Model Evaluation and Selection

Daniel J. Navarro & In Jae Myung
Department of Psychology
Ohio State University
1827 Neil Avenue
Columbus OH 43210, USA.
{navarro.20, myung.1}@osu.edu

March 2, 2004

Abstract

Assessing the viability of a quantitative model given observed data is generally done in three steps: model fitting, testing, and selection. In model fitting, we find the parameter values that best fit observed data. This is followed by model testing in which null hypothesis significance tests allow us to evaluate the descriptive adequacy of a model's fit to data. Model selection concerns the issue of choosing the model, among a set of competing models, that generalizes best to as yet unseen data samples.

Keywords: Parameter estimation, model testing, Akaike information criterion, cross validation, Bayesian model selection, minimum description length.

Introduction

The main reason for building models is to link theoretical ideas to observed data, and the central question that we are interested in is "Is the model any good?" When dealing with quantitative models, we can at least partially answer this question using statistical tools. Before going into detail, there is a touchy, even philosophical, issue that one cannot ignore. A naive view of modeling is to identify the underlying process (truth) that has actually generated the data. This is an ill-posed problem, meaning that the solution is non-unique. The finite data sample rarely contains sufficient information to lead to a single process and also, is corrupted by unavoidable random noise, blurring the identification. An implication of noise-corrupted data is that it is not in general possible to determine with complete certainty that what we are fitting is the regularity, which we are interested in, or the noise, which we are not. A model that assumes a certain amount of error is present may be worse than a yet to be postulated model which can explain more of what we thought of as error in the first model. In short, identifying the true model based on data samples is an unachievable goal. Furthermore, the "truth" of any phenomenon is likely to be rather different from any proposed model. Ultimately, it is crucial to recognize that *all* models are wrong, and a realistic goal of modeling is to find a model that represents a "good" approximation to the truth in a statistically defined sense.

In what follows, we assume that we have a model $\mathcal{M}$ with k free parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, and a data set that consists of n observations $\boldsymbol{X} = (X_1, \ldots, X_n)$. Quantitative models generally come in two main types: They either assign some probability to the observed data $f(\boldsymbol{X}|\boldsymbol{\theta})$ (probabilistic models), or they produce a single predicted data set $\boldsymbol{X}^{prd}(\boldsymbol{\theta})$ (deterministic models). We should

note that most model testing and model selection methods require a probabilistic formulation, so it is commonplace to define a model as $\mathcal{M} = \{f(\boldsymbol{X}|\boldsymbol{\theta})|\ \boldsymbol{\theta} \in \Omega\}$ where Ω is the parameter space. When written in this form, a model can be conceptualized as a family of probability distributions.

Model Fitting

At a minimum, any reasonable model needs to be able to mimic the structure of the data: It needs to be able to "fit" the data. When measuring the goodness of a model's fit, we find the parameter values that allow the model to best mimic the data, denoted $\hat{\boldsymbol{\theta}}$. The two most common methods for this are *maximum likelihood estimation* (for probabilistic models) and *least squares estimation* (for deterministic models). In the maximum likelihood approach, introduced by Sir Ronald Fisher in the 1920s, $\hat{\boldsymbol{\theta}}$ is the set of parameter values that maximizes $f(\boldsymbol{X}|\boldsymbol{\theta})$, and is referred to as the maximum likelihood estimate (MLE). The corresponding measure of fit is the maximized log-likelihood $\hat{L} = \ln f(\boldsymbol{X}|\hat{\boldsymbol{\theta}})$. See [3] for a tutorial on maximum likelihood estimation with example applications in cognitive psychology.

