Automated adaptive inference of dynamical phenomenological models in systems biology

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

***Short Abstract* — Dynamical models of cellular regulation often consist of large and intricate networks of interactions at the molecular scale. Since individual interaction parameters are usually difficult to measure, these parameters are often estimated implicitly, using statistical fits. This can lead to overfitting and degradation in the quality of models' predictions. Here we study phenomenological models that adapt their level of detail to the amount of available data, leading to accurate predictions even when microscopic details are not well understood. 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, sigmoidal networks, phenomenological models, prediction**

I. Introduction

An important goal of any modeling effort is to make predictions regarding experimental conditions that have not yet been observed. In order to make such predictions, large intricate models of microscopic cellular processes require similarly large datasets of microscopic experimental data. However, it is more common to have data that describes aggregate input/output properties (e.g., total receptor phosphorylation) rather than the detailed microscopic dynamics (e.g., site-specific phosphorylation kinetics data). In such situations, fitting parameters to a highly complex model can cause overfitting and poor predictions, even if the model structure is known well. How should modeling be done when the microscopics are unknown?

We propose a systematic procedure to find predictive phenomenological models based on any amount of available time series data. We use an ordered hierarchy of dynamical models that can account for dynamics with arbitrary complexity. To select the best model within the hierarchy, we use a modification of the Bayesian Information Criterion [1] that accounts for the typical "sloppiness" of the Fisher information matrix in large dynamical systems [2].

Multiple machine learning methods have recently been proposed to infer networks with an observed input/output relationship [3,4]. Our approach has the advantages of scaling in complexity with the amount of available data and remaining computationally efficient.

[1]Center for Complexity and Collective Computation, Wisconsin Institute for Discovery, UW-Madison. E-mail: bdaniels@discovery.wisc.edu

[2]Departments of Physics and Biology, Emory University, Atlanta, Georgia. E-mail: ilya.nemenman@emory.edu

II. Methods

A. Model hierarchy

A hierarchy of models is guaranteed to produce, on average, a statistically consistent model selection and accurate inference and predictions if: 1) models are nested, such that each next model includes all parts of the previous model, 2) the nesting is ordered, so that, for any two models, one always completely includes the other, and 3) the hierarchy is complete, so that data of arbitrary complexity can be fit by some sufficiently complex model in the hierarchy [5]. We find that choosing a model hierarchy that represents typical behavior of cellular interactions leads to better performance. For instance, while a polynomial fit of increasing degree produces simple models, formalisms such as S-systems [6] or recurrent sigmoidal networks [7] match biological behavior better. Further, these formalisms allow for construction of model hierarchies that obey the three above conditions, can approximate arbitrary dynamical nonlinearities, and can account for unobserved variables.

B. Modified Bayesian Information Criterion

The Bayesian Information Criterion (BIC) is a natural approach for making a tradeoff between model complexity and fit to in-sample data. It produces the maximally predictive models in the limit of well-constrained parameters [1]. In the case of large dynamical network models, parameters typically remain unconstrained even for large amounts of data, evident in the so called "sloppy" spectrum of the Fisher information matrix [2]. We thus modify BIC to account for unconstrained directions in parameter space.

C. Test cases

We test our inference procedure by fitting simulated output from various typical biochemical models, including an n-site phosphorylation model and a model of oscillations in yeast glycolysis [8]. In the undersampled regime, the inferred phenomenological models outperform microscopically accurate models of the processes.

References

[1] Bialek W, Nemenman I, Tishby N (2001) *Neural Comput* **13**, 2409.
[2] Gutenkunst RG, et al. (2007) *PLOS Comput Biol* **3**, e189.
[3] Schmidt MD, et al. (2011) *Phys Biol* **8**, 055011.
[4] Francois P, Siggia ED (2008) *Phys Biol* **5**, 026009.
[5] Nemenman I (2005) *Neural Comput* **17**, 2006.
[6] Savageau MA, Voit EO (1987) *Math Biosci* **87**, 83.
[7] Beer RD (2006) *Neural Comp* **18**, 3009.
[8] Ruoff P, et al. (2003) *Biophys Chem* **106**, 179.