Supporting Information

Mattingly et al. 10.1073/pnas.1715306115

Model Selection from Data

The usual discussion of model selection takes place after observing data x. If we wish to compare some models* labeled by d, each with some prior $p_d(\theta)$, then one prescription is to choose the model with the largest $p(x)$. Labeling this explicitly, we write

$$p(x|d) = \int_{\Theta_d} d\theta\, p(x|\theta)\, p_d(\theta), \quad p_d(\theta) > 0 \text{ on } \Theta_d \subset \Theta. \quad \textbf{[S1]}$$

If the Bayes factor $p(x|d)/p(x|d')$ is larger than one, then (absent any other information) d is preferred over d' (1).† In the usual asymptotic limit $m \to \infty$, this idea leads to minimizing the BIC (2),

$$-\log p(x|d) \approx -\log p(x|\hat{\theta}_d) + \frac{d}{2}\log m + \mathcal{O}(m^0),$$

where $-\log p(x|\hat{\theta}_d) = \frac{1}{2}\chi^2 = \frac{1}{2}\sum_{i=1}^{m}\left[x_i - y_i(\hat{\theta}_d)\right]^2/\sigma^2 \sim \mathcal{O}(m)$, and $\hat{\theta}_d$ is a maximum-likelihood estimator for x, constrained to the appropriate subspace:

$$\hat{\theta}_d(x) = \underset{\theta\in\Theta_d}{\operatorname{argmax}}\, p(x|\theta).$$

The term $d\log m$ penalizes models with more parameters, even though they can usually fit the data more closely. Despite the name this procedure is not very Bayesian: One chooses the effective model (and hence the prior) after seeing the data, rather than simply updating according to Bayes' theorem.‡

Related prescriptions can be derived from MDL ideas. To allow reconstruction of the data we transmit both the fitted parameters and the residual errors, and minimizing the (compressed) length of this transmission drives a tradeoff between error and complexity (5–7). A convenient version of this is NML (8, 9) and chooses the model d which maximizes

$$p_d^{\mathrm{NML}}(x) = \frac{p\big(x|\hat{\theta}_d(x)\big)}{Z_d}, \quad Z_d = \int_{\Theta_d} dx'\, p\big(x'|\hat{\theta}_d(x')\big). \quad \textbf{[S2]}$$

This is not Bayesian in origin and does not depend on the prior on each effective model d, only its support Θ_d. The function $p_d^{\mathrm{NML}}(x)$ is not expected data in the sense of $p(x)$—it is not the convolution of the likelihood with any prior.§ In the asymptotic limit $p_d^{\mathrm{NML}}(x)$ approaches $p(x)$ from the Jeffreys prior, and this criterion agrees with BIC (6), but away from this limit they differ.

In Fig. S1 we apply these two prescriptions to the exponential example treated in the main text. At each $\vec{x} \in X$ we indicate which one of a list of models is preferred.¶ Fig. S2 instead draws the distributions being used.

- Fig. S1*A* compares three models: the complete model (with the Jeffreys prior), the optimal model described by discrete prior $p_\star(\vec{y})$, and an even simpler model with weight only on the three vertices $\vec{y} = (0,0), (\frac{1}{2},\frac{1}{2}), (1,1)$.
- Fig. S1*B* instead compares the complete model to three different one-parameter models (along the three boundaries of the allowed region of the $\vec{y}$ plane) and a zero-parameter model (one point $\vec{y}$, in no particularly special place). In terms of decay rates the three lines are limits $k_1 = k_2$, $k_1 = 0$, and $k_2 = \infty$.

Different effective models are preferred for different values of data x. At a given point x, if several models contain the same $\hat{\theta}(x)$, then the simplest among them is preferred, which in the NML case means precisely the one with the smallest denominator Z_d. In fact, a trivial model consisting of just one point $\Theta_0 = \hat{\theta}(x)$ would always be preferred if it were among those considered—there is no automatic preference for models which can produce a wide range of possible data.

By contrast, our prior selection approach aims to be able to distinguish as many possible outcomes in X as possible. Applied to the same list of models as in Fig. S1, this gives the following fixed scores (base e):

$$I_{\text{full}} = 1.296, \qquad I_\star = 1.630, \qquad I_{\text{corners}} = 1.098$$

and

$$\begin{aligned} I_{\text{upper}} &= 0.852, & I_{\text{lower-left}} &= 0.845, \\ I_{\text{lower-right}} &= 1.418, & I_{\text{one-point}} &= 0. \end{aligned} \quad \textbf{[S3]}$$

By definition $p_\star(\theta)$ has the highest score. In second place is the line along the lower edge (corresponding to $k_1 = k_2$). The shorter lines are strongly disfavored because they cover a much smaller range of possible data.