Alternatively, the least squares estimate of $\hat{\boldsymbol{\theta}}$ is the set of parameters that minimizes the *sum of squared errors* (SSE) and the minimized SSE value is denoted by $\hat{E}$:

$$\hat{E} = \sum_{i=1}^{n} (X_i - X_i^{prd}(\hat{\boldsymbol{\theta}}))^2.$$

When this approach is employed, there are several commonly-used measures of fit. They are *mean squared error* $\text{MSE} = \hat{E}/n$, *root mean squared deviation* $\text{RMSD} = \sqrt{\hat{E}/n}$, and *squared correlation* (also known as proportion of variance accounted for) $r^2 = 1 - \hat{E}/\text{Var}(\boldsymbol{X})$. In the last formula $\text{Var}(\boldsymbol{X})$ is the variance of the data set $\sum_{i=1}^{n}(X_i - \bar{\boldsymbol{X}})^2$, where $\bar{\boldsymbol{X}}$ denotes the mean of $\boldsymbol{X}$. There is a nice correspondence between maximum likelihood and least squares, in that for a model with independent, identically and normally distributed errors, the same set of parameters is obtained as the one that maximizes the log-likelihood L but also minimizes the sum of squared errors SSE.

Model fitting yields goodness-of-fit measures, such as $\hat{L}$ or $\hat{E}$, that tell us how well the model fits the observed data sample but by themselves are not particularly meaningful. If our model has a minimized sum squared error of 0.132, should we be impressed or not? In other words, a goodness of fit measure may be useful as a purely descriptive measure, but by itself is not amenable to statistical inference. This is because the measure does not address the relevant question: "Does the model provide an *adequate* fit to the data, in a defined sense?". This question is answered in model testing.

Model Testing

Classical null hypothesis testing is a standard method of judging a model's goodness-of-fit. The idea is to set up a null hypothesis that "the model is correct," obtain the P-value, and then make a decision about rejecting or retaining the hypothesis by comparing the resulting P-value with the alpha level.

For discrete data such as frequency counts, the two most popular methods are the Pearson chi-square (χ^2) test and the log-likelihood ratio test (G^2), which have test statistics given by

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(X_i - X_i^{prd}(\hat{\boldsymbol{\theta}})\right)^2}{X_i^{prd}(\hat{\boldsymbol{\theta}})};\ G^2 = -2\sum_{i=1}^{n} X_i \ln \frac{X_i^{prd}(\hat{\boldsymbol{\theta}})}{X_i},$$

where ln is the natural logarithm of base e. Both of these statistics have the nice property that they are always non-negative, and are equal to zero when the observed and predicted data are in full agreement. In other words, the larger the statistic, the greater the discrepancy. Under the null hypothesis, both are approximately distributed as a χ^2 distribution with $(n - k - 1)$ degrees of freedom, so we would reject the null if the obtained P-value is larger than some critical value obtained by setting an appropriate α level.

For continuous data such as response time, goodness-of-fit tests are a little more complicated, since there are no general-purpose methods available for testing the validity of a single model, unlike the discrete case. Instead, we rely on the *generalized likelihood ratio test* that involves two models. In this test, in addition to the theoretically motivated model, denoted by $\mathcal{M}_r$ (reduced model), we create a second model, $\mathcal{M}_f$ (full model), such that the reduced model is obtained as a special case of the full model by fixing one or more of $\mathcal{M}_f$'s parameters. Then, the goodness of fit of the reduced model is assessed by the following G^2 statistic:

$$G^2 = 2\left(\ln \hat{L}_f - \ln \hat{L}_r\right),$$

recalling that $\hat{L}$ denotes the maximized log-likelihood. Under the null hypothesis that the theoretically motivated, reduced model is correct, the above statistic is approximately distributed as χ^2 with degrees of freedom given by the difference in the number of parameters $(k_f - k_r)$. If the hypothesis is retained (rejected), then we conclude that the reduced model $\mathcal{M}_r$ provides (does not provide) an adequate description of the data (see [4] for an example application of this test).

Model Selection

What does it mean that a model provides an adequate fit of the data? One should not jump to the conclusion that one has identified the underlying regularity. A good fit merely puts the model on a list of candidate models worthy of further consideration. It is entirely possible that there are several distinct models that fit the data well, all passing goodness of fit tests. How should we then choose among such models? This is the problem of model selection.

In model selection, the goal is to select the one, among a set of candidate models, that represents the closest approximation to the underlying process in some defined sense. Choosing the model that best fits a particular set of observed data will not accomplish the goal. This is because a model can achieve a superior fit to its competitors for reasons unrelated to the model's exactness. For instance, it is well known that a complex model with many parameters and highly nonlinear form can often fit data better than a simple model with few parameters even if the latter generated the data. This is called *overfitting*.

