科技部補助專題研究計畫成果報告
期末報告

依時變動函式的學習：實徵資料與理論模型

計畫類別：個別型計畫
計畫編號：MOST 104-2410-H-004-048-
執行期間：104年08月01日至106年01月31日
執行單位：國立政治大學心理學系

計畫主持人：楊立行

計畫參與人員：碩士級-專任助理人員：鄭惟尹

報告附件：出席國際學術會議心得報告

中華民國 106 年 05 月 04 日

中文摘要：本次研究主題為預測（forecasting）的心理機制。由於預測是人類很基本的認知功能，且幾乎在各個環境中都需要預測，探究人類究竟如何能對未來事件進行預測很具有基礎科學研究的價值。在過去這類的研究較常見於經濟學的研究，像是對股市的預測。然而，一般預測研究的作法是提供實驗參與者所有的歷史資料，再進行未來的預測，例如，預測下週開盤的股價。這樣的預測似乎很需要專業知識才能進行；然而，廣大的投資群眾並非都具有商業的專業知識，但他們就算只觀察一小波段的股價走勢圖，也能大致預測股價的漲跌。顯然這樣的預測有更為先天不需依賴專業知識的成份。因此，本研究以實證實驗針對實驗參與者，測量他們在動態預測作業中的表現。為求精確並排除專業知識涉入的可能，本研究實驗要求參與者以滑鼠點擊他們認為標靶會出現的位置。標靶出現的位置，則是由不同的函式定義。研究結果發現，只要前後兩次標靶出現的位置具有高相關，參與者便能正確學會預測函式。同時，本研究發展了一個簡單的類神經網路說明人類是如何習得預測。僅管如此，本研究並非否定專業知識對預測的重要。這樣的結果不僅延伸了函式學習的範圍，也替預測找到心理運作機制。同時，也對未來的預測研究開啟了新的研究方向。

中文關鍵詞：預測、函式學習、類神經網路、知識分化

英文摘要：Forecasting is referred to predicting the future status of a variable according to a series of its historical statuses. Normally, forecasting is thought to be more relevant to decision making, specifically in the field of economics. For instance, the trend of the stock price in the past six months often is used as an index to forecast the current price. Although forecasting the price of a stock market requires domain knowledge (e.g., economics), to a normal person, it can be realized as a job just to predict the future value according to the previous values. In psychology, this is a case of function learning in that yt = f(yt?1,yt?2,···). In this study, with this position held, it is demonstrated that forecasting can have a simple associative account, just like function learning, and can be directed by a partially relevant context cue, regardless of the true forecasting function, just like knowledge partitioning.

英文關鍵詞：Forecasting, Function learning, Neural Network Model, Knowledge Partitioning

Psychological Foundation of Forecasting

Lee-Xieng Yang

Department of Psychology and Research Center for Mind, Brain, and Learning

National Chengchi University

Abstract

Forecasting is referred to predicting the future status of a variable according to a series of its historical statuses. Normally, forecasting is thought to be more relevant to decision making, specifically in the field of economics. For instance, the trend of the stock price in the past six months often is used as an index to forecast the current price. Although forecasting the price of a stock market requires domain knowledge (e.g., economics), to a normal person, it can be realized as a job just to predict the future value according to the previous values. In psychology, this is a case of function learning in that $y_t = f(y_{t-1}, y_{t-2}, \cdots)$. In this study, with this position held, it is demonstrated that forecasting can have a simple associative account, just like function learning, and can be directed by a partially relevant context cue, regardless of the true forecasting function, just like knowledge partitioning.

In this study, the main idea is to explore the psychological foundation of forecasting. Different from the concern in business research such as how to improve the accuracy of forecasting or the concern in information technology such as how to make a machine more accurate on forecasting outcomes, the focus of this study is more put on exploring what the underling mechanism for people to forecast is. To this end, the forecasting task with dynamic presentation of time series was adopted. As shown in the study of Kusev, van

Schaik, Tsaneva-Atanasova, Juliusson, and Chater (in press), in the prediction task with dynamic presentation of time series, the stimuli are a time series presented to participants one after one for predicting the immediate next one. This is actually a forecasting task and different from the usual way of examining forecasting with all historical data displayed as reference, this kind of forecasting is dynamic forecasting. Due to the similarity between dynamic forecasting and function learning, it was assumed that forecasting, at least dynamic forecasting, shares the same mechanism of function learning. Based on this assumption, a neural network model was proposed by adapting the well-known neural network model in function learning (Add model's name). Also, two experiments were conducted to examine a special phenomenon in associative learning, namely knowledge partitioning (Yang & Lewandowsky, 2003, 2004). Shall dynamic forecasting be a special case of function learning and have an associative account, knowledge partitioning could occur in dynamic forecasting.

Function Learning

Our cognitive system is good at detecting the relationship between variables and further generalizing this relationship to predict one variable according to another. For instance, we can predict how long we need to mow a lawn according to the weather temperature. Obviously, there must be some relationship between the mowing time mt and weather temperature wt. We can collect the data of these two variables and summarize the relationship between them as a mathematical function such as $wt = f(mt)$. However, it is unnecessary to know this function in advance so as to mow a lawn. One can capture the relationship between mowing time and weather temperature via simply observing a series of pairs of mowing time and weather temperature.

