When a good fit can be bad

Mark A. Pitt and In Jae Myung

How should we select among computational models of cognitive architecture? Although it is commonplace to measure how well each model fits the data, this is insufficient. Good fits can be misleading because they can result from properties of the model that have nothing to do with it being a close approximation to the cognitive process of interest (e.g. overfitting). Selection methods are introduced that factor in these properties when measuring fit. Their success in outperforming standard goodness-of-fit measures stems from a focus on measuring the generalizability of a model's data-fitting abilities, which should be the goal of model selection.

The explosion of interest in modeling cognitive processes over the past 20 years has fueled the cognitive sciences in many ways. Not only has it opened up new ways of thinking about research problems and possible solutions, but it has also enabled researchers to gain a better understanding of their theories by simulating a computational instantiation of it. Modeling is now sufficiently mainstream that one can get the impression that the models themselves are replacing the theories from which they evolved.

What has not kept pace with the advances and interest in modeling is the development of methods for evaluating and testing the models themselves. A model is not interchangeable with a theory, but only one of many possible quantitative representations of it. A thorough evaluation of a model requires methods that are sensitive to its quantitative form. Criteria used for evaluating theories [1], such as testing their performance in an experimental setting, do not speak to the quality of the choices that are made in building their quantitative counterparts (i.e. choice of parameters, how they are combined) or their ramifications. The paucity of such model selection methods is surprising given the centrality of the problem itself. What could be more fundamental than deciding between two alternative explanations of a cognitive process?

How not to compare models

Mathematical models are frequently tested against one another by evaluating how well each fits the data generated in an experiment or simulation. Such a test makes sense given that one criterion of model performance is that it reproduce the data. A goodness-of-fit measure (GOF; see Glossary) is invariably used to measure their adequacy in achieving this goal. What is measured is how much a model's predictions deviate from the observed data [2,3]. The model that provides the best fit (i.e. smallest deviation) is favored. The logic of this choice rests on the assumption that the model that provides the best fit to all data must be a closer approximation to the cognitive process under investigation than its competitors [4].

Such a conclusion is reasonable if measurements were made in a noise-free (i.e. errorless) system. One of the biggest challenges faced by cognitive scientists is that human and animal data are noisy. Error arises from several sources, such as the imprecision of our measurement tools, variation in participants and their performance over time. The problem of random
**Glossary**

**Complexity**: the property of a model that enables it to fit diverse patterns of data; it is the flexibility of a model. Although the number of parameters in a model and its functional form can be useful for gauging its complexity, a more accurate and intuitive measure is the number of distinct probability distributions that the model can generate by varying its parameters over their entire range. Details of this ‘geometric’ complexity measure can be found in [a].

**Functional form**: the way in which the parameters (θ) and data (x) are combined in a model’s equation: y = αx + β, y = α + βx, and y = α + βx + γ have the same number of parameters but different functional forms (multiplicative versus additive).

**Generalizability**: the ability of a model to fit all data samples generated by the same cognitive process, not just the currently observed sample (i.e., the model’s expected GOF with respect to new data samples). Generalizability is estimated by combining a model’s GOF with a measure of its complexity.

**Goodness of fit (GOF)**: the precision with which a model fits a particular sample of observed data. The predictions of the model are compared with the observed data. The discrepancy between the two is measured in a number of ways, such as calculating the root mean squared error between them.

**Minimum Description Length (MDL)**: a versatile measure of generalizability. MDL was developed within algorithmic coding theory in computer science [b], where the goal of model selection is to choose the model that permits the greatest compression of data in its description. Regularities in the data are assumed to imply redundancy. The more the data can be compressed by the model by extracting this redundancy, the more that is learned about the cognitive process.

**Overfitting**: the case where, in addition to fitting the main trends in the data, a model also fits the microvariation from this main trend at each data point. Compare the middle and right graph inserts in Fig. 1.

**Parameters**: variables in a model’s equation that represent mental constructs or processes; they are adjusted to improve a model’s fit to data. For example, in the model y = αx + β, α and β are parameters.

**Probability density function**: a function that specifies the probability of observing each outcome of a random variable given the value of a model’s parameter.

