CellPress

Opinion

The Importance of Falsification in Computational Cognitive Modeling

Stefano Palminteri,[1,2,*,‡] Valentin Wyart,[1,2,*,‡] and
Etienne Koechlin[1,2,*]

In the past decade the field of cognitive sciences has seen an exponential growth in the number of computational modeling studies. Previous work has indicated why and how candidate models of cognition should be compared by trading off their ability to predict the observed data as a function of their complexity. However, the importance of falsifying candidate models in light of the observed data has been largely underestimated, leading to important drawbacks and unjustified conclusions. We argue here that the simulation of candidate models is necessary to falsify models and therefore support the specific claims about cognitive function made by the vast majority of model-based studies. We propose practical guidelines for future research that combine model comparison and falsification.

Trends

Computational modeling has grown exponentially in cognitive sciences in the past decade.

Model selection most often relies on evaluating the ability of candidate models to predict the observed data.

The ability of a candidate model to generate a behavioral effect of interest is rarely assessed, but can be used as an absolute falsification criterion.

Recommended guidelines for model selection should combine the evaluation of both the predictive and generative performance of candidate models.

Complementary Roles of Comparison and Falsification in Model Selection

Computational modeling has grown considerably in cognitive sciences in the past decade (Figure 1A). Computational models of cognition are also becoming increasingly central in neuroimaging and psychiatry as powerful tools for understanding normal and pathological brain function [1–5]. The importance of computational models in cognitive sciences and neurosciences is not surprising; because the core function of the brain is to process information to guide adaptive behavior, it is particularly useful to formulate cognitive theories in computational terms [6,7] (Box 1). Similarly to cognitive theories, computational models should be submitted to a selection process. We argue here that the current practice for model selection often omits a crucial step: **model falsification** (see Glossary).

One universally recognized heuristic for theory selection is Occam’s law of parsimony: ‘*pluralitas non est ponenda sine necessitate*’ (plurality is never to be posited without necessity). This principle dictates that among ‘equally good’ explanations of data, the less complex explanation should be held as true. More formally, a trade-off exists between the complexity of a given model (which specifically grows with its number of ‘free’ and adjustable parameters) and its goodness-of-fit (the likelihood of the observed data given the model). Different quantitative criteria (e.g., the Bayesian information criterion, Bayes factor, and other approximations of the model evidence) have been proposed to take **model parsimony** into account when comparing different models. These criteria are based on the **predictive performance** of a model, in other words its ability to predict the observed data [8–11]. We refer to them as relative comparison criteria because they imply no absolute criterion for model selection or rejection. Following these criteria, the ‘winning’ (or ‘best’) model is the model with the strongest evidence (i.e., trading off goodness-of-fit with complexity) compared to rival models [8,12]. Various statistical methods can then be used to test whether there is significantly stronger evidence in favor of the winning model than rival models.

[1] Laboratoire de Neurosciences Cognitives, Institut National de la Santé et de la Recherche Médicale, Paris, France
[2] Institut d’Étude de la Cognition, Departement d’Études Cognitives, École Normale Supérieure, Paris, France
[‡] These authors contributed equally.

*Correspondence:
stefano.palminteri@ens.fr (S. Palminteri),
valentin.wyart@ens.fr (V. Wyart),
etienne.koechlin@ens.fr (E. Koechlin).

http://dx.doi.org/10.1016/j.tics.2017.03.011

© 2017 Elsevier Ltd. All rights reserved.

Box 1. Delineating Computational Modeling Approaches

In cognitive sciences, computational models can be used either as analytical tools for analyzing empirical data or as instantiations of cognitive hypotheses. In the first case, the typical results consist of comparing model parameters across conditions or subjects [27], in other words computational models are treated as statistical models, similar to multiple regressions. In this approach, model comparison is not crucial because the models are not instantiations of cognitive theories.

