# A STATE-SPACE MODEL FOR PROTOTYPE LEARNING

In Jae Myung and Jerome R. Busemeyer

Department of Psychological Sciences
Purdue University

## ABSTRACT

A general state-space model of prototype learning was formulated in terms of a set of internal states and nonlinear input-output mappings. The general model includes several previous models as special cases such as Hintzman's (1986) multiple trace model, Metcalf's (1982) holographic model, and two parallel distributive memory models (Knapp & Anderson, 1984; McClelland & Rumelhart, 1985). Two basic properties common to the three models were defined in terms of this general model--additivity and time invariance. An experiment was conducted to test the basic properties using random spectral patterns as stimuli allowing possible nonlinear input and output distortions. Especially, ordinal tests of additivity were performed with few assumptions about internal features that subjects may use to encode the stimulus information. The results support additivity but time-invariance was clearly violated. Implications of these findings for models of the human memory system are discussed.

## INTRODUCTION

One of the most intriguing questions about the structure and organization of human memory is how new experience interacts with old memory to compose an abstraction. For example, when we meet a new person, we form a first impression, and later, this impression is changed and modified with subsequent meetings with the same person. Somehow, later impressions interact with previous experience in memory to establish the current revised impression. What underlying learning processes enable humans to do such an abstraction? Recently, we have witnessed a surge of adaptive neuro-network models of this dynamic learning process. Interestingly, many of the models have a common core of fundamental assumptions. It would be worthwhile to empirically test the validity of these assumptions before we move on to further development of the models.

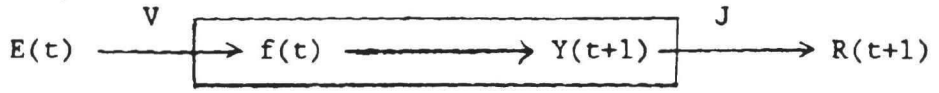
The purpose of this study was to empirically test these common assumptions. Specifically, the present experiment was designed to test two basic properties of memory structure assumed by several memory models--additivity and time invariance of memory system. The memory models were Hintzman's multiple trace model (1986), Metcalfe's holographic memory model (1982), and parallel distributed memory models (Knapp & Anderson, 1984; McClelland & Rumelhart, 1985). In order to test the basic properties, we chose to study prototype learning using a new experimental paradigm called prototype production (Busemeyer & Myung, 1988). In the prototype production task subjects are shown a sequence of exemplars (e.g., a series of pictures or sounds) generated from one or more prototypes with category labels. Then subjects are given a category label and are asked to produce their prototype estimate of the category (e.g., draw a picture or vocalize a sound that best represents the category). Note that in the prototype production task, abstraction is a task requirement and so the major question is "how does abstraction occur?"

The present article is organized as follows. First, A general state-space model of prototype formation will be presented, followed by definitions

of the two basic properties. Then we will discuss the three memory models in relation to the general model, and we will show that all three models satisfy the two basic properties. Finally, we present experimental tests of additivity and time-invariance followed by discussion of implications of the experimental findings.

## STATE-SPACE MODEL OF PROTOTYPE EVOLUTION

We begin by distinguishing between representations of images formed by the subject and by the experimenter. On each trial, denoted t, an exemplar image is presented visually (e.g., a photograph) or auditorily (e.g., a tone sequence). We assume that the experimenter records the exemplar image by obtaining a set of physical measurements. This record is represented by a vector denoted E(t). Another vector, denoted f(t), is used to represent the subject's perceptual image of the corresponding physically defined exemplar image E(t). The values of elements of f(t) represent feature strengths. In prototype learning, an exemplar ensemble consists of two components, an image f(t) and category label denoted g(t) as a vector (e.g., the title of a picture). Then the exemplar ensemble can be represented by a vector $h(t)=g(t)|f(t)$ where the symbol | indicates concatenation of two vectors. Any memory task involves some type of retrieval cue which is used to probe memory and retrieve an image. The retrieval cue is denoted by a finite vector v(t) and the output image retrieved by the cue is represented by the finite vector Y(t). Finally, the output mapping of the internal image Y(t) into an observable response R(t) in the experimenter's coordinates is symbolized by a monotonically increasing function J. The diagram below illustrates the relationship among the inputs and outputs. The square box represents the unobservable memory system which is described next. The two functions, V and J, represent nonlinear input and output response functions, respectively.