Algorithms

The standard algorithm for maximizing channel capacity (of discrete memoryless channels) was written independently by Blahut (11) and Arimoto (12). This aspect of rate-distortion theory is mathematically the same as the problem we consider, of

---

*The word model unfortunately means several things in the literature. We mean parameter space Θ_d always equipped with a likelihood function $p(x|\theta)$ and usually with a prior $p_d(\theta)$. When this is a subspace of some larger model Θ_D (whose likelihood function agrees, but whose prior may be unrelated), then we term the smaller one an effective model, or a reduced model, although we do not always write the adjective. The optimal prior $p_\star(\theta)$ defines an effective model in this sense. Its support will typically be on several boundaries of Θ_D. If the boundaries of Θ_D (of all dimensions) are regarded as a canonical list of reduced models, then $p_\star(\theta)$ is seldom a submodel of any one of them.

†If one of the priors is improper, say $\int d\theta\, p_d(\theta) = \infty$, then $p(x|d)$ will also be infinite. In this sense the Bayes factor behaves worse than the posterior $p(\theta|x)$, which can still be finite.

‡Terms penalizing more complex models can be translated into shrinkage priors, which concentrate weight near simpler models (3). Perhaps the shrinkage priors closest to the ones in this paper are the penalized complexity priors of ref. 4. Those are also reparameterization invariant and also concentrate weight on a subspace of Θ, often a boundary. However, both the subspace (or base model) and the degree of concentration (scaling parameter) are chosen by hand, rather than being deduced from $p(x|\theta)$.

§This relevant optimization problem can be described as minimizing worst-case expected regret, written (8) as

$$p_d^{\mathrm{NML}} = \underset{q}{\operatorname{argmin}}\, \max_x \log \frac{p(x|\hat{\theta}_d(x))}{q(x)}, \qquad \hat{\theta}_d(x) \in \Theta_d.$$

Perhaps the closest formulation of our maximum MI problem is that our $p_\star(x)$, the distribution on X and not the prior, can be found as (10)

$$p_\star = \underset{q\in B}{\operatorname{argmin}}\, \max_\theta \int_X dx\, p(x|\theta) \log \frac{p(x|\theta)}{q(x)},$$

where $q(x)$ is constrained to be a Bayes strategy, i.e., to arise from some prior $p_\star(\theta)$. Note the absence of $\hat{\theta}_d(x)$ and the presence of an integral over X, corresponding to the fact that this maximization takes place without being given a subspace Θ_d or seeing data x. The resulting distributions on X are also different, as drawn in Fig. S2. If plotted on Fig. 5*B*, $p_2^{\mathrm{NML}}(x)$ from the full model would be somewhere between the two expected data $p(x)$ lines there. But it is not a comparable object; its purpose is model comparison as in Fig. S1.

¶Recall that $\vec{x}$ is $\vec{y}$ corrupted by Gaussian noise, and $\vec{y}$ is constrained to the area shown in Fig. 4*A* because it arises from decay rates k_μ via Eq. **5**. We may regard either y_t or k_μ as being the parameters, generically θ.

maximizing MI by choosing the prior. The algorithm starts with $p_0(\theta) = \text{const.}$ and then at each time step updates this by

$$p_{\tau+1}(\theta) = \frac{1}{Z_\tau} e^{f_{\mathrm{KL}}(\theta)} p_\tau(\theta), \qquad \textbf{[S4]}$$

where $Z_\tau = \int d\theta'\, e^{f_{\mathrm{KL}}(\theta')} p_\tau(\theta')$ maintains normalization, and $f_{\mathrm{KL}}(\theta) = D_{\mathrm{KL}}[p(x|\theta) \,||\, p(x)]$ is computed with $p_\tau(\theta)$. Since this is a convex optimization problem, the algorithm is guaranteed to converge to the global maximum. This makes it a good tool to see discreteness emerging.

Fig. 2 and Fig. S3 show the progress of this algorithm for the 1D and 2D models in the main text. We stress that the number and positions of the peaks which form are unchanged when the discretization of θ is made much finer. Note also that the convergence to delta functions happens much sooner near the boundaries than in the interior. The convergence to the correct value of MI, and toward the optimum distribution on data space $p(x)$, happens much faster than the convergence to the correct number of delta functions.

