$$

$$\beta_y = \frac{50}{-1 + exp(-(V_m + 0.0499)/0.005} \tag{58}$$

K_M channel:

$$z_\infty = \frac{1.0}{1.0 + exp(-(V_m + 0.035)/0.01)} \tag{59}$$

$$\tau_z = \frac{1.0}{6.6 * exp((V_m + 0.035)/0.02) + exp(-(V_m + 0.035)/0.02)} \tag{60}$$

Appendix D: Fixed Parameters of the Model

$$cell\ area = 4\pi * (10\ \mu m)^2 = 1.2566 * 10^{-9}\ m^2 \tag{61}$$

$$C = CM * area = (0.008\ \frac{F}{m^2}) * (1.2566 * 10^{-9}\ m^2) = 1.00531 * 10^{-11}\ F \tag{62}$$

$$R_m = RM * area = (2.0\ \frac{ohm}{m^2}) * (1.2566 * 10^{-9}\ m^2) = 1.59155 * 10^{9}\ ohm \tag{63}$$

$$E_m = -0.077\ V \tag{64}$$

$$E_{N_a} = 0.055\ V \tag{65}$$

$$E_K = -0.090\ V \tag{66}$$

$$G_{Na} = \overline{G}_{Na}\ m^2\ h \tag{42}$$

$$G_{K_{DR}} = \overline{G}_{K_{DR}}\ n \tag{43}$$

$$G_{K_A} = \overline{G}_{K_A}\ x\ y \tag{44}$$

$$G_{K_M} = \overline{G}_{K_M}\ z \tag{45}$$

The parameters $\overline{G}_C$, where C represents an ionic channel species, are the variable parameters of the models. The gating variables m, h, n, x, y, and z all can be described by the Hodgkin-Huxley formalism ([13]) as:

$$\tau_X \frac{dX}{dt} = X_\infty - X \tag{46}$$

$$\tau_X = \frac{1}{\alpha_X + \beta_X} \tag{47}$$

$$X_\infty = \frac{\alpha_X}{\alpha_X + \beta_X} \tag{48}$$

for a given gating variable X, where τ_X, X_∞, α_X, and β_X are all functions of membrane potential V_m. α_X is the forward rate constant and β_X is the reverse rate constant for the channel opening reaction ([13]). These functions are listed below. In some cases α and β constants were used to compute X_∞ and τ values, while in other cases the latter values were calculated directly. Note that V_m is in volts. The channel models are adapted from those in [35] except for the K_M channel which is adapted from [19]. To get the cells to spike in a manner similar to that of regular-spiking cortical pyramidal cells the kinetics of the potassium channels were altered relative to those in the references.

$\underline{Na}$ channel:

$$\alpha_m = \frac{-15008 - 32000\ V_m}{-1 + exp(-(V_m + 0.0469)/0.004)} \tag{49}$$

$$\beta_m = \frac{5572 + 280000\ V_m}{-1 + exp((V_m + 0.0199)/0.005)} \tag{50}$$

$$\alpha_h = \frac{128}{exp((V_m - 0.043)/0.018)} \tag{51}$$

$$\beta_h = \frac{4000}{1 + exp(-(V_m - 0.020)/0.005)} \tag{52}$$

delayed-rectifier K channel:

Bayes' theorem is fundamental because it allows inversion: interchanging conditioning and non-conditioning propositions. In a sense, it embodies learning since it describes exactly how to update our degree of belief $p(X|I)$ in X, in light of the new piece of information provided by Y to obtain the new $p(X|Y,I)$. $p(X|I)$ is also called the prior probability (or "prior" for short) and $p(X|Y,I)$ the posterior probability with respect to Y (or "posterior" for short). This rule can obviously be iterated as information becomes available.

Appendix B: Error Bars

For illustration purposes, consider a modeling situation with one parameter θ, and a uniform prior. Let $f(\theta) = -\log p(D|\theta)$ be the negative log-likelihood of the data. Under mild differentiability conditions, a maximum likelihood estimator θ_0 satisfies $f'(\theta_0) = 0$. Therefore in the neighborhood of θ, we can expand $f(\theta)$ in a Taylor's series:

$$f(\theta) \approx f(\theta_0) + \frac{1}{2} f''(\theta_0)(\theta - \theta_0)^2 \tag{38}$$

or

$$p(D|\theta) = e^{-f(\theta)} \approx C e^{-\frac{1}{2} f''(\theta_0)(\theta-\theta_0)^2} \tag{39}$$

where $C = e^{f(\theta_0)}$. So the likelihood, as well as the posterior $p(\theta|M)$, locally behave like a Gaussian, with a standard deviation $1/\sqrt{f''(\theta_0)}$, associated with the curvature of f.

Appendix C: Model Equations

The basic equation for the model is the current-balance equation:

$$C\frac{dV_m}{dt} + \frac{V_m - E_m}{R_m} + I_{ionic} = I_{stim} \tag{40}$$

where C is the membrane capacitance, V_m is the membrane potential, R_m is the membrane resistance, E_m is the "leak current" reversal potential (an ill-defined and yet nearly universal convention among modelers [19], [7]), I_{ionic} is the sum of currents through voltage-dependent ionic conductances, and I_{stim} is the input current. I_{ionic} can be further subdivided as follows:

$$\begin{aligned} I_{ionic} &= I_{Na} + I_{K_{DR}} + I_{K_A} + I_{K_M} \\ &= G_{Na}(V_m - E_{Na}) + G_{K_{DR}}(V_m - E_K) \\ &\quad + G_{K_A}(V_m - E_K) + G_{K_M}(V_m - E_K) \end{aligned} \tag{41}$$

where G_X is the conductance value for channel X, E_{Na} is the reversal potential for Na, and E_K is the reversal potential for K. These conductance values are time- and voltage-dependent according to the following equations:

$$\pi(\bar{X}|I) = F[\pi(X|I)] \tag{33}$$

The third axiom is only slightly more complex. Consider this time two neurons A and B and the corresponding 4 possible joint states. Then our degree of belief that A is on and B is off, for instance, naturally depends on our degree of belief that neuron A is on, and our degree of belief that neuron B is off, *knowing that A is on.* Again, it is sensible that this relationship be independent of the neurons considered and the type of background information I. Thus, in mathematical terms, the third axiom states that there exists a function G such that:

$$\pi(X, Y|I) = G[\pi(X|I), \pi(Y|X, I)] \tag{34}$$

So far, we have not said much about the information I. I is a proposition corresponding to the conjunction of all the available pieces of information. I can represent background knowledge, such as general anatomical or physiological information in neuroscience. I can also include specific experimental or other data. When it is necessary to focus on a particular corpus of data D, we can write $I = (I, D)$. In any case, I is not necessarily fixed and can be augmented with, or replaced by, any number of symbols representing propositions, as already seen in the right hand side of equation (34). When data is acquired sequentially, for instance, we may write $I = (I, D_1, ..., D_n)$. Inside a discussion where I is well defined and fixed, it can be dropped altogether from the equations.

The three axioms above determine entirely, up to scaling, how to calculate degrees of belief. In particular, one can prove that there is always a rescaling w of degrees of beliefs such that $p(X|I) = w(\pi(X|I))$ is in $[0, 1]$. Furthermore, p is unique and satisfies all the rules of probability. Specifically, if degrees of belief are restricted to the $[0, 1]$ interval, then the function F and G must be given by: $F(x) = 1 - x$ and $G(x, y) = xy$. The corresponding proof will not be given here and can be found, for instance, in ([9], [16]). As a result, the second axiom can be rewritten as the Sum Rule of probability:

$$p(X|I) + p(\bar{X}|I) = 1 \tag{35}$$

And the third axiom as the Product Rule:

$$p(X, Y|I) = p(X|I)p(Y|X, I) \tag{36}$$

From here on, we can then replace degrees of confidence by probabilities. Notice that if uncertainties are removed, that is if $p(X|I)$ is 0 or 1, then equation (35) and equation (36) yield, as a special case, the two basic rules of deduction or Boolean algebra, for the negation and conjunction of propositions [(1) "X or $\bar{X}$ "is always true; (2) "X and Y" is true if and only if both X and Y are true].

The immediate consequence of equation (35) and equation (36) is the so-called Bayes' theorem:

$$p(X|Y, I) = \frac{p(Y|X, I)p(X|I)}{p(Y|I)} = p(X|I)\frac{p(Y|X, I)}{p(Y|I)} \tag{37}$$

complex one composed of the conjunction of many more elementary propositions. A model M can also be viewed as an hypothesis. The difference is that models tend to be very complex hypotheses involving a large number of parameters. In discussions where parameters are important, we will consider that $M = M(\theta)$ where θ is the vector of all parameters. A complex model M can easily be reduced to a binary proposition in the form: "model M accounts for data D with an error level ϵ" (this is a vague statement that will be made more precise in the following discussion). But for any purpose, in what follows there is no real distinction between models and hypotheses.