**References**


---

In the field of cognitive sciences, there is a growing interest in understanding the generalizability of models and how they fit data. A model’s fit to a particular sample of observed data can be assessed using various GOF measures, such as the Root Mean Squared Error (RMSE), which is insensitive to the different sources of variation in the data, whether it is random error or due to the cognitive process of interest. This is particularly important because many models are capable of fitting a dataset reasonably well. However, the set of such candidate models is potentially quite large, which can only ever infer the likelihood with which each model under consideration generated the data. In this regard, generalizability is a statistical inference problem. It is no different conceptually from estimating the replicability of an experimental result using inferential statistics or inferring the characteristics of a population from a sample.

A handful of generalizability measures have been proposed in the last 30 years [6]. Typically, all include a measure of GOF to assess a model’s fit to the data (see Box 2). Terms encoding information about the model itself are included to level the playing field among models so that one model, by virtue of its design choices (i.e. mathematical instantiation) does not have an inherent advantage over its competitors in fitting the data best, and thus being selected. By nullifying such model-specific properties, the model that is the best approximation to the cognitive process under study, and not simply the one that absorbs the most variation in the data, will be selected.

**Measures of generalizability**

Early measures of generalizability such as the Akaike information criterion (AIC) [7,8] and Bayesian information criterion (BIC) [9] addressed the most salient differences among models: the number of free parameters. As is generally well known, a model with many free parameters can provide a better fit to a data sample than a model with fewer parameters, even if the latter generated the data. The second term in these measures includes a count of the number of parameters (k). AIC and BIC penalize a model more as the number of parameters increases. To be selected, the model with more parameters must overcome this penalty and provide a substantially better fit than a model with fewer parameters. That is, the superior fit obtained with the extra parameters must justify the necessity of those parameters in fully capturing the cognitive process.

An equally salient but much less tangible dimension along which models also differ is in their functional form, which refers to the way in which the parameters are combined in a model’s equation. More sophisticated selection methods, such as Bayesian model selection (BMS) [10] and MINIMUM DESCRIPTION LENGTH (MDL) [11,12], are sensitive to a model’s functional form as well as to the number of parameters. Functional form is taken into account in the third term in the MDL measure. In BMS, both are hidden in the integral. (Cross Validation, although not listed, is another selection method that is thought to be sensitive to both dimensions of model complexity. It involves applying GOF in a non-standard way.)

The second and third terms in MDL together provide a measure of a model’s complexity. Conceptually, complexity refers to that characteristic
Box 1. Why GOF alone is a poor measure of model selection

The ability of a model to fit data is a necessary condition that all models must satisfy to be taken seriously. GOF measures such as RMSE are inappropriate as model selection methods because all they do is assess fit. This myopic focus is problematic when variation in the data can be due to other factors (e.g., random sampling, individual differences) as well as the cognitive process under study.

The severity of the problem is shown in Table I, which contains the results of a model recovery simulation using RMSE. Four datasets were generated from a combination of the two models (MA and MB), defined as follows:

$M_a: y = (a - 1) \cdot y + \epsilon$

$M_b: y = b \cdot \epsilon + \epsilon$

where $a, b > 0$. Datasets were generated from each in the four conditions (rows). In the first condition, all 100 samples were generated by $M_a$ with $a = 0.4$ and only sampling error introduced as noise. In the second condition, variation due to individual differences was also added to the data by using a different parameter value ($a = 0.6$) half of the time. In the third condition, half of the data were generated by model $M_a$ and half by $M_b$. Condition four is the reverse of condition one, with the data plus sampling error being generated by $M_b$. Models $M_a$ and $M_b$ were then fitted to the data in each condition using RMSE. The mean RMSE fits, along with the percentage of time each model provided the best fit, are shown on the two right-most columns of the Table.

A good model selection method must be able to ignore irrelevant variation in the data (e.g., sampling error, individual differences that are not being modeled) and recover the model that generated the data. That is, the selection method must be capable of differentiating between the variation that the model was designed to capture and the variation due to noise. RMSE fails miserably at this task. $M_a$ was chosen 100% of the time in all four conditions and the mean fit is substantially better than that of $M_b$. $M_a$ never provided a better fit than $M_b$, even when some or all of the data were generated by $M_b$ (conditions 1–3). This is why a good fit can be bad.

Readers who have some familiarity with modeling might not be surprised by the results given that $M_b$ has two more parameters than $M_a$. A model with more parameters will always provide a better fit, all other things being equal [a]. The typical solution to this problem is to control for the number of parameters, but there are at least two reasons why this fix is unsatisfactory. The most obvious is that it limits model comparison to those situations in which the number of parameters is equated across models. The diversity of models in the cognitive sciences can make this a significant impediment in doing research. Less obvious although even more important is that a model’s data-fitting abilities are also affected by other properties of the model, such as its functional form [b, c, d]. Unless they are taken into account by the selection method, simply equating for number of parameters will not place the models on an equal footing.