Of course, not all functions are equally easy to learn. With the experiments adopting feedback-learning paradigm, the past studies summarized a number of characteristics of function learning, such as that the linear function is easier to learn than the nonlinear function and that the interpolation of a learned function is more accurate than the extrapolation when generalizing predictions to novel stimuli (see DeLosh, Busemeyer, &

McDaniel, 1997; Koh & Meyer, 1991). Although in earlier literature, the representation for the learned function was assumed to be a rule in a polynomial like format, the rule-based account seems to overestimate the performance of people when extrapolating the learned function to the stimuli outside the range of training stimuli. However, the neural network model proposed by Busemeyer, Byun, Delosh, and McDaniel (1997) can account for the miss resulting from extrapolating the learned function in virtue of the similarity between stimuli. Further, Kalish, Lewandowsky, and Kruschke (2004) proposed the POLE model, which consists of many modules and each of which is a simple neural net, in charge of learning a linear relationship between two variables. According to the POLE model, a complex function is not learned as a whole, but is partitioned by the input value to smaller segments to learn. This idea was actually inspired by a series of studies addressing knowledge partitioning in category learning that people learned to rely on context to generate partial rules for categorization, instead of learning the true categorization rule (Yang & Lewandowsky, 2003, 2004) and in function learning that people learned to rely on context to generate partial functions instead of the true function (Lewandowsky, Kalish, & Ngang, 2002). These authors provided behavioral evidence that people do rely on the input value, as context, to partition the to-be-learned function in order to simplifying the learning difficulty. The POLE model can accommodate this result.

Neural Network Model for Dynamic Forecasting

Dynamic forecasting refers to predicting the values in a time series one after one. Apparently, when predicting the value y_t at time t, the previously seen values y_{t-1}, y_{t-2}, $\cdots$, and y_1 are the only source of prediction. Thus, forecasting can be formulated as an autoregressive problem f, where $\hat{y}_t = f(y_{t-1}, y_{t-2}, \cdots, y_{t-n+1})$. In fact, when the correlation between successive values is high, forecasting can be done by relying on only the last value (Yang & Lee, 2015). Thus, the forecasting function can have the format of $y_t = f(y_{t-1})$. Recall that the function designed for examining function learning always has the format of $y = f(x)$. Due to the similarity between these two types of functions, it is reasonable

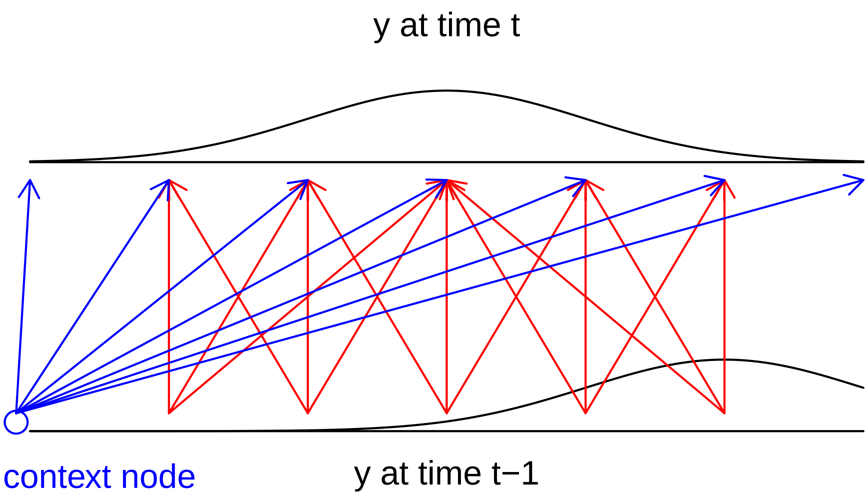

Figure 1. The Neural Network Model for Forecasting.

to regard forecasting as a kind of function learning with the input and the output as the values at successive moments. Therefore, it is straightforward to treat the neural network model proposed by Busemeyer et al. (1997) as a candidate model for accounting for dynamic forecasting.

The neural network model of Busemeyer et al. (1997) is a two-layered neural net, with the input and the output layers respectively represent for the input dimension and the output dimension. However, in the current study, the input and the output respectively represent for the values of a time series at time $t-1$ and t. See Figure 1.

There are n input nodes as well as n output nodes. In no matter the input or output layer, each node corresponds to a value s on the stimulus dimension and functions as the exemplar in the exemplar-based categorization model. When the stimulus at time $t-1$, x_{t-1}, is received, the input nodes are activated to the extent of the closeness of them to that stimulus, which is computed as

$$A_{t-1}^T = exp(-||x_{t-1} - s||^2/\sigma), \tag{1}$$

where A_{t-1}^T is an n-element vector containing the activation of input nodes. The parameter σ is used to tune the size of the receptive field of each input node. When σ is large, the system is less able to tell the difference between stimuli and vice versa. In addition to the input nodes, the context node is used to represent the context on observing the stimulus value. In the situation where no particular context should be considered to predict the stimulus value in a time series, the context node can be omitted. In the experiments in the latter sections where context was the color of the stimulus, the activation of context node is set up as 1 and -1 for two different colors. The activation of context node is weighted by an attention weight $0 \leq \alpha \leq 1$.

The activation of input nodes is normalized to 0 and 1 and then weighted by the associative weights W and summed as the activation of output nodes A_t^T as

$$A_t^T = WA_{t-1}^T. \tag{2}$$

If a context cue is considered as input signal as well, here A_{t-1}^T is changed to $A_{in}^T = [A_{t-1}, A_c]^T$, where A_c is the activation of context node. Therefore, in this circumstance, the output activation is computed as

$$A_t^T = WA_{in}^T. \tag{3}$$

When generating the prediction, only the activation of the output nodes around the winning output node is considered, in order to counterbalance the privilege for the central value due to blending all output activation. This is done by setting as 0 the activation of output nodes outside the receptive field of the winning node. The system output is the expected value of the stimulus values S, whose corresponding output nodes are within the neighborhood of the winning output node M. The probability of each candidate stimulus value to be chosen as the system output is computed as

$$p(s_\kappa) = \frac{A_{\kappa,t}}{\sum_{k \in M} A_{k,t}}. \tag{4}$$

Of course, the larger activation the output node has, the more likely its corresponding value is to be chosen. The system output then is

$$\hat{y_t} = \sum_{k \in M} p(s_k) s_k. \tag{5}$$

This neural network model adopts the error-driven learning algorithm to adjust the associative weights. First, the target value observed at time t will be transformed to the the activation of output nodes using Equation 1 and also be normalized to 0 and 1, just like what is done for the input activation. Second, the activation of output nodes now is called the target activation T. Therefore, the error on each output node k is $T_{k,t} - A_{k,t}$. Therefore, the associative weight between the kth output node and the ith input node will be changed by summing

$$\Delta w_{i,k} = \eta \beta (T_{k,t} - A_{k,t}) A_{i,t-1}. \tag{6}$$

where η is the learning rate, a small positive number, and $\beta = exp(-\xi(t-1))$ is the decay rate to attenuate the learning step according to the elapsed time with ξ is a freely estimated parameter. The larger ξ the more quickly the learning is halted.