Whereas propositions are either true or false, we wish to reason in the presence of uncertainty. Therefore the next step is to consider that, given a certain amount of information I, we can associate with each hypothesis a degree of plausibility or confidence (also called in the literature degree or strength of belief). Let us represent it by the symbol $\pi(X|I)$. While $\pi(X|I)$ is just a symbol for now, it is clear that in order to have a scientific discourse one should be able to compare degrees of confidence. That is, for any two propositions X and Y, either we believe in X more than in Y, or we believe in Y more than in X, or we believe in both equally. Let us use the symbol "$>$" to denote this relationship, so that we write $\pi(X|I) > \pi(Y|I)$ if and only if X is more plausible than Y. It would be very hard not to agree that in order for things to be sensible, the relationship "$>$" should be transitive. That is if X is more plausible than Y, and Y is more plausible than Z, then X must also be more sensible than Z. More formally this is the first axiom:

$$\pi(X|I) > \pi(Y|I) \quad \text{and} \quad \pi(Y|I) > \pi(Z|I) \Rightarrow \pi(X|I) > \pi(Z|I) \tag{32}$$

This axiom is trivial; it has, however, an important consequence: "$>$" is an ordering relationship, and therefore degrees of belief can be expressed by real numbers. That is, from now on $\pi(X|I)$ represents a number. This does not mean that such a number is easy to calculate, but merely that such a number exists, and the ordering among hypotheses is reflected in the ordering of real numbers. To proceed any further and stand a chance of calculating degrees of beliefs we need additional axioms or rules for relating numbers representing strengths of belief. The amazing fact is that only two additional axioms are needed to constrain the theory entirely. This axiomatic presentation is usually attributed to Cox and Jaynes ([9],[17]). To better understand these two remaining axioms, the reader may imagine a world of very simple neurons, where at each instant in time a given neuron can either be on or off. Thus, all the elementary hypotheses or propositions in this world, at a given time, have the simple form: "neuron A is on" or "neuron A is off". Clearly, the more confident we are that neuron A is on, the less confident we are that neuron A is off. So, for any given proposition X, there should be a relationship between $\pi(X|I)$ and $\pi(\bar{X}|I)$. Without assuming anything about this relationship, it is sensible to consider that, everything else being equal, the relationship should be the same for all neurons and for all types of background information, i.e. for all propositions X and I. Thus, in mathematical terms, the second axiom states that there exists a function F such that:

a model such as the variable-rate Poisson model can already be used immediately, in conjunction with the most complex existing CMs, to estimate likelihoods of large data sets and compare models pairwise.

Finally, there is an entirely different application of Bayesian ideas to CNS that dates at least all the way back to Helmholtz. This is the idea that brains in general, and perceptual systems in particular, are inference machines that must function in a Bayesian way. Perceptual systems [2] must continually build and update a statistical model of the outside world, using data provided by the sensory ports, and priors provided by internal expectations that include a possibly strong genetic component. Strong priors at each time are essential in order for the update to be tractable on short time scales compatible with survival. In its stronger form, this general approach concludes that neurons in the brain are involved in population coding of conditional probabilities ([1]).

This is one of the current ideas that emphasizes the role of internal expectations in perception and, for instance, the putatively associated role of massive feedback connections in the visual or olfactory systems. For instance, in ([14], [10]) a recognition model is associated with the feed-forward pathways, and a generative model is associated with the feed-back pathways. A key point is the existence of unsupervised alternating learning algorithms for each pathway. Connections in each pathway are updated in alternation using the activity and current status of the opposite pathway. Such alternating algorithms can be viewed as a sort of clock for Bayesian update, and provide a suggestive hypothetical interpretation for some of the rhythms and oscillations frequently found in nervous tissues ([3]). Needless to say, considerable experimental work remains to be done to establish or invalidate any such vague ideas.

In any case, because Bayes dictates THE proper way of conducting inference and because inference is inherently computationally intensive, it is hard to see how the brain and its perceptual systems could implement anything other than some approximation to Bayesian inference. The real question, of course, is not whether the brain is Bayesian or not, but rather to which extent and how? That is, how close are the approximations and the details of their implementations?. In the quest for an answer, however, it appears likely to us that Bayesian ideas can provide both a solid foundation for detailed low level models, such as CMs, as well as useful guidance for more global theories of brain organization and function.

Appendix A: The Cox-Jaynes axioms

The objects we deal with in inference are propositions about the world. For instance, a typical proposition X is: "neuron A is firing an action potential at time t". A proposition is either true or false, and we denote by $\bar{X}$ the complement of a proposition X. A hypothesis H about the world is a proposition, albeit a possibly

[2] In this direction, an even broader formal foundation to motor control and behavior can be provided by Decision Theory. There exist axiomatic versions of Decision Theory that encompass Bayesianism as a sub-theory ([5], [27]).

discourse on models, to determine how well they fit the data, and ultimately to compare models and hypotheses. We have demonstrated three specific ways for constructing a probabilistic CM out of its deterministic component. Third, the Bayesian approach clarifies how to proceed with inference *i.e.* how to compare models with each other and with the data objectively. In particular, it provides unambiguous unique answers to well-posed questions. It defines the set of rules required to play a fair modeling game.

The Bayesian framework is entirely open to, and actually encourages, questioning of any piece of information. It deals with the subjectivity inherent in the modeling process not by sweeping it under the rug, but rather by incorporating it up-front in the modeling process. It is fundamentally an iterative process in which models are progressively refined. New ideas are also likely to emerge from the marriage of CMs and Bayesian methods, as has been the case in other fields ([21], [22]), [26], and references therein). For instance, the variable-rate Poisson model was in part suggested by the Bayesian approach, and the difficulties encountered with the Gaussian ISI model. The Bayesian approach can lead to a better understanding of the weaknesses of a model, and thereby help in generating better models. In addition, having an objective way of comparing models, and making predictions based on models, will become more important as the number, scope and complexity of CMs representing the same brain structures increases. It is our belief that CMs are still at a very early stage of development: for instance, there are not many Purkinje cell CMs in the world. Thus issues of model comparison and prediction are still somewhat hidden in the background, but will progressively become more central as neurobiological databases grow in size and complexity.

In the present case, the two main statistical models for the spike timings we used have complementary advantages and disadvantages. The variable-rate Poisson model handles extra or missing spikes more gracefully than the Gaussian ISI model, since it independently computes the probability of getting a spike in every time interval. The Gaussian ISI has a very ad-hoc and inelegant procedure for dealing with missing or extra spikes which requires the estimation of (at least) two new parameters. However, the Poisson model underpenalizes missing spikes (it can be shown that the penalty for a missing spike is roughly 1 log unit, in contrast to much higher values for the Gaussian ISI model on average). Thus neither model is an ideal model of the spike-generating process in this particular data set, and probably more models need to be developed in the future. The variable-rate Poisson model, however, is extremely simple and flexible and can provide at least a first approximation in a wide range of situations.

The main drawback of the Bayesian approach is that it can be computationally intensive, especially when averages need to be computed over high-dimensional distributions. For the largest CM models, it is unlikely that one may be able to carry out a complete Bayesian integration on currently available computers. But continuing progress in Monte Carlo ([33], [6]) and other approximation techniques, as well as steady increases in raw computing power in workstations and parallel computers, is very encouraging and will hopefully make these methods feasible for compartmental neural models at a reasonable level of complexity. It should be clear, however, that

	variable-rate Poisson model		Gaussian ISI model	
	C_4	C_3	C_4	C_3
$\log P(D\|M)$	-86.04	-136.86	-41.02	-81.45
$\log P(M)$	-0.69	-0.69	-0.69	-0.69
$\log P(M\|D)$	-86.73	-137.55	-41.71	-82.14

Table 5: The likelihoods ($\log P(D|M)$), priors ($\log P(M)$, and posteriors ($\log P(M|D)$) (up to constants) of the whole-class comparison using both the variable-rate Poisson model and the Gaussian ISI model to determine the likelihoods.

class likelihood for the 3-conductance class is of the same order as the likelihood of the "optimal" model in this class (in fact, since the likelihood of the optimal model is an upper bound, the class likelihood is most likely greater than that of the optimal model). This is a consequence of the fact that the estimate of y for the class likelihood was much less accurate than for the optimal model (where it came from running the model with noise 1000 times) and therefore may have underestimated the penalty for missing/extra spikes for the most probable models.

Having a mapping of the space of likelihoods can be useful for other purposes as well. For instance, one could determine whether there is one unique maximum in the likelihood function or several, and how peaked the likelihoods are around each local maximum. One can also use this data for prediction problems, such as the problem of predicting when the first spike is most likely to occur given the class. Note also that if one has the resources to do a brute-force mapping of the parameter-space, using an optimization method for a parameter search is unnecessary. In general, this will not be possible, but Markov Chain Monte Carlo (MCMC) methods are being developed for more efficient sampling of high-dimensional distributions.

