Journal of Mathematical Psychology 113 (2023) 102755

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

# Length of the state trace: A method for partitioning model complexity

F. Gregory Ashby

*University of California, Santa Barbara, United States of America*

ARTICLE INFO

*Article history:*
Received 1 July 2022
Received in revised form 31 January 2023
Accepted 8 February 2023
Available online 20 February 2023

*Keywords:*
Model complexity
State-trace analysis
Signal-detection theory
Generalized context model
Prototype model

ABSTRACT

A novel and easy-to-compute measure is proposed that compares the relative contribution of each parameter of a mathematical model to the model's mathematical flexibility or complexity, with respect to accounting for the results of some specific experiment. When the data space is a two-dimensional plot of the type used in standard state-trace analysis, then the model complexity contributed by a single parameter equals the length of the state trace (LOST) that results when that parameter is varied and all other parameters are held constant. For the normal, equal-variance, signal-detection model, the average LOST when the response-criterion parameter $X_C$ is varied is about four times greater than the average LOST when the sensitivity parameter $d'$ is varied. As a result, applying the signal-detection model to random data almost always leads to the conclusion that all the points share the same value of $d'$ but were generated under different values of $X_C$. Parameters that have non-monotonic effects on performance, such as the attention-weight parameter that is used in popular exemplar and prototype models of categorization, tend to have large LOSTs, and therefore contribute to model flexibility more than parameters that have monotonic effects on performance. Comparing LOSTs for exemplar and prototype models also leads to some deep new insights into the structure of both models.



## 1. Introduction

This article describes a novel, and exceedingly simple method for investigating the structure of a mathematical model, and specifically for determining how much each parameter of the model contributes to the model's mathematical flexibility or complexity. The popular model-selection statistics AIC and BIC assume that all parameters contribute equally to a model's complexity, but this is far from true. For example, as the new method will show, in the normal, equal-variance model of signal-detection theory, the contribution of the criterion $X_C$ to model complexity is about four times greater than the contribution of the sensitivity parameter $d'$. It is critical to understand how different parameters contribute to complexity when interpreting results of model fitting. For example, as data become noisier, it becomes more and more likely that a version of the model in which a high-complexity parameter is varied will fit better than a version in which a low-complexity parameter is varied – regardless of the psychological structure of that data.

This article is organized as follows. The next section briefly reviews the literature on model complexity and describes the new proposed measure. The third section applies the new method to the standard normal, equal-variance, signal-detection model, Section 4 considers the effects on model complexity of parameters that predict non-monotonic changes in performance, and Section 5 applies the method to exemplar and prototype models of categorization. Finally, the last section closes with a general discussion and conclusions.

## 2. Model complexity

Consider an experiment with $N_t$ trials or observations in which the results are described by recording $N_d$ data values or dependent variables. If each recorded value describes the outcome of a single trial or observation then $N_d = N_t$, whereas if our $N_d$ recorded values are summary statistics then $N_d < N_t$. Now consider a space with one dimension for every one of these dependent variables. This is the experiment's *data space* **D**, and any point in **D** can be indexed by the ordered $N_d$-tuple $\underline{\boldsymbol{d}} = [DV_1, DV_2, \ldots, DV_{N_d}]$, where $DV_i$ denotes the $i$th of the dependent variables and $N_d \leq N_t$. If $N_d = N_t$, we say that the data space is *saturated*. Note that the outcome of our experiment is represented by one single point in saturated data space, and collectively, all the points in saturated data space represent all possible outcomes of our experiment.

Next, consider some mathematical model of this experiment. Suppose the model includes $r$ free parameters $\theta_1, \theta_2, \ldots, \theta_r$. The space of all possible values of these parameters, $\boldsymbol{\Theta}$, defines the model's *parameter space*, and any point in $\boldsymbol{\Theta}$ can be indexed by the ordered $r$-tuple $\underline{\boldsymbol{\theta}} = [\theta_1, \theta_2, \ldots, \theta_r]$. Any specific combination of numerical values chosen for these $r$ parameters is represented by one single point in parameter space.

*E-mail address:* fgashby@ucsb.edu.

https://doi.org/10.1016/j.jmp.2023.102755

Note that if specific numerical values for each parameter are inserted into the model's equations, the model will make specific numerical predictions about the outcome of the experiment. If the data space is saturated, then most mathematical models in psychology will not predict the actual outcome of each of the $N_t$ trials or observations in the experiment, but instead will predict a probability distribution over the possible outcomes on each trial. For example, the model might predict that the response time on trial $i$ is a random sample from an ex-Gaussian distribution with mean $\mu_i$, variance $\sigma_i^2$, and rate $\lambda_i$. When the data space is saturated, the set of all possible model predictions defines the model's *statistical manifold* **M**.[1]

On the other hand, if the data space is limited to summary statistics, the model might predict exact values for each of the $N_d$ dependent variables. For example, for any given value of $d'$ and $X_C$, the normal, equal-variance, signal-detection model predicts single values for the observed proportion of hits and false alarms. In this case, note that the space of all possible model predictions has the same structure as data space – that is, for any point in parameter space, the model predicts a single numerical value for each of the $N_d$ dependent variables that define data space (rather than a probability distribution). As a result, the model's equations could be interpreted as mapping the point in parameter space that indexes the set of parameter values that we chose to a point in data space. If that point in data space corresponds to exact values of all our recorded summary statistics, then the model provides a perfect fit to those statistics. Note that if we change the numerical values of any parameters, the predictions of the model will change. The model's equations therefore map different points in parameter space to different points in the data space defined by our recorded summary statistics. Finally, suppose we identify all points in this data space that result from systematically iterating through all possible values of all $r$ parameters. This is the model's image in data space and the points in this image denote the set of all possible experimental outcomes that the model can fit perfectly. When $N_d > r$; that is, when there are more summary statistics than free parameters, then we expect that there will be possible experimental outcomes that the model cannot perfectly fit. In these cases, the image of the model is a proper subset of **D** and it defines the model's *topological manifold* (called the model manifold by Ashby & Bamber, 2022).[2]

Many mathematical models in psychology are fit to and tested against summary statistics, rather than single trial-by-trial response data. For example, this is true in almost any application in which parameter estimates are obtained via the method of least squares. More importantly for the present purposes however, summary statistics, rather than trial-by-trial data, are typically used to guide model design and development. For example, the normal, unequal-variance, signal-detection model was originally developed because of the observation that empirical ROC curves are skewed rather than symmetric, and not because of any consideration of some possible sequence of YES and NO responses a participant might make in a YES-NO detection task. Similarly, a choice of what response-time model to adopt might depend on whether correct mean response times are expected to be faster or slower than incorrect mean response times, rather than on a consideration of specific response times that might be observed over some possible sequence of trials. The goal of this article is to provide a new model-development tool (rather than a new model-selection statistic), and for this reason, this article focuses on applications in which each dependent variable that defines the data space is a summary statistic. In other words, unless specifically noted otherwise, by data space I will mean a space in which the $N_d$ dimensions denote the values of our $N_d$ recorded summary statistics, and the possible outcomes of the experiment are the possible values of these $N_d$ summary statistics.