Because θ must be discretized for this procedure, it is poorly suited to high-dimensional parameter spaces. However, once we know that $p_\star(\theta)$ is discrete it is natural to consider algorithms exploiting this. With K atoms, we can adjust their positions $\vec{\theta}_a$ and weights λ_a using gradients

$$\frac{\partial \mathrm{MI}}{\partial \theta_a^\mu} = \lambda_a \int dx\, \frac{\partial p(x|\vec{\theta})}{\partial \theta^\mu} \log \frac{p(x|\vec{\theta})}{p(x)} \bigg|_{\vec{\theta}=\vec{\theta}_a}$$
$$\frac{\partial \mathrm{MI}}{\partial \lambda_a} = f_{\mathrm{KL}}(\vec{\theta}_a) - 1. \qquad \textbf{[S5]}$$

Figs. 1 and 3*A* and the square plot points in Fig. 4 were generated this way. This optimization is not a convex problem (there is some tendency to place two atoms on top of each other and thus use too few points of support) but it can often find the optimum solution. We can confirm this by calculating $f_{\mathrm{KL}}(\theta)$ everywhere—any points for which this is larger than its value at the atoms indicate that we do not have the optimal solution and should add an atom.

Monte Carlo algorithms for this problem have been investigated in the literature (refs. 13 and 14 and especially ref. 15). (Incidentally, we observe that ref. 13's table 1 contains a version of scaling law Eq. **4** with $\zeta \approx 1/2$. No attempt was made there to use the optimal number of atoms, only to calculate the channel capacity to sufficient accuracy.)

1. Kass RE, Raftery AE (1995) Bayes factors. *J Am Stat Assoc* 90:773–795.
2. Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464.
3. Bhadra A, Datta J, Polson NG, Willard B (2016) Default Bayesian analysis with global-local shrinkage priors. *Biometrika* 103:955–969.
4. Simpson D, Rue H, Riebler A, Martins TG (2017) Penalising model component complexity: A principled, practical approach to constructing priors. *Stat Sci* 32:1–28.
5. Wallace CS, Boulton DM (1968) An information measure for classification. *Comput J* 11:185–194.
6. Rissanen J (1978) Modeling by shortest data description. *Automatica* 14:465–471.
7. Grünwald PD, Myung IJ, Pitt MA (2009) *Advances in Minimum Description Length: Theory and Applications* (MIT Press, Cambridge, MA).
8. Myung JI, Navarro DJ, Pitt MA (2006) Model selection by normalized maximum likelihood. *J Math Psychol* 50:167–179.
9. Grünwald PD (2007) *The Minimum Description Length Principle* (MIT Press, Cambridge, MA).
10. Haussler D (1997) A general minimax result for relative entropy. *IEEE Trans Inf Theory* 43:1276–1280.
11. Blahut R (1972) Computation of channel capacity and rate-distortion functions. *IEEE Trans Inf Theory* 18:460–473.
12. Arimoto S (1972) An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans Inf Theory* 18:14–20.
13. Chang C-I, Davisson LD (1988) On calculating the capacity of an infinite-input finite (infinite)-output channel. *IEEE Trans Inf Theory* 34:1004–1010.
14. Lafferty J, Wasserman L (2001) Iterative Markov chain Monte Carlo computation of reference priors and minimax risk. *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann Publishers, Inc., San Francisco), pp 293–300.
15. Dauwels J (2005) Numerical computation of the capacity of continuous memoryless channels. *Proceedings of the 26th Symposium on Information Theory in the Benelux*. Available at www.dauwels.com/Papers/memoryless.pdf. Accessed May 1, 2017.