Avoiding overfitting is what every model selection method is set to accomplish. The essential idea behind modern model selection methods is to recognize that, since data are inherently noisy, an ideal model is one that captures only the underlying phenomenon, not the noise. Since noise is idiosyncratic to a particular data set, a model that captures noise will make poor predictions about future events. This leads to the present-day "gold standard" of model selection, *generalizability*. Generalizability, or predictive accuracy, refers to a model's ability to predict the statistics of future, as yet unseen, data samples from the same process that generated the observed data sample.

The intuitively simplest way to measure generalizability is to estimate it directly from the data, using cross-validation (CV; [10]). In cross-validation, we split the data set into two samples, the calibration sample $\boldsymbol{X}_c$ and the test sample $\boldsymbol{X}_t$. We first estimate the best-fitting parameters by fitting the model to $\boldsymbol{X}_c$ which we denote $\hat{\boldsymbol{\theta}}(\boldsymbol{X}_c)$. The generalizability estimate is obtained by measuring the fit of the model to the test sample at those original parameters, that is, $\hat{\boldsymbol{\theta}}(\boldsymbol{X}_c)$,

$$\mathrm{CV} = \ln f(\boldsymbol{X}_t|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_c)).$$

The main attraction of CV is its ease of implementation (see [4] for its application example for psychological models). All that is required is a model fitting procedure and a resampling scheme. One concern with CV is that there is a possibility that the test sample is not truly independent of the calibration sample: Since both were produced in the same experiment, systematic sources of error variation are likely to induce correlated noise across the two samples, artificially inflating the CV measure.

An alternative approach is to use theoretical measures of generalizability based on a single sample. In most of these theoretical approaches, generalizability is measured by suitably combining goodness-of-fit with model complexity. The practical difference between them is the way in which complexity is measured. One of the earliest measures of this kind was the Akaike information criterion (AIC; [1]), which treats complexity as the number of parameters k:

$$\mathrm{AIC} = -\ln f(\boldsymbol{X}|\hat{\boldsymbol{\theta}}) + k,$$

The method prescribes that the model minimizing AIC should be chosen. AIC seeks to find the model that lies "closest" to the true distribution, as measured by the Kullback-Leibler [8] discrepancy. As shown in the above criterion equation, this is achieved by trading the first, minus goodness-of-fit (lack of fit) term of the right hand side for the second complexity term. As such, a complex model with many parameters, having a large value of the complexity term, will not be selected unless its fit justifies the extra complexity. In this sense, AIC represents a formalization of the principle of Occam's razor, which states "Entities should not be multiplied beyond necessity" (William of Occam, ca. 1290 - 1349).

Another approach is given by the much older notion of Bayesian statistics. In the Bayesian approach, we assume that a priori uncertainty about the value of model parameters is represented by a prior distribution $\pi(\boldsymbol{\theta})$. Upon observing the data $\boldsymbol{X}$, this prior is updated, yielding a posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{X}) \propto f(\boldsymbol{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. In order to make inferences about the model (rather than its parameters), we integrate across the posterior distribution. Under the assumption that all models are a priori equally likely (because the Bayesian approach requires model priors as well as parameter priors), Bayesian model selection chooses the model $\mathcal{M}$ with highest marginal likelihood defined as:

$$f(\boldsymbol{X}|\mathcal{M}) = \int f(\boldsymbol{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

The ratio of two marginal likelihoods is called a *Bayes factor* (BF; [2]), which is a widely used method of model selection in Bayesian inference. The two integrals in the Bayes factor are non-trivial to compute unless $f(\boldsymbol{X}|\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ form a conjugated family. Monte Carlo methods are usually required to compute BF, especially for highly parameterized models. A large sample approximation of BF yields the easily-computable Bayesian information criterion (BIC; [9])

$$\mathrm{BIC} = -\ln f(\boldsymbol{X}|\hat{\boldsymbol{\theta}}) + \frac{k}{2}\ln n.$$

The model minimizing BIC should be chosen. It is important to recognize that the BIC is based on a number of restrictive assumptions. If these assumptions are met, then the difference between two BIC values approaches twice the logarithm of the Bayes factor as n approaches infinity.