If two competing models provide equally good fits to a data set, then a basic principle of model selection is to favor the less mathematically flexible model over the more flexible model. The basic idea is that a rigid model makes strong predictions, whereas a flexible model makes weak predictions. The rigid model is stating that only a few possible outcomes of the experiment are possible, whereas the flexible model is stating that many outcomes are possible. So, if one of the few possible outcomes predicted by the rigid model actually occurs, then that model should be given credit for correctly predicting this outcome. In the model selection literature, the mathematical flexibility of a model is known as its complexity, so in the remainder of this article I will use the term complexity to mean mathematical flexibility.

Model complexity has been studied most extensively within the statistical field of model selection. The goal here is to identify, among a set of competing models, the one model that provides the most parsimonious account of an existing data set. In other words, during model selection, one begins with a single observed point in saturated data space and a set of alternative models of the experiment, and the goal is to identify the single model in this set that provides the best account of that single point in data space. The goal of this article is very different. Specifically, the goal here is to develop an easy-to-use tool for assessing how strong or weak a model's a priori predictions are for the chosen experiment, and specifically, in how much each model parameter contributes to the complexity of these predictions. Therefore, in contrast to model selection, no single point in data space will be privileged and only one model at a time will be considered. As a result, these goals require a measure of a model's complexity that is defined relative to the experiment, but does not depend on any observed outcome of that experiment, and that is easy to compute.

Within the field of model selection, many different measures of model complexity have been proposed. These are usually applied as penalties to some goodness-of-fit statistic, such as minus log likelihood. One option therefore, is to strip the complexity term off of one of these goodness-of-fit measures and use this term to examine the relative contribution of different parameters to the model's overall complexity. Unfortunately, this approach fails for a variety of reasons. The most widely known statistics that penalize for model complexity are AIC and BIC. However, both of these define model complexity solely in terms of the number of the model's free parameters, and therefore they both assume that all parameters contribute equally to complexity. Adding more parameters to a model will almost always increase its complexity, but it is well known that models with the same number of free parameters are not necessarily equally complex.

A variety of other goodness-of-fit statistics – more sophisticated than AIC or BIC – include complexity terms that apply different penalties to different parameters. Furthermore, many of these have the attractive property that they are sensitive to experimental design (e.g., Pitt, Myung, & Zhang, 2002). This seems promising, but there are problems here too. For example, several of these require the model's Fisher information matrix. This includes, for example, stochastic complexity measures that arise as asymptotic expansions of the complexity terms from goodness-of-fit statistics used in normalized maximum likelihood and Bayesian model selection (Myung, Navarro, & Pitt, 2006).

[1] A statistical manifold is a Riemannian manifold in which the metric is the Fisher information metric.

[2] Technically, to be a topological manifold, the model's equations must be one-to-one and continuous, and the inverse of the equations must also be continuous (i.e., the mapping from the model's image in data space back to parameter space). However, almost all mathematical models within psychology satisfy these conditions.

Unfortunately, the Fisher information matrix is analytically unavailable for most models in psychology, and as a result, applications that use these measures typically construct the matrix from estimates of the variance of various partial derivatives of the log-likelihood function evaluated at the maximum likelihood estimates. As a result, this approach is inappropriate for the present purposes because it depends on the single actual outcome that occurred in the experiment. Similarly, the complexity measure in the negative free energy statistic requires the variance–covariance matrix of the maximum likelihood estimates (Ashby, 2019), which also depends on the outcome of the experiment. Other sophisticated complexity measures, which are sensitive to experimental design and do not depend on any single experimental outcome, fail to meet the easy-to-compute criterion. For example, the complexity measure in one form of normalized maximum likelihood (called stochastic complexity 1 by Myung et al., 2006) is computed by taking the logarithm of the sum of the maximum likelihood fits that would occur for all possible outcomes of the experiment. In other words, for every point in saturated data space, (1) maximum likelihood estimates of all model parameters are computed, (2) these are used to compute a goodness-of-fit score if that data outcome actually occurred in an experiment, (3) these goodness-of-fit scores are all added together, and (4) model complexity is defined as the log of this sum.

An alternative approach comes from the field of information geometry, which defines the complexity of a model by the volume of its statistical manifold (Amari, 2016). Recall that the points of a model's statistical manifold are probability distributions, and that the distances needed to compute volume are defined by the Fisher information metric, which in effect defines the distance between two probability distributions by the amount of information lost if one distribution is replaced by the other (Amari, 2016). If one model tends to predict that only a few possible outcomes of our experiment are possible, then all predicted probability distributions will be similar, and therefore close together in the model's statistical manifold according to the Fisher metric, with the result that the volume of the model's statistical manifold will be small. Information geometry judges this model to have low complexity. In contrast, if another model predicts that many different outcomes of the experiment are possible, then many different probability distributions will be predicted, and therefore some will be far apart in the statistical manifold, with the result that the volume of the model's statistical manifold will be large. Information geometry judges this model to have high complexity. The volume of the model's statistical manifold will almost always increase with the addition of a new parameter, but models with the same number of parameters are not typically associated with the same volumes. For example, a two-parameter power function model is more complex than a two-parameter log function model (Myung, Balasubramanian, & Pitt, 2000; i.e., because a power function is more flexible or "bendy" than a log function).

Despite its intuitive appeal, this measure of complexity has not become popular in psychology for at least three reasons. First, computing the volume of the model's statistical manifold for anything but the simplest possible models is an exceedingly difficult and tedious computational challenge. Second, statistical manifolds are usually impossible to visualize because of their high dimensionality and because their points are probability distributions. Third, the volume of a model's statistical manifold is a property of the whole model and does not allow one to examine the contribution of individual parameters to this volume.

All three of these problems disappear when the dependent variables that define the data space are just a few summary statistics, and especially when this number is limited to $N_d = 2$. In this case, the data space is two dimensional, and the model predicts a single numerical value for each dependent variable, rather than a probability distribution. Thus, the model's statistical manifold reduces to a topological manifold embedded in a two-dimensional data space and therefore is easy to visualize. In addition, the volume is straightforward to compute since the relevant metric is just familiar Euclidean distance. In fact, this is exactly the model complexity measure proposed by Veksler, Myers, and Gluck (2015). Even so, their goals were quite different from the goals of this article. They made no attempt to examine how different model parameters contributed to this volume. Instead, their focus was on computing the proportion of the volume of data space covered by the model manifold, and then interpreting this proportion as a type of *p*-value for the null hypothesis that a good fit of the model to the data is because of chance. Note that this use of the model's volume requires the extra assumption that all data outcomes (i.e., all points in data space) are equally likely.

The present goal is instead to examine how each parameter in a model contributes to the model's overall complexity. As we will shortly see, when only one of the model's parameters are varied and all others are held constant, then the model's topological manifold is a one-dimensional curve, and therefore the volume of the manifold equals the length of that curve, which is simple to compute. Therefore the goal here is to compare these lengths for different model parameters as a method to understand how each parameter contributes to model complexity.

It might seem that these conditions – namely, that there are only two dependent variables and both are summary statistics – are too restrictive to be of much interest. However, it turns out that such cases are quite popular in psychology and have been so for many decades. For example, these are exactly the conditions required for *state-trace analysis* (STA; Ashby & Bamber, 2022; Bamber, 1979).