Experiment

As shown in the discussion in the previous sections that forecasting can be viewed as a special case of function learning, in this experiment, a time series was defined by $y_t = 7\sin(0.17(t-1) - 7.2)$ as a forecasting function for participants to learn. See Figure 2. The abscissa shows the position of the values in the time series. The coordinate shows the value of y. In the experiment, the participants were asked to predict by moving the mouse cursor the position of a target trial by trial. The target would be presented on a horizontal line with the left and right ends corresponding to the minimum and maximum of y. When a response was made, the correct position would be indicated with an arrow printed in color red or color green. Here the color is partially predictive of the target position, as it can only predict the moving direction of the target not the moving distance. In fact, the true forecasting function does not include the context cue. That is, it is unnecessary to attend to the color in order to make a forecast. In the past research, it was found that people might rely on an irrelevant context cue to decide which rule should be applied for current categorization (Yang & Lewandowsky, 2003, 2004) and to decide which function should be applied for generating the current response (Lewandowsky et al., 2002). This phenomenon is called knowledge partitioning. Due to that forecasting is highly similar to function learning, it is reasonable to expect that knowledge partitioning should occur in this experiment.

Method

Participants and Apparatus

Thirty-seven undergraduate students in National Chengchi University participated in this experiment. The whole experiment was conducted on an IBM-compatible PC in a quiet booth. The processes of stimulus displaying and response recording were under the control of a computer script composed by PsychoPy (Peirce, 2007). After testing, each participant was reimbursed with NT$ 120 ($\simeq$ US$ 4) for their time and traffic expense.

Procedure

The participants were instructed to predict (by moving the mouse cursor to) where the target would appear on a horizontal line on the computer screen. Before the onset of experiment, every participant had gone through four trials for practicing how to move the mouse cursor to a given position. There were two sessions in this experiment, each of which

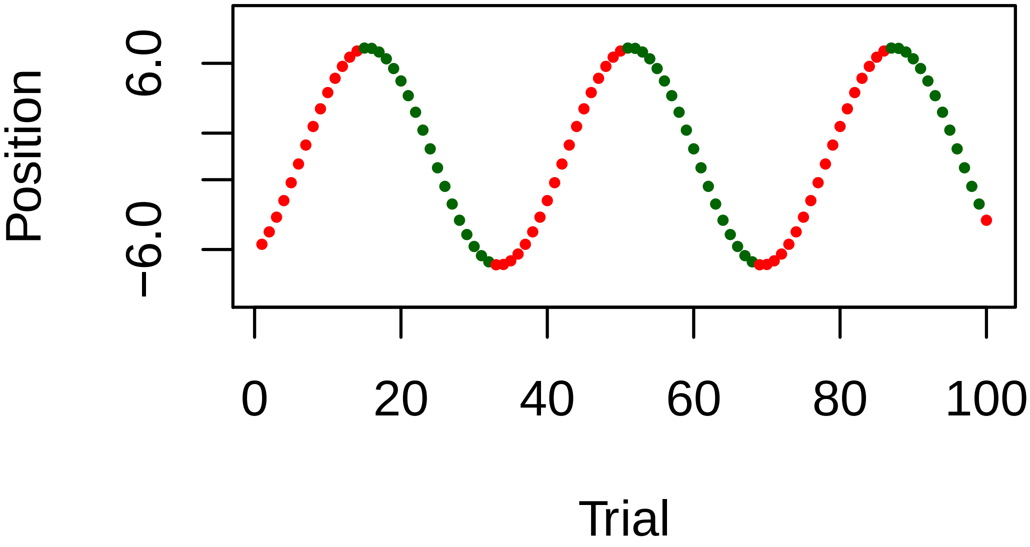

Figure 2. The Forecasting Function.

consisted of a learning phase and a transfer phase. The second session was actually the repetition of the first one.

In the learning phase, the target moved back and forth in a sine function as shown in Figure 1. Note the final learning trial in the first session was by accident paired with a color for going right, but it instead provided a chance to check for knowledge partitioning in the learning phase. See the increment on the prediction for this trial in Figure 2. The pairing between the color (i.e., red or green) and the moving trend of the target (i.e., moving left or right) was counterbalanced across the odd-numbered and even-numbered participants. Once the prediction was made, the correct position would be indicated with a colored arrow together with a textual message Hit or Miss message. If the different between the predicted position and the correct position is close enough, a Hit message would be presented and otherwise a Miss message would be presented.

In the transfer phase, the participants were instructed to predict the next three positions of the target, given its current position as a cue. There were four transfer positions paired with two colors. Thus, there were eight transfer trials in total. The cue positions were randomly presented. There was a self-paced break between the two sessions.

Results

Learning Performance

The averaged predictions across all participants are shown as gray dots in Figure 3. The correct answers are denoted by red or green crosses. In this case, red means going right and green means going left. Apparently, the participants learned very well to forecast the target positions.