5 Discussion

We have briefly presented the Bayesian approach to modeling, and how it could be applied to CNS in general, and to CMs in particular. The main advantage of a Bayesian approach is obvious: it provides a principled approach to inference, with a strong foundation in probability theory. In fact one of the most compelling reasons in favor of Bayesian induction is its unicity, under a very small set of common sense axioms. We grant that mathematicians may be more receptive to such an argument than biologists.

We have also shown that it would not be too difficult to make current CMs more Bayesian, at least at the most basic level. The main advantage of Bayesian methods for CMs is the clarification of a number of issues, at three different levels at least. First, a Bayesian approach forces one to clarify the prior knowledge, the data, and the hypotheses. Second, and this is perhaps the main lesson here, CM modelers *must* come to grip with the issues of noise and variability in the data, and in a quantifiable way. Without this step, it is impossible to have a rigorous scientific

entire parameter space by dividing the range of each parameter into 16 segments, running a model for each parameter set, and calculating the likelihood for each model. Since this requires $16^4 = 65536$ simulations for the 4-parameter case and $16^3 = 4096$ simulations for the 3-parameter case, we cannot run each one of the corresponding simple noise models 1000 times to estimate the means and standard deviations of the spike timings or the ISIs. Therefore we choose $\vec{\mu} = \vec{\mu}(\theta)$ to be the position of the spike in the corresponding deterministic component with parameters θ, both for the variable-rate Poisson and the Gaussian ISI models. The σ values used in the whole-class comparisons were fixed and estimated by running 5 models from each class (with parameters randomly chosen from within the allowable range) 1000 times with noise. Although the value of σ differs between models and between currents in the same model, in general the σ for the ISIs is roughly 1 $msec$ while that for the actual spike times is roughly 2 $msec$. The spike times have a larger σ since the variation in the timing of a current spike is not only due to the variation in the last interspike interval but also to the variation in all previous interspike intervals for that trace. Since we are sampling the entire parameter space, regimes where the candidate models produce extra or missing spikes are very common. This poses a problem for the Gaussian ISI model because we cannot estimate the x or y parameters directly for each model, since that would require running each of the thousands of models with noise a large number of times. Instead we used the 5 random models for each class, ran each of them 1000 times with noise, calculated an average value of x for each class and estimated y using equation (19). The average found was $x = 0.968$ (resp. 0.921) for the 4-conductance case (resp. 3-conductance). So, in the variable-rate Poisson case, we are comparing the classes $C_i(\vec{\mu}(\theta), \sigma = 2, C = 1/2\sqrt{2\pi})$ for $i = 3$ or 4. In the Gaussian ISI case, the classes are $C_4(\vec{\mu}(\theta), \sigma = 1, x = 0.968, y = (1-x)/2)[(1-x) \quad \text{when} \quad M = 0])$ and $C_3(\vec{\mu}(\theta), \sigma = 1, x = 0.921, y = (1-x)/2))[(1-x) \quad \text{when} \quad M = 0])$.

We approximate the likelihooods using our samples by

$$\int_{\theta \in C} p(D|\log(\theta), C) d\log(\theta) \approx \sum p(D|\log(\theta), C) \Delta\log(\theta) \tag{29}$$

and (after each parameter range has been divided into 16 increments):

$$\Delta\log(\theta) = (1/16 * 2\ log(4))^4 \tag{30}$$

(or $(1/16 * 2log(4))^3$ for the 3-parameter case) which gives:

$$\begin{aligned} p(D|C) &\approx 1/(2\log(4)))^i * (1/16 * 2\ log(4))^i * \sum(p(D|\log(\theta))) \\ &\approx 1/(16^i) * \sum(p(D|\log(\theta))) \end{aligned} \tag{31}$$

with $i = 3$ for the 3-conductance class and $i = 4$ for the four-conductance class (this is just the average of the likelihood over the class). The results are shown in Table 5. It can be seen that the class of 3-conductance models is quite inferior to the class of 4-conductance models, regardless of which statistical model (variable-rate Poisson or Gaussian ISI) was used. Note also that with the Gaussian ISI model the

parameter	target	4-conductance	3-conductance
Na	800.000	776.247	755.462
K	100.000	100.168	101.522
K_A	50.000	38.116	not applicable
K_M	100.000	102.119	104.974

Table 3: The final parameter values for a four-conductance and a three-conductance model, compared with the parameter values for the target model. The parameters were determined by a parameter-search algorithm and represent maximal conductances of the ionic channel types (in S/m^2). Note that the four-parameter model is different from the two four-parameter models used in the within-class comparison.

	variable-rate Poisson model		Gaussian ISI model	
	M_4^{opt3}	M_3^{opt}	M_4^{opt3}	M_3^{opt}
$\log P(D\|M)$	-69.31	-77.15	-3.24	$<$ -81.45
$\log P(M)$	-4.08	-3.06	-4.08	-3.06
$\log P(M\|D)$	-73.39	-80.21	-7.32	$<$ -84.51

Table 4: The likelihoods ($\log P(D|M)$), priors ($\log P(M)$, and posteriors ($\log P(M|D)$) (up to a constant) of the between-class comparison using both the variable-rate Poisson model and the Gaussian ISI model to determine the likelihoods.

in each of the 1000 runs, so we can only estimate an upper bound 0.001 for the parameter y in the Gaussian ISI model (i.e. no spike will occur in less than one in a thousand trials). A further step could consist in putting a prior on y also, and estimating its posterior. The results of the Bayesian analysis are shown in Table 4.

Note that the log probability for the 3-conductance model is much smaller than that for the 4-conductance model, both for the variable-rate and Gaussian ISI models. The magnitude of the difference is larger for the Gaussian ISI model than for the variable-rate Poisson model, since, as discussed below, the Poisson model does not penalize missing spikes as harshly as does the Gaussian ISI model. These results are interesting in that our intuition was shown to be wrong; the K_A current appears to have a significant function in the context of this model. Note also that even though the prior for the M_3 model gives a smaller penalty than for the M_4 model (penalizing a more complex model), this has little effect on the final results, since the likelihood terms are much larger. The results of the Bayesian analysis are shown in Table 4, both for the variable-rate Poisson model ($M_i(\vec{\mu}, \vec{\sigma}, \vec{C})$) and the Gaussian ISI model ($M_i(\vec{\mu}, \vec{\sigma}, \vec{x}, \vec{y})$) ($i = 3$ or 4), with parameters determined as described above (see also Table 6).

4.7 Comparing Entire Model Classes

Here we use equation (11) to calculate the likelihood values in order to compare the C_3 and C_4 model classes. To approximate the integral (11), we sample the

	variable-rate Poisson model		Gaussian ISI model	
	M_4^{opt3}	M_3^{opt}	M_4^{opt3}	M_3^{opt}
$\log P(D\|M)$	-73.83	-90.38	-8.81	-18.84
$\log P(M)$	-4.08	-4.08	-4.08	-4.08
$\log P(M\|D)$	-77.91	-94.46	-12.89	-22.92

Table 2: The likelihoods ($\log P(D|M)$), priors ($\log P(M)$, and posteriors ($\log P(M|D)$) (up to a constant) of the within-class comparison, using both the variable-rate Poisson model and the Gaussian ISI model.

given the data set, regardless of which statistical model is used. This is not surprising since the parameter values in model 1 are closer to the "target" parameters than those of model 2 (with the exception of parameter 3, corresponding to the K_A conductance, which is slightly closer to the correct value in model 2). Looking at the spike outputs of the corresponding deterministic components of the models in Figure 5(a), we would agree that model 1 matches the data better than model 2. Bayesian methods have allowed us to determine this in a precise quantitative manner.

4.6 Comparing Individual Models from Different Classes

When comparing models from different classes, one may take the likelihood of the classes into account, so that Bayes' theorem can be written as:

$$p(M, C|D) = \frac{p(D|M, C)p(M|C)p(C)}{p(D)} \tag{28}$$

In the present example, we will assume a uniform prior over the two model classes $p(C_3) = p(C_4) = 1/2$, so that in fact we can ignore the term $p(C)$. The term $p(M|C)$ is the prior on the model given the class and is calculated as in section 3.1. Thus the equation is the same as that used in the previous section, except for a multiplicative constant which is irrelevant to model comparison.

We again ran two parameter searches (one for each class) to determine optimal models in each class. In Table 3, we show the parameters describing the target model as well as those of the two models found by the parameter search. Note that the final conductance values obtained by the parameter-search method are extremely close to the "correct" values. Note also that even though the three-conductance model has no K_A conductance, the other parameter values are very close to the original values as well. In Figure 5(b), we show the spike outputs of the two final models without noise as raster plots, compared to the original target model without noise.

The four-parameter model matches the spike times (and hence the ISIs) of the target data set almost perfectly, while the three-parameter model also generates an excellent match, with the exception of a spurious spike at the lowest level of current injection. In order to calculate likelihoods, we ran the simple noise version of both models 1000 times and estimated the means and standard deviations of each spike time and ISI. The three-parameter model with noise generated a spurious spike

parameter	target	model 1	model 2
Na	800.000	530.37	2546.44
K	100.000	119.74	135.07
K_A	50.000	25.51	31.64
K_M	100.000	93.40	37.38

Table 1: The final parameter values for two models from the same class, compared with the parameter values for the target model. The parameters were determined by a parameter-search algorithm and represent maximal conductances of the ionic channel types (in S/m^2).

function used is related to the likelihood of the corresponding Gaussian ISI model. M_4^{opt1}, M_4^{opt2}... denote different 4-compartment models derived by this procedure on different runs (Table 6).