As instantiations of cognitive theories, computational models can target different levels of description. Clearly identifying the target level should precede a model comparison analysis. A key distinction is between aggregate versus mechanistic models [9]. Aggregate models aims to describe average behaviors using a synthetic mathematical model, such as an exponential learning curve [28]. Mechanistic models aim to explain how behaviors are generated, such as the 'delta rule' in reinforcement learning [29]. Because these two types of models do not target the same level of description, there is no reason to arbitrate between aggregate versus mechanistic models. For example, an aggregate exponential learning curve could be derived formally from a 'delta rule' such that both models are equivalent. The distinction between aggregate and mechanistic models has been further developed by Marr [6], who proposed three distinct levels of description. The 'computational' level corresponds to the goal of the model. The 'representational' or 'algorithmic' level corresponds to a computational model formulated in terms of the mathematical operations (algorithms) that transform inputs into outputs (representations). Finally, the 'physical' or 'implementational' level corresponds to the biological implementation of a computational model in the brain (or an artificial device). Again, there is no reason to arbitrate between models across levels of description. In addition, the comparison of models has different meanings at the 'computational', 'algorithmic', and 'physical' levels. At the 'computational' level, model comparison informs about the actual task that subjects realize, whereas at the 'algorithmic' level model comparison informs about the way subjects perform this task [30]. Because simulating a model requires an algorithm to be specified, it is essential to clearly mention whether the model reflects a hypothesis at the computational or algorithmic level.

However, contemporary epistemology recognizes that parsimony is not the heuristic required for selecting theories. Proposing a new theory requires researchers to report experimental data that contradict (or 'falsify') an existing theory, whereas the new theory is able to account for these data (along with previous ones) [13,14]. Falsifying a cognitive model relies on showing that it is unable to account for a specific behavioral (or neural) effect of interest. We propose to define the inability to account for a specific effect of interest as an absolute rejection criterion during model selection [15]. The ability of a cognitive model to reproduce (or not) the effect of interest – which we refer to as its **generative performance** – needs to be assessed by simulating the model and comparing the simulated data to the observed data. Various statistical approaches – both frequentist (e.g., *t*-tests, analyses of variance) and Bayesian – can then be used to test whether the simulated and observed effects are different, in which case the simulated model can be rejected outright irrespective of its comparison to other models.

Relative comparison criteria are inappropriate for falsifying models because (i) they focus on relative evidence in favor of the winning model and against rival models, and (ii) they are blind to the ability of candidate models to produce any specific effect of interest found in the data.

To illustrate the complementary roles of model comparison (based on model fitting) and model falsification (based on model simulations), we sketch two recent examples taken from the learning and decision-making literature [16,17].

In the first study, the authors studied the origin of human choice variability in a canonical decision-making task involving the categorization of sequences of visual stimuli of variable lengths (Figure 2A) [16]. They compared a standard model in which variability arose from a noisy response selection process to a new model in which variability arose from errors in the inference process. In this example, the two models had the same complexity – in other words one variability parameter located either at the inference or response selection stages. The authors first assessed the predictive performance of the two models, which provided substantial evidence in favor of the 'noisy inference' model. Then, to determine why the 'noisy inference'

Glossary

Generative performance: the ability of a given model to generate the data. The generative performance is evaluated by comparing model simulations to the actual data. For this comparison both frequentist and Bayesian statistics can be used.

Model falsification: showing through model simulations that a given model is not able to generate a specific behavioral effect of interest. The simulated data should be generated using the best-fitting parameter values. Ideally, this 'model falsification' step should include two related results: (i) the behavioral phenomenon is not detectable in the simulated data, and (ii) a significant difference between observed and simulated data should be detected. Statistical tests used in model falsification could belong to both Bayesian and frequentist statistical traditions.

Model generalizability: evaluating the ability of the best-fitting model and the best-fitting parameters to predict the data out-of-sample.

Model parsimony: the opposite of model complexity, which is classically indexed by the number of free/adjustable parameters of a given model.

Model recovery: a procedure consisting of generating synthetic data from a known candidate model and subsequently verifying the ability of a relative model comparison criterion to correctly identify the model used to generate the synthetic data.

Predictive performance: the ability of a given model to predict the data. Typically the predictive performance is instantiated by the likelihood of observing the experimental data given the model. The predictive performance of models is used to calculate various approximations of the model evidence (e.g., BIC, AIC, and others).

Download English Version:

https://daneshyari.com/en/article/4762189

Download Persian Version:

https://daneshyari.com/article/4762189

Daneshyari.com