

An Introduction to Model Selection: Tools and Algorithms

Sébastien Hélie

Université du Québec À Montréal

Model selection is a complicated matter in science, and psychology is no exception. In particular, the high variance in the object of study (i.e., humans) prevents the use of Popper's falsification principle (which is the norm in other sciences). Therefore, the desirability of quantitative psychological models must be assessed by measuring the capacity of the model to fit empirical data. In the present paper, an error measure (likelihood), as well as five methods to compare model fits (the likelihood ratio test, Akaike's information criterion, the Bayesian information criterion, bootstrapping and cross-validation), are presented. The use of each method is illustrated by an example, and the advantages and weaknesses of each method are also discussed.

The main goal of scientific investigation is to explain and predict empirical phenomena. The former is usually handled by formulating *theories* which explain the observations using one or several abstract concepts that are causally related to the experience (Humes, 1888). However, explaining observed phenomena is not sufficient: the tentative explanation, expressed as a particular theory, must also account for new observations (prediction). This is often achieved by operationalizing the theory into a *model*. A model is a specification of a theory¹ which makes the prediction of new phenomena possible. As a result, the predictions of a given model can be empirically tested in order to assess its desirability. In particular, Karl Popper argued in favour of a simple way to determine the scientific

validity, and thus the desirability, of a theory: the *falsification principle* (Popper, 1959). This principle states that, in order to be deemed scientific, there must exist an empirical way of showing the falsity of the theory.

The main consequence of the falsification principle is that a scientific theory must be translatable into a model which generates falsifiable predictions. Also, a model which makes finer-grained predictions is easier to falsify and thus more desirable. This second consequence is responsible for the gradual shift from qualitative models (e.g., the Atkinson-Shiffrin memory model; Atkinson & Shiffrin, 1968) to quantitative models (e.g., the Context Model; Medin & Schaffer, 1978). The former type of models allows predictions such as "task A is harder than task B" while the latter permits more precise predictions such as "task A is two times more difficult than task B".

In psychology, the shift from qualitative to quantitative models has been particularly difficult (Cousineau, 2005). Several reasons might explain this difficulty but chief among them is the complexity of human behaviour: Many variables influence the performance of human participants, some of them being external (e.g., group membership, the weather, etc.; Feldman, 1998; Rosenthal, 1993), while others are internal (e.g., the limits of cognitive processes, background knowledge, etc.; Simon, 1972; Tanaka & Taylor, 1991). While this list is far from exhaustive, it does make a clear point about the impossibility for experimenters to control all variables affecting human behaviour. Therefore,

This research was supported by scholarships from Le fonds québécois de la recherche sur la nature et les technologies and the Natural Sciences and Engineering Research Council of Canada. The author would like to thank Gyslain Giguère, Catherine Mello, and an anonymous reviewer for their useful comments on an earlier draft. Requests for reprints should be addressed to Sébastien Hélie, Laboratoire d'Étude en Intelligences Naturelle et Artificielle, Département de psychologie, Université du Québec À Montréal, C. P. 8888, Succ. Centre-ville, Montréal (Québec) H3C 3P8, CANADA, or using e-mail at Helie.Sebastien@courrier.uqam.ca.

¹ Several models can be used to operationalize a single theory.

contradicting results are common in psychology and clear falsification of a model is almost impossible. The desirability of quantitative models is thus often assessed by some error measure of the model's prediction on the observed data. However, in most cases, the asymptotic distribution of the error measure is unknown, which prevents the use of Fisherian statistics to distinguish between models that fit the data equally well and models which truly differ as to their error measure.

Many techniques have been developed to cope with the difficulty of distinguishing between true statistical difference and noise in the absence of Fisherian statistics. This paper surveys some of these methods from the fields of mathematical statistics (Larsen & Marx, 2001), Bayesian statistics (Jeffreys, 1961), information theory (Goldman, 1953), and machine learning (Bishop, 1995; Hastie, Tibshirani & Friedman, 2001). The remaining of this tutorial is organized as follows. First, the concept of *likelihood* is introduced. This measure is used throughout the text to compare models. Second, the likelihood ratio test is presented (Chernoff, 1954; Wilks, 1938). This is the only statistical test that leads to a clear conclusion concerning the difference between the likelihood of two models. However, its use is limited by the strong assumption that the models are nested. Next, two methods, the Akaike Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Bishop, 1995; Hastie et al., 2001), are presented. The aim of these methods is to penalize the error of the model on the training data (internal error) to estimate the error on unseen cases (generalization error). Finally, simulation methods (bootstrapping and cross-validation), which directly estimate the generalization error without using the internal error, are presented.

An important measure: The likelihood

The likelihood of a model is defined as the joint probability that each datum was generated by the model. More precisely, if the data are independent:

$$l[P(X|\theta)] = \prod_{i=1}^n P(x_i|\theta) \quad (1)$$

where $l[\bullet]$ is the likelihood measure, X is the data set, n is the number of elements in X , x_i is a particular datum and θ is the list of the model's parameters. Because, this measure is a probability, it is bounded between $[0, 1]$, one representing the certainty that a particular model generated the data set and zero representing the impossibility that the model generated the data. While this measure is easily interpretable, it is seldom used because the multiplication of a list of numbers between zero and one can never increase. Hence, the likelihood is always a very small number which results in underflow in modern computers (the smallest representable number is usually of the order of 1×10^{-17}).