STA is a method for determining the complexity of a set of data in which one or more independent variables are manipulated across experiments or conditions and, most typically, two separate dependent variables are measured (e.g., performance in two tasks). A state-trace plot is a graph that plots performance on two dependent variables against each other as some independent variable(s) or model parameter(s) changes. The dependent variables may come from the same or different tasks. For example, an STA could be performed on an ROC curve, which plots the probability of a hit against the probability of a false alarm from a YES-NO detection task. Alternatively, the STA could be performed on data collected from two different categorization tasks, where the two dependent variables are the proportion of correct responses in each task. In other words, a state-trace plot is a data space with multiple points in which the number of dimensions $N_d = 2$.

Ashby and Bamber (2022) showed that when the dimensionality of data space is at least as large as the number of free parameters (i.e., $N_d \geq r$) then the model's topological manifold is $r$-dimensional. Therefore, a model in which a single parameter varies always predicts a one-dimensional state-trace plot (Bamber, 1979). As a result, the volume of the model's topological manifold in this case equals the length of its state trace (LOST). This volume – that is, the LOST – does not suffer from any of the problems associated with the information geometric complexity measure (i.e., the volume of the model's statistical manifold), because the LOST is easy to compute, state traces are simple to visualize, and it is straightforward to compute LOSTs separately for each parameter of a model.

The LOST associated with the parameter $\theta_i$ can be quickly computed as follows. Suppose our state-trace curve plots values of $DV_1$ against values of $DV_2$, where $DV_1$ and $DV_2$ are two summary statistics that describe the outcome of our experiment. The first

step is to compute the state-trace curve predicted by the model when $\theta_i$ varies and all other parameters are held constant at some fixed values. In practice, all such state-traces are computed by selecting an ordered set of $M$ different values of $\theta_i$, denoted by $\theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,M}$, and then computing the $M$ predicted ordered pairs $[DV_1(\theta_{i,j}), DV_2(\theta_{i,j})]$ for $j = 1...M$ that result when each of these values of $\theta_i$ are used to generate model predictions. Given this discrete state-trace curve, the LOST for $\theta_i$ is approximately

$$LOST(\theta_i) \doteq \sum_{j=1}^{M-1} \sqrt{\left[DV_1(\theta_{i,j+1}) - DV_1(\theta_{i,j})\right]^2 + \left[DV_2(\theta_{i,j+1}) - DV_2(\theta_{i,j})\right]^2}, \quad (1)$$

and the approximation improves as $M$ increases. In other words, to compute the LOST, we compute the Euclidean distance between each successive pair of points that define the state trace, and then add all these distances together.

This article explores the benefits of comparing LOSTs for different parameters of the same model. We will see that even the most common models include parameters that differ greatly in their LOSTs, and that these disparities have profound implications for the results of fitting the model to noisy data. In particular, the version of the model in which the free parameter is the one with the greatest LOSTs is highly likely to fit noisy data better than versions of the model in which the free parameter has smaller LOSTs, regardless of the psychological structure of the data. As a result, knowing the relative LOSTs of the various model parameters can greatly benefit the interpretation of goodness-of-fit analyses.

## 3. Signal-detection theory

We begin with one of the most widely used models in all of psychology – namely, the standard normal, equal-variance model from signal-detection theory. In this case, the obvious state-trace plots are the standard ROC curves that plot the probability of a hit [$P$(H)] against the probability of a false alarm [$P$(FA)]. It is well known that the normal, equal-variance, signal-detection model can fit any single observed combination of $P$(H) and $P$(FA) values perfectly (e.g., Ashby & Wenger, 2023). As a result, an exceedingly common analysis, used in scores of studies dating back at least to the seminal work of Green and Swets (1966), is to attempt to account for a scatter plot of [$P$(FA), $P$(H)] pairs by assuming that all pairs were generated by a version of the model in which one of the two model parameters varies – either the response criterion $X_C$ or the sensitivity $d'$ – and the other parameter remains fixed at some constant value. The most common version of the curves that result from this approach is the isosensitivity curve, in which $X_C$ is varied while $d'$ is held constant. The less common alternative is an isobias curve that is generated by varying $d'$ while holding $X_C$ at some fixed constant value. Examples of both types of curves are shown in Fig. 1, along with the lengths of each state-trace and the means of these lengths.

Note that the lengths of the iso-sensitivity curves are, on average, almost four times longer than the lengths of the isobias curves. As a result, most of the complexity of the normal, equal-variance model of signal-detection theory comes from the criterion parameter $X_C$ – much more so than from $d'$.

Fig. 1 shows that varying $X_C$ contributes to the complexity of the signal-detection model much more than varying $d'$ because all isosensitivity curves vary between the points (0,0) (when $X_C = \infty$) and (1,1) (when $X_C = -\infty$). The shortest path between these two points, which occurs when $d' = 0$, has a length of $\sqrt{2}$ and the longest path, which occurs when $d' = \infty$ has a length of 2. In contrast, all isobias curves are vertical line segments that begin on the major diagonal (when $d' = 0$) and end at $P$(H) = 1 (when $d' = \infty$). So note that the shortest isobias curve has length 0 and the longest has length 1. As a result, isosensitivity curves are substantially longer than isobias curves.

**Table 1**
Simulation results (i.e., proportion of 100,000 simulations in which each model provided the better fit).

| Model | 5 Data points | 10 Data points |
|---|---|---|
| $X_C$ varies across conditions | .857 | .971 |
| $d'$ varies across conditions | .143 | .029 |

The fact that isosensitivity curves are longer than isobias curves suggests that a set of randomly generated points are collectively all more likely to fall near some single isosensitivity curve than near some single isobias curve. For example, consider an empirical ROC with multiple points that were generated under different experimental conditions, or perhaps by different participants. Suppose that we fit two different versions of the normal, equal-variance, signal detection model to these data. In one version, $X_C$ varies across points but $d'$ remains constant, whereas in the second version $d'$ varies across points but $X_C$ remains constant. In other words, in both cases, we are fitting a model with one free parameter to the data. If the points were randomly generated, or if they are characterized by high levels of noise, then the much longer LOSTs associated with $X_C$ predict that the version with the free $X_C$ parameter should be much more likely to provide the better fit than the version with the free $d'$ parameter.

To test this prediction, I simulated 100,000 replications of two different experiments – one with 5 separate conditions (or participants) and one with 10. In both experiments, the data from each condition were an ordered pair [$\hat{P}$(FA), $\hat{P}$(H)] that was generated by randomly sampling a point (from a uniform distribution) in the upper left-half region of ROC space – that is, in the region corresponding to $d' \geq 0$. The random sampling produced a scatter plot of either 5 or 10 points (in simulated Experiments 1 and 2, respectively), all that fell on or above and to the left of the main diagonal in ROC space. Next, I fit two normal, equal-variance, signal-detection models to each scatter plot. Both models had one free parameter. One model assumed that all conditions shared the same $d'$ but differed in their value of $X_C$. So to fit this model, I estimated the single value of $d'$ that produced the isosensitivity curve that best fit the data. The second model assumed that all conditions shared the same value of $X_C$ but differed in the value of $d'$. To fit this model, I estimated the single value of $X_C$ that produced the isobias curve that best fit the data.