A third approach is minimum description length (MDL; [5]), which originates in algorithmic coding theory. In MDL, a model is viewed as a code that can be used to compress the data. That is, data sets that have some regular structure can be compressed substantially if we know what that structure is. Since a model is essentially a hypothesis about the nature of the regularities that we expect to find in data, a good model should allow us to compress the data set effectively. From an MDL standpoint, we choose the model that permits the greatest compression of data in its total description: That is, the description of data obtainable with the help of the model plus the description of the model itself. A series of papers by Rissanen expanded on and refined this idea, yielding a number of different model selection criteria (one of which was identical to the BIC). The most complete MDL criterion currently available is the *stochastic complexity* (SC; [7]) of the data relative to the model,

$$\mathrm{SC} = -\ln f(\boldsymbol{X}|\hat{\boldsymbol{\theta}}) + \ln \int f(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}(\boldsymbol{Y}))\, d\boldsymbol{Y}.$$

Note that the second term of SC represents a measure of model complexity. Since the integral over the sample space is generally non-trivial to compute, it is common to use the Fisher-information approximation (FIA; [6]): Under regularity conditions, the stochastic complexity asymptotically approaches

$$\mathrm{FIA} = -\ln f(\boldsymbol{X}|\hat{\boldsymbol{\theta}}) + \frac{k}{2}\ln\left(\frac{n}{2\pi}\right) + \ln \int_{\boldsymbol{\Theta}} \sqrt{\det I(\boldsymbol{\theta})}\, d\boldsymbol{\theta},$$

where $I(\boldsymbol{\theta})$ is the expected Fisher information matrix of sample size one, consisting of the covariances between the partial derivatives of L with respect to the parameters. Once again, the integral can still be intractable, but it is generally easier to calculate than the exact SC. As in AIC and BIC, the first term of FIA is the lack of fit term and the second and third terms together represent a complexity measure. From the viewpoint of FIA, complexity is determined by the number of free parameters (k) and sample size (n) but also by the "functional form" of the model equation, as implied by the Fisher information $I(\boldsymbol{\theta})$, and the range of the parameter space $\boldsymbol{\Theta}$.

When using generalizability measures, it is important to recognize that AIC, BIC and FIA are all *asymptotic* criteria, and are only guaranteed to work as n becomes arbitrarily large, and when certain regularity conditions are met. The AIC and BIC in particular can be misleading for small n. The FIA is safer (i.e., the error level generally falls faster as n increases), but it too can still be misleading in some cases. The SC and BF criteria are more sensitive, since they are exact rather than asymptotic criteria, and can be quite powerful even when presented with very similar models or small samples. However, they can be difficult to employ, and often need to be approximated numerically. The status of CV is a little more complicated, since it is not always clear what CV is doing, but its performance in practice is often better than AIC or BIC, though it is not usually as good as SC, FIA, or BF.

Conclusion

When evaluating a model, there are a number of factors to consider. Broadly speaking, statistical methods can be used to measure the descriptive adequacy of a model (by fitting it to data and testing those fits), as well as its generalizability and simplicity (using model selection tools). However, the strength of the underlying theory also depends on its interpretability, its consistency with other findings, and its overall plausibility. These things are inherently subjective judgements, but they are no less important for that. As always, there is no substitute for thoughtful evaluations and good judgement. After all, statistical evaluations are only one part of a good analysis.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds), *Second International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

[2] Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773-795.

[3] Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 47*, 90-100.

[4] Myung, I. J & Pitt, M. A. (2002). Mathematical modeling. In J. Wixten (ed.), *Stevens' Handbook of Experimental Psychology (Third Edition)*, pp. 429-459. New York, NY: John Wiley & Sons.

[5] Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific.

[6] Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory 42*, 40-47.

[7] Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory 47*, 1712-1717.

[8] Schervish, M. J. (1995). *Theory of Statistics*. New York: Springer.

[9] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464.

[10] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of Royal Statistical Society, 36*, 111-147.