A good solution to the underflow problem is to use the logarithmic transformation. Because the Log is a monotonic increasing function, this transformation does not affect the order of the measures (e.g., if $a > b$, $\text{Log}[a] > \text{Log}[b]$). Also, in the logarithmic scale, the products become sums. Hence,

$$\begin{aligned} \text{Log}(l[P(X|\theta)]) &= \text{Log}\left[\prod_{i=1}^n P(x_i|\theta)\right] \\ &= \sum_{i=1}^n \text{Log}[P(x_i|\theta)] \\ &= ll[P(X|\theta)] \end{aligned} \quad (2)$$

where $ll[\bullet]$ is the log-likelihood. This new function is bounded between $]-\infty, 0]$, minus infinity representing the incapacity of a model to generate the data set and zero representing the certitude that a model generated the data. However, this function is always negative, which is counterintuitive for an error function. Therefore, the negative of the log-likelihood is usually used ($-ll[\bullet]$). This function varies between $[0, \infty[$, infinity representing the certitude that a model did not generate the data set (maximum error) and zero representing the absolute certainty that a model generated the data (minimum error).

Example

One of the most important applications of the likelihood measure is estimating the best-fitting parameters of a given model. For example, it is known that the distribution of Intelligence Quotients (IQs) is Gaussian (because of the central limit theorem). Still, the free parameters of the Gaussian distribution must be estimated in order to generate precise predictions. This ubiquitous model is described by Equation 3.

$$P(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (3)$$

where $\theta = \{\mu, \sigma\}$ are the free parameters to be estimated. These parameters could be estimated by numerically minimizing $-ll[P(X|\mu, \sigma)]$ (Box, Davies & Swann, 1969; Dion & Gaudet, 1996). Here, the negative log-likelihood is defined by:

$$\begin{aligned} -ll[P(X|\mu, \sigma)] &= -\sum_{i=1}^n \text{Log}\left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right] \\ &= n\text{Log}[\sqrt{2\pi}\sigma] + \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \end{aligned} \quad (4)$$

To accelerate computation time, only the second term of Equation 4 needs to be minimized, because the first term is not affected by the data set. However, in the particular case of the Gaussian distribution, one does not need to minimize the function because there is a well-known exact solution. The following system of equations can be solved for μ and σ :

$$\begin{aligned}
\frac{\partial}{\partial \mu} \left[n \text{Log} \left[\sqrt{2\pi} \sigma \right] + \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right] &= \\
\frac{\partial}{\partial \sigma} \left[n \text{Log} \left[\sqrt{2\pi} \sigma \right] + \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right] &= 0 \\
\Rightarrow \mu &= \frac{1}{n} \sum_{i=1}^n x_i \\
\Rightarrow \sigma &= \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2} \quad (5)
\end{aligned}$$

Thus, the Maximum Likelihood Estimator (MLE) of the parameter μ is the sample mean and the MLE of the parameter σ is the biased sample standard deviation.

To sum up, $-ll[\bullet]$ can be directly interpreted as the internal prediction error of the model. In particular, this measure is equivalent to the well known mean squared error used in linear regression: the same solution is obtained by minimizing any of these measures (Bishop, 1995; Ruck et al., 1990). Therefore, in addition to being useful to estimate best-fitting parameters, the likelihood can be used to quantify a model's desirability.

The likelihood ratio test

The likelihood ratio test is the only existing statistical test which allows to directly compare the adequacy of two models. However, one of its assumptions is stringent: the two compared models must be nested. In other words, this implies that a second model must be obtainable by constraining one or several of the free parameters in the first one. When this assumption is met, only the likelihood measures for both models and the number of free parameters are needed. This test is based on an important result by Chernoff (1954), which states that: if a first model (M1) has n free parameters, a second (M2) has $n + m$ free parameters and $M1 \subset M2$, then:

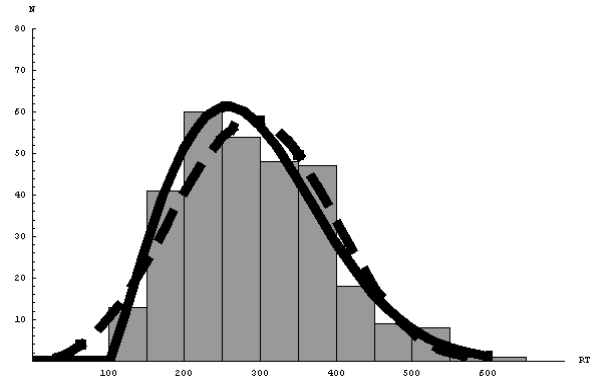
$$\begin{aligned}
-2 \text{Log} \left[\frac{l(M1)}{l(M2)} \right] &\sim \chi^2(m) \\
\Rightarrow 2 \{ -ll(M1) \} - \{ -ll(M2) \} &\sim \chi^2(m) \quad (6)
\end{aligned}$$

where $\chi^2(m)$ is the centered chi-square distribution with m degrees of freedom. In words, twice the difference between the models' respective $-ll[\bullet]$ is subject to a standard chi-square test.