I used the method of least squares for parameter estimation and sum of squared errors (SSEs) to evaluate goodness-of-fit. The SSEs can be directly compared since both models have the same number of free parameters (i.e., one), so the best model is the one with the smallest SSE. The whole process was repeated 100,000 times for each experiment. The results are shown in Table 1. Note that, as predicted, the model that assumed all conditions shared the same $d'$ but had different values of $X_C$ was the clear winner, and its dominance increased with the number of data points. With 10 randomly generated data points, this model fit better than the model that assumed all conditions shared the same $X_C$ but had different values of $d'$ more than 97% of the time.[3]

The results summarized in Table 1 assume no limits on the estimated values of either $X_C$ or $d'$. In practice, extreme values of $X_C$ are rarely questioned, whereas extreme values of $d'$ are often

[3] As the number of random data points increases, it becomes less and less likely that they will all have an approximate vertical alignment of the type needed for the $d'$-varying model to outperform the $X_C$-varying model.

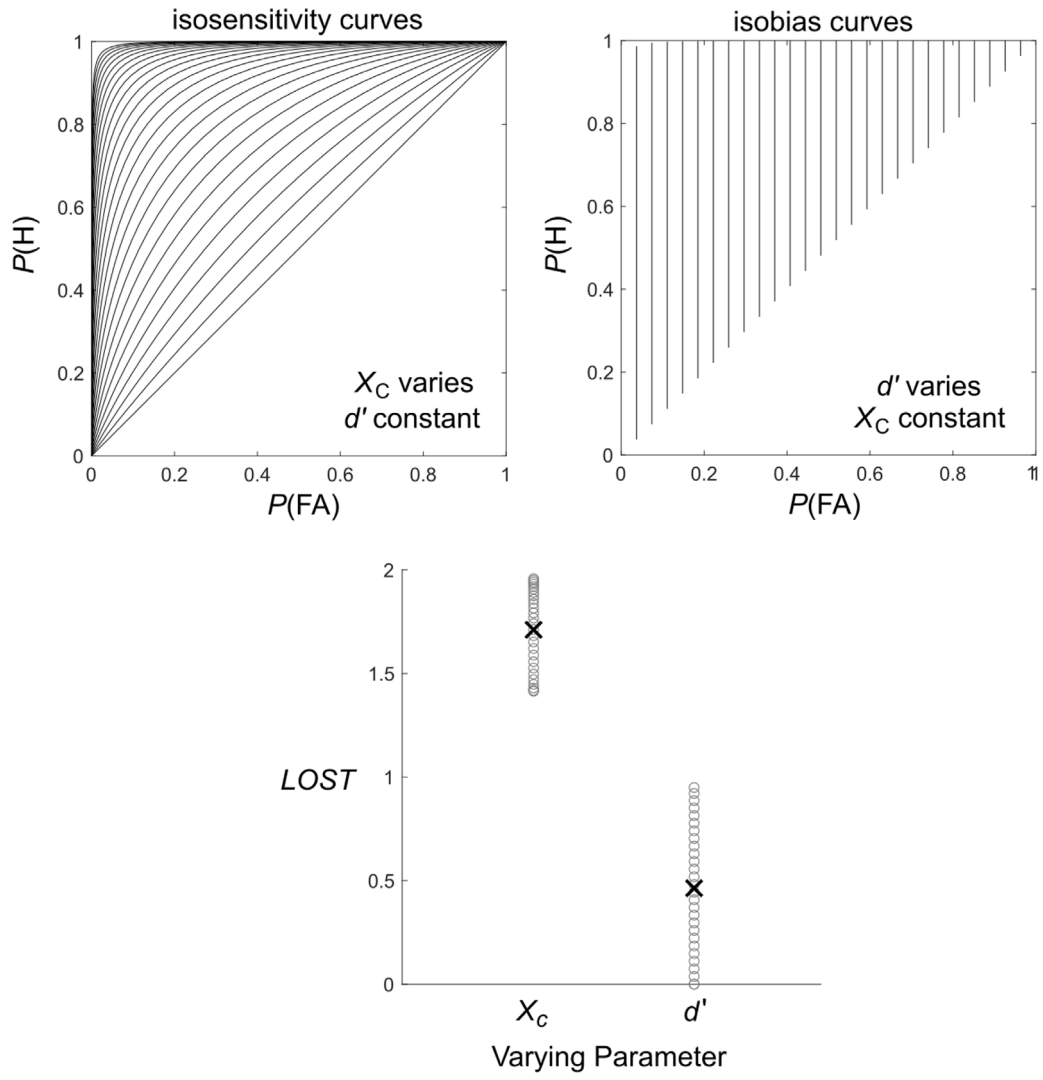


**Fig. 1.** The top two panels show ROC curves predicted by the normal, equal-variance, signal-detection model. The top left panel shows 27 isosensitivity curves that result from varying $X_C$ (from $-8$ to +8, in increments of 0.1). Each curve has a different fixed value of $d'$ (from 0 to 4, in increments of 0.15). The top right panel shows isobias curves that result from varying $d'$ (from 0 to 8, in increments of 0.1) for 28 different fixed values of $X_C$ (from $-\infty$ to $+\infty$, chosen to partition the standard normal distribution into equal areas). The bottom panel plots the lengths of the various ROCs from each panel and the X shows the mean length for each parameter.

viewed skeptically. For example, in most applications of signal-detection theory, experimental conditions are arranged so that errors are expected. As a result, any analysis of data collected from such an experiment that reported $d'$ estimates greater than 3 or 4 would be viewed suspiciously. In contrast, any estimated value of $X_C$ could be plausible. A Bayesian approach would implement these beliefs by placing a narrower prior distribution on the $d'$ parameter than on $X_C$. The Discussion section describes a formal method of incorporating prior distributions into the LOST computation.

Limiting the upper value of $d'$ reduces the mean *LOST*s of both isosensitivity and isobias curves. Even so, the effect is greater on isobias curves than on isosensitivity curves. Fig. 2 plots the ratio

$$\frac{\text{mean}(LOST_{\text{isosensitivity}})}{\text{mean}(LOST_{\text{isobias}})} \tag{2}$$

for different upper limits on the allowable value of $d'$. Note that this ratio drops sharply as the maximum allowable value of $d'$ increases. The limit as the maximum $d'$ approaches 0 is infinity, and the curve asymptotes at 4 for large values of the maximum.[4] Table 1 assumes no upper limit on $d'$. Fig. 2 suggests that if an upper limit much less than 2 is placed on $d'$ then the results of Table 1 would be even more extreme, in the sense that, under the same conditions, the conclusion that $X_C$ was varying across conditions, rather than $d'$, would be even more likely.[5]

These results suggest that a conclusion that the response criterion varies across conditions while sensitivity remains constant must be interpreted cautiously if based solely on goodness-of-fit. This section showed that such a conclusion is virtually inevitable in randomly generated data.

[4] When $d' = 0$, the $LOST_{\text{isosensitivity}} = \sqrt{2}$, whereas $LOST_{\text{isobias}} = 0$. In contrast, when $d' = \infty$, the $LOST_{\text{isosensitivity}} = 2$, and mean$LOST_{\text{isobias}} = .5$.

[5] Thanks to Donald Bamber for pointing out this consequence of limiting the range of $d'$.