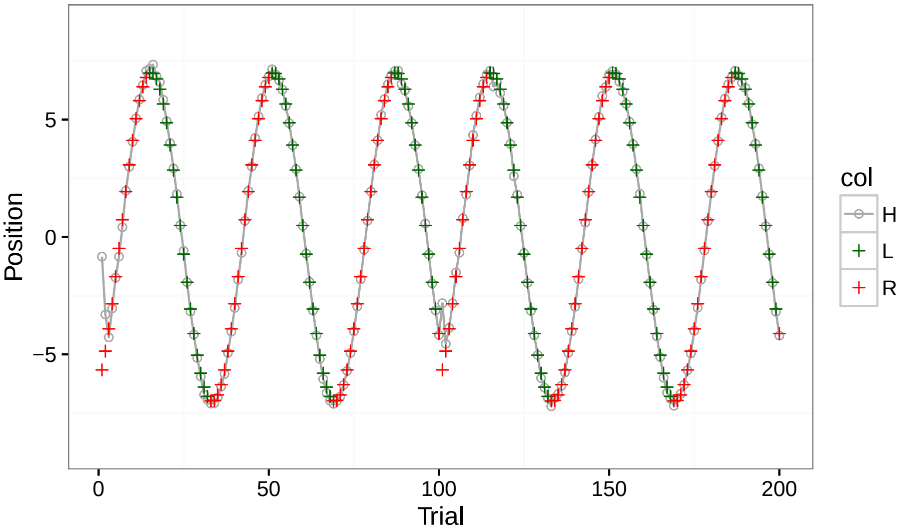

Figure 3. The participants' predictions and correct answers in the learning phases. H: Human, L: Left, and R: Right.

For the convenience of data analysis, the items in each cycle of the sine wave are aggregated as one block. Thus, there are three blocks in each session. The learning performance gets better, in respect of the accuracy and prediction error, as the learning trials increase. Then the accuracy is computed as the proportion of Hit received in each block. The prediction error is the squared difference between the predicted and correct positions. A Context (2) × Block (3) × Session (2) within-subject ANOVA shows a higher accuracy on the left context than the right context ($M_{left} = .40$ vs. $M_{right} = .35$), $F(1, 36) = 26.99$, MSe = 0.01, $p < .01$. This is reasonable, as the left context appeared later in the learning phase.

The low accuracies result from a rigorous threshold to get a Hit message. Also, the accuracy gets better along the learning blocks (from .31 to .44), $F(2, 72) = 36.88$, MSe = 0.02, $p < .01$ and the learning accuracy is higher in the second session as well ($M_1 = .34$ vs. $M_2 = .42$), $F(1, 36) = 26.44$, MSe = 0.03, $p < .01$. The interaction effect between Context and Block is significant, $F(2, 72) = 3.38$, MSe = 0.02, $p < .05$. This is because the accuracy in the right context catches up that in the left context in the last block. There is no other significant interaction effect [for Context × Session, $F(1, 36) = 2.48$, MSe = 0.02, $p = .12$; for Block × Session, $F(2, 72) < 1$; for Context × Block × Session, $F(2, 72) = 1.30$, MSe = 0.01, $p = .28$].

For the prediction error, again a Context (2) × Block (3) × Session (2) within-subject ANOVA is conducted. The results show a significant main effect for Context [$F(1, 36) = 22.88$, MSe = 7.49, $p < .01$], Block [$F(2, 72) = 33.95$, MSe = 8.6, $p < .01$], and Session [$F(1, 36) = 13.67$, MSe = 3.32, $p < .01$]. The prediction error decreases as the block number increases (from 2.59 to 0.11) and as the session number increases (from 1.25 to 0.61). All interaction effects are significant: Context × Block [$F(2, 72) = 22.51$, MSe = 7.49, $p < .01$], Context × Session [$F(1, 36) = 9.72$, MSe = 3.56, $p < .01$], Block × Session [$F(2, 72) = 11.91$, MSe = 3.42, $p < .01$], and the three way interaction effect [$F(2, 72) = 10.43$, MSe = 3.57, $p < .01$].

Transfer Performance

If knowledge partitioning occurs, it should be expected that the target is predicted to move right in the right context and left in the left context. The difference between successive positions can tell us the information of moving direction. Thus, the differences between the predicted positions as well as the difference between the first predicted position and the cue position are computed. The mean of these differences is the dependent variable in data analysis. If the mean difference is positive, the target is predicted to move right; whereas if the mean difference is negative, it is predicted to move left.

The transfer performance can be seen in Figure 4. It looks like that the predicted moving direction differs in different contexts. Specifically, for the two middle cue positions (i.e., the ordinate values as -1.5 and 1.5 in Figure 2), this context-gated response pattern is more salient. However, for the cue positions on the two sides (i.e., the ordinate values as -6 and 6 in Figure 1), this pattern is not that clear. As the side positions are at around the edges of the position scale in the learning phase, this might display the characteristic of function learning that the extrapolation prediction is less accurate than the interpolation prediction.

A Session (2) × Context (2) × Stimulus (4) within-subject ANOVA is conducted for these mean differences. The results show that the mean difference is different in different contexts [$F(1, 36) = 23.43$, MSe = 29.60, $p < .01$] and for different stimuli [$F(3, 108) = 23.86$, MSe = 18.6, $p < .01$], and there is a marginally significant difference on the mean difference in different sessions [$F(1, 36) = 3.59$, MSe = 6.98, $p = .07$]. The observed context-dependent pattern on the mean difference is more clear in Session 2 than Session 1, that is supported by the significant interaction effect between Context and Session [$F(1, 36) = 9.20$, MSe = 7.50, $p < .01$]. The marginally significant interaction effect between Context and Stimulus [$F(3, 108) = 2.38$, MSe = 7.99, $p = .07$] supports the observation that the predictions are more context dependent for the middle cue positions than the side cue positions. The predictions for the stimuli are not different in different sessions, $F(3, 108) < 1$. The three-way interaction effect is also not significant [$F(3, 108) = 1.01$, MSe = 9.13, $p = .39$].