Example

In cognitive psychology, response time (RT) distributions of participants in simple tasks are often used to test assumptions about the relations between the processes responsible for a given response (Cortese & Dzhaferov, 1996; Cousineau, Goodman & Shiffrin, 2002; Dzhaferov & Cortese, 1996; Hockley, 1984; Logan, 1992; Luce, 1986). In particular, one possible relation is that of a race (Cousineau, Lacroix & Hélie, 2004). In a race model, many processes are individually computing a response and the first process to

Figure 1. Data used to illustrate the application of the likelihood ratio test. The histogram shows the data, the solid line is the best-fitting Weibull distribution with three free parameters (M2) and the dashed line the best-fitting Weibull distribution with two free parameters (M1).



produce an output wins the race and is responsible for the answer. If this hypothesis turns out to be correct, the RTs should follow a Weibull distribution (Weibull, 1951). For instance, after 300 trials in a given psychological task, the RTs obtained might look like those presented by the histogram shown in Figure 1. The Weibull hypothesis allows two different models described by Equations 7 and 8:

$$P(x | \beta, \gamma) = \gamma \beta^{-\gamma} x^{\gamma-1} e^{-\left(\frac{x}{\beta}\right)^\gamma} \quad (7)$$

$$P(x | \alpha, \beta, \gamma) = \gamma \beta^{-\gamma} (x - \alpha)^{\gamma-1} e^{-\left(\frac{x-\alpha}{\beta}\right)^\gamma} \quad (8)$$

where α is the position parameter and represents the minimum RT, β is the spread parameter (proportional to variance) and γ is the shape parameter (proportional to skewness).

In the present example, Equation 7 is obtained by setting α to zero in Equation 8. Thus, if M1 is defined as Equation 7 and M2 is defined as Equation 8, M1 is a nested model of M2 and the likelihood ratio test can be applied to decide if M2 is a better model of the data than M1. It is important to note that because M2 has more free parameters than M1 and M1 is nested in M2, M2's fit will always be better or equal than M1's. The likelihood ratio test aims at deciding between these two alternatives.

The best fitting models have been numerically estimated using Mathematica (also shown in Figure 1) with the following results: $-ll(M1) = 1795.03$ and $-ll(M2) = 1783.88$. The hypotheses of the likelihood ratio test are the following:

H_0 : M1 and M2 fit the data equally well.

H_1 : M2's fit is better than M1's.

Next, the probability of committing a Type I error is chosen. In the present case, .01 was chosen as the decision criterion. The test can now be applied (Equation 6): $2(1795.03 - 1783.88) = 22.3$. By looking at a chi-square table with $3 - 2 = 1$ degree of freedom, the cutoff score is found to be 6.6349.

Thus, M2 fits the data better than M1 ($\chi^2(1) = 22.3$, $p < .01$). The addition of a position parameter is therefore required to explain the present data.

Estimation of the generalization error by penalization of the internal error

The likelihood ratio test is a simple yet efficient way to compare nested models. The problem is that nested models are very rarely compared. Therefore, other methods are needed to compare models which have nothing in common except the data set on which they are to be tested. The following methods are based on another kind of assumptions: when a model is chosen, the goal is not only to explain *observed* data but also to generalize and explain *future* data. However, the modeller only has access to the internal error of the model: this error is biased and underestimates the generalization error because the same data have been used to find the best-fitting parameters and test the model's performance. Thus, the internal error always underestimates the generalization error (Hastie et al., 2001). In particular,

$$Err_{gen} = Err_{int} + Optimism \quad (9)$$

where Err_{int} is the internal error, Err_{gen} is the generalization error and $Optimism$ is a term representing the expected difference between the true error of the model (Err_{gen}) and the observed error (Err_{int}). Following Hastie et al. (2001), it can generally be shown that:

$$Optimism \propto \text{Covariance}(y, \hat{y}) \quad (10)$$

where \hat{y} is the model's estimates and y is the data set. In words, the magnitude of the underestimation is proportional to the covariance between the model's estimates and the observations. That is, the better the fit, the higher the optimism (underestimation).

An obvious way of estimating the true error of a model is to estimate the optimism and add this estimation to the internal error. While the covariance between the training set and the model's estimations can be directly computed, this measure is dependent on the variance of each of these individual variables. The variance in the data set is constant for all candidate models and can be ignored. However, the variance resulting from a model should not handicap its adequacy. Therefore, a different estimation of the *Optimism*, which is invariant to a model's variance, is needed. This is exactly what the AIC (Akaike, 1973) and the BIC (Bishop, 1995; Hastie et al., 2001) are.

Akaike's Information Criterion

The AIC (Akaike, 1973) estimates the covariance between the model's predictions and the observations by simply using a measure of the model's complexity. As with the

likelihood ratio test, only the error measure ($-ll[\bullet]$) and the number of free parameters in the model are needed. More precisely,

$$AIC = 2[-ll(\text{data})] + 2d \quad (11)$$

where d is the number of free parameters. One simply chooses the model with the smallest AIC. The advantage of choosing this measure instead of the smallest $-ll[\bullet]$ is that the latter method always select the more elaborate model while the former penalizes the models for their complexity.

The Bayesian Information Criterion

The BIC (Bishop, 1995; Hastie et al., 2001) is similar to the AIC (Akaike, 1973) except that it is motivated by the Bayesian model selection principle (Hastie et al., 2001). In addition to the information required to compute the AIC (number of free parameters and $-ll[\bullet]$), the computation of the BIC requires the number of observations. The computation and origins of this criterion are now described.