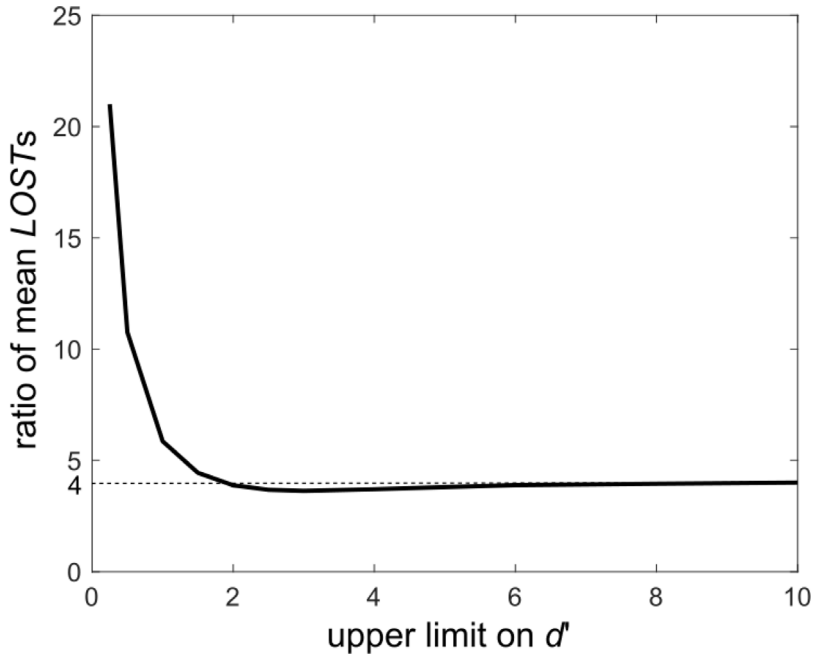


**Fig. 2.** The ratio of mean *LOST*s as a function of the maximum possible value of $d'$ [i.e., mean($LOST_{\text{isosensitivity}}$)/mean($LOST_{\text{isobias}}$)].

## 4. Parameters that predict non-monotonic changes in performance

The parameter $X_C$ contributes to the complexity of the signal-detection model much more than the $d'$ parameter because isosensitivity curves fill more of the ROC space than isobias curves (i.e., the former are longer than the latter). Note that the ROC space is constrained to the unit square since the values on each dimension are constrained to the interval [0,1]. In such constrained spaces, curved state traces will tend to be longer than linear state traces. In fact, for any two state traces that begin and end at the same two points, the one with more curvature will tend to be longer.

Technically, it is possible for a state trace produced by varying a single parameter to have enough curvature to fill an entire two-dimensional state-trace space (Ashby & Bamber, 2022). With more common models, such as the normal, equal-variance, signal-detection model, this requires two free parameters that simultaneously vary (Ashby & Bamber, 2022). For example, as previously noted, by allowing both $X_C$ and $d'$ to vary, the signal-detection model can perfectly fit any observed [$P$(FA), $P$(H)] pair. According to information geometry, a model with a single free parameter that fills the entire state-trace space has the same complexity as a two-parameter model that also fills the whole space.

Although I know of no current models in the psychological literature that include any single parameter that allows the model to fit any point in a two-dimensional state-trace plot perfectly, such models are mathematically possible. In particular, as noted by Ashby and Bamber (2022), it is possible to construct single-parameter, well-behaved models[6] that produce state-trace curves that, in practice, would be statistically impossible to distinguish from a two-parameter model that fills the whole space. Most space-filling curves that have been proposed are constructed by taking the limit of a sequence of simpler curves, each of which is a one-to-one mapping from the unit interval to the unit square, with the property that each successive curve in the sequence more closely approximates the area-filling limit. So a model in which varying one parameter produces a curve that is late in such a sequence, but before the limit, will fill much of the state-trace plot. Most importantly, because of statistical error (measurement, perceptual, cognitive, or individual difference), such a model would be impossible to discriminate from a model that fills the whole space.

Although no current models include any single parameter that produces space-filling predictions, there are many models, besides the normal, equal-variance, signal-detection model, with parameters that predict curved state traces, and as we have seen, the more curvature they predict, the longer the resulting state traces, and therefore, the more model complexity added by that parameter. One way to produce high levels of curvature is via parameters that when increased, have non-monotonic effects on predicted performance. If increases in a parameter always cause the model to predict monotonic changes in performance, then the state-trace plot is guaranteed to be monotonic (Ashby & Bamber, 2022; Dunn & Kirsner, 1988). Such a plot is limited in its maximum length by the limits on each dependent measure. On the other hand, if increases in a parameter cause the model to predict non-monotonic changes in performance (e.g., increasing then decreasing) in one of the two dependent variables used to construct the state-trace plot, but not both, then the resulting state-trace plot is guaranteed to be a single, non-monotonic curve (Ashby & Bamber, 2022; Bamber, 1979). In general, non-monotonic state trace curves will display more curvature than monotonic curves, and therefore tend to have greater length.

Many popular models include parameters that when increased have non-monotonic effects on predicted performance. First, many models include a selective-attention parameter that is defined as the proportion of attention allocated to one of two stimulus dimensions. In tasks where both dimensions carry diagnostic information, such models predict improvements in performance with increases in the parameter up until attention is divided optimally between the two dimensions and then decreases in performance as the parameter increases beyond this point. In contrast, in tasks where only one dimension is relevant, the models predict monotonic changes in performance as the

[6] By well behaved, I mean models in which the mapping from parameter space to data space is one-to-one and continuous and with a continuous inverse mapping (Ashby & Bamber, 2022).

parameter is increased. For example, prototype, exemplar, and decision-bound models of categorization all include such a parameter (e.g., see Ashby & Maddox, 1993). Second, decision bound models predict that accuracy is maximized in any categorization task when the decision bound has some specific intermediate intercept and curvature. Thus, decision bound models predict that accuracy will increase to some peak value and then decrease as the intercept increases from $-\infty$ to $\infty$. A similar prediction occurs for the amount that the decision bound curves (e.g., from negative to positive). Third, the COVIS rule-learning model predicts that accuracy will increase with the parameter that is sensitive to brain dopamine levels and then decrease when these levels pass an optimal value (Ashby, Paul, & Maddox, 2011). Fourth, all connectionist and neural network models predict that for any given amount of training, accuracy increases with the value of the learning-rate parameter up to a point, and then performance will deteriorate if the learning rate becomes too large. Thus, if any one of these parameters vary across one of the tasks or conditions, then the resulting state-trace plot could be a single, non-monotonic curve.

In summary, many models include parameters that when increased, predict non-monotonic changes in performance, and such models will often predict non-monotonic state-trace curves. Non-monotonic state-trace curves tend to be longer than monotonic curves. Therefore, in this class of models, the parameter predicting non-monotonic changes in performance should contribute to model complexity more than parameters predicting monotonic changes.

## 5. Exemplar and prototype models of categorization

To test the prediction that parameters predicting non-monotonic changes in performance should contribute more to model complexity than parameters predicting monotonic changes in performance, I computed the LOSTs for the three most commonly used parameters of the popular generalized context model of categorization (GCM; Nosofsky, 1986) and for a prototype model that had the identical mathematical structure (e.g., Homa, Sterling, & Trepel, 1981; Smith, 2002; Smith & Minda, 2001).

Both models were applied to the two categorization tasks shown in Fig. 3. In the rule-based (RB) task, perfect performance requires deciding whether the value of the stimulus on dimension 1 is small or large, while the value of the stimulus on dimension 2 is irrelevant. In the information-integration (II) task, both dimensions are equally important to the categorization decision. The state-trace analysis from these tasks plots the predicted probability of responding correctly on the II task (on the ordinate) against the predicted probability correct on the RB task (on the abscissa). Many studies have examined state-trace plots of exactly this type, and the goal was typically to decide whether the resulting plot was consistent with a model in which only one parameter was varying (e.g., Ashby, 2014, 2019; Ashby & Bamber, 2022; Dunn, Newell, & Kalish, 2012; Newell, Dunn, & Kalish, 2010; Stephens, Matzke, & Hayes, 2019).