4.5 Comparing Individual Models from the Same Class

As the most basic test of the Bayesian method, we chose two independent models from the class of four-conductance models by running two separate parameter searches. Since the parameter search algorithm is stochastic, this produced two different candidate models M_4^{opt1} and M_4^{opt2}, both of which matched the spike timings of the data set within a fairly small error. The spike outputs of the deterministic component of the candidate models are shown as raster plots in Figure 5(a). In this case, the two corresponding simple noise models produced the same number of spikes per trace almost always, so no penalty for missing/extra spikes had to be imposed. Specifically, in the Gaussian ISI model, the parameter x is always very close to 1 and is estimated by running the corresponding simple noise model 1000 times. The parameter y can be estimated using (19), but is not needed because $M = K$ and only (26) is used for the likelihoods.

The parameters of the resulting models are shown in Table 1, along with the original parameters of the target model. It is important to notice that because of the approximations we make and the search procedure, the two models derived have a good data likelihood, but do not necessarily correspond to a global maximum of the likelihood.

To compare two models, one needs to compare their posterior probabilities probabilities $p(M|D)$, using (4). Using (9) for the priors, and (15) or (20) for the likelihood, it is easy to compute the log-posterior probabilities of each model to within a normalizing factor. Notice that the term associated with the probability of the data ($P(D)$) is the same for both models, and therefore can be ignored in the comparison. The prior term could also have been ignored, since the prior used is uniform over the class. The results of the Bayesian analysis are shown in Table 2, using both the variable-rate Poisson model ($M_4(\vec{\mu}, \vec{\sigma}, \vec{C})$) and the Gaussian ISI model ($M_4(\vec{\mu}, \vec{\sigma}, \vec{x}, \vec{y})$) to estimate the likelihoods, with parameters determined as described above (see also Table 6).

The analysis shows that model 1 is significantly more probable than model two

60%) cluster in a single bin, corresponding to standard deviations ranging from 0 to 0.55 ms. Thus, to a first approximation, we see that:

1. The various ISIs have a Gaussian distribution and are independent (the covariance matrix terms are typically close to 0 away from the main diagonal).

2. The mean of each Gaussian can be approximated by using the corresponding value in the same model run deterministically, i.e. without any noise.

3. The standard deviations of the Gaussians are typically in the $[0, 5]$ ms range. Most often they are close to 1 ms, so that whenever necessary we can reasonably use $\sigma_i = 1$ ms.

The procedure to obtain the parameters x and y varied depending on the comparison being performed. In sections 4.5 and 4.6, we ran the corresponding simple noise models 1000 times and used the incidence of missing/extra spikes to determine x and y separately for each level of current injection. In the case of the class comparison (section 4.7), this was not feasible, so we chose several models in each class with parameters randomly distributed in the allowable range. We used these to generate fixed estimates for x and y, which we then used for all models and all currents. The estimates were $x = 0.968$ (resp. $x = 0.921$) for the 4-conductance (resp. 3-conductance) class, with $y = (1-x)/2$ ($y = (1-x)$ when $M = 0$) (Table 6), as described above. We also used a fixed value of $T = 400$ $msec$, twice the length of the current stimulus.

While the general question of the approximation of the simple noise models by the corresponding variable-rate Poisson models or Gaussian ISI models could receive a detailed Bayesian treatment, it is not the focus of the present study. From now on, the focus is on how well these models fit the given data set. Finally, notice that another underlying assumption we use in the comparisons is that the level of the current injection in each trace is known. This is a standard assumption in real life situations, related to how the data is acquired under experimental control conditions.

4.4 Parameter Estimation: Determining Optimal Parameter Values

For the "within-class" and "between-class" comparisons, we would like to choose reasonably optimal parameter sets for models in one or both classes. In this case, we use an automated parameter-search procedure to find a good match between the candidate models and the target model. The details of the search procedure itself are irrelevant to the Bayesian comparison, but are briefly described here for completeness. To speed up parameter searches, a modified target data set is taken by running the target model without noise. We use simulated annealing, with a modified simplex gradient-descent procedure ([29]), for the optimization itself. We have previously compared a number of parameter-searching algorithms for neural models ([36]), and found this one to be highly efficient. The optimization algorithm systematically varies the values of the adjustable parameters (four parameters for class 1, three parameters for class 2) until a best match is found. The matching

We will use subscripts to refer to the number of parameters of the deterministic component of a model or a class. Thus, for instance, $C_4(\vec{\mu}, 2, 1/(2\sqrt{2\pi}))$ denotes the sub-class of variable-rate Poisson models, with 4 variable conductance parameters in the underlying deterministic component, with $\vec{\mu}$ determined by the deterministic component, and $\sigma = 2msec$, and $C = 1/(2\sqrt{2\pi})$. In what follows, we do not distinguish between classes and sub-classes, as things should be clear from the context.

The second model class (C_3) is identical to the first one, except that we remove the K_A conductance from the model. This represents the case where the evidence for the presence of this conductance in the real cell is inconclusive. We remove the K_A conductance specifically because of our intuition that this conductance has a very small effect on the ability of the model to generate outputs that match the target data set. The Bayesian methods allow us to test this intuition in a quantitative and precise manner. Thus each model in the second class is characterized by three parameters.

4.3.2 Probabilistic Component

As discussed in the previous section, we use both the variable-rate Poisson model and the Gaussian ISI model to estimate likelihoods.

For the variable-rate Poisson model, we used a bin size of 1 $msec$, on the assumption that there is a vanishingly low probability of having more than one spike in the bin. The height of the Gaussian centered on each spike was set to be equal to $1/(\sqrt{2\pi}\sigma)$, as described above. The only other parameters needed are the means and standard deviations of the Gaussian distributions centered at each spike. For the comparisons of individual models, we ran the models with noise 1000 times and calculated the means and standard deviations of all spikes. We used the means and standard deviation of each particular spike (*e.g.* the second spike at current level 3) to set these parameters. For the class comparison (section 4.7) this was not computationally feasible, so we ran the corresponding deterministic component and used the spike times to determine the μ's, with a standard deviation of 2 $msec$. We used a slightly larger standard deviation than for the Gaussian ISI model (see below) because we observed that the standard deviations of the spike times were typically somewhat larger than those for the ISIs (data not shown) (see Table 6).

For the Gaussian ISI model, we picked several models from both the 3- and 4-parameter classes with conductance parameters randomly chosen within the prior range and ran each of them 1000 times with random Gaussian noise with mean 0 nA and variance 0.1 nA. For each model, we computed the covariance matrices of the ISIs across runs and across different current injections. Several sample ISI histograms are shown in Figure 2. The line through the distributions represent the ISI of the deterministic component of the model. Notice that this line is very close to the mean. A sample (normalized) covariance matrix is shown using a grayscale colormap in Figure 3, with white representing high covariance, black representing anticovariance, and gray representing no covariance. The covariance is very close to zero away from the main diagonal. In Figure 4 we show a histogram of standard deviations for the ISIs of a typical model. Most of the standard deviations (over

(K_A), and slow non-inactivating K or "M-current" (K_M). The kinetics of the channels are adapted from models of hippocampal conductances for the first three channels ([35]), and from a model of bullfrog sympathetic ganglion cell in the case of K_M ([19]). These particular conductances can reproduce a number of realistic behaviors. In particular, they can produce spiking outputs reasonably similar to those of regular-spiking cortical pyramidal cells ([24]). Furthermore, with the inputs we used (see below) a range of behaviors, from not spiking to rapidly spiking, can be observed, with significant spike frequency adaptation, similar to what is seen in real pyramidal cells.

The mathematical details of the model are given in appendices C and D. We refer to this model as the "target" since we seek to generate a model that replicates the data produced by this model. The inputs to the target are a series of six constant current injections, with intensities ranging from 0.1 nA to 1.0 nA. The duration of the input stimuli is 200 ms, and the stimuli start 50 ms after the beginning of the simulation. The 50 ms pre-stimulus period was sufficient to cause the target model without noise to settle into a steady state, where "steady state" is defined as being the state where all variables (e.g. membrane potential) are constant. Variability in the data is simulated by adding a certain amount of independent Gaussian noise to the injected current at each time step. The mean amplitude of the injected noise is 0 nA, with standard deviation 0.1 nA. Note that this represents a fairly large amount of added noise: the standard deviation of the noise ranges from 10% to 100% of the base current. The simulation were performed using the Crank-Nicolson implicit integration method ([19]), with time steps of 20 μs.