First, the fundamental theorem in Bayesian statistics is that of reverse probabilities:

$$P(M | Z) = \frac{P(Z | M)P(M)}{P(Z)} \quad (12)$$

where M is a model and Z is the data set. This theorem states that the probability of a model, given a specific data set, is equal to the probability of the data set given the model (likelihood) times the probability of the model. The denominator is just a normalizing term which ensures that the result is between zero and one. Therefore, without loss of generality, this term can be dropped and Equation 12 becomes:

$$P(M | Z) \propto P(Z | M)P(M) \quad (13)$$

When two candidate models are compared, it is common to use the *odd*:

$$\frac{P(M1 | Z)}{P(M2 | Z)} = \frac{P(M1) P(Z | M1)}{P(M2) P(Z | M2)} \quad (14)$$

where M1 and M2 are the candidate models. If the odd is greater than one, M1 is more probable than M2 and if it is smaller than one, M2 is more probable than M1. If no prior knowledge is available and there is no reason to favour a model over the other, the first term equals one and Equation 14 simplifies as:

$$\frac{P(M1 | Z)}{P(M2 | Z)} = \frac{P(Z | M1)}{P(Z | M2)} \quad (15)$$

Equation 15 is also called the Bayesian Factor and is fundamental to the Bayesian approach to statistics (Jeffreys, 1961; Kass & Raftery, 1995). Each $P(Z | Mi)$ is an integral which needs to be computed. By a Laplace approximation, and after some simplification, this integral is equal to (Hastie et al., 2001):

$$\begin{aligned}
\text{Log}[P(Z|M)] &= \text{Log}[P(Z|M, \theta)] - \frac{d}{2} \text{Log}[n] \\
&= 2[-ll(Z)] + \text{Log}[n]d \\
&= \text{BIC}
\end{aligned} \tag{16}$$

where n is the number of observations. Equation 16 is also equivalent to twice Schwartz's criterion (Schwartz, 1979).

The first thing to be pointed out is that the BIC (Equation 16) is similar to the AIC (Equation 11), except that the factor multiplying the number of free parameters is $\text{Log}[n]$ instead of two. Therefore, when the number of observations is greater than e^2 (≈ 7.4), the BIC's estimation of the optimism term is higher. More generally, Equations 12-16 shows that choosing the model with the smallest BIC is equivalent to choosing the model with the highest posterior probability. As a result, the BIC can be used not only to choose the best model (smallest BIC) but also to assess the relative merit of each of the tested models:

$$P(Z|M_j) = \frac{e^{-2\text{BIC}_{M_j}}}{\sum_{i=1}^l e^{-2\text{BIC}_{M_i}}} \tag{17}$$

where M_j is a given model and l is the number of models considered. Also, the BIC is asymptotically consistent; that is, if a set of models is tested (including the true model), as n tends towards infinity, the probability that the true model has the smallest BIC tends towards one. In contrast, when n tends towards infinity, the AIC (Akaike, 1973) chooses models that are too complex. However, for a finite n , the BIC chooses models that are too simple. The use of both these methods is next illustrated by an example.

Example

Nowadays, many models of categorization allow quantitative predictions. One of the most popular models in categorization is the Generalized Context Model (GCM; Nosofsky, 1986), which is an implementation of the exemplar theory of categorization (Medin & Schaffer, 1978). Nosofsky (1987) has applied the GCM to five experimental conditions involving integral stimuli (Garner, 1970; Hélie et al., 2002). Ten years later, Nosofsky's data were fit with another categorization model, the Exemplar-Based Random Walk (EBRW; Nosofsky & Palmeri, 1997). In this second paper, the performance of the EBRW was compared to the performance of the GCM using the $-ll[\bullet]$ measure. The obtained results are shown in Table 1. The $-ll[\bullet]$ columns show the results reported by Nosofsky and his colleague. As seen, the EBRW is a better explanation of the data in the first three conditions, as well as in the Pink – Brown condition. However, the EBRW has more free parameters than the GCM. Because the likelihood, the number of free parameters and the number of observations is provided, the calculation of the AIC (Akaike, 1973) and the BIC (Bishop, 1995; Hastie

et al., 2001) is straightforward. As seen, the penalization of the internal error by the AIC does not change the order of the models in most conditions. However, according to the AIC, the GCM is a better explanation of the Pink – Brown condition: the advantage of the EBRW in this condition was a result of its additional free parameter.

Table 1 also shows the BIC in each condition. Because there are twelve data points and $\text{Log}[12] \approx 2.485 \approx 2$, the BIC gives the same results as the AIC. If more data points had been involved, the difference between the AIC and the BIC would have been more important.

Estimation of the generalization error using simulations

Another method used to assess a model's desirability is to directly estimate its generalization error using *Monte Carlo* simulations (Metropolis et al., 1953). Here, two such methods will be detailed: bootstrapping and cross-validation (Bishop, 1995; Hastie et al., 2001). Unlike the previously presented methods, bootstrapping and cross-validation possess no constraints or assumptions whatsoever: all that is needed is a data set. Any error function can be used, but $-ll[\bullet]$ is used in the following example for simplicity.