In summary, it may make perfect sense to use GOF to determine whether a given model can even pass the test of fitting a dataset reasonably well (i.e. capturing the main trends), but going beyond this and comparing such fits between models, although intuitive, is risky.

References


Box 2. Model selection criteria as measures of generalizability

Listed in Table II are two GOF measures (RMSE, PVAF) and four generalizability measures (AIC, BIC, BMS, MDL). Except for BMS, in which the likelihood function is integrated over the parameter space, the measures of generalizability use the maximized log-likelihood, that is, $ln f(y|\theta_0)$, as a GOF measure, the minus of which represents lack of fit. This fit index is combined with a measure of model complexity to yield a generalizability measure. The four generalizability criteria differ from one another in the conceptualization of model complexity. In AIC, the number of parameters (k) is the only dimension of complexity that is considered, whereas BIC also considers sample size (n). BMS and MDL go one step further and also take into account the functional form of a model’s equation. In MDL, this is reflected through the third term of the criterion equation whereas in BIC it is hidden in the integral. These selection methods, except for PVAF, prescribe that the model that minimizes the given criterion should be chosen.

Reference


Table I. Results of a model recovery simulation in which a GOF measure (RMSE) was used to discriminate models when the source of the error was varied.

<table>
<thead>
<tr>
<th>Condition (sources of variation)</th>
<th>Model the data were generated from</th>
<th>Model fitted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M_a$ $a = 0.4$</td>
<td>$M_b$ $a = 0.6$</td>
</tr>
<tr>
<td>(1) Sampling error</td>
<td>100</td>
<td>–</td>
</tr>
<tr>
<td>(2) Sampling error + individual differences</td>
<td>50 50</td>
<td>– 50</td>
</tr>
<tr>
<td>(3) Different models</td>
<td>– 50 50</td>
<td>0.075 (0%)</td>
</tr>
<tr>
<td>(4) Sampling error</td>
<td>– 50 100</td>
<td>0.079 (0%)</td>
</tr>
</tbody>
</table>

Table II. Two GOF Measures, four generalizability measures, and the dimensions of complexity to which each is sensitive

<table>
<thead>
<tr>
<th>Selection method</th>
<th>Criterion equation</th>
<th>Dimensions of complexity considered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root Mean Squared Error</td>
<td>RMSE = (SSE/N)^0.5</td>
<td>None</td>
</tr>
<tr>
<td>Percent Variance Accounted For</td>
<td>PVAF = 100(1-SSE/SST)</td>
<td>None</td>
</tr>
<tr>
<td>Akaike Information Criterion</td>
<td>AIC = -2 ln(\hat{f}(\hat{\theta})) + 2k</td>
<td>Number of parameters</td>
</tr>
<tr>
<td>Bayesian Information Criterion</td>
<td>BIC = -2 ln(\hat{f}(\hat{\theta})) + k ln(n)</td>
<td>Number of parameters, sample size</td>
</tr>
<tr>
<td>Bayesian Model Selection</td>
<td>MDL = ln \left( \int \hat{f}(\hat{\theta}) d\theta \right) + (k2) ln(n2n2) + \text{det}((\hat{\theta}))</td>
<td>Number of parameters, sample size, functional form</td>
</tr>
</tbody>
</table>

In the equations above, $y$ denotes observed data, $\hat{\theta}$ is the model’s parameter, $\hat{\theta}_0$ is the parameter value that maximizes the likelihood function $f(y|\theta)$, $k$ is the number of parameters, $n$ is the sample size, $N$ is the number of data points fitted, SSE is the minimized sum of the squared errors between observations and predictions, SST is the sum of the squares total, $n(\theta)$ is the parameter prior density, $\hat{f}(\hat{\theta})$ is the Fisher information matrix in mathematical statistics $[a]$, $\text{det}$ denotes the determinant of a matrix, and $ln$ denotes the natural logarithm of base e.
of a model that makes it flexible and easily able to fit diverse patterns of data, usually by a small adjustment in one of its parameters. In Fig. 1, it is what enables the model (wavy line) in the lower right graph to provide a better fit to the data (dots) than that in the middle and left graphs. Both functional form and the number of parameters contribute to model complexity.