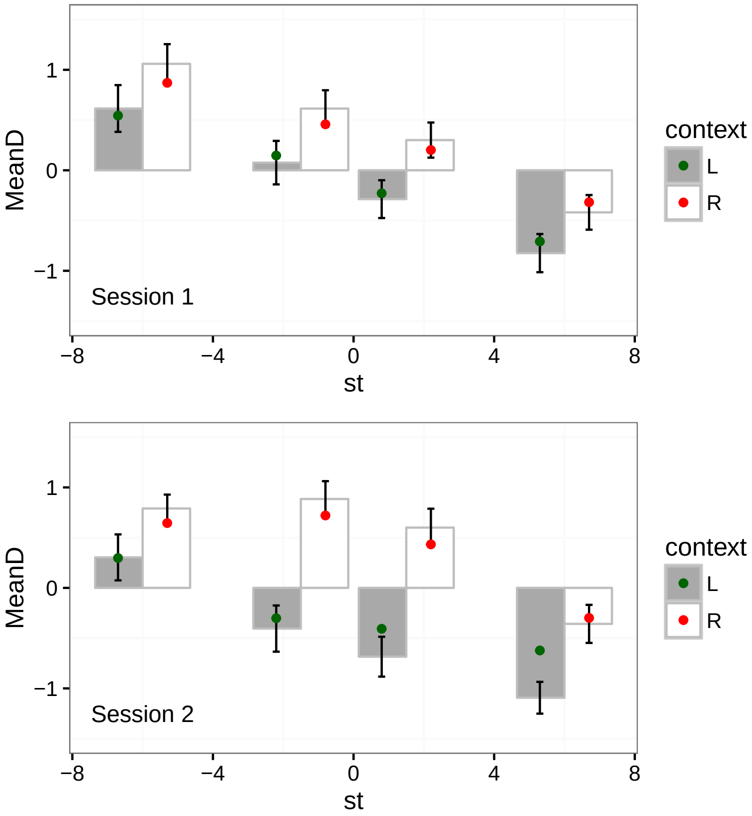

Figure 4. The mean difference of predictions for cue positions made by all participants. L: Left, and R: Right. Bar: Human data, Dot: Model prediction.

Individual Differences

The response pattern in Figure 4 is partly consistent with the expectation for knowledge partitioning (i.e., for the middle cue positions). One reasonable suspicion is that there might be individual differences behind the averaged response. We first compute the squared mean difference on the transfer predictions for each participant and preclude from data analysis those participants whose squared mean difference is smaller than 75% of the participants. Ten participants are precluded.

For the rest twenty-seven participants, we compute the distance from their prediction patterns to the ideal KP pattern and the ideal sine wave pattern (SW). As we care about the moving direction not the moving distance, the ideal pattern on the eight trials in each transfer phase can be coded as a 16-element vector with 1 as moving right and -1 as moving left. Thus, in the KP pattern, 1 and -1 are coded respectively for the trials in the right and left contexts. However, in the SW pattern, 1s and -1s are coded respectively for the trials with the left and right two cue positions. The participants' mean differences are recoded as 1 and -1 for a positive and negative mean difference on each of 16 trials. A participant would be classified as the KP group, if his/her response pattern has a shorter Euclidean distance to the KP pattern and vice versa. Consequently, there are 14 KP participants and 11 SW participants and 2 participants equidistant to these two groups.

The transfer response pattern of KP group can be seen in Figure 5. Comparing with Figure 4, the context-gated prediction gets more clear even for the cue positions on the left and right sides.

A Session (2) × Context (2) × Stimulus (4) within-subject ANOVA is conducted for the mean differences of these 14 participants. The results show that the pattern of mean differences is not different in different sessions [$F(1, 13) = 2.81$, MSe $= 1.04$, $p = .12$], but it is in different contexts [$F(1, 13) = 51.01$, MSe $= 2.19$, $p < .01$] and for different stimuli [$F(3, 39 = 11.5$, MSe $= 1.08$, $p < .01$]. None of the interaction effects is significant [for Session × Context, $F(1, 13) = 2.76$, MSe $= 1.45$, $p = .12$; for Context × Stimulus, $F(3, 39) < 1$; for Session × Stimulus, $F(3, 39) = 2.34$, MSe $= 0.97$, $p = .09$, and for Session × Context × Stimulus, $F(3, 39) = 1.26$, MSe $= 1.01$, $p = .30$].

The mean differences of the SW group ($N = 11$) can be seen in Figure 6. Different from the KP group, these participants do not go with context, but go with the cue position: predicting the target to move right for the left cue-positions and left for the right cue-positions. A Session (2) × Context (2) × Stimulus (4) within-subject ANOVA is conducted for the mean differences. Among all effects, only the main effect of Stimulus is significant, $F(3, 30) = 20.84$, MSe $= 2.94$, $p < .01$. The main effect of Session is not significant [$F(1, 10) = 2.21$, MSe $= 1.04$, $p = .17$] and nor is the main effect of Context [$F(1, 10) < 1$]. The interaction effects are not significant [for Session × Context, $F(1, 10) = 2.58$, MSe $= 0.51$, $p = .14$; for Session × Stimulus, $F(3, 30) < 1$; for Context × Stimulus, $F(3, 30) < 1$;