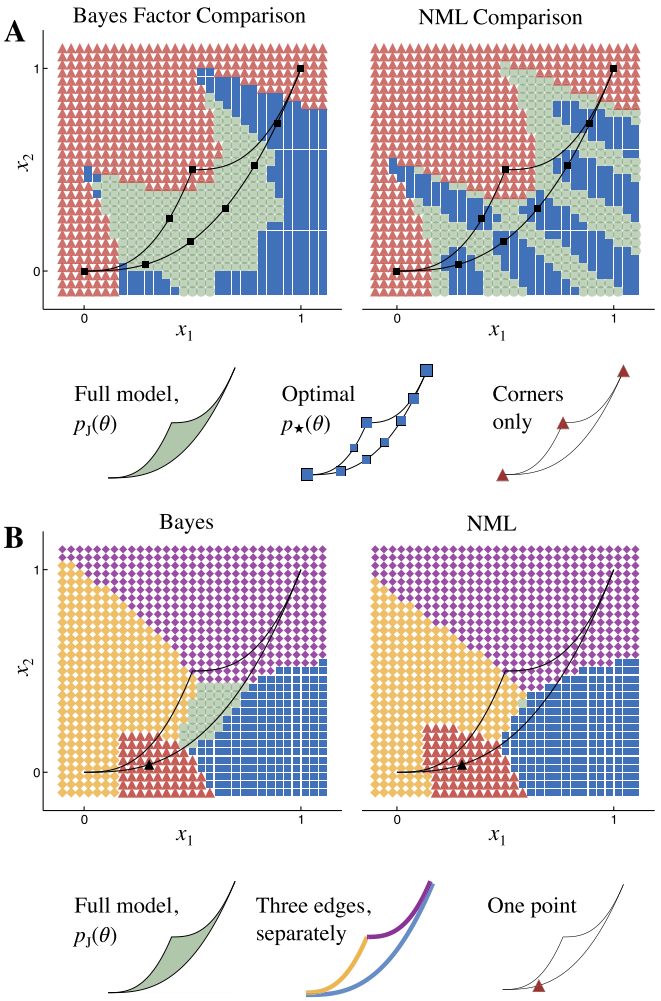

Fig. S1. Model selection from data point x. Each large plot shows which one of a list of effective models is preferred after observing data $x \in X$. In *A* and *B*, *Left* the criterion is maximizing Eq. **S1**, and in *A* and *B*, *Right* the criterion is Eq. **S2**. We study the same exponential model considered above, with $\sigma = 1/10$. *A* compares our optimal effective model with prior $p_\star(\theta)$ to the full model (with the Jeffreys prior) and to an even simpler model whose prior is just three delta functions (these are shown in *A*, *Lower*). *B* compares the full model to three different one-dimensional models, each allowing only one edge of Θ (with a uniform prior along this, i.e., the one-dimensional Jeffreys prior), and also to a trivial model (just one point), again with colors as indicated in *B*, *Lower*.

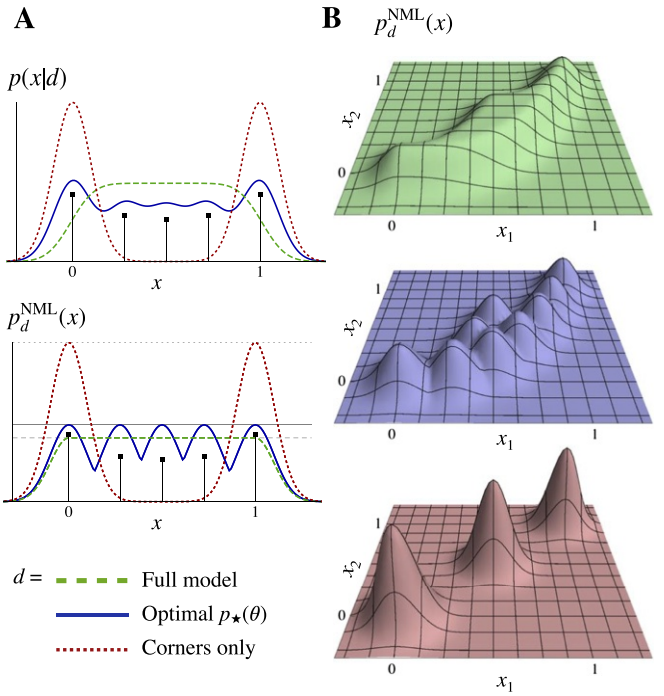

Fig. S2. Distributions over X. *A* shows $p(x|d)$ (Eq. **S1**) and $p_d^{\mathrm{NML}}(x)$ (Eq. **S2**) for the one-parameter Gaussian example Eq. **2** with $\sigma = 1/10$. The three models being compared are the full model (with a flat prior), the simpler model defined by $p_\star(\theta)$, and a model with just the endpoints of the line. Under the Bayes factor comparison, the $p_\star(\theta)$ model would never be preferred here. *B* draws $p_d^{\mathrm{NML}}(x)$ for the two-parameter exponential model (Eq. **5**), for the complete model, for the effective model defined by the support of $p_\star(\theta)$, and for an even simpler model allowing only three points—the same three models compared in Fig. S1*A*. Note that $p_d^{\mathrm{NML}}(x)$ is always a constant on the allowed region $x \in y(\Theta_d)$.

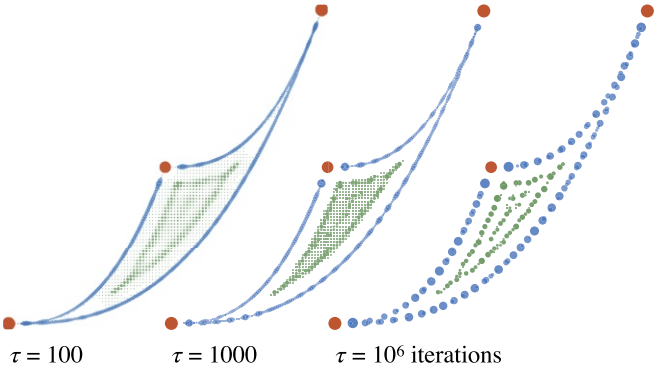

Fig. S3. Convergence of the BA algorithm (Eq. **S4**) for the exponential model. This shows $p_\star(\vec{y})$ for the case $\sigma = 1/50$, with $\vec{y}$ discretized on a grid of spacing $1/100$ in the bulk and $1/200$ along the boundaries of the allowed region.