Bootstrapping

The bootstrap method is used with small data sets (Hastie et al, 2001). The following algorithm is used for the simulations:

Input: A data set of size n and a model

Output: A list of m error measures

- 1) Pose the data set as the population;
- 2) Repeat m times:
 - 1) Define a sample by randomly performing n draws with replacement from the population;
 - 2) Numerically estimate the best-fitting parameters of the model using the sample;
 - 3) Compute and keep the model's error on the population.

Once the simulations are complete, one can use any of the available statistical tools (Hays, 1973) to analyze the list of error measures.

Bootstrapping has two major inconvenients. First, like all simulation methods, it only considers the error measures of the models. Therefore, models with more free parameters have a head start on simpler models. Second, because the data used to form and test the models are the same, the postulate of independence is violated. Therefore, the generalization error is underestimated. Hastie and his colleagues (2001) provide a simple example to illustrate this point. First, when drawing a particular sample, the probability that a particular datum is drawn at least once is

binomial and described by:

$$P(\text{datum } i \in \text{sample } Z) = 1 - \left(1 - \frac{1}{n}\right)^n \quad (18)$$

where Z is a particular sample, i is a particular data point and n is the number of data points. If n is large, this binomial distribution can be approximated by the Poisson (Ross, 1998):

$$P(\text{datum } i \in \text{sample } Z) \approx 1 - e^{-1} \approx 0.632 \quad (19)$$

In words, the probability that datum i is part of sample Z is 0.632. Accordingly, the probability that sample Z does not contain datum i is: $1 - 0.632 = 0.368$. If the data set is composed of two equiprobable categories (“A” and “B”) and no other information is available, the best thing a model can do is constantly answer “A”². Clearly, the generalization error of this model is 0.5. However, if datum i is part of category “A”, the model is correct each time i is drawn. Moreover, when datum i is not drawn, the model is incorrect only half the time (remember that $P(A) = P(B) = 0.5$). As a result, the estimated generalization error by bootstrapping is $0.5 \times 0.368 = 0.184$, which clearly underestimates the true generalization error (0.5). This limitation of bootstrapping is solved by the next simulation technique.

Cross-validation

The difference between cross-validation and bootstrapping is that the former splits the data set into several independent subsamples, one of which is used to test the model while the remaining are used to estimate the

best-fitting parameters (Bishop, 1995; Hastie et al., 2001). As a result, the training set and the test data are independent (unlike in bootstrapping). However, because the data set must be split, more data points are necessary to use this method. Cross-validation is done by performing the following algorithm:

Input: A data set of size n and a model

Output: A list of k error measures

- 1) Split the data set into k subsamples of equal size;
- 2) Repeat k times, each time using a different test set:
 - 1) Numerically estimate the best-fitting parameters of the model using $k - 1$ subsamples;
 - 2) Compute and keep the model’s error on the unused subsample (test set);

As for bootstrapping, any statistical tool can be used on the list of error measures. However, the choice of k is critical. Before discussing the consequence of this choice, another decomposition of the error term must be introduced (Bishop, 1995; Geman, Bienenstock & Doursat, 1992; Hastie et al., 2001):

$$Error = Bias + Variance \quad (20)$$

The first term, the bias, typically decreases with increasing model complexity and is not affected by the data. The second term, variance, increases with model complexity and is affected by the data. While this short explanation is sufficient to understand the following discussion, the interested reader is referred to Geman et al. (1992) for a particularly clear derivation and discussion of this decomposition.

Table 1. Fit of the GCM and the EBRW to data collected in Nosofsky (1987)

Conditions ($n = 12$)	GCM (2)			EBRW (3)		
	$-ll[\bullet]$	AIC	BIC	$-ll[\bullet]$	AIC	BIC
Saturation A	46	96	97	41.1	88.2	89.7
Saturation B	58.8	121.6	122.6	43.4	92.8	94.3
Brightness	60.3	124.6	125.6	45.1	96.2	97.7
Criss-Cross	50.4	104.8	105.8	52.9	111.8	113.3
Pink-Brown	70.9	145.8	146.8	70.6	147.2	148.7
Diagonal	99.6	203.2	204.2	106.7	219.4	220.9

Note. Grey cells are the chosen model according to a particular measure. Numbers in parenthesis are the number of free parameters.

² The exact same argument applies if “B” is constantly chosen.

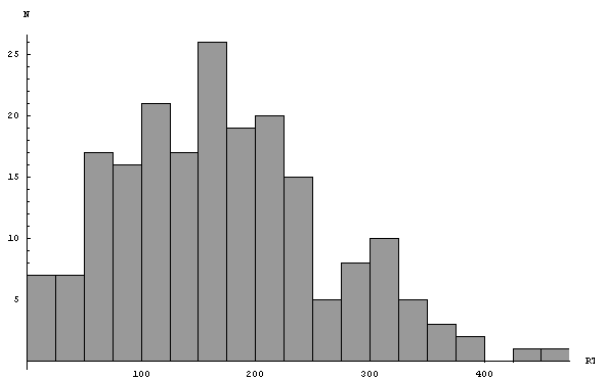
The choice of k in cross-validation affects each term of the preceding decomposition (Equation 20) differently (Hastie et al., 2001). Choosing a large k results in small bias because, in each iteration, the estimate of the free parameters is stable (each training set is large). However, large k s also results in high variance, because the k training sets are strongly overlapping. On the other hand, choosing a small k increases the model's bias but diminishes the variance. Typical choices for k are five, ten, or n . The last case is called *leave-one-out* and produces (almost) unbiased estimates of the model's error (accompanied by extremely high variance). A useful rule of thumb is to look at the effect of sample size (n) on the error estimate. In all cases, a large n improves the error estimation but, if the slope of the effect of n is large (a small change in n strongly improves the estimate), a small k can be chosen (five or ten). Otherwise, a large k must be chosen and the computation time increases.