The GCM is an exemplar model because it assumes that stored representations of all category exemplars contribute to the categorization decision. In a task with two categories, A and B, the GCM assumes that the probability of responding A on a trial when stimulus $k$ is presented equals

$$P(\mathrm{A}|k) = \frac{\beta \sum_{i \in C_A} \eta_{ik}}{\beta \sum_{i \in C_A} \eta_{ik} + (1-\beta) \sum_{i \in C_B} \eta_{ik}}, \tag{3}$$

where $C_A$ and $C_B$ are sets containing the stimuli in categories A and B, respectively, $\eta_{ik}$ is the similarity between stimuli $i$ and $k$, and $\beta$ is a parameter that reflects the participant's bias toward responding A. Similarity is assumed to be inversely related to the weighted Euclidean distance between the perceptual representations of the stimuli. More specifically, the distance between the perceptual representations of stimuli $i$ and $k$, denoted $\delta_{ik}$, equals:

$$\delta_{ik} = \sqrt{w(x_{i1} - x_{k1})^2 + (1-w)(x_{i2} - x_{k2})^2}, \tag{4}$$

where $w$ is the proportion of attention allocated to dimension 1, and $x_{ij}$ is the coordinate value of stimulus $i$ on the $j$th perceptual dimension. Similarity is inversely related to distance via:

$$\eta_{ik} = \exp(-c\delta_{ik}^2) \tag{5}$$

where $c$ is a parameter that increases with the overall discriminability of the stimuli.[7]

The GCM predicts that accuracy is maximized in the Fig. 3 RB task if the proportion of attention allocated to dimension 1 – that is $w$ – equals 1. As a result, it predicts that accuracy will increase monotonically as $w$ increases from 0 to 1. In contrast, in the II task, the GCM predicts that accuracy is maximized when $w = .5$. As $w$ increases from 0, predicted accuracy increases to a maximum at $w = .5$ and then continuously decreases until $w = 1$. In fact, the GCM predicts that accuracy will be a non-monotonic function of $w$ in all tasks, except those in which only one of the two stimulus dimensions is relevant. If dimension 1 is relevant and dimension 2 is irrelevant (as in the Fig. 3 RB task), then predicted accuracy increases monotonically with $w$, whereas if the only relevant dimension is dimension 2, then accuracy will monotonically decrease with $w$.

The left column of Fig. 4 shows the LOST results for the three parameters of the GCM in the state-trace curves that result when II accuracy is plotted against RB accuracy (ignore the right column for now). The top three panels show state-trace plots predicted by the GCM for these two tasks and the bottom panel shows the LOSTs. Each curve in the top row is a state-trace generated by varying $w$. The different curves show predictions for different fixed values of $c$ and $\beta$. The second row shows the same thing, except these curves were generated by varying $c$, while holding $w$ and $\beta$ constant, and the state traces in the third row were generated by varying $\beta$, while holding $c$ and $w$ constant. In all cases, $w$ varied from 0 to 1, $c$ varied from 400 to 3600, and $\beta$ varied from .05 to .5. Values of $c$ below 400 were excluded because both models predict that performance on both tasks is near chance for all $c < 400$. Similarly, values of $\beta$ above .5 were excluded because both models make identical predictions in both tasks when $\beta = p$ and $\beta = 1 - p$, for all values of $p$. Note that, as expected, the state traces that are generated by varying $w$ are all non-monotonic, whereas the state traces that result from varying either $c$ or $\beta$ are all monotonic.

The bottom row of Fig. 4 plots the lengths of the various state traces from each panel and the X shows the mean length for each parameter. Note that the LOSTs for the $w$ parameter are more than twice as great as the LOSTs for the $c$ and $\beta$ parameters. In fact, given how much curvature the $w$ state traces display and that the values on each dimension are constrained to the interval [.5, 1], it is difficult to imagine how one could construct another model of these tasks that could include a parameter with significantly greater LOSTs than the $w$ parameter of the GCM. Therefore, as predicted, the non-monotonicity effects that $w$ have on II accuracy cause changes in $w$ to increase the complexity of the model much more than the other model parameters. As a

[7] Other versions of the GCM are common. For example, sometimes city-block distance is used rather than Euclidean, and sometimes an exponential similarity function is used rather than a Gaussian. However, Ashby and Bamber (2022) showed that the four different versions of the GCM created by taking all possible combinations of city-block versus Euclidean distance and exponential versus Gaussian similarity function produce highly similar state-trace plots.

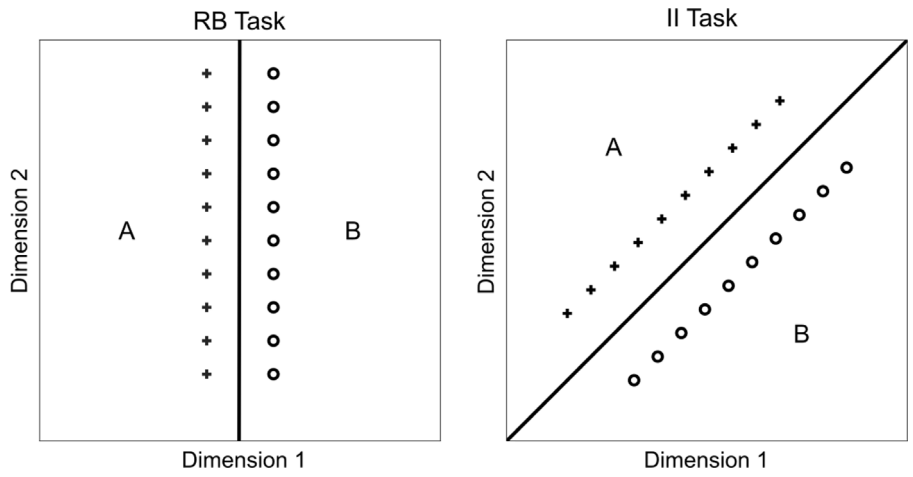


**Fig. 3.** Two categorization tasks. RB = rule based, II = information integration.

result, if different versions of the GCM are fit to a set of noisy or random points in a state-trace plot of II versus RB categorization accuracy, or if the points were generated by a model that is qualitatively different from the GCM, then the most likely conclusion will be that $w$ varies across the points, but both $c$ and $\beta$ are constant.

The GCM assumes that categorization decisions are made by computing the similarity of the current stimulus to stored representations of all previously seen exemplars from each category, adding all similarities to exemplars from the same category together, and then inserting these summed similarities into the Eq. (3) Luce-Shepard choice model (Luce, 1963; Shepard, 1957) to determine the various response probabilities. A natural question to ask is to what extent the contributions that $w$, $c$, and $\beta$ make to model complexity depend on the psychological assumptions the model makes about how categorization decisions depend on accessing memory traces of previously seen exemplars versus the mathematical assumptions the model makes – specifically, about how distance and similarity are computed and that the Luce-Shepard choice model is used to compute response probabilities (which are not considered core assumptions of the theory).

