Predict choice: A comparison of 21 mathematical models

Eric Schulz (eric.schulz@cs.ucl.ac.uk)
Department of Computer Science, University College London, London WC1E 6BT

Maarten Speekenbrink (m.speekenbrink@ucl.ac.uk) and David R. Shanks (d.shanks@ucl.ac.uk)
Cognitive, Perceptual and Brain Sciences, University College London, London WC1H 0AP

Abstract

How should we choose a model that predicts human choices? Two important factors in this choice are a model's predictive power and a model's flexibility. In this paper, we compare these aspects of models in a large set of models applied to an experiment in which participants chose between brands of fictitious chocolate bars and a quasi-experiment predicting movies' gross revenue. We show that there is a trade-off between flexibility and predictive power, but that this trade-off appears to lie more towards the "flexible" side than what was previously thought.
Keywords: Choices; Forecasting; Overfitting.

Introduction

Choosing a good model to predict choices is an important task for both researchers of decision making and statisticians. One crucial debate within this area concerns the flexibility of the model used to predict human choices. On the one hand, there is the belief that more flexible models should be preferred as they potentially capture the underlying psychological phenomenon well. Proponents of this approach try to show how more flexible models can outperform simpler models in many different predictive tasks (Chater et al., 2003). On the other hand, there are researchers who argue that models which are too flexible tend to overfit the data, capturing unimportant noise in the training set which results in suboptimal generalization to the test set. One example for overfitting is that by increasing the degrees of a polynomial regression to capture the average temperature of one year, one will reach a point where the models' predictive performance for the next year goes down (Gigerenzer & Brighton, 2009). Both sides at least implicitly assume that if a Model A makes more correct predictions than a Model B, Model A somehow captures the underlying process better than Model B, an assumption that can be argued against from various points (Salmon, 1971). However, given an equal amount of evidence for both Model A and Model B, it is common practice to accept as better the model that makes more accurate experimental predictions. In the past, researchers such as Gigerenzer & Brighton (2009) showed that in tasks such as predicting city sizes or professors' salary, simple models can outperform more sophisticated models such as Multiple Regression or Naive Bayesian Classifiers. Later, Chater et al. (2003) showed that other models such as Decision Trees or Feedforward Networks can perform at least as well as simple models such as Take The Best. However, both studies only used a limited amount of models applied to a somewhat artificial data set. The paper at hand puts the aforementioned performance-flexibility trade off to a test in the potentially more interesting area of human choices. In doing so, we will introduce an empirical measurement of a model's flexibility based on its ability to recover and predict data generated by other models. Furthermore, we will use this flexibility measurement to rank and compare the 21 models' performance in predicting human choices in a two-alternative forced-choice task, and predicting movies' gross revenue and a different key variable. We conclude that, even though there clearly is a trade-off between flexibility and predictive performance, the point at which more flexibility diminishes predictive performance is at a higher level of flexibility than was previously expected.

Assessing model flexibility

To shed more light on the debate about various degrees of flexibility and performance, one needs to introduce a reliable measurement of flexibility. Different measurements have been suggested, such as Kolmogorov complexity (Chater & Vitányi, 2003) and a model's degrees of freedom. The flexibility measure proposed here is defined in terms of the average ability of a model to capture and predict observations that have been generated from a different model. Importantly, the generating model has itself been fitted to a random set of data, and the best fitting parameters are then used to generate the learning and test sets. We used randomly generated data sets for the initial model fits in order to not bias the final recovery result in any systematic direction. While we could have used real world data sets for this initial simulation stage, our main concern was to assess a model's ability to recover data generated by different models, and not their ability to recover systematic characteristics of particular data sets. The averaged overall predictive performance then is rank-transformed as an indicator for a model's relative flexibility within the set of models under consideration. The models used here and their performance and obtained flexibility ranks are presented in Table 1[1].

In more detail, we obtained the relative flexibility measure as follows. We first generated 100 values for four independent variables X_j by sampling each value independently from a Normal distribution with a mean of

[1]All models were fitted using Matlab R2011B.