Example

The following example is an extension of the one presented in the likelihood ratio test section. A data set of 200 RTs was collected and distribution models are used to infer the relationship between the processes implied in the generation of the responses. The collected data are shown in Figure 2. As in the preceding example, the race hypothesis is considered, so Equation 7 is the first model. Three other models are also introduced.

Another type of processing often considered in cognitive psychology is the random walk (RW; Nosofsky & Palmeri, 1997; Ratcliff, 1978). In RWs, evidence in favour of one of two alternatives is accumulated over time until a criterion is reached. When the criterion is reached, the associated response is given. Therefore, RWs are akin to race models (Cousineau et al., 2003) except that in the former, and not the latter, evidence in favour of one alternative is necessarily against the other. If the processes involved in a participant's

Figure 2. Data used to illustrate the application of the Monte Carlo simulation methods



response are RWs, the resulting RTs follow a Wald distribution (Burbeck & Luce, 1982; Wald, 1947), which is described by:

$$P(x | \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} e^{-\frac{\lambda(x-\mu)^2}{2x\mu^2}} \quad (21)$$

where λ represents the spread and μ is related to the distribution's mean.

For many years, psychologists have postulated that cognitive processes were conducted by a series of modules and that the resulting RTs were a function of each individual module's processing time (e.g., Atkinson & Shiffrin, 1968). If processes are a series of modules, two of the possible ways they can interact are hermetically and permeably. In the former, each module completes its computation before passing the result to the next. As a result, the final RT is simply the sum of each individual module's RT. In particular, if an exponentially distributed module is additively affected by Gaussian noise, the resulting RT distribution is the Ex-Gaussian (Heathcote, Popiel & Mewhort, 1991; Hockley, 1984; Luce, 1986):

$$P(x | \mu, \sigma, \tau) = \frac{1}{2\tau} e^{\frac{\sigma^2}{2\tau^2} + \frac{\mu-x}{\tau}} \left[1 + \text{Erf} \left(\frac{1}{\sqrt{2}} \frac{x-\mu}{\sigma} - \frac{\sigma}{\tau} \right) \right] \quad (22)$$

where μ and σ are the free parameters of the Gaussian noise, τ is the free parameter of the exponential process and $\text{Erf}(\bullet)$ is the error function encountered in integrating the Gaussian distribution.³

In the case of permeability, the modules are leaky and a given module can start its computation before the preceding module has finished. With leaky processes, the model is multiplicative and the resulting RT distribution is the Lognormal (Ulrich & Miller, 1993).

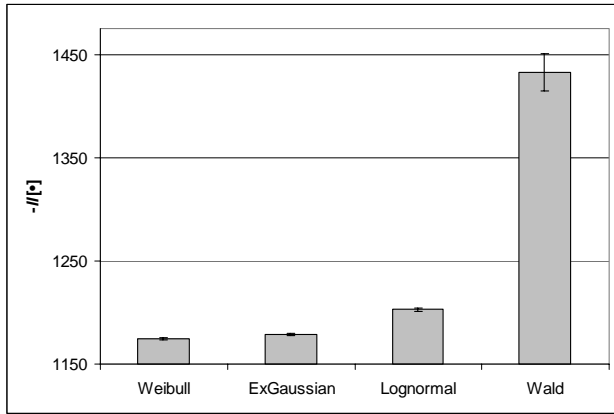
$$P(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi x\sigma}} e^{-\frac{(\text{Log}(x)-\mu)^2}{2\sigma^2}} \quad (23)$$

where μ and σ are related to the mean and spread respectively.

These four models of RTs (Equation 7, Equations 21 - 23) were used to explain the data in Figure 2 using bootstrapping and cross-validation (Bishop, 1995; Hastie et al., 2001). The results of bootstrapping for $m = 50$ is shown in Figure 3. As seen, the Weibull distribution constitutes the best model for the current data set with an error measure of 1174 ± 1.12 . This result is good news, because the data were artificially generated by a Weibull random generator. The Ex-Gaussian distribution, which uses one more free parameter than the other models, comes as a close second

³ $\text{Erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$

Figure 3. Estimation of the model's error using bootstrap ($m = 50$)



with an error measure of 1179 ± 1.12 . However, the error bars do not cross, which suggests that the Weibull distribution is significantly better than the Ex-Gaussian. The other two models' predictions clearly differ from the data set.