While the data itself is generated here by a simple noise CM, from now on this fact will be entirely forgotten, and the data studied as if it had originated in a true recording session. The data consists of a sequence of interspike intervals. The first interspike interval is operationally defined as the time between current onset and the first spike. For speed purposes, we chose to use interspike intervals (ISIs) as the basis for our model comparisons, rather than directly comparing membrane potential (V_m) traces. The Bayesian framework can of course be applied also to V_m traces. This would be necessary, for instance, if the data consisted of voltage recordings from a non-spiking cell. Figure 1a shows the voltage output of the target model without noise for the six currents used, while Figure 1b shows the output of the same model with noise for five different current injections of 0.8 nA.

4.3 The models

4.3.1 Deterministic Component

Two candidate model classes are used to illustrate the Bayesian approach to model comparison. The first model class is identical to the target model, except that the maximal conductances of the four ionic conductances are not known. This corresponds to the case where the electrotonic structure of a cell, and the identity and kinetics of all voltage-dependent ionic channels, has been established, but the maximal conductance values of these channels are unknown. Thus the deterministic component of each model in the first class (C_4) is characterized by four parameters.

next section). In the following simulations, we use the Gaussian ISI model in two forms (see Table 6):

- $M(\vec{\mu}, \vec{\sigma}, \vec{x}, \vec{y})$ where all the parameters are essentially estimated from the corresponding simple noise model (used in the individual model comparisons; see below);
- $M(\vec{\mu}, 1, x = 0.968 \quad \text{or} \quad 0.921, y = (1-x)/2 \quad \text{or} \quad y = (1-x) \quad \text{for} \quad M = 0)$(used in the whole-class comparisons).

4 Examples

4.1 Objectives and methods

The previous sections illustrate the feasibility of using Bayesian methods on CMs in principle. In this section we demonstrate this approach using actual compartmental model simulations. The first question we ask is: given a data set and two candidate models, which model is more likely? We do this in two different cases. In the first case (the "within-class" case), we choose two models from the same class (we will define what we mean by class below). In the second case (the "between-class" case), we choose one model from each of two different classes and compare them.

The second question we ask is: which *class* is more probable given the available data? This is the "whole-class" case. It serves to demonstrate that Bayesian methods are not limited to comparing individual models, but can be used to address any inference question.

We use both variable-rate Poisson models $M = M(\vec{\mu}, \vec{\sigma}, \vec{C})$, and Gaussian ISI models $M = M(\vec{\mu}, \vec{\sigma}, \vec{x}, \vec{y})$ to calculate all the likelihoods. It would have been easy to carry out all the comparisons with a fixed setting of the parameters. Instead, we have tried to explore and demonstrate the space of possibilties. Whenever feasible, we have estimated these parameters from runs of the corresponding simple noise model. We have resorted to fixed values only in the most computational intensive class comparison tasks. The resulting inhomogeneity of models is not a problem as long as the rules are clear within a comparison task, and that likelihoods obtained with different methods are not compared directly.

The simulations in this section were carried out on Unix workstations running Linux and Solaris 2.3, using the neural simulation program GENESIS version 2.0 ([7]), with an extension library to implement the parameter-search methods. Data analysis was performed using Matlab (The Mathworks Inc.).

4.2 The Data

For simplicity and optimal control, we use an artifical data set generated using a simple one-compartment model. The model contains four voltage-dependent ionic conductances: fast Na, delayed rectifier K (K_{DR}), inactivating K or "A-current"

$$p(J) = p(K) = x \tag{24}$$

which is not very different from the original equation 16, since x is normally close to 1. We shall denote such type of model by $M = M(\vec{\mu}, \vec{\sigma}, \vec{x}, \vec{y})$.

The corresponding log-likelihoods are given by

$$\log p(I_1, ..., I_K) = (M - K) \log y + \sum_{j=1}^{K} \log \frac{1}{\sqrt{2\pi}\sigma_j} - \sum_{j=1}^{K} \frac{(I_j - \mu_j)^2}{2\sigma_j^2} \tag{25}$$

if $K < M$,

$$\log p(I_1, ..., I_M) = \log x + \sum_{j=1}^{M} \log \frac{1}{\sqrt{2\pi}\sigma_j} - \sum_{j=1}^{M} \frac{(I_j - \mu_j)^2}{2\sigma_j^2} \tag{26}$$

if $K = M$, and

$$\begin{aligned} \log p(I_1, ..., I_M) \quad = \quad & (K - M) \log y + \sum_{j=1}^{M} \log \frac{1}{\sqrt{2\pi}\sigma_j} - \sum_{j=1}^{M} \frac{(I_j - \mu_j)^2}{2\sigma_j^2} \\ & + \log(K - M)! - (K - M) \log(T - t_M) \end{aligned} \tag{27}$$

if $K > M$. In the case of missing spikes ($K < M$), a fixed penalty equal to $\log y$ is added for each missing spike. In the case of extra spikes ($K > M$), the penalty for each extra spike is slightly more complex. It consists of a penalty term of $\log y - \log(T - t_M)$ for each additional spike, and an additional $\log(K - M)!$ for $(K - M)$ extra spikes. In the following experiments, $(K - M)$ is usually very small (typically one or two spikes at most) and the $\log(K - M)!$ term can be ignored. Likewise a time horizon T can be selected so that the terms $\log(T - t_M)$ can be neglected. As a result, for most of the following, we use a fixed penalty $\log y$ for each additional or missing spike. [Alternatively, one could use a more complex parametrisation of the density $p(K)$ to achieve a similar effect.] One exception is with the class comparisons in the next section, where integrals over the entire space of parameters lead to parameter regions with large numbers of missing or extra spikes. In this case we use the above equations without modification.

In summary, if the CM generates M spikes and a data trace contains K spikes, to compute the likelihood we let $J = \inf(M, K)$. We then pair the first J spikes of each trace to each other using a linear ordering without gaps. The corresponding J ISIs are compared using quadratic penalties. Finally, a fixed penalty is used for each extra or missing spike. This is obviously not always optimal. In particular, optimal pairing of ISIs between traces could have a more complex structure. As in the case of the previous variable-rate model, such a generative model can be coupled to a CM by the way in which the parameters are determined. Again these parameters can be estimated by running the corresponding simple noise model many times. Another possibility is to determine the μ_i by taking the ISI observed in the corresponding deterministic CM trace (we discuss the validity of this last approximation in the

can further model this distribution with a second parameter $0 \leq y < 1$, by setting $p(M+k) = y^k$ for $k = 1, 2, 3...$, and likewise $p(M-k) = y^k$ for $k = 1, 2, ..., M$ (where $p(i)$ denotes the probability of observing a total of i spikes in a given trace). This simple model assumes a symmetric fall-off from the mode, according to a geometric progression with rate y. Obviously, more complex profiles for $p(i)$ could be selected if necessary. Since $\sum_i p(i) = 1$, x and y must be related by

$$x + (y + y^2 + \cdots + y^M) + (y + y^2 + \cdots) = 1 \tag{17}$$

where the second term on the left-hand side represents the probability of getting less than M spikes, while the third term represents the probability of getting more than M spikes. This simplifies to

$$x + \frac{2y - y^{M+1}}{1 - y} = 1 \tag{18}$$

This relation can be solved exactly, at least for small values of M. For instance, when $M = 0$, $y = (1-x)/(2-x)$. When x is close to 1, we have the approximations (these are also upper bounds)

$$y \approx 1 - x \quad \text{for} \quad M = 0 \quad \text{and} \quad y \approx \frac{1-x}{2} \quad \text{for} \quad M > 0 \tag{19}$$

We can now adapt equation (16) to the case where the number of spikes K is not necessarily equal to M, by multiplying the right-hand side by a term $q(J)$ as follows:

$$p(t_1, ..., t_J) = p(I_1, ..., I_J) = q(J) \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(I_j - \mu_j)^2}{2\sigma_j^2}} \tag{20}$$

where $J = \inf(K, M)$ *i.e.* J is the smaller of K and M. If $K < M$ (the data is missing spikes which are normally present in the model) we use

$$q(J) = p(K) = y^{M-K} \tag{21}$$

whereas if $K > M$ ($K = M + k, k > 0$; this means that the data contains extra spikes not normally present in the model), then

$$q(J) = p(K) \frac{(K-M)!}{(T - t_M)^{K-M}} = y^{K-M} \frac{(K-M)!}{(T - t_M)^{K-M}} \tag{22}$$

The extra factor $(K-M)!/(T-t_M)^{K-M}$ corresponds to the assumption that the extra k spikes are selected uniformly at random over the time interval $[t_M, T]$, where T is some fixed horizon. This term is needed to ensure proper normalisation, that is $p(K)$ is the probability of observing $K > M$ spikes, or

$$\int_0^T \int_{t_{M+1}}^T ... \int_{t_{M+1}+...+t_{K-1}}^T dI_{M+1} ... dI_K = \frac{(T - t_M)^{K-M}}{(K-M)!} \tag{23}$$

Finally, if $K = M$ we now use:

and for similar reasons it should not be too large either. A reasonable range for σ is 1 to 10 ms. In the following simulations, whenever we have used a fixed σ, we have used the value $\sigma = 2$ ms. Likewise, C cannot be too large if we do not want to replace an isolated spike with a burst. It can be shown ([2]) that if $\sigma > 1$ and $C \approx 1/\sqrt{2\pi}\sigma$, then the number of spikes generated by a variable-rate model associated with a single Gaussian has essentially a Poisson distribution with mean 1. For simplicity, in the following simulations, we have always used $C_i = 1/\sqrt{2\pi}\sigma_i$. Ultimately, models with variable C's and σ's can be expected to be more flexible and suitable for complex spike trains.