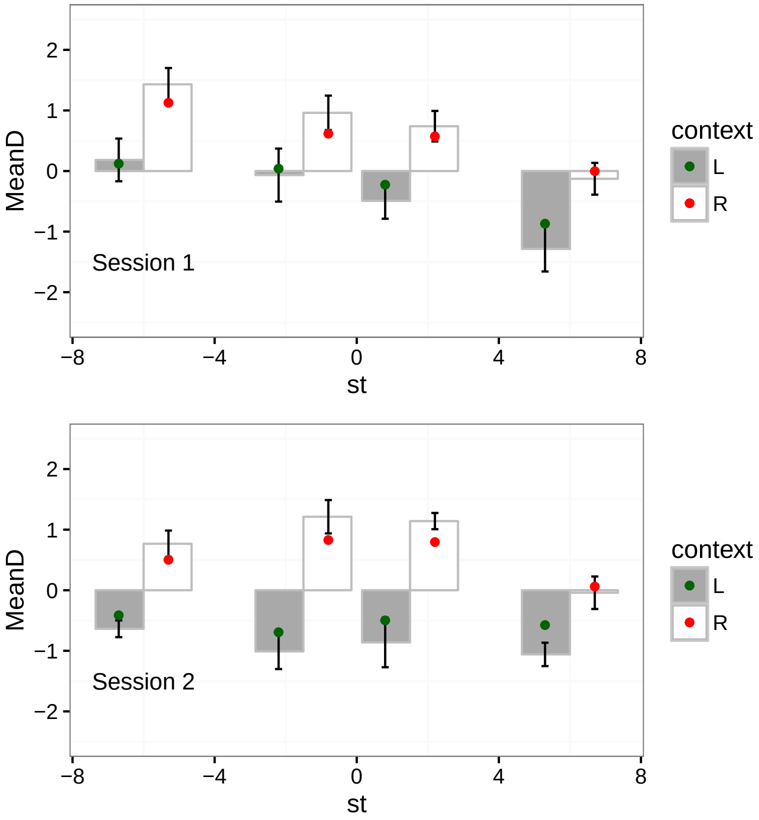

Figure 5. The mean difference of predictions made by the KP group. Bar: Human data, Dot: Model prediction.

for the three-way interaction effect, $F(3, 30) < 1$].

We also examine the confidence of these two groups with their predictions for different cue positions (middle vs. side) in different contexts. With the absolute mean differences as the dependent variable, a Group (2) × Context (2) × Stimulus-Type (2) between-within ANOVA shows that only the interaction effect Group and Stimulus-Type is significant [$F(1, 23) = 10.58$, MSe $= 0.81$, $p < .01$]. This is because the SW group is more confident than the KP group with their predictions for the side positions (1.53 vs. 1.00) [$F(1, 23) = 7.19$, MSe $= 1.93$, $p < .05$], but not for the middle positions (1.26 vs. 1.20) [$F(1, 23) < 1$]. The SW group knows well that the target will turn to another direction at the two sides, according to the sine-wave moving trend. However, the KP group hesitates due to the

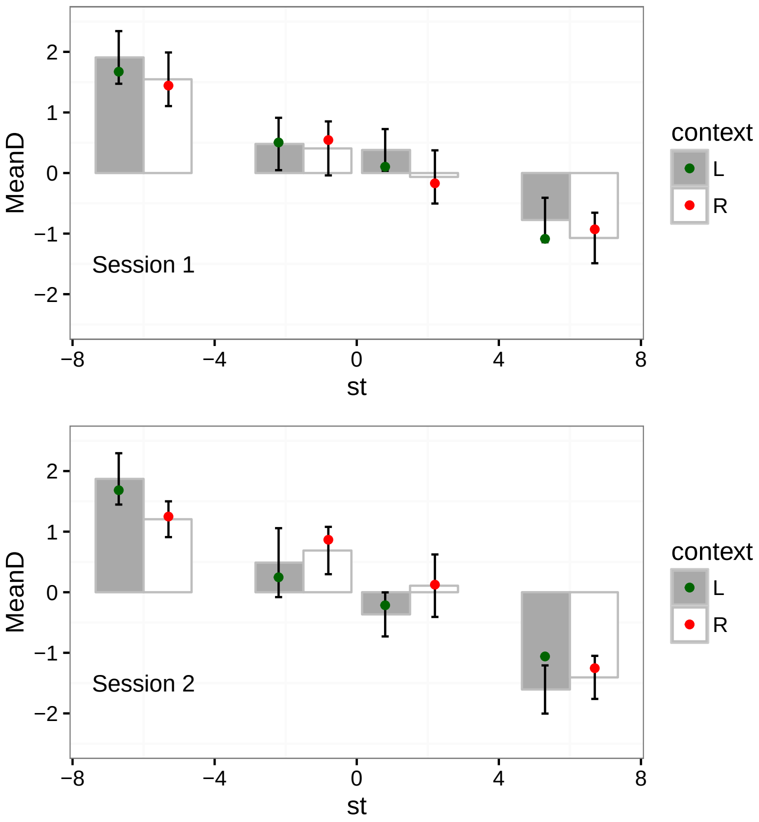

Figure 6. The mean difference of predictions made by the SW group (i.e., the group learning the true sine-wave function). Bar: Human data, Dot: Model prediction.

considering the context.

Modeling

In order to understand the mental mechanism to forecast the target position in this experiment, we propose a simple neural network model[1]. In this model, the target position is not a scalar but is represented as the activation distribution of 100 position nodes. It is assumed that each of the position node activates the most to a real location on the horizontal line, where the target will appear. If a stimulus appears at x and the i-th position node

[1] The MoE architecture is not adopted, as context is involved in generating predictions in weak knowledge partitioning.

activates the most to the location l_i, the activation of this node S_i for the stimulus is computed as $S_i = exp^{-(x-l_i)^2/\sigma}$, where σ is a freely estimated parameter, representing for the size of the receptive field of the position node. The activation of all position nodes S is then normalized to 0 and 1. Of course, the closer the stimulus to the location corresponded to by a node, the stronger the activation of that node is.

The input layer consists of 100 position nodes as well as a context node[2], which is weighted by an attention weight α. Each trial starts with the correct target position on the preceding trial, which is presented by a colored arrow on the screen[3]. Thus, the input layer represents the perceived location of the target and its color as a 101-element vector, A_{in}.

The output layer consists of 100 position nodes only, as the participants were not asked to predict the target color. The output vector A_{out} is the model prediction for the target position, which is computed as $A_{out} = WA_{in}$, where W is the associative weights between these two layers. The associative weights are set up as 1 for the connections between the nodes in different layers corresponding to the same location and 0 for the other connections. With this design, the model starts learning from copying the observed target position as its prediction.