Figure 4 shows estimations of the error for each model using cross-validation. For $k = 5$, the best models are the Weibull and the Ex-Gaussian with estimated errors of 236 ± 1.77 and 237 ± 1.99 respectively. However, in this case, even though the two models are better than the remaining, they are not statistically distinguishable. Therefore, the modeller must keep in mind that the Ex-Gaussian model is more complex, and should parsimoniously prefer the Weibull distribution in the present case. The same pattern of results is found when $k = 10$ and $k = n = 200$. The Weibull and Ex-Gaussian models are better than the remaining and are statistically indistinguishable with estimated errors of 118 ± 0.80 and 118 ± 0.91 respectively when $k = 10$ and estimated errors of 5.89 ± 0.05 and 5.91 ± 0.05 when $k = n = 200$.

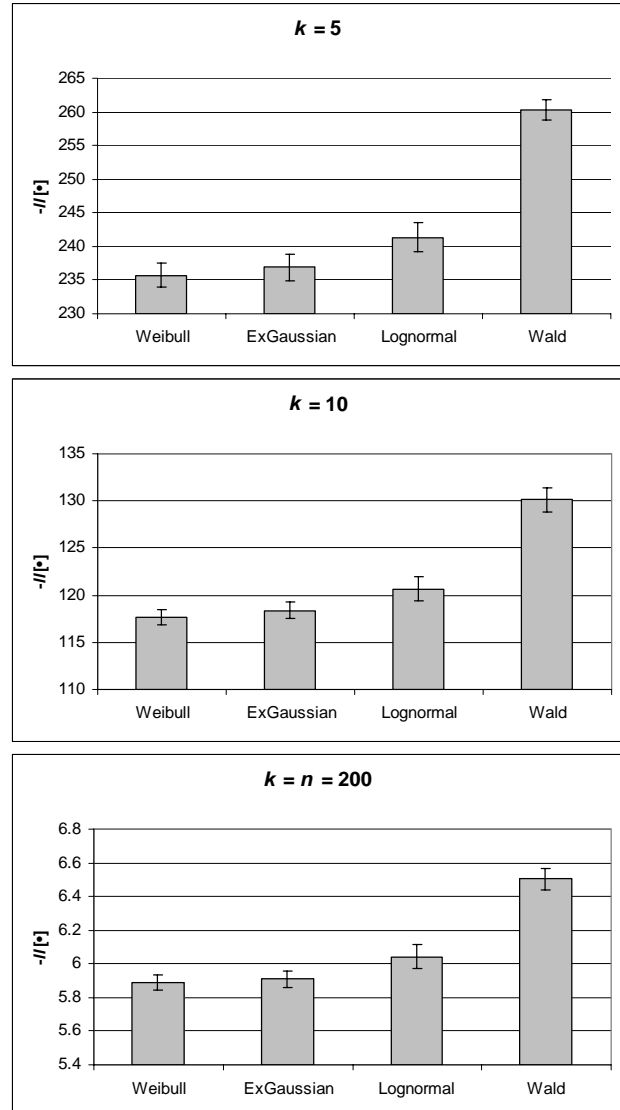
Conclusion

In the first section of this paper, the likelihood measure was presented as an interesting alternative to estimate a model's error on the training data. Using this error function, the best-fitting parameters can be estimated and several models can be compared to assess their relative desirability. Following the presentation of the error measure, several methods used to compare the models' adequacy were presented. First, in the ideal case where the models are nested, the likelihood ratio test can be applied (Chernoff, 1954; Wilks, 1938). This is the only statistical test (in the Fisherian sense) that allows the comparison of likelihoods. In the remaining cases, two different approaches were used to estimate the models' generalization error: penalization of the internal error and Monte Carlo simulations. In the former, two methods were presented, the AIC (Akaike, 1973) and the BIC (Bishop, 1995; Hastie et al., 2001). The AIC is based on information theory (Goldman, 1953) while the

BIC is based on Bayesian statistics (Jeffreys, 1961; Kass & Raftery, 1995). Both methods are justified by the fact that the internal error of a model always underestimates the generalization error (because the same data are used to estimate the parameters' values and compute the models' error).

The second approach, Monte Carlo simulations, does not use the internal error measure to estimate the generalization error: this error is directly estimated by splitting the data in training and test sets. A different separation is made during each iteration and several estimations of the generalization error are obtained. Following these estimations, standard statistics can be computed. The difference between the two presented methods, bootstrapping and cross-validation (Bishop, 1995; Hastie et al., 2001), is that the former uses the

Figure 4. Estimation of the model's error using cross-validation for $k = 5, 10$, and n .



same data to train and test the model while the latter uses independent data sets.

Overall, the bootstrap method always underestimates the generalization error and the use of cross-validation requires a lot of data points and much computation time⁴. On the other hand, the AIC and the BIC are biased estimators because they only consider the model's complexity in the penalization of the internal error: the data set is ignored (Hastie et al., 2001). The Monte Carlo simulation methods are less biased but they result in high error variance (Equation 20), because their estimation only relies on the data set (the model is ignored).