The general memory model that describes the dynamics of the memory system (the square box in above diagram) can be elegantly expressed by the discrete time state space representation of system theory (Csaki, 1977). The model is based on a system of three equations:

$$z(t)=\Theta[h(t)] \quad (1)$$

$$X(t+1)=\Psi[t,X(t),z(t)] \quad (2)$$

$$Y(t+1)=U[X(t+1),v(t+1)] \quad (3)$$

In the first equation, $\Theta$ specifies how category label features, g(t), and exemplar image features, f(t), are associated to produce a memory trace, z(t). In other words, the two types of information in h(t) are somehow combined or associated to form a single memory trace, which is subsequently fed into the memory system to preserve an experience. In general, the memory trace, z(t), is some matrix function $\Theta$ of h(t). We may interpret the matrix function $\Theta$ as the memory encoding process. Later, we will show how the precise form of $\Theta$ varies depending on each specific memory model. In the second equation, $\Psi$ is a matrix function that specifies how the memory system is organized and updated. In this sense, $\Psi$ may be interpreted as the learning process used to preserve an experience. Each input z(t) contributes to update the present state of knowledge, represented by the real valued

state matrix X(t). In the state space representation, the state matrix X(t) retains all the relevant information obtained from a sequence of exemplars presented up to trial t-1. Thus X(t) is interpreted as the memory of the system.

When subjects are asked to respond to the experimenter's instruction after observing a sequence of exemplar patterns, somehow they have to transform the internal state X(t) into a proper image for output. This process to build the retrieved image Y(t) from the preserved knowledge X(t) upon a given retrieval cue v(t) is characterized by a function, U. The function U can be interpreted as the memory retrieval process.

## DEFINITIONS OF THE TWO BASIC PROPERTIES

The two basic properties can be defined in terms of functional characteristics of the updating function $\Psi$ and the retrieval function U.

### Additivity

Additive systems are defined by Equation 5 below, which states that the retrieved image can be expressed as a weighted sum of the effects of each of the input memory traces. Equation 5 can be derived from two separate assumptions regarding the functions $\Psi$ and U. The first assumption is that $\Psi$ is a linear dynamic system:

$$X(t+1) = \Psi[t,X(t),z(t)] = \Phi(t)X(t) + H(t)z(t) \qquad (4)$$

where $\Phi(t)$ and H(t) are, in general, time dependent matrix functions, which can be interpreted as the system matrix and the weight matrix for new information, respectively.

The second assumption is that the retrieval function U is a linear transformation with respect to the first argument. Then we can derive the retrieved image,

$$Y(t)=U[\Phi(t-1)\cdot\cdot\Phi(0)X(0),v(t)] + \Sigma\, U[Q(t,k)H(k)z(k),v(t)]. \qquad (5)$$

Thus, assuming that both $\Psi$ and U are linear, one can express the retrieved image, Y(t), as a weighted sum of the effects of the memory traces z(k) for trials k = 1, ..., t-1 as in Equation 5.

### Time-invariance

Time-invariant systems are systems with updating functions, $\Psi$, that are not an explicit function of time coordinate, t:

$$X(t+1) = \Psi[X(t),z(t)] \qquad (6)$$

where $\Psi$ can be any linear or nonlinear function. If the system defined by Equation 6 is in the same state at two different points in time, and the same input is applied at these two time points, then the same output will be generated at these two time points. In other words, the system does not change solely as a function of time.

## MEMORY MODELS

In this section the three memory models will be briefly described and interpreted in terms of the general state-space model. More rigorous derivations will be given elsewhere (Myung & Busemeyer, manuscript under preparation).

### Multiple Trace Model

Hintzman's (1986) schema abstraction model assumes that each exemplar presentation produces a separate memory trace, a retrieval cue contacts all traces simultaneously, activating each according to its similarity to the cue, and information retrieved from memory reflects the summed content of all

Table 1: Characteristic functions in equations (1), (3), & (4) assumed by each memory model. The last column only applies to the blocked prototype production task.

| Memory Model | $\theta[h(t)]$ | $\Phi(t)$ | $H(t)$ | $U(X,v)$ | $w(t-k)$ |
|---|---|---|---|---|---|
| Multiple Trace | $1(t)h(t)'$ | $\alpha$ | $\gamma$ | $X'(Xv)^3$ | $[\gamma\alpha^{t-k}(g'g)]^3$ |
| Holographic Memory | $g(t)*f(t)$ | $\alpha$ | $\gamma$ | $v\#X$ | $\gamma\alpha^{t-k}$ |
| Hebb Rule | $g(t)f(t)'$ | $\alpha$ | $\gamma$ | $X'v$ | $\gamma\alpha^{t-k}(g'g)$ |
| Delta Rule | $g(t)f(t)'$ | $I-\gamma g(t)g(t)'$ | $\gamma$ | $X'v$ | $\gamma g'\Phi^{t-k}g$ |

activated traces responding in parallel.