$\mu = 0$ and a standard deviation of $\sigma = 2.5$, i.e.

$$X_{ij} \sim \mathcal{N}(0, 6.25), \quad i = 1, \ldots, 100, j = 1, \ldots, 4 \qquad (1)$$

We used a normal distribution for the initial simulation as the variables in the actual tasks were generated by a normal distribution as well. Of course, the choice of this initial distribution is rather artificial, but we believe that a normal distribution is more likely (as for example compared to a uniform distribution) to represent the actual distributions in the data sets later on. In addition, values of a dependent variable Y were generated as independent draws from a Bernoulli distribution with a probability of $p = .5$, i.e. $Y_i \sim \text{Bern}(p = 0.5)$, $i = 1, \ldots, 100$. A given model was then fitted to this data set of 100 observations. The fitted model was used to generate 2 new data sets of 100 observations each by drawing new values of the independent variables according to Equation 1 and then generating new values of the dependent variable as the model prediction for these values of the independent variable. One of these new data sets was treated as a learning set and the other was used as a test set. An alternative model was then fitted to the learning set and used to predict the dependent variable in the test set. This procedure was repeated 100 times for every possible model-model combination and the average number of correct predictions over these 100 replications was calculated at the end. For a set of 21 models, this means that every model produced 21 averaged values of how well it recovered and predicted all the other models (including the model itself[2]). These 21 values were then averaged across each model to get an overall measurement of a model's flexibility. This measurement was rank-transformed to obtain relative flexibility values (for the question at hand, the exact differences do not matter as much as the fact that one model is more flexible than another). The ranked values are an unbiased estimate of each model's position in the whole population of models, reflecting the probability that a randomly chosen model is less flexible.

We acknowledge that our proposed flexibility measurement can only be seen as an approximation to the de facto flexibility of a given model as it only checks for the ability to recover structure within a limited domain. In addition, it can be argued that including more models that are either more or less flexible could change the ranking completely; a concern that we tried to address by including a set of models that –in our opinion– can be seen as representative for models that are normally used within standard data mining tasks. Our set up fits into the more common set up of model mimicry in that models that are able to mimic the behavior of a large set of different models tend to have a higher flexibility score. However, it is not completely the same as a model's complexity Wagenmakers et al. (2004). As complexity is normally defined as rather being model-specific and our flexibilty measurement is a combination of the situation and the model, complexity is not the same as predictive accuracy a priori. Flexible models can be not very complex and vice versa Spiegelhalter et al. (2002). We have chosen to use random data at the first simulation stage to avoid systematic biases in the flexibility comparison. This of course means that some of the original model fits are very weak, a behavior that can be changed in future studies to see whether more systematic relations at that stage might shift the measurement into a different direction.

In general, our flexibility measurement is a first attempt to quantify flexibility in the psychological domain and the resulting ranking seems to be prima facie plausible. Additionally, our experiments show that it can be used to produce reliable and reproducible results.

Experiment 1: Choices over time

The first experiment confronted participants with two-alternative forced choices between fictitious chocolate bars that were described on 4 different scales. The descriptions of these scales were generated from a pilot study (n = 21, bars= 12), where participants had to describe real chocolate bars on 30 different scales. The evaluations then were entered into a factor analysis without rotation, forcing the total number of factors to be equal to 4. The resulting factor structure explained 82% of the variance and from each resulting dimension one scale was chosen so that all the 4 scales were slightly positively correlated with each other[3]. The final scales were described as "Design", "Calories", "Crunchiness" and "Richness of Taste". The main experiment was programmed in HTML and hosted online on the Unipark survey platform. Participants were recruited via university email lists. Within the experiment, participants were presented with pairs of fictitious chocolate bars that were described by a value on the aforementioned scales. Their task was to choose which of the two bars they would prefer. The values describing the bars were generated at random to be distributed as $\mathcal{N}(5, 6.25)$ between the range of 0 and 10. As such, participants revealed how they integrate the presented information in order to make a final choice between chocolate bars. Participants were randomly assigned to one of five inter-correlations between the dimensions, $r \in \{0, 0.2, 0.4, 0.6, 0.8\}$. Different levels of inter-correlation were used in order to make the results more generalizable across different choice environments. The experiment was spread over 6 days; participants were presented with 50 pairs of chocolate bars on the first day, and 20 pairs of chocolate bars on each of the following days. Different time-points were

[2]One of our earlier questions was about specificity. The ability to capture data well that was produced by the same model class.

[3]We wanted the factors to be slightly correlated to allow for a range of different inter-correlations in the actual experiment.