To conclude, it seems that there is no absolute best way to compare non-nested models. Particular attention should be allotted to experimental control in order to reduce the variance in empirical data. Therefore, the falsification principle could be re-established in psychology. Also, when creating or evaluating a model, it must be kept in mind that any data set can be fit with arbitrary accuracy if model complexity is not an issue. As a result, simple models in which each axiom refers to a psychological process should be preferred over models with smaller estimation errors on the data. After all, the interpretations given to a model's axioms are what distinguish a good model from a mere mathematical equation.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov & F. Csáki (Eds.) *2nd International Symposium on Information Theory* (pp. 267-281). Tsahkadsov, Armenia, USSR.
- Atkinson, R.C. & Shiffrin, R.M. (1968). Human memory: A proposed system and its control processes. In K.W. Spence & J.T. Spence (Eds.) *The Psychology of Learning and Motivation: Advances in Research and Theory. Volume 2* (pp. 89-195). New York: Academic Press.
- Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.
- Box, M.J., Davies, D., & Swann, W.H. (1969). *Non-linear Optimization Techniques*. Edinburgh: Oliver & Boyd.
- Burbeck, S.L. & Luce, R.D. (1982). Evidence from auditory simple reaction times for both change and level detectors. *Perception and Psychophysics*, 32, 117-133.
- Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, 25, 573-578.
- Cortese, J.M. & Dzhafarov, E.N. (1996). Empirical recovery of response time decomposition rules II. Discriminability of serial and parallel architectures. *Journal of Mathematical Psychology*, 40, 203-218.
- Cousineau, D. (2005). The rise of quantitative methods in psychology. *Tutorials in Quantitative Methods for Psychology*, 1, 1-3.
- Cousineau, D., Goodman, V.W., & Shiffrin, R.M. (2002). Extending statistics of extremes to distributions varying in position and scale and the implications for race models. *Journal of Mathematical Psychology*, 46, 431-454.
- Cousineau, D., Lacroix, G.L., & Hélie, S. (2003). Redefining the rules: Providing race models with a connectionist learning rule. *Connection Science*, 15, 27-43.
- Dion, J.-G. & Gaudet, R. (1996). *Méthodes d'Analyse Numérique: De la Théorie à l'Application*. Mont-Royal, QC: Modulo.
- Dzhafarov, E.N. & Cortese, J.M. (1996). Empirical recovery of response time decomposition rules I. Sample-level decomposition tests. *Journal of Mathematical Psychology*, 40, 185-202.
- Feldman, R.S. (1998). *Social Psychology. 2nd Edition*. Upper Saddle River, NJ: Prentice Hall.
- Garner, W.R. (1970). The stimulus in information processing. *American Psychologist*, 25, 350-358.
- Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4, 1-58.
- Goldman, S. (1953). *Information Theory*. New York: Dover.
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.
- Hays, W.L. (1973). *Statistics for the Social Sciences. 2nd Edition*. New York: Holt, Rinehart and Winston Inc.
- Heathcote, S., Popiel, S.J., & Mewhort, D.J.K. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin*, 109, 340-347.
- Hélie, S., Cousineau, D., Charbonneau, D., & Lefebvre, C. (November 2002). Stimulus processing and task dependency. *43th Annual Meeting of the Psychonomic Society*. Kansas City, KS.
- Hockley, W.E. (1984). Analysis of response time distributions in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10, 598-615.
- Hume, D. (1888/1967). *A Treatise of Human Nature*. Oxford: Clarendon Press.
- Jeffreys, H. (1961). *Theory of Probability. 3rd Edition*. Glasgow: Oxford University Press.
- Kass, R.E. & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.
- Larsen, R.S. & Marx, M.L. (2001). *An Introduction to Mathematical Statistics and Its Applications. 3rd Edition*. Upper Saddle River, NJ: Prentice Hall.
- Logan, G.D. (1992). Shapes of reaction-time distributions

⁴ The simulation results presented in the bottom panel of Figure 4 took about ten hours of computing time using Mathematica 5 on a Athlon 1600+ computer.

- and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 883-914.
- Luce, R.D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*. New York: Oxford University Press.
- Medin, D.L. & Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review*, 85, 207-238.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., & Teller, A.H. (1953). Equation of state calculations with fast computing machines. *The Journal of Chemical Physics*, 21, 1087-1092.
- Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 39-57.
- Nosofsky, R.M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 87-108.
- Nosofsky, R.M. & Palmeri, T.J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, 104, 266-300.
- Popper, K. (1959/2004). *The Logic of Scientific Discovery*. New York: Routledge.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85, 59-108.
- Rosenthal, R.N. (1993). *Winter Blues*. New York: Guilford Press.
- Ross, S.M. (1998). *A First Course in Probability. Fifth Edition*. Upper Saddle River, NJ: Prentice-Hall.
- Ruck, D.W., Rogers, S.K., Kabrisky, M., Oxley, M.E., & Suter, B.W. (1990). The multilayer Perceptron as an approximation to a Bayes optimal discriminate function. *IEEE Transactions on Neural Networks*, 1, 296-298.
- Schwartz, G. (1979). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Simon, H.A. (1972). Theories of bounded rationality. In C.B. McGuire & R. Radner (Eds.) *Decision and Organization: A in Honor of Jacob Marschak* (pp. 161-176). Minneapolis, MN: University of Minnesota Press.
- Tanaka, J.W. & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23, 457-482.
- Ulrich, R. & Miller, J. (1993). Information processing models generating lognormally distributed reaction times. *Journal of Mathematical Psychology*, 37, 513-525.
- Wald, A. (1947). *Sequential Analysis*. New York: John Wiley and Sons.
- Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, 18, 292-297.
- Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9, 60-62.

Received August 22, 2005

Accepted September 7, 2005.