This model can be represented by the general state-space model as follows. The state matrix X(t) would be a Nxp matrix with a very large N. The $\theta$ function is given by $\theta(h(t))=1(t)h(t)'$ where 1(t) is a Nx1 row vector with zeros on all locations except row t and an apostrophe represents the transpose of a vector and matrix. The state matrix X(t) is updated according to the following time-invariant linear system (X(0) = 0):

$$X(t+1)=\alpha X(t) + \gamma z(t) \tag{7}$$

where $\Phi(t)=\alpha > 0$ and $H(t)=\gamma > 0$ are scalars. Note that X(t) has nonzero elements only up to row t-1 and all zeros afterwards. Therefore, as shown in above equation, each exemplar h(t) is being separately preserved in the state matrix as a distinct row vector. The retrieved image Y(t) can be computed from the state matrix X(t) and the retrieval cue v(t) by the following function:

$$Y(t)=U[X(t),v(t)]=X(t)'[x(t)v(t)]^3 \tag{8}$$

where $A^n$ symbolizes the element-by-element power, $(A^n)_{ij}=(A_{ij})^n$. In general, the retrieval function in above equation is nonlinear for arbitrary Nxp matrices X. But U does satisfy linearity for the special form of X(t) defined in this model.

Holographic Memory Model

Metcalfe's (1982) holographic memory model (CHARM) is an associative memory model based on convolution and correlation algebra. The holographic memory model represents the interactive association between the category label and exemplar features, denoted g(t) and f(t), in the memory encoding step as the convolution of the two vectors, $z(t)=\theta(g(t),f(t))=g(t)*f(t)$.

The resulting memory trace z(t) is used to update the state vector, X(t), according to the same linear time-invariant system as Equation 7.

The retrieved image Y(t) is a correlation of the state matrix X(t) with the cue v(t),

$$Y(t)=U[X(t),v(t)]=v(t)\#X(t) \tag{9}$$

Note that the correlation operation '#' is a linear retrieval U function.

Parallel Distributed Memory Models

Parallel distributed memory models (Knapp & Anderson, 1984; McClelland

& Rumelhart, 1985), assume that each trial involves three events--first an input is presented to the memory system, this input generates an output, and finally this output is compared to a target as a desired output for that trial. Learning is viewed as a gradual change of connectivity strength among basic memory units.

In this model, the associative memory trace on trial t between the i-th input feature $g_i(t)$ and the j-th target feature $f_j(t)$ is the product of the two feature elements, $z_{ij}(t) = g_i(t)f_j(t)$. The collection of $z_{ij}(t)$'s forms a matrix $z(t) = \theta(g(t),f(t)) = g(t)f(t)'$. Then the memory trace $z(t)$ is used to update the state matrix, $X(t)$, called the connection matrix, which represents the present state of connection strengths between the i-th input feature and the j-th output feature. The connection matrix $X(t)$ as a state matrix is assumed to be updated according to either a Hebb rule or a delta rule and the image retrieved by a cue $v(t)$ is a matrix product of $X(t)$ and $v(t)$.

Table 1 summarizes the relations between each of the memory models and the general state-space model. As can be seen in the Table, all of the memory models does satisfy additivity and all but one (the delta rule) satisfy time-invariance. However, for the experimental procedure used in the present study the delta rule also obeys time-invariance (see next section).

## APPLICATION OF THE THREE MODELS TO THE PROTOTYPE PRODUCTION TASK

The experiment reported below used a blocked procedure in conjunction with the prototype production task. In the blocked procedure, subjects learn a sequence of exemplar images associated with a single category label within a block of trials, and after completing the block, they move to another block of trials with an unrelated category label. In this situation, the models are greatly simplified. Within each block of trials, the category label features $g(t)$ of the exemplar ensemble $h(t)$ are fixed, $h(t) = g(t) | f(t) = g | f(t)$. Furthermore, the retrieval cue is also fixed to the same category label within a block. Finally, the category labels across blocks are completely unrelated (i.e., orthogonal vectors). For this condition, it can be shown that all three models are consistent with the following special case of (assuming $X(0) = 0$ and $f(0) = 0$) :

$$Y_j(t+1) = \Sigma\, w(t-k)f_j(k), \quad \text{for} \quad k = 1, \ldots, t, \qquad (10)$$

where the weight $w(t-k)$ is a scalar function of the lag $(t-k)$ and is shown in Table 1 for each memory model.