Table 1: Model description, flexibility performance and assigned rank value

Model	**Description**	**Performance**	**Rank Score**
Coin Flip	No fitting involved, assumes that each possible outcome is equally likely	50.0%	0.00
Pick cue at random	Picks one of the 4 predictors at random and predicts that the item with a higher value wins	59.2%	0.05
Tallying	Unit weight strategy, sums up all the values of the cues per item and predicts that the item with the higher sum wins	70.5%	0.10
Minimalist	Take The Best-algorithm with randomly determined cue order	71.9%	0.15
Biased Coin	Calculates the mean of the dependent variable and flips a coin with $p(y) = \mu(Y_{\text{learning sample}})$	73.4%	0.20
Response Bias	Calculates the mean of the independent variable and predicts that all values will be what the majority of the items were in the learning sample	73.8%	0.25
Context Model	Uses the weighted distance to every win and loss for predictions	74.1%	0.30
Take The Best	Standard Take The Best (Gigerenzer & Goldstein, 1999)	75.9%	0.35
Nearest Neighbour	Predicts that the same is going to happen as in the most similar case in the learning sample	76.7%	0.40
Classification Trees	Builds a classification tree to classify items	78.7%	0.45
Linear Discriminant Analysis	Draws a linear function within the dimensions to separate losses from wins	82.2%	0.50
Logistic Regression	Standard Logistic Regression	82.7%	0.55
Cascade Network	Neuronal Network with cascade backwards propagation, MSE-learning, 5 neurons	82.9%	0.60
Multi-Adaptive Regression Splines	Non-parametric regression that uses a weighted sum of linear basis functions	83.9%	0.65
Naive Bayesian Classifier	Estimates the conditional probability to win or lose given the data to classify items, assumption of no covariance	84.4%	0.70
Generalized Regression Network	Uses a Radial function to approximate the underlying data structure	85.2%	0.75
Polynomial Logistic	Polynomial regression with up to 3 degrees of freedom, determined by AIC	85.8%	0.80
Pattern Recognition Network	Network with a Tansig backward propagation and MSE training function	85.9%	0.85
Support Vector Machine	Support Vector classifier with a linear Kernel	86.8%	0.90
Random Forest	Ensemble classifier based on decision trees, prediction is the mode of the outputs of all trees	87.2%	0.95
Feed Forward Network	Unsupervised Net with 2 layers, number of knots determined by cross-validation of training set	89.2%	1.00

used in order to asses to what extent participants' preferences are consistent over a (relatively short) period of time. 15 participants ($\mu_{age} = 24.5$, $\sigma_{age} = 1.4$, 7 females) participated in the study and received a real chocolate bar as a reward.

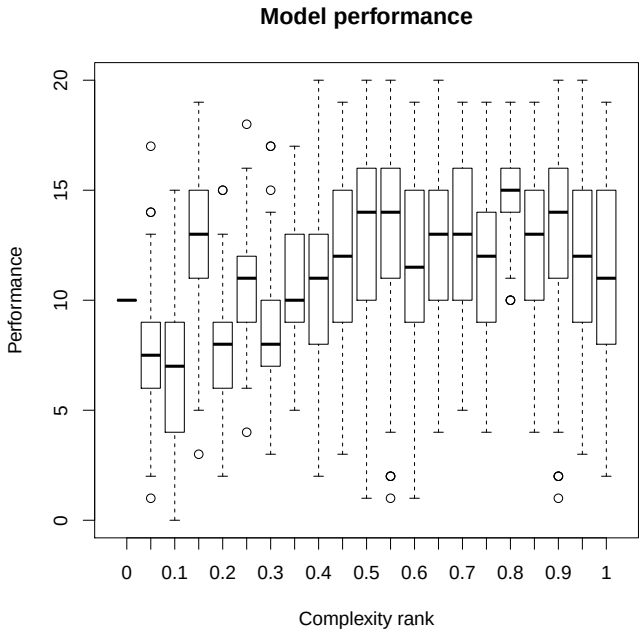
After the experiment was completed, each of the 21 models in Table 1 was fitted to the first 50 choices of each participant individually by calculating the differences between the corresponding scales[4] and treating the choices as binary. Afterwards, every model was used to predict the following 100 choices and the percentage of correct predictions for every day was calculated. Based on the argument above, we hypothesized the following:

1. More flexible models will, on average, perform better than less complex models.
2. There is a point after which an increase in flexibility will reduce predictive performance.

Results and discussion

The predictive performance of the models is shown in Figure 1. The overall correlation between flexibility and

Figure 1: Bar chart of model performance



performance was $r = 0.53$, $p < 0.01$. This significant positive correlation means that, on average, more flexible models indeed performed better than less flexible models. In order to check for a potential turning point, we analyzed the data by using a generalized polynomial regression with a logit link function and mean-centering the complexity (the resulting scale of flexibility was between -0.5 and 0.5). As can be seen in Table 2, the Cubic form was found to be the best according to Akaike's "An Information Criterion" (AIC; Akaike (1974)). This means that the final model is of the form presented in Equation 1.