To answer this question, I repeated the LOST analysis described in Fig. 4 exactly, except swapping out the psychological assumptions the GCM makes about individual exemplars for a prototype model. Specifically, the prototype model was identical to the model described in Eqs. (3) - (5), except the summed similarities in Eq. (3) were replaced by the similarity of the current stimulus (i.e., stimulus $k$) to the category prototypes. So, for example, $\sum_{i \in C_A} \eta_{ik}$ in the numerator of Eq. (3) was replaced with $\eta_{iA_p}$, where $A_p$ is the prototype (i.e., the mean) of category A. A large literature confirms that this model makes very different psychological assumptions than the GCM (e.g., Homa et al., 1981; Smith, 2002; Smith & Minda, 2001). So the two models have exactly the same three parameters and use the exact same distance, similarity, and Luce-Shepard choice model equations to compute response probabilities (i.e., Eqs. (3)–(5)). They differ only in their assumptions about the nature of the stored category-relevant representations that are used to make categorization decisions.

Results are shown in the right column of Fig. 4. Each panel in this column can be compared directly to the analogous panel in the left column since both state traces from the two models were generated using the exact same values of all three model parameters (i.e., $w$, $c$, and $\beta$). Note that the state traces for the two models are quite different, but that the LOSTs are all highly similar. The most important point here is that in both models, the mean LOSTs for $w$ are more than twice as large as the mean LOSTs for either other parameter. Therefore, the much greater relative contribution to complexity provided by $w$ is due to its role in Eqs. (3)–(5), and not because of the GCM assumption that categorization decisions depend on accessing the memory representations of all previously seen exemplars.

Although the parameters of the GCM and prototype model have similar LOSTs, Fig. 4 shows that the state traces for the prototype model are restricted to a much smaller region of state-trace space than the traces for the GCM. For example, note that none of the one-parameter versions of the prototype model can account for II accuracy below about 0.6 or for the combination of high II accuracy and low RB accuracy. In contrast, all three versions of the GCM can account for either of these outcomes. These results suggest that overall, the GCM might be more complex than the prototype model that is based on the exact same Eqs. (3)–(5) (e.g., according to the model complexity measure proposed by Veksler et al., 2015). To investigate this possibility, I computed the GCM and prototype model manifolds for the RB and II tasks of Fig. 3. Specifically, for both models I simultaneously varied $c$ from 0.1 to 4,500, $w$ from 0 to 1, and $\beta$ from 0 to .5.[8] For each combination of these three parameters, I then plotted the predicted accuracy in the two tasks. Results are shown in Fig. 5. The shaded regions include all experimental outcomes that each model can fit perfectly by simultaneously varying all of its three parameters – that is, the shaded regions denote the model manifolds for these two tasks.

Note that the GCM has a substantially larger model manifold than the prototype model, and therefore is considerably more complex than the prototype model for these two tasks. The prototype model predicts a positive correlation between performance on the two tasks, whereas the GCM can account for almost any possible experimental outcome. In fact, the model manifold of the prototype model is contained completely within the model manifold of the GCM. As a result, there is no possible experimental outcome that the prototype model could account for that the GCM could not also fit perfectly. In contrast, there are many outcomes that the GCM could fit perfectly that are incompatible with all versions of the prototype model. The prototype model makes strong a priori predictions about the outcomes of these tasks – specifically, it predicts that performance on the two tasks should be similar. In contrast, the GCM makes almost no a priori

[8] Note that the ranges for $c$ and $\beta$ here are slightly greater than the ranges that were used to generate the state traces shown in Fig. 4. This had virtually no effect on any GCM predictions, but the small values of $c$ allowed the prototype model to account for some slightly smaller accuracies on the II task (i.e., below .63).

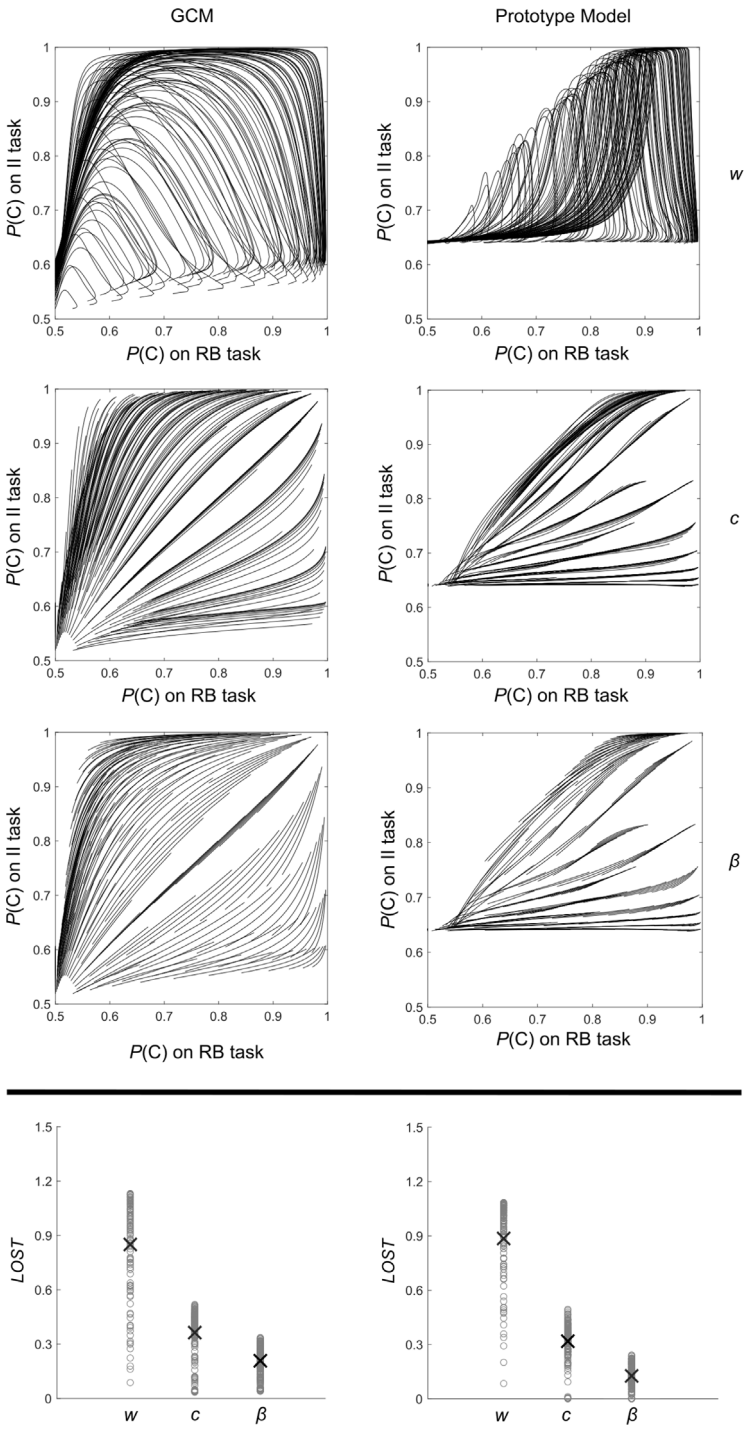


**Fig. 4.** State traces and LOSTs predicted by the GCM (left column) and the prototype model (right column) for the RB and II categorization tasks described in Fig. 3. Each curve in the first three rows is a state-trace generated by varying the single parameter indicated on the right. The bottom row plots the lengths of the various state traces from each panel and the X shows the mean LOST for each parameter.

predictions. According to the GCM, almost any outcome is possible. In consequence, there are many actual outcomes that could falsify the prototype model and almost no outcomes that could falsify the GCM. Furthermore, this GCM complexity advantage occurs despite the fact that the two models both have the same three free parameters, make the same assumptions about how distance and similarity are computed, and use the same Luce-Shepard choice model to convert the model-relevant similarities to response probabilities.

