Statistical model selection and prediction of systems' responses to exogenous perturbations

Bryan C. Daniels[1,2] and Ilya Nemenman[2]

***Short Abstract* — Typical models of cellular regulation strive for a detailed, microscopic description, faithfully reproducing the nuances of a myriad of molecular interactions. However, when experimental data is limited, this can lead to overfitting and a degraded quality of the model's predictions. We argue that, when a system's microscopic interactions are not well-resolved by the available data, a better option might be to create a phenomenological model that is complex enough to fit the data, yet simple enough to avoid overfitting. We implement this idea by defining a hierarchy of increasingly complex molecular interaction models and by using Bayesian model selection to choose the best among them. We test the method on synthetic data and find that phenomenological models inferred this way often outperform detailed, correct molecular models in making predictions about responses of the system to signals yet unseen.**

***Keywords* — Bayesian model selection, s-systems, power law models, phenomenological models, prediction**

I. Motivation

A central goal of any modeling effort is to make predictions regarding experimental conditions that have not yet been observed. Overly simple models will not be able to fit the original data well, but overly complex models are likely to overfit the data and thus produce bad predictions. An example of a model with many states and many parameters — measuring the net phosphorylation of a protein species with n phosphorylation sites — demonstrates the danger of inferring the parameters of a complex model from a small amount of data: Even if we know the exact form of the underlying microscopic kinetics that created a set of (synthetic) data, the best fit model parameters produce predictions that are often absurd.

II. Methods

A. Calculating the Bayesian posterior probability

When faced with limited data to constrain a complex, nonlinear system with uncertain topology and/or parameters, the best model is likely to be a phenomenological one in which we gradually add complexity until a balance is struck to best fit the data without overfitting. This balance can be quantified by the Bayesian posterior probability that a given model produced the data. In certain limits, this posterior probability is the sum of the maximum likelihood error (how well the model fits the data) and a term measuring the volume of parameter space that adequately fits the data (which penalizes for overly complex models); this posterior probability is directly related to the expected prediction error [1], and it forms the core of the model selection approach known as Bayesian Information Criterion (BIC). The criterion needs modification for biochemical systems since the estimation of the penalty term is complicated by "sloppiness" [2], the inability of the data to constrain certain directions in parameter space.

B. Model hierarchy

The process of Bayesian model selection requires a series of models of increasing complexity that 1) are nested (each model includes all parts of the previous model), 2) form a one-dimensional hierarchy, and 3) are guaranteed, as the model complexity grows, to approximate any data set arbitrarily well. Under these conditions, on average, a single model in the hierarchy is guaranteed to achieve the best posterior probability. We have constructed a hierarchy of interactions among species that satisfies these conditions.

What type of models should we use in this series? The simplest might be a polynomial fit of increasing degree, but this does not incorporate the fact that molecular systems are governed by continuous ODE dynamics, with interactions among nodes taking typical sigmoidal shapes. These constraints are better captured by the s-system power-law network [3] or sigmoidal network [4] formalisms, where we increase the complexity by adding dynamical variables.

III. Results

We test the selection process by fitting simulated output from various typical biochemical models, including the above phosphorylation model and a 3-gene transcription network. This allows us to explicitly verify that models with higher posterior probability do in fact make better predictions. Models that incorporate more details of the underlying dynamics, but are not unnecessarily complex, typically perform better in making predictions.

[1] Laboratory of Atomic and Solid State Physics, Physics Department, Cornell University, Ithaca, NY. E-mail: bdaniels@physics.cornell.edu

[2] Computer, Computation and Statistical Sciences Division and Center for Nonlinear Studies, Los Alamos National Lab, Los Alamos, NM. E-mail: ilya@menem.com

References

[1] Bialek W, Nemenman I, Tishby N. *Neural Comput* **13**, 2409 (2001).
[2] Gutenkunst, RG, et al. *PLOS Comput Biol* **3**, e189 (2007).
[3] Savageau, MA, Voit EO. *Math Biosci* **87**, 83 (1987).
[4] Beer, RD. *Neural Comp* **18**, 3009 (2006).