Table 2: Estimates of the different polynomials within the generalized linear regression. * = best model.

Form	AIC
Linear	7392.9
Quadratic	7367.2
Cubic	7351.2*
Quartic	7352.7

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \quad (2)$$

The estimated parameters for this model are presented in Table 3, alongside their standard errors. From the cubic regression model, it is straightforward to calculate the flexibility value with maximum predictive performance[5] as follows:

Table 3: Parameters estimates of the best fitting generalized polynomial linear regression model in Experiment 1.

	β_i	$\sigma(\beta_i)$
β_0	-0.04	0.01
β_1	0.71	0.08
β_2	-0.59	0.11
β_3	-1.88	0.44

$$f(x) = -0.04 + 0.71x - 0.59x^2 - 1.88x^3$$
$$\frac{d}{dx} f(x) = 0.71 - 1.18x - 5.64x^2$$
$$\frac{d}{dx} f(x_{\max}) = 0 \rightarrow x_{\max} = 0.77$$

Thus, the maximum point is at a normalized relative flexibility level of about 0.77[6]. The model closest to this point is the generalized regression network, which is also the model that performs best overall with an average of 80% correct predictions. Interestingly, the Minimalist heuristic performed surprisingly well in the task too, which could indicate that the way participants integrated information might have changed over time. However, when we explored this possibility, we did not find an effect of time on model performance; this may be because the time period was relatively short.

Summarizing Experiment 1, more flexible models seem to perform better on average, but there is a flexibility-performance trade-off, which occurs in our experiments

[4] Resembling the same structure as in the rank generation.

[5] The point after which more flexibility starts reducing the predictive performance of the model.

[6] Checking that $\frac{d^2}{dx^2} f(x) < 0$, which is true in our case

at a normalized relative rank of 0.77. This is further towards the side of flexibility than those proposing simple heuristics may have expected. As this experiment contained a limited number of subjects, and choices were made between fictitious products, we sought to replicate these findings with a different data set.

Experiment 2: Movies' gross revenue

The second (quasi-)experiment had a similar design as the first experiment, but this time we used publicly available data from the internet on movies' gross revenue (how much money a given movie made whilst running in the cinemas). Notice that this can still be seen as a choice scenario, where a movie with a higher gross revenue was preferred by more people than a movie with a lower gross revenue. As predictors of revenue, we included the costs of the movie, the number of google hits received, its IMDB-score, as well as the number of likes on facebook (as of July 2011). All the models were fitted to 80 randomly-drawn pairs of movies from the IMDB Top 100 of the year 2000 and used to predict 20 randomly drawn pairs from the Top 100 of each of the following years between 2001 and 2010 (100 predictions in total). The proportion of correct predictions for each model and year were calculated as before. The hypotheses tested in Experiment 2 were as follows:

1. There will be again a trade-off between flexibility and predictive accuracy.
2. The point of this trade-off will be close to the point found in Experiment 1.

Results and Discussion

Replicating the findings of Experiment 1, flexibility was again positively correlated with overall performance, $r = 0.71$, $p < 0.01$. A similar logistic regression analysis as before, where predictive success is regressed on model flexibility, revealed a cubic polynomial without the quadratic term as the best model. The final form

Table 4: Estimates of the different polynomials within the generalized linear regression. * = best model.

Form	**AIC**
Linear	918.4
Quadratic	920.4
Cubic	914.4
Cubic (without quadratic term)	912.4*

of this model is presented in Equation 3 and the parameter estimates are presented in Table 5, alongside their standard errors.

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^3 \quad (3)$$

Table 5: Parameters estimates of the best fitting generalized polynomial linear regression model in Experiment 2.

	β_i	$\sigma(\beta_i)$
β_0	-0.49	0.05
β_1	0.9	0.21
β_2	-3.3	0.31

Again, it is possible to calculate the point of maximum performance (where increasing flexibility further reduces predictive performance) through the following calculations:

$$f(x) = -0.49 + 0.9x - 3.3x^3$$
$$\frac{d}{dx} f(x) = 0.9 - 9.9x^2$$
$$\frac{d}{dx} f(x_{\max}) = 0 \rightarrow x_{\max} = 0.8$$

The maximum point thus lies roughly at the same point as in Experiment 1.

Experiment 2 tried to replicate the overall findings from Experiment 1 within a different setting. Again, we found a trade-off between flexibility and predictive performance and the optimal level of flexibility was close to that found in Experiment 1. However, this time the Random Forest algorithm (flexibility=0.95) performed best overall, even though the fitted model showed the smooth maximum to be at around 0.8. Interestingly, for this data, the Minimalist heuristic only achieved a performance predicting 60% of the choices correctly.