As shown in Equation 10, the general definitions of two basic properties given in the earlier section can be reinterpreted in the present task in terms of the relationships between the input and retrieved feature vectors, $f(t)$ and $Y(t)$. This equation states that the exemplar image features from different trials are combined according to an additive composition rule to produce the prototype image. Time-invariance follows from the assumption that the weight $w(t-k)$ depends upon only on the lag or recency $(t-k)$ of the exemplar image.

## METHOD

The experiment was conducted on a microcomputer with all the procedures preprogrammed. The stimuli were mass spectra of fictitious chemical samples as shown in Figure 1, where chemical names correspond to category labels $g(t)$ and mass spectra correspond to exemplar patterns $f(t)$. For a given category label, subjects received four different exemplar patterns of the category and

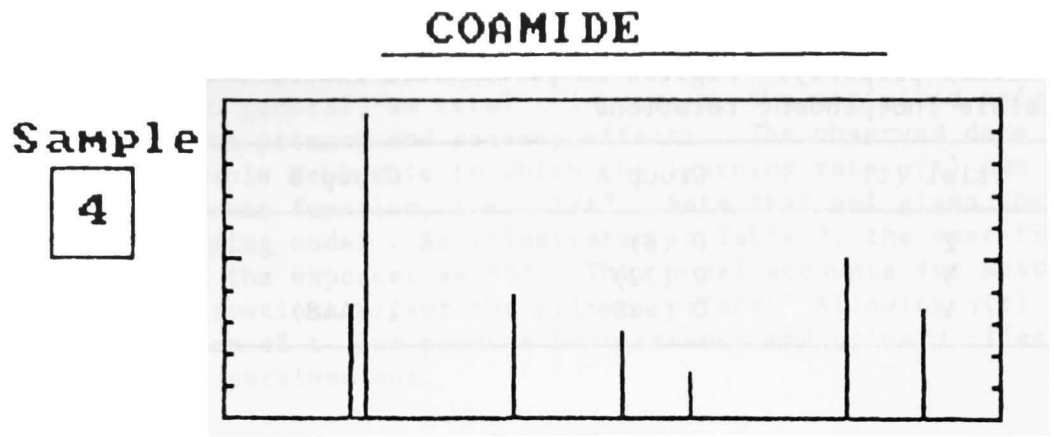



Figure 1. A typical stimulus pattern shown on a video screen.

they were asked to estimate the true pattern for each category based on the four distorted patterns. On each trial, subjects were shown stimulus patterns in the upper half of a video screen, then the pattern was erased and the subjects were asked to draw their estimate of a prototypic spectrum in the lower half of the same screen. After finishing the fourth pattern of a category, they moved to another four trials of a different category. There were two different groups of subjects. One group (Group A) was instructed to provide their estimate after each trial and the other group (Group B) was asked to provide a drawing only at the end of the fourth trial. Each subject received 100 categories (400 exemplar patterns). The subjects were 16 students attending Purdue University. Eight subjects were randomly assigned to each group.

## RESULTS

The following results were based on the observed responses averaged across category labels and eight subjects in each group.

### Test of Additivity

Additivity (Equation 5) was assessed by testing joint independence properties among patterns. Roberts (1973, p. 210) has described sufficient conditions for an additive system. However, joint-independence is the only property that is empirically testable in the present experiment. Thus, the following joint-independence condition was tested to support or refute additivity.

$$R_{pqrs} > R_{mnrs} \iff R_{pqol} > R_{mnol} \tag{11}$$

where $R_{pqrs}$ represents the prototype estimate after observing a sequence of exemplar patterns, $(P_p, P_q, P_r, P_s)$. Both the prototype estimate and exemplar pattern are recored as 7x1 column vectors where the j-th element is the height of the j-th vertical bar. This relationship should hold for all seven elements of the prototype estimate vector as well as for all trials of prototype production. Note that the test is relatively free of assumptions about how an exemplar pattern is transformed into the subject's internal memory representation. Therefore, additivity across exemplars was tested without mentioning anything about the internal state representation (i.e., the state vector, X(t)) that the memory system actually uses to encode the exemplar information. In this sense, the test of additivity can be considered

Table 2: Test of additivity by counting the number of violations of independence property. Figures in parenthesis indicate the total number of possible independent relations.

| Trial (t) | Group A | Group B |
|---|---|---|
| 2 | 0 (8) | - |
| 3 | 0 (72) | - |
| 4 | 0 (448) | 1 (448) |

a feature-free test.