In summary, in the following simulations we use the Poisson variable-rate model in two forms (see Table 6):

- $M(\vec{\mu}, \vec{\sigma}, (C_i = 1/\sqrt{2\pi}\sigma_i)$, with $\vec{\mu}$ and $\vec{\sigma}$ estimated from running the corresponding simple noise model 1000 times (used in the individual model comparisons; see below);
- $M(\vec{\mu}, 2, 1/2\sqrt{2\pi})$, with $\vec{\mu}$ determined by the times of the spikes in the underlying deterministic component of the model (used in the whole-class comparisons).

3.2.3 Gaussian ISI Model

In this model, the spiking times of the neurons are treated as Gaussian random variables. More precisely, for the sake of argument assume that both the data trace(s) and the deterministic CM trace(s) have the same number M of spikes, for a given level of current injection. The first assumption is that all ISIs (interspike intervals) are independent, and each ISI I_i has a Gaussian distribution with a mean μ_i and a standard deviation σ_i. Then the model generates spikes at times $t_1, ..., t_M$, and the corresponding ISIs $I_1, ..., I_M$ with probability:

$$p(D|M) = p(t_1, ..., t_M) = p(I_1, ..., I_M) = \prod_{j=1}^{M} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(I_j - \mu_j)^2}{2\sigma_j^2}} \qquad (16)$$

which is just a product of the (independent) Gaussian distributions for each spike. The only problem now is to estimate the parameters of the model (*i.e.* the μ_j's and the σ_j's), which again can be done from the data or by running the simple noise model a large number of times.

The difficulty with the above method is that in many cases models with noise will not always generate the same numbers of spikes for a given current injection, and the number they do generate may be different from that seen in the data set. Therefore we must modify the above equation for the likelihood to take into account extra/missing spikes. Thus the second assumption is as follows. Suppose that under a fixed set of conditions, a deterministic CM yields a spike train with M spikes at times $\tau_1, ..., \tau_M$. We can imagine that with added noise the probability of observing M spikes has a certain probability x ($0 \leq x \leq 1$), possibly close to 1. It is reasonable to assume that the probability of the number of spikes has a maximum at M. We

3.2.2 Variable-Rate Poisson models

One of the most basic models for point processes, such as spike trains, is a Poisson process with rate λ. If we subdivide time into very small intervals Δt (the "bin size"), a Poisson process can be seen as a sequence of independent Bernoulli trials (*i.e.* trials with two possible outcomes, one of which has a small constant probability λ, and the other probability $1-\lambda$). Over a period of time T, the number of successes, or spikes, has a binomial distribution which can be approximated by a Poisson distribution with average λT. The main problem with such a model is that spike train rates are usually not stationary. Thus the natural generalisation is a Poisson model with *variable* rate $\lambda(t)$. To generate such a process, we can imagine a large sequence of Bernouilli trials, where the probability of success in the trial indexed by time t is given by $\lambda(t)$. Clearly the rate function must satisfy $0 \leq \lambda(t) \leq 1$. The rate function can be chosen in different ways. A natural possibility is to decompose it as a sum of Gaussians

$$\lambda(t) = \sum_{i=1}^{M} C_i e^{-(t-\mu_i)^2/2\sigma_i^2} \tag{13}$$

We denote such a model by $M = M(\vec{\mu}, \vec{\sigma}, \vec{C})$. In any case, under any variable-rate Poisson model M, the likelihood of any spike train, with spiking times $(t_1, ..., t_K)$, is easily computed by

$$p(t_1, ..., t_K) = \prod_{i=1}^{K} \lambda(t_i) \prod_{t \neq t_i} (1 - \lambda(t)) = \prod_{i=1}^{K} \frac{\lambda(t_i)}{1 - \lambda(t_i)} \prod_{t} (1 - \lambda(t)) \tag{14}$$

Notice that $Q = \prod_t (1 - \lambda(t))$ is the probability of having no spikes at all. The log-likelihood is given by:

$$\log p(t_1, ..., t_K) = \sum_{i=1}^{K} \log \lambda(t_i) - \sum_{i=1}^{K} \log(1 - \lambda(t_i)) + \log Q \tag{15}$$

When $\lambda(t)$ results from a single Gaussian curve around an isolated spike, the term $\log \lambda(t_i)$ gives rise to the usual quadratic error term. The mathematical properties of the variable-rate model are studied in detail in ([2]).

As it stands, the variable-rate model is a generative model which has nothing to do with CMs. However, it is easy to couple such a model with an underlying deterministic component by the way in which the parameters μ, σ and C are determined. In the case of $\vec{\mu}$, one possibility is to run the associated deterministic component, and then superimpose a Gaussian model on each one of the spike times. If the deterministic component of the model produces spikes at times $\tau_1, ..., \tau_M$, then we set $\mu_i = \tau_i$. Another possibility is to run the corresponding simple noise model many times and use the expected spike times to set $\mu_i = E(\tau_i)$. In the case of σ, we can again run the corresponding simple noise model many times and use $\sigma_i = E(\sigma_i)$. Another faster possibility is just to use a fixed value for σ. For this, σ should not be too small (i.e. below 1 ms) in order to avoid overfitting single isolated spikes,

likelihoods. One example is when a modeler is interested in the particular shape of an action potential (see [20] for a related Bayesian treatment of spike classification). Here we shall concentrate on the case where a modeler is attempting to reproduce the characteristics of an entire spike train, while injecting currents of different amplitudes into the soma of the CM. Thus the data consists of a series of numbers representing spike times. Several criteria have been used to compare spike trains, but whatever criterion is used, it comes with an underlying probabilistic model that needs to be clarified and is amenable to Bayesian analysis. Indeed, if the fit of a model $M = M(\theta)$ with parameters θ is measured by some error function $f(\theta, D) \geq 0$ to be minimized, one can always define the associated likelihoods to be:

$$p(D|M(\theta)) = \frac{e^{-f(\theta,D)}}{Z} \tag{12}$$

where $Z = \int_\theta e^- f(\theta, D)$ is a normalizing factor (the "partition function" in statistical mechanics) which makes the probabilities integrate to 1. As a result, minimizing the error function is equivalent to maximum likelihood estimation. In particular, when the sum squared difference in spike timings (or interspike interval timings), is used to compare two spike trains or two other data sets D and D', a rather common practice, this implies an underlying Gaussian model (see below). Thus the Bayesian point of view clarifies the probabilistic assumptions that *must* underly any criteria for comparing two data sets D and D'.

We shall now consider three possible ways of building a truly probabilistic model of spike trains, using an underlying deterministic CM, for the computation of likelihoods. The underlying deterministic CM will be referred to as the "deterministic component" of the model.

3.2.1 Simple Noise Model

Real neurons have a number of sources of noise, such as stochastic transitions of ion channels between open and closed states, stochastic release of neurotransmitter vesicles from presynaptic terminals, etc. etc. ([13]). A natural and simple way to model this noise is by incorporating Gaussian white noise in the amplitudes of injected currents, at least in the compartment(s) associated with the soma. Adding Gaussian noise to the current input to the model is appropriate since the sum of a large number of independent stochastic processes tends to a Gaussian distribution by the central limit theorem. This gives us a model capable of producing stochastic outputs which can be compared to the data. We refer to this as the "simple noise model". While this is perhaps the most simple and natural model, it does not by itself provide a computationally efficient way of estimating likelihoods. Data likelihoods can in principle be estimated by running the simple noise model a very large number of times with different noise instantiations, collecting the spike timings of the model, and computing how often any given data set D was generated by the model. This, however, is usually impractical. We will now describe two alternative statistical models for determining likelihoods of spike events, which we use extensively in the simulations described below.

results can be derived with different priors. To these arguments, Bayesians can offer at least four different answers.

1. Even when priors are not mentioned explicitly, they are used implicitly. The Bayesian approach forces a clarification of one's assumptions without sweeping the problem of priors under the rug.

2. The effects of priors diminishes as the amount of data increases.

3. There exist situations where objective criteria (maximum entropy) can be used to determine non-informative priors (for instance [12])

4. Finally, and most importantly, the effects of different priors, as well as different models and model classes, can be assessed within the Bayesian framework by comparing the corresponding probabilities.