While one could argue that predicting a movies' gross revenue is not a very psychological problem in itself, or that variables such as facebook likes or IMDB scores are directly caused by how many people watch a movie, so that this is more a problem of backwards prediction, the data analysed here closely resembles those used in similar studies of model performance. Importantly, the replication of the flexibility-performance trade-off tells us that there seems to be some truth behind the fact that more flexible models do not always lead to better predictive performance.

General Discussion

In two experiments we found a flexibility-performance trade-off that occurred at an assigned relative rank value of about 0.8. This result nicely brings together both opinions mentioned in the introduction, according to which either more flexibility or more simplicity should be preferred. At least according to our findings, there exists a point where more flexibility reduces a model's predictive performance, but this points occurs rather far on our generated scale. This means that –in some sense– both sides of the argument seem to be right (and wrong).

On the one hand, flexible models should not always be preferred if one wants to make good predictions, as there is a point at which increasing flexibility reduces predictive performance. But the point where flexibility starts penalizing predictive performance lies more towards the flexible side than what some might have expected. For example, all of the included simple heuristics were far less flexible than the optimal level of flexibility.

To our knowledge, we compared more mathematical models of choice than ever before in a single study. In addition, we proposed a relative flexibility measure that was useful to investigate the trade-off between flexibility and predictive performance. The paper at hand can be seen as a first attempt to capture real psychological choices with a large set of different models, whereas past work has mainly focused on rather artificial data sets such as city sizes or professors' salaries (e.g., Gigerenzer & Goldstein, 1999; Chater & Vitányi, 2003). Of course, there are limitations to the interpretation of our findings. First of all, our focus was on the flexibility-performance trade-off, whereas some might argue that the real trade-off is between complexity and performance. Model simplicity is an ambiguous concept. For example, if a heuristic had happened to be the best within our simulations, then the heuristic would have been assigned the highest flexibility rank, even though one might not consider heuristics as very complex models. But whether a model is intuitively "simple" is, loosely put, language dependent (Speekenbrink, 2003). While we admit a model's ability to recover data generated by other models is not a direct indication of a model's complexity, flexibility as defined here captures one of the main reasons why overly complex models have poor predictive performance: their ability to fit random noise in a training set. Additionally, recovery ability has been used to show superior model performance in the literature before (Pitt & Myung, 2002). Another problem is that rank-transforming flexibility values allows for only relative positioning. Introducing many even more complicated models would arbitrarily push the rankings towards 0 and the found trade-off point might have been closer to the less flexible side. A main reason for rank-transforming the values is to make them less task-dependent. The finding that the optimal trade-off point is relatively more towards the more flexible side remains, even if we used the actual percentages of correct predictions Our proposed method to assess model flexibility involved fitting models to a test data set generating by pairing random values of a dependent variable (e.g., choices) to plausible values of the independent variables (e.g., product dimensions). This is similar to using permutation methods in non-parametric statistics. In future work, we plan to explore this link further. Another future step could include simulations of different environments in order to find out the necessary conditions for less flexible models to outperform other models. For example, modelling changing environments between the learning and the test set could give us a better understanding of the driving forces behind a model's performance. By doing so, one could then catalogue the specific attributes of environments that lead to a superior performance of a certain model class. Another step could be to model even more realistic scenarios with our approach. Modelling real choice scenarios could shed more light on the flexibility required to make good predictions in naturalistic situations.

Only by focusing on real psychological phenomena as well as using computationally rigorous approaches can we actually try to answer the question of how we as scientists should actually predict human behavior, but this choice is up to us.

Acknowledgments

This work was supported by a DAAD grant and by the UK Centre for Doctoral Training in Financial Computing & Analytics

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716–723.

Chater, N., Oaksford, M., Nakisa, R., & Redington, M. (2003). Fast, frugal, and rational: How rational norms explain behavior. *Organizational behavior and human decision processes*, *90*(1), 63–86.

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in cognitive sciences*, *7*(1), 19–22.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143.

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: take the best and its relatives. *Simple Heuristics that Make Us Smart. Oxford University Press, New York*, (pp. 75–95).

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in cognitive sciences*, *6*(10), 421–425.

Salmon, W. C. (1971). *Statistical explanation and statistical relevance*. University of Pittsburgh Press.

Speekenbrink, M. (2003). The hierarchical theory of justification and statistical model selection. In *New Developments in Psychometrics*, (pp. 331–338). Springer.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*(1), 28–50.