When the colored arrow, as feedback, is provided to the participants, the location of it is represented as an activation distribution over 100 position nodes, which is then normalized to 0 and 1 and used as the target vector T for the model to learn. In fact, the target vector on trial t is the input vector without the element for the context node on trial $t + 1$.

The associative weights are adjusted by error-driven learning, as $\Delta W = \eta\beta(T - A_{out})A_{in}^T$, where η is the learning rate. The parameter $\beta = exp^{-\xi(t-1)}$ is used to gradually attenuate the learning rate along trials and the parameter ξ is estimated when modeling.

As we are particularly interested in how the context-gated forecasting occurs, we fit the model to each participant's data with the goodness-of-fit as the RMSD (Root-Mean-

[2] -1 for the left context and 1 for the right context.

[3] Thus, the first guess made by the participants is not included and the first correct position of the target launches the whole process of learning.

Square Deviation) between the participant's predictions in the transfer phase and the model predictions. The model prediction is a scalar representing for the predicted position, which is converted from the activation of output nodes. That is, we focus on the nodes C within the receptive of the winning node and normalize their activation to 0 and 1. Then the activation of each node $i \in C$ is turned to a probability $P_i = A_{out,i} / \sum A_{out,j \in C}$. The model prediction is the weighted sum of $\sum P_i l_i$, $i \in C$. When modeling, the model goes through all trials in each phase in each session, which the participants went through as well. Four parameters are freely estimated to optimize the model's performance: the learning rate $[0.0001 \leq \eta \leq 0.9]$, the size index for receptive field $[0.01 \leq \sigma \leq 0.25]$, the decay rate $[0 \leq \xi \leq 2]$, and the attention weight on context $[0 \leq \alpha \leq 1]$.

Results

The transfer predictions of the model can be seen as the colored dots in the figures from Figure3 to Figure 5. The means of goodness of fit and the parameter values providing a best fit are listed in Table 1. Apparently, the model accounts for well the performance of each group.

Table 1: The mean GOF and parameter values providing a best fit for different groups.

	RMSD	η	σ	ξ	α
All	0.09	0.30	0.08	0.22	0.56
KP	0.08	0.32	0.06	0.11	0.69
SW	0.15	0.41	0.19	0.31	0.46

Table 1 shows the statistics of modeling data. The GOFs are good in all modeling situations. However, the model is relatively better at accounting for the KP pattern than the SW pattern, $t(23) = -2.58$, $p < .05$. This might be because context in weak knowledge partitioning still can predict something about the outcome. The comparison between the KP and SW groups suggest that their starting learning rates are not different [$t(23) = -0.75$,

p = .46], but the SW group has a larger size of receptive field [$t(23) = -3.04$, $p < .01$], more quickly halts learning [$t(23) = -3.96$, $p < .01$], and puts less attention on context (marginally significant) [$t(23) = 1.82$, $p = .08$]. Nonetheless, it is suggested that knowledge partitioning in forecasting can have an associative-based account.

Conclusions

It is clear that the participants can learn to forecast the values in a time series generated by a complex sine function. Also, it is found that some participants learn to perform weak knowledge partitioning in forecasting, whereas some others learn to ignore context when making a forecast. A simple two-layered neural network model can provide good accounts for the performance of different groups. This is consistent with the past studies about function learning and knowledge partitioning.

References

Busemeyer, J. R., Byun, E., Delosh, E., & McDaniel, M. A. (1997). *Learning functional relations based on experience with input-output pairs by humans and artificial neural networks* (K. Lamberts & D. R. Shanks, Eds.). Cambridge, MA, US: The MIT Press.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Jounral of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 968-986.

Kalish, M., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and funciton leanring. *Psychological Review*, *111*, 1072-1099.

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 811-836.

Kusev, P., van Schaik, P., Tsaneva-Atanasova, K. T., Juliusson, A., & Chater, N. (in press). Adaptive anchoring model: how static and dynamic presentation of time series influence judgments and predictions. *Cognitive Science*.

Lewandowsky, S., Kalish, M., & Ngang, S. K. (2002). Simplified learning in complex situations: Knowledge partitioning in function learning. *Journal of Experimental Psychology: General*, *131*, 163-193.

Peirce, J. W. (2007). Psychopy-psychophysics software in python. *Journal of Neuroscience Methods*, *162*, 8-13.

Yang, L.-X., & Lee, T.-H. (2015, July). Learning of time varying functions is based on association between successive stimuli. In D. C. Noelle et al. (Eds.), *Porceedings of the 37th annual conference of the cognitive science society* (p. 2727-2732). Austin, TX: Cognitive Science Society.

Yang, L.-X., & Lewandowsky, S. (2003). Context-gated knowledge partitioning in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 663-679.

Yang, L.-X., & Lewandowsky, S. (2004). Knowledge partitioning in categorization: Constraints on exemplar models. *Journal of Experimental Psychological: Learning, Memory and Cognition*, *30*, 1045-1064.