3.2 Data Likelihood for Compartmental Models

In order to define $p(D|M)$, one must come to grips with how a model M could also give rise to a different observation set D': in a Bayesian framework, compartmental models *must* be probabilistic. In contrast, with the exception of models that use random stimuli, such as input synapses that are activated randomly (*e.g.* [30], [34]), most current CMs are deterministic. A deterministic model assigns a probability zero to all the data, except the one it can produce exactly. This is clearly inadequate in biology. In fact, recently many experimentalists and theorists have been investigating the nature, sources, and functions of noise in real neurons (see, for instance [34], [32], [23]). However, because of the complexity inherent to CMs, modelers tend to be absorbed exclusively by the structural and deterministic components of their models (*e.g.* the number, morphology and conductances of the compartments) while ignoring the equally important probabilistic aspects. This is perhaps one of the major lessons to be derived from the Bayesian point of view, and an area where CMs need to be improved. A scientific discourse on CMs, how well they fit the data and how they can be compared to each other, is *impossible* if the likelihood issue is not addressed honestly.

The likelihood question is thus clearly related to issues of variability and noise. It is also important to notice that results from very different sources and experiments (*e.g.* different cells, different organisms, different experimental preparations) are routinely combined in CMs. For instance, channel kinetic data obtained from patch-clamp analyses performed at room temperature may be combined with intracellular current- or voltage-clamp data performed in a brain-slice preparation, at a different temperature, in a different lab, and with anatomical data from yet another lab. Current CMs in general attempt to capture the behavior of an "average" cell, or piece of circuitry. Thus mismatches between the behavior of specific cells and "average" models are bound to occur, and the probabilistic connection between CMs and the underlying neurobiological system being modeled is even stronger.

Since CMs of different types are used to study an array of very different phenomena, it is impossible to give a completely general prescription for computing

a time constant, c occupies some value in an interval $[a, b]$ and we do not have any additional specific information, it is reasonable to define the prior density

$$p(k < \log c < k + \Delta k) = \frac{\Delta k}{\log b - \log a} \tag{7}$$

for any $[k, k + \Delta k]$ in $[\log a, \log b]$, and 0 otherwise. So, if c is equally likely to be 1, 10 or 100, $\log_{10} c$ will have a flat distribution between 0 and 2. Starting from equation (4), it is easy to calculate the prior term $\log p(M)$ by assuming that all the basic parameters are independent. That is if a model is characterized by a set of conductances c_i satisfying equation (7) over an interval $[a_i, b_i]$, then:

$$\log(p(M)\Delta\vec{k}) = \sum_i \log \frac{\Delta k_i}{\log b_i - \log a_i} \tag{8}$$

where $\Delta\vec{k} = \Delta k_1 \Delta k_2 ....$ In terms of probability densities, this becomes:

$$\log(p(M)) = \sum_i \log \frac{1}{\log b_i - \log a_i} \tag{9}$$

Another useful prior for conductances, or their logarithms, is a gamma prior. The gamma density ([15]) with parameters α and λ, is given by

$$\Gamma(x; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \tag{10}$$

for $x > 0$, and 0 otherwise. By varying α and λ and translating x, the gamma density allows for a wide range of priors, with more mass concentrated in one specific region of parameter space.

It is of course also possible in Bayesian modeling to tie together different sets of parameters. For instance, one could assume that conductances are linked, or have the same values or ranges, in different regions of a large dendritic tree. An example of this in a CM with four types of regions can be found in ([30]). This approach is a form of hierarchical modeling, where hyperparameters are used to control the priors on the model parameters. A typical example is when there are many parameters with the same prior Gaussian distribution. In this case, the mean and standard deviation of the common Gaussian can be used as hyperparameters. Hierarchical levels can of course be iterated. Hierarchical modeling allows for the introduction of a certain amount of structure on the model parameters, and also for parameter reduction since the model prior can be calculated from a (usually smaller) number of hyperparameters. In symbolic form

$$p(\theta) = \int_\alpha p(\theta|\alpha) p(\alpha) d\alpha \tag{11}$$

where α represents hyperparameters with prior $p(\alpha)$.

The use of priors is a strength of the Bayesian approach, since it allows incorporation of prior knowledge into the modeling process. It has also been considered a weakness in the past, on the ground that priors are subjective, and that different

many fewer compartments. To compare C_1 and C_2 in the Bayesian framework, one must compute $p(C_1|D)$ and $p(C_2|D)$. Using Bayes' theorem again: $p(C|D) = p(D|C)p(C)/p(D)$. In addition to the prior $p(C)$, one must calculate the evidence $p(D|C)$, and this is obtained by averaging over the entire model class:

$$p(D|C) = \int_{\theta \in C} p(D, \theta|C) d\theta = \int_{\theta \in C} p(D|\theta, C) p(\theta|C) d\theta \tag{6}$$

Naturally, in cases where the likelihood $p(D|\theta, C)$ is very peaked around its maximum, such expectations can be approximated using the mode (*i.e.* the value with the highest probability). But in general, such integrals require better approximations, for instance using Monte Carlo sampling methods ([33], [25], [6]). Available computer power and its progress are of course important issues to be considered in this context.

As a final point raised in the introduction, it does not make sense to choose a simple model on the basis that available data is scarce. *Everything else being equal*, however, it is true that one should prefer a simple hypothesis to a complex one. This is Ockham's razor. As pointed out by several authors, Ockham's razor is automatically embodied in the Bayesian framework ([18], [22]) in at least two different ways. In the first trivial way, one can introduce priors that penalize complex models. But even without such priors, parametrized complex models will tend to be consistent with a larger volume of data space. Since a likelihood $p(D|M)$ must sum to 1 over the space of data, if $p(D|M)$ covers a larger expanse of data space the likelihood values for given data sets will be smaller on average. Therefore, everything else being equal, complex models will tend to assign a correspondingly smaller likelihood to the observed data.

In summary, Bayesian inference provides a rigorous and provably unique framework for inference. The basic step is to compute model plausibilities with respect to the available data, and the associated expectations, using the rules of probability theory and possibly numerical approximations. We can now turn to the specific application of the general Bayesian framework to CMs.

3 Bayesian Compartmental Modeling in Neuroscience

What then are the implications of a Bayesian approach for modeling in CNS? For any type of model class, it is clear from equation (2) that in order to evaluate $p(M|D)$ the first step consists in making the prior $p(M)$ and the likelihood $p(D|M)$ explicit.

3.1 Prior Probabilities for Compartmental Models

A typical class of CMs $M(\theta)$ has a large number of parameters θ. Different types of parameters can be used, but for single-cell models the majority of these parameters correspond to the conductances, and possibly the kinetic time constants, of ionic channels in the different compartments. Both conductance values and time constants are *scale* parameters. Thus, unless any additional information is available, a reasonable prior is a uniform logarithmic prior ([21]). That is, if a conductance, or

$$\log p(M|D) = \log p(D|M) + \log p(M) - \log p(D) \tag{4}$$

To apply equation (4) to CMs, or to any other class of models, one must specify the prior $p(M)$, and the data likelihood $p(D|M)$. Once the prior and data likelihood terms are made explicit, the initial modeling effort is completed. All that is left is cranking the engine of probability theory.

Two specific compartmental models M_1 and M_2 can be compared by comparing their probabilities $p(M_1|D)$ and $p(M_2|D)$. More generally, one objective is often to find or approximate the "best" model within a class, that is to find the set of parameters θ maximizing the posterior $p(M|D)$ or $\log p(M|D)$), and the corresponding error bars (see appendix B). This is called MAP (maximum a-posteriori) estimation. In order to deal with positive quantities, this is also equivalent to minimizing $-\log p(M|D)$. Notice that in this calculation, the term $p(D)$ in (4) plays the role of a normalizing constant that does not depend on the parameters θ, and is therefore irrelevant for this optimization. If the prior $p(M)$ is uniform over all the models considered, then the problem reduces to finding the maximum of $p(D|M)$ (or $\log p(D|M)$). This is just maximum likelihood (ML) estimation. In most interesting models, the function being optimized is complex and its modes cannot be solved analytically. Thus one must resort to iterative and possibly stochastic methods such as gradient descent or simulated annealing. A comparative study of such algorithms in the case of a simple CM can be found in ([36]).

Whereas finding the optimal model is common practice, it is essential to notice that this is really useful only if the distribution $p(M|D)$ is sharply peaked around a unique optimum. In situations characterized by a high degree of uncertainty and relatively small amounts of available data, this is often not the case. Thus a Bayesian is really interested in the function $p(M|D)$ over the entire space of models, rather than in its maxima only, and more precisely in evaluating expectations with respect to $p(M|D)$. This is the case in prediction problems, marginalisation of nuisance parameters, and class comparisons.