All possible joint-independence relations were tested to assess additivity. The result is shown in Table 2, which summarizes the number of significant violations of the joint independence using the confidence interval of $\alpha$=.05 level. As can be seen in Table 2, no significant violations for Group A and only a single violation for Group B were observed. Considering the fact that the total number of independence relations was 528 for Group A and 448 for Group B, it can be concluded that additivity holds quite well for both conditions. The percentage of violations was still reasonably small even when zero confidence interval was used (5.3% for Group A and 10.7% for Group B).

Test of Time-invariance

Time-invariance was tested by fitting the following model:

$$R(t+1)=J[\Sigma\, w(t,k)f(k)] \quad \text{for } k=1, \ldots, t. \tag{12}$$

If time-invariance holds, then we should have $w(t,k)=w(t-k)$ for all t and k. Thus, the magnitude of the effect of each exemplar depends only on the lag, (t-k). Both the output response function J and the weights were estimated using a powerful estimation technique called the B-spline method (see DeBoor, 1978). The estimated response function J turned out to be a slightly nonlinear S shaped (not reported in this article). Table 3 contains the estimated weights for different t and k values.

Time-invariance implies that the weight w(t,k) should be solely a function of the lag (t-k), not depending upon the number of exemplars that subjects have seen, that is, trial t. As can be seen in Table 3, time-

Table 3: Test of time-invariance by estimating the weights w(t,k). Figures in parenthesis are predictions from the Hebb rule with time-variable parameters, $\alpha(t)=1-1/t^a$ & $\gamma(t)=1/t^a$ in Equation 7, where the least squares estimate of the exponent was a=.953.

| Group Condition | Trial (t) | Lag (t-k) 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|
| A | 2 | .49 (.52) | .48 (.48) | - | - |
| A | 3 | .41 (.35) | .30 (.34) | .29 (.31) | - |
| A | 4 | .28 (.27) | .24 (.26) | .22 (.25) | .26 (.23) |
| B | 4 | .26 (.27) | .23 (.26) | .23 (.25) | .26 (.23) |

invariance is clearly violated (for example, see the second column at the lag (t-k)=1). In general, as trial t increases, the estimated weights decrease with both primacy and recency effects. The observed data was fit with a time-variable Hebb rule in which the learning rate $\gamma(t)$ can vary according to a power function, i.e., $1/t^a$. Note that a=1 gives the simple arithmetic averaging model. As illustrated in Table 3, the best-fit model was the one with the exponent a=.953. This model accounts for most of qualitative observations except the primacy effect. Allowing $\gamma(t)$ to be an arbitrary function of t can produce both recency and primary effects though it would be less parsimonious.

## CONCLUSIONS

The goal of this study was to explore how abstraction occurs in human memory system. Specifically the present experiment was designed to empirically test two basic properties of prototype evolution using the prototype production paradigm-- additivity across exemplars and time-invariance of the memory system. The results indicate that additivity held reasonably well but time-invariance was clearly violated. The additivity result is somewhat surprising because it provides evidence for a linear dynamic memory system. It indicates that we don't have to resort to complex nonlinear dynamic models of memory for understanding prototype abstraction. The violation of time-invariance suggests that adaptive network models need to include a time-varying learning rate parameter of the form $1/t^a$ to simulate the abstraction process.

## ACKNOWLEDGEMENTS

This work was supported by NSF Grant BNS # 8710103.

## REFERENCES

Busemeyer, J. R., & Myung, I. J. (1988). A new method for investigating prototype learning. Journal of Experimental Psychology: Learning, Memory and Cognition, 14, 3-11.

Csaki, F. (1977). State-space Methods for Control System. Akademiai Kiado, Budapest.

DeBoor, C. (1978). A Practical Guide to Splines. New York: Springer-Verlag.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace model. Psychological Review, 93, 411-428.

Knapp, A. G., & Anderson, J. A. (1984). Theory of categorization based on distributed memory storage. Journal of Experimental Psychology: General, 10, 616-637.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. Journal of Experimental Psychology: General, 114, 159-188.

Metcalfe, J. (1982). A composite holographic associative recall model. Psychological Review, 89, 627-661.

Myung, I. J., & Busemeyer, J. R. (1989). A general theory of prototype learning and experimental test of ordinal properties. (manuscript under preparation)

Roberts, F. S. (1979). Measurement theory. Readings, M.A.: Addison-Wesley.