出國報告

本人於 2016 年赴美國波士頓參加 2016 annual meeting of Psychonomics，並以海報發表研究成果。本次報告的研究內容主題是”A Reference Point Explanation for XOR Extrapolation in Categorization with Kernel Methods”。這個主題主要動機來自於 Conaway 和 Kurtz（2015）於認知科學年會中發表之論文。這些作者主張以歧義類別結構（ambiguous category structure）測驗實驗參與者所習得的分類策略時，發現有 1/3 的參與者自行演生出一個對稱的 XOR 結構，而另外 2/3 的參與者則滿足一般範例為基礎的分類模型的預測，以相似性進行分類。這些作者並同時提供電腦模擬的結果說明，以範例為基礎的分類模型，例如 ALCOVE，無法解釋那 1/3 使用 XOR 策略的參與者的表現。因為在測驗階段中新出現的刺激竟被判斷為與之較不相似的類別。同時，這些作者也提出證據說明，DIVA 模型可以同時解釋這兩種分類策略。

本次研究主要在於證明，以範例為基礎的分類模型，確實可以預測 XOR 的分類策略，只需要將原先的距離相似性改成向量相似性即可。研究同樣使用電腦模擬並以機器學習中的 2 次方多項式核心（polynomial kernel）作為計算相似性的公式。結果顯示，即使是範例為基礎的分類模型，也一樣可以預測 XOR 分類策略。

會議中除了與許多國外學者討論本研究的內容，也和美國雪城大學 Mike Kalish 教授討論後續可能的研究，並且討論可以如何進行工作記憶廣度與分類學習之相關的研究。Kalish 教授並允諾將於 2018 年 4 月來台灣，商討進一步的合作研究可能。此外，會議中也與南澳阿得雷得大學的 John Dunn 教授一同討論關於 state trace theory 的可能延伸議題。

會後並接獲 Ken Kurtz 的來信，其為 DIVA 模型的開發者，並於信中表示對本研究的高度興趣。同時也與本人分享他最新的論文，一同切搓關於分類策略的心得。

科技部補助計畫衍生研發成果推廣資料表

日期:2017/05/03

科技部補助計畫	計畫名稱: 依時變動函式的學習：實徵資料與理論模型
	計畫主持人: 楊立行
	計畫編號: 104-2410-H-004-048- 學門領域: 實驗及認知心理學

無研發成果推廣資料

104年度專題研究計畫成果彙整表

計畫主持人：楊立行	**計畫編號**：104-2410-H-004-048-
計畫名稱：依時變動函式的學習：實徵資料與理論模型	

成果項目					量化	單位	質化 （說明：各成果項目請附佐證資料或細項說明，如期刊名稱、年份、卷期、起訖頁數、證號...等）
國內	學術性論文	期刊論文			0	篇	
		研討會論文			1		Yang &amp; Lee （2015） Learning of Time Varying Functions is Based on Association Between Successive Stimuli. In Gabriella Airenti， et al. （Eds.）， Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science （p.722-727）. Torino， Italy.
		專書			0	本	
		專書論文			0	章	
		技術報告			0	篇	
		其他			0	篇	
	智慧財產權及成果	專利權	發明專利	申請中	0	件	
				已獲得	0		
			新型/設計專利		0		
		商標權			0		
		營業秘密			0		
		積體電路電路布局權			0		
		著作權			0		
		品種權			0		
		其他			0		
	技術移轉	件數			0	件	
		收入			0	千元	
國外	學術性論文	期刊論文			0	篇	
		研討會論文			0		
		專書			0	本	
		專書論文			0	章	
		技術報告			0	篇	
		其他			0	篇	
	智慧財產權及成果	專利權	發明專利	申請中	0	件	
				已獲得	0		
			新型/設計專利		0		
		商標權			0		
		營業秘密			0		

		積體電路電路布局權	0		
		著作權	0		
		品種權	0		
		其他	0		
	技術移轉	件數	0	件	
		收入	0	千元	
參與計畫人力	本國籍	大專生	0	人次	
		碩士生	0		
		博士生	0		
		博士後研究員	0		
		專任助理	1		負責協助實驗進行、執行報帳相關等行政工作。
	非本國籍	大專生	0		
		碩士生	0		
		博士生	0		
		博士後研究員	0		
		專任助理	0		
其他成果 （無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。）					

科技部補助專題研究計畫成果自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現（簡要敘述成果是否具有政策應用參考價值及具影響公共利益之重大發現）或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以100字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形（請於其他欄註明專利及技轉之證號、合約、申請及洽談等詳細資訊）

論文：■已發表　□未發表之文稿　□撰寫中　□無

專利：□已獲得　□申請中　■無

技轉：□已技轉　□洽談中　■無

其他：（以200字為限）

本研究成果已初步完成一篇論文，並已在2015年於義大利都靈召開之歐洲亞太認知科學聯合會議中發表。另，第二篇會議論文已投稿至第37屆美國認知科學年會（Society of Cognitive Science），雖然第一階段審查未獲通過，但主編建議以海報形式發表，故目前正在撰寫文稿以備審查。此外，也將再撰寫一篇論文，擬投稿至cognitive science期刊發表。

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性，以500字為限）

本次研究主題為預測（forecasting）的心理機制。由於預測是人類很基本的認知功能，且幾乎在各個環境中都需要預測，探究人類究竟如何能對未來事件進行預測很具有基礎科學研究的價值。在過去這類的研究較常見於經濟學的研究，像是對股市的預測。然而，一般預測研究的作法是提供實驗參與者所有的歷史資料，再進行未來的預測，例如，預測下週開盤的股價。這樣的預測似乎很需要專業知識才能進行；然而，廣大的投資群眾並非都具有商業的專業知識，但他們就算只觀察一小波段的股價走勢圖，也能大致預測股價的漲跌。顯然這樣的預測有更為先天不需依賴專業知識的成份。因此，本研究以實證實驗針對實驗參與者，測量他們在動態預測作業中的表現。為求精確並排除專業知識涉入的可能，本研究實驗要求參與者以滑鼠點擊他們認為標靶會出現的位置。標靶出現的位置，則是由不同的函式定義。研究結果發現，只要前後兩次標靶出現的位置具有高相關，參與者便能正確學會預測函式。同時，本研究發展了一個簡單的類神經網路說明人類是如何習得預測。這樣的結果不僅延伸了函式學習的範圍，也替預測找到心理運作機制。對未來的預測研究開啟了新的研究

方向。
4. 主要發現 本研究具有政策應用參考價值：■否　□是，建議提供機關 （勾選「是」者，請列舉建議可提供施政參考之業務主管機關） 本研究具影響公共利益之重大發現：■否　□是 說明：（以150字為限）