How complexity is related to generalizability and GOF: the dilemma of Occam’s razor

The notion of model complexity is a good vehicle with which to illustrate further the goal of generalizability and distinguish it from GOF. As demonstrated in Box 1, fit is maximized with GOF. Because additional complexity will improve fit, the two are positively related; this is depicted by the top function in Fig. 1. Generalizability, on the other hand, is not related to complexity so straightforwardly. Rather, its function follows the same trajectory as that of GOF up to a certain point, after which fit declines as complexity increases.

Why do the two functions diverge? The data being fitted have a certain degree of complexity that reflects the operation of the cognitive process. This point corresponds to the peak of the generalizability function. Any additional complexity beyond that needed to capture the underlying process (to the right of the peak) will cause the model to overfit the data by also capturing the microvariation due to random error, and thus reduce generalizability. The reason both functions overlap to the left half of the peak is that the model itself must be sufficiently complex to fit the data well. The model will underfit the data if it lacks the necessary complexity (i.e. it is too simple), as illustrated in the lower left graph in Fig. 1. The dilemma that is faced in trying to maximize generalizability should be clear: a delicate balance must be struck between sufficient complexity on the one hand and good generalizability on the other. MDL achieves this balance by choosing the model whose complexity is most justified by considering the complexity of the data relative to the complexity of the model.

Selection methods at work: the proof is in the pudding

Generalizability measures like BMS and MDL have thus far been developed for testing only those models that can be described in terms of a probability density function. Examples of such statistical models are models of categorization and models of information integration [14,15].

The following model-recovery tests demonstrate the relative performance of the three classes of selection methods shown in Table 1. The three models described (see Table footnote) were compared. In each simulation, 1000 datasets were generated from each model. Each selection method was then tested on its ability to determine which of the three models did in fact generate the 3000 datasets. A good selection method should be able to discern correctly the model that generated the data. The ideal outcome is one in which each model generalizes best only to data...
generated by itself. In the $3 \times 3$ matrices in Table 1, this corresponds to perfect recovery in the diagonal going from the upper left to the lower right. Errors (cells in the off diagonals) reveal a bias in the selection method toward either the more or less complex model.

The top matrix shows the results using PVAF, with the percentage of correct recoveries in each cell. The problem with using a GOF measure shows up in the first column of the matrix, where the data were generated by the one-parameter model $M_1$. $M_2$, never recovered its own data, with one of the two-parameter models always fitting the data best. Comparison of these data with the middle matrix shows that using AIC rectifies this problem. Because of its sensitivity to the number of parameters in a model, it does a reasonably good job of distinguishing between data generated by $M_1$ or $M_2$. By contrast, note how model recovery performance remains constant across these two matrices in columns 2 and 3. This is not surprising because AIC, like PVAF, is insensitive to functional form, which is the dimension along which $M_2$ and $M_3$ differ. Only when MDL is used is an improvement in model recovery performance observed in these columns (bottom matrix).

Why did MDL not perform perfectly, and why did it perform slightly worse than AIC and PVAF when the data came from $M_2$ (middle column)? Recall that, like a statistical test, model selection is an inference problem. The quality of the inference depends on the information that the selection method uses. Even though MDL makes use of all of the information available (data and the model), this does not guarantee success, but it greatly improve the chances of the inference being correct [16]. MDL will outperform selection methods such as AIC and PVAF most of the time, but neither it nor its Bayesian counterpart, BMS, are infallible.

The inferential nature of model selection makes it imperative to interpret model recovery results in the context of the data that were used in the test itself. Poor performance might not indicate a problem with the selection method, but rather a constraint on the resolving power of the selection method in discriminating which model could have generated the data. Conversely, biases in the selection method itself can masquerade as good model-recovery performance. In a recent comparison of BMS and RMSD, we demonstrated both of these outcomes [17].

Conclusion

Methods like AIC and MDL can give the impression that model selection can be automated and require minimal thought. However, choosing between competing models is no easier or less subjective than choosing between competing theories. Selection methods should be viewed as a tool that researchers can use to gain additional information about the models under consideration. These tools, however, are ignorant of other properties of the models, such as their plausibility or the quality of the data, so it is inappropriate to decide between them solely on what is learned from a test of generalizability.

References

5 Bishop, C.M. (1995) Neural Networks for Pattern Recognition, (Chapters 1 and 9), Oxford University Press