Consider, for instance, a prediction problem, where we are trying to predict the output value y of an unknown parametrized function f_θ, given an input x. It is easy to show that the optimal prediction is given by the expectation

$$E(y) = \int_\theta f_\theta(x) p(\theta|D) d\theta \tag{5}$$

This integral is just the average of the predictions made by each possible model f_θ, weighted by the plausibility $p(\theta|D)$ of each model. Another example is the process of marginalisation, where integration of the posterior parameter distribution is carried only with respect to a subset of the parameters, the so-called nuisance parameters ([11]). In a frequentist framework, the notion of distribution over the parameters is not defined and therefore nuisance parameters cannot be integrated out easily. In CMs, one is also often lead to the problem of comparing two model classes C_1 and C_2. As an example, these could be compartmental models with two very different numbers of compartments, possibly representing a highly anatomically detailed model with a large number of compartments versus a simplified model with

The axiomatic presentation of the Bayesian approach is given in appendix A for completeness. For brevity, we do not present any proofs or any historical background for the Bayesian approach, nor do we discuss any controversial issues regarding the foundations of probabilities, such as the frequentist versus Bayesian point of view. All of these can be found in various books and articles such as ([4], [5], [8], [28], [16], and references therein).

While the Bayesian approach is universal and applies to any problem or model, here we focus on CMs for the reasons given above. In section 2, we briefly review Bayesian inference. In section 3, we analyse the steps compartmental modelers need to take for a proper Bayesian treatment of their models. In particular, we provide specific suggestions for the computation of CM priors and likelihoods. In section 4, we apply the methodology in detail to a specific class of CMs. We summarize the benefits and drawbacks of the method in the discussion session, together with some possible extensions.

2 Bayesian Inference

As described in appendix A, the immediate consequence of the Cox-Jaynes axioms is Bayes' theorem:

$$p(X|Y,I) = \frac{p(Y|X,I)p(X|I)}{p(Y|I)} = p(X|I)\frac{p(Y|X,I)}{p(Y|I)} \tag{1}$$

where X and Y are propositions and I represents background information. $p(X|Y,I)$ denotes the conditional probability of X knowing Y and I. We will apply this equation to a parametrized model $M = M(\theta)$ and a corpus of data D. For simplicity, we will drop the background information I from the following equations, or equivalently incorporate it into the data. From Bayes theorem we have immediately,

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)} = p(M)\frac{p(D|M)}{p(D)} \tag{2}$$

The term $P(M)$ is referred to as the prior probability or simply the prior. The term $P(M|D)$ is referred to as the posterior probability or posterior. The term $P(M|D)$ is referred to as the likelihood or the evidence. The prior represents our estimate of the probability that model M is correct before one has obtained any data, while the posterior represents our updated belief in the probability that model M is correct given that we have observed the data set D. For data obtained sequentially, one has

$$p(M|D_1,...,D_n) = p(M|D_1,...,D_{n-1})\frac{p(D_n|M,D_1,...,D_{n-1})}{p(D_n|D_1,...,D_{n-1})} \tag{3}$$

In other words the old posterior $p(M|D_1,...,D_{n-1})$ plays the role of the new prior. Bayesian modeling is entirely open to both new data and new models. Bayesian modeling is iterative by nature, as both data and models evolve.

For technical reasons, it is often easier to work with the corresponding logarithms, so that:

When reasoning in the presence of certainty, one uses deduction. This is how the most advanced theories in information-poor sciences, such as physics or mathematics, are presented in an axiomatic fashion. Deduction is not controversial. The overwhelming majority of people agree on how to perform deductions in a unique way: if A implies B, and A is true, then B must be true. This is the essence of Boole's algebra, and is at the root of all digital computers. When reasoning in the presence of uncertainty, one uses induction and inference: if A implies B, and B is true, then A is *more plausible*. An amazing and still poorly known fact is that there exists a simple and *unique* consistent set of rules for induction, model selection and comparison: it is called Bayesian induction. The Bayesian approach has been known for quite some time, but only recently has it started to infiltrate different areas of science and technology systematically with useful results ([12]), [11], [21], [22]). To our knowledge, however, it has not been applied systematically to CNS. Our purpose here is to introduce the Bayesian point of view and the advantages that may derive from it to computational neuroscientists, especially those involved in building CMs. In a companion paper ([2]) we use a Bayesian framework to derive new algorithms for the analysis of neuronal multi-unit recordings.

The Bayesian point of view has a simple intuitive description. The Bayesian approach assigns a degree of plausibility to any proposition, hypothesis or model [throughout the paper, hypotheses or models are essentially synonymous: models tend to be complex hypotheses with many parameters]. More precisely, in order to properly carry out the induction process, one ought to proceed in three steps:

1. Clearly state what the hypotheses or models are, along with *all* the background information and the data.

2. Use the language of probability theory to assign prior probabilities to the hypotheses.

3. Use probability calculus for the inference process, in particular to evaluate posterior probabilities (or degrees of belief) for the hypotheses in light of the available data, and derive *unique* answers.

Such an approach certainly seems a reasonable one [1]. But why should it be so compelling? Why use the language of probability thory, as opposed to any other method? The surprising answer to this question is that it can be proved, in a strict mathematical sense, that this is the only consistent way of reasoning in the presence of uncertainty. Specifically, there exists a small set of very simple common-sense axioms, the Cox-Jaynes axioms, under which it can be shown that the Bayesian approach is the unique consistent approach to inference and induction. Under the Cox-Jaynes axioms, degrees of plausibility satisfy exactly all the rules of probabilities. Probability calculus is then all the machinery that is required for inference, model selection and model comparison.

[1] Notice that the Bayesian approach is not directly concerned with the creative process, how to generate new hypotheses or models. It is only concerned with assessing the value of models with respect to the available knowledge and data. This assessment procedure, however, may of course be very helpful in generating new ideas.

data is still sparse, and probabilistic aspects are often ignored. As a tutorial, we demonstrate the Bayesian approach on a class of one-compartment models with three and four conductances.

* To whom all correspondence should be addressed.

1 Introduction

Computational modeling is being increasingly used in neuroscience, over vastly different spatio-temporal scales, from molecules to entire brain areas, and at different levels of complexity. While available neurophysiological and neuroanatomical data is rapidly increasing, what is known today constitutes only a tiny fraction of what remains to be known. Thus in computational neuroscience (CNS) in particular, and more generally in biology and all other information-rich sciences, scientists must reason in the presence of a high-degree of uncertainty: many facts are missing, and some of the facts are wrong. Computational neuroscientists are then constantly faced with *inference* problems related to building models from available data. What is the right class of models and level of complexity? Which details are important and which should be discarded? How can one compare different models, and select the best one, in light of available knowledge and sometimes limited data? Or, in short, how do we know if a model is a good model?

Within CNS, compartmental models (CMs) have become an increasingly important and standard approach ([19], [7]). CMs are detailed neuronal models, where neurons are subdivided into a number of interconnected compartments , the precise number depending on the degree of anatomical detail available or desired (typically this number has increased substantially as computing resources and available data have increased). The basic simplification, in general, is that each compartment is treated as being isopotential, with uniform spatial properties. Currently, one of the largest single neuron compartmental models is a Purkinje cell model ([30], [31]) with a total of 4550 compartments and 8021 ionic channels. It is precisely because CMs have a large number of parameters and available data is sometimes limited that inference issues become fundamental. In particular CMs have sometimes been criticized on the ground that they may be capable of accomodating almost any behavior with different parameter settings and that, when little data is available, simpler models with fewer parameters should be preferred. CM practitioners know that there exist many *implicit* constraints that emerge from the structure of the models and which, in fact, render arbitrary behavior very difficult, if not impossible, to reproduce. More importantly, as pointed out in ([26]), choosing simpler models because little data is available does not make much sense. While it is a widespread practice and occasionally a useful heuristic, it is clear that the amount of data collected and the complexity of the underlying source are two completely different things. It is not hard to imagine situations characterized by a very complex source, and relatively little data (in fact, this would encompass most of biology at the present time). Thus it is clear that questions of inference are particularly central to CNS in general, and to CMs in particular.

On the Use of Bayesian Methods for Evaluating Compartmental Neural Models

Pierre Baldi*
Department of Computation and Neural Systems
California Institute of Technology
Pasadena, CA 91125
pfbaldi@ccosun.caltech.edu
(818) 395-6818 (818) 449-0679 (FAX)

Michael C. Vanier
Department of Computation and Neural Systems
California Institute of Technology
Pasadena, CA 91125
mvanier@bbb.caltech.edu
(818) 395-6818

James M. Bower
Department of Computation and Neural Systems
California Institute of Technology
Pasadena, CA 91125
jbower@bbb.caltech.edu
(818) 395-6818

"A model should be as simple as possible, but no simpler."
(attributed to Einstein)

Abstract

Computational modeling is being increasingly used in neuroscience. In deriving such models, *inference* issues, such as model selection, model complexity, and model comparison, must constantly be addressed. Here, we briefly present the Bayesian approach to inference. Under a simple set of common sense axioms, there exists essentially a unique way of reasoning under uncertainty by assigning a degree of confidence to any hypothesis or model, given the available data and prior information. Such degrees of confidence must obey all the rules governing probabilities, and can be updated accordingly as more data becomes available. While the Bayesian methodology can be applied to any type of model, as an example we outline its use for an important, and increasingly standard, class of models in computational neuroscience: compartmental models. Induction issues are particularly relevant for compartmental models. While their parameter spaces are typically very large, neurophysiological and neuroanatomical