What causes the GCM to be more complex than the prototype model? As a first step in addressing this question, consider Fig. 6, which shows the regions of data space that are compatible with some version of the GCM and incompatible with all versions of the prototype model. Note that there are two such regions – a lower region in which accuracy is worse in the II task than predicted by the prototype model, and an upper region where accuracy on the II task is better than predicted by any prototype model. These two regions are associated with qualitatively

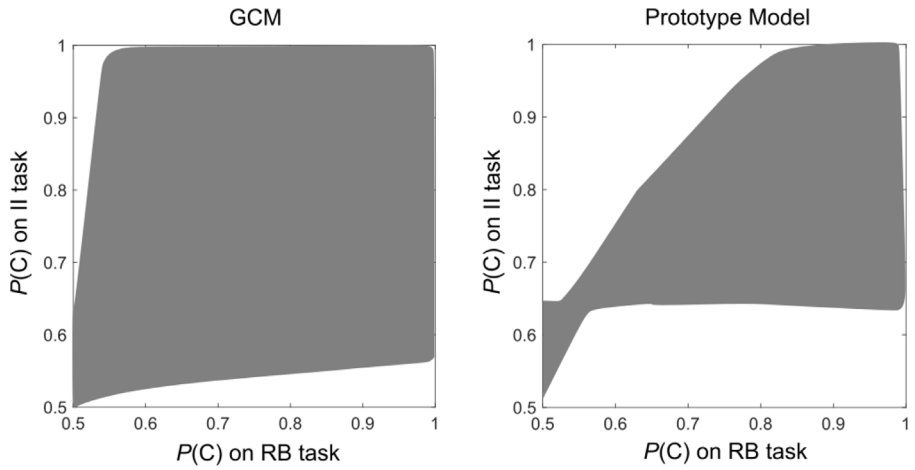


**Fig. 5.** The model manifolds of the GCM and the prototype model for the RB and II tasks shown in Fig. 3.

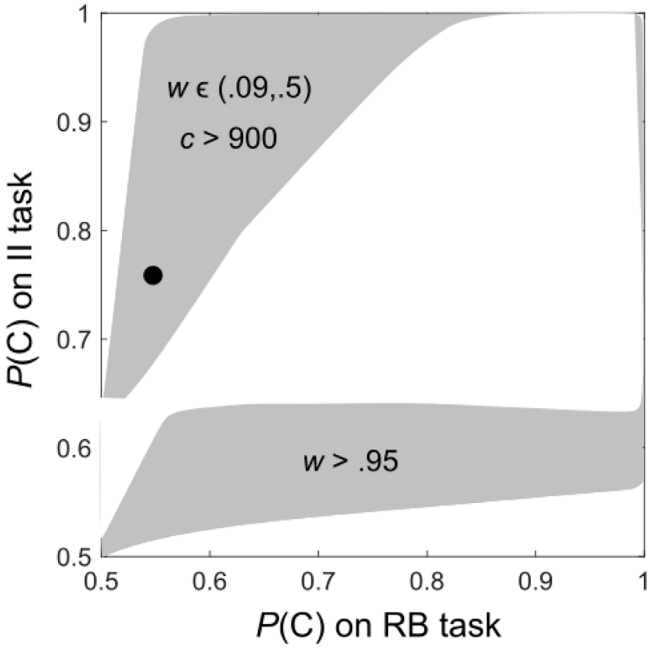


**Fig. 6.** The shaded regions show areas of data space that denote possible experimental outcomes that can be fit perfectly by the GCM, but are incompatible with all versions of the prototype model (for the RB and II tasks shown in Fig. 3).

different sets of GCM parameter values. GCM predictions fall in the lower region when the attention weight is large (i.e., $w >$ .95), whereas GCM predictions fall in the upper region when $w$ is small [i.e., $w \in (.09, .5)$] and the discriminability parameter $c$ is large (i.e., $c > 900$). The psychological assumptions made by the GCM are very different in these two regions, so the predictions of the model in these two regions need to be examined separately.

First, consider the lower region. In this case, the attention weight $w$ is large. Recall that in the RB task, the optimal strategy is to allocate all attention to the single relevant dimension – that is, to dimension 1, in which case $w = 1$. In contrast, in the II task, the optimal strategy is to set $w = .5$, because both dimensions are equally important in this task. When $w = 1$, stimulus values on dimension 2 are ignored, which is equivalent to collapsing both category structures onto dimension 1. The resulting category representations are shown in Fig. 7. Note that there is considerable overlap of the category A and B exemplars in the II task, whereas the A and B exemplars are perfectly partitioned and widely separated in the RB task. As a result, the GCM predicts poor performance in the II task and good performance in the RB task. On the other hand, the bottom row of Fig. 7 shows that the category prototypes are approximately the same in the two tasks, so the prototype model predicts roughly equal RB and II performance. The lower region of Fig. 6 conforms to our common understanding of the GCM – performance depends on all exemplars and when exemplars from contrasting categories overlap, performance suffers.

The rationale behind the GCM predictions in the upper region is very different – primarily because of the large value of $c$. First, however, note that because $w < .5$ in this region, the allocation of attention is suboptimal for both tasks, but much worse (i.e., further from optimal) for the RB task than for the II task. As a result, we expect predicted performance to suffer from these suboptimal values more in the RB task than in the II task, and as this analysis predicts, all points in this upper region are associated with higher II than RB accuracy. Second, and more importantly, what is the effect of such large values of $c$? This question is answered in Fig. 8, which shows the GCM predicted similarities between the bottom-most stimulus in category B and all individual exemplars in both categories. The similarities were computed from Eq. (5) with $w = .2$ and $c = 1000$. These values cause the GCM to predict accuracies in the two tasks denoted by the black dot in Fig. 6. Points above this dot tend to be associated with even larger values of $c$, so these parameter values are less extreme than many in this upper region.

As required by Eq. (5), the GCM always predicts that all self-similarities equal 1 and all other similarities are less than 1. This is true for all values of $c$. The interesting result is that because $c$ is so large in this upper region of data space, the model predicts that almost all other similarities are effectively 0 (i.e., all similarities denoted as 0 in Fig. 8 are less than .001). Recall that the GCM response probabilities are computed by comparing the sum of all these similarities to category A and category B exemplars. In the II task this sum equals 1.0067 for category B and 0 for category A, whereas in the RB task the sums are 1 and .3355 for categories A and B, respectively. Because the difference between the sums is considerably larger in the II condition, the GCM predicts that II performance will be much better than RB performance, given these parameter values.

Fig. 8 illustrates why the GCM is able to account for experimental outcomes that fall in this upper region of data space, but note that the psychological interpretation of this performance strongly violates the tenets of exemplar theory – namely, that all exemplars contribute to the categorization decision. In this portion of data space, the GCM is operating as a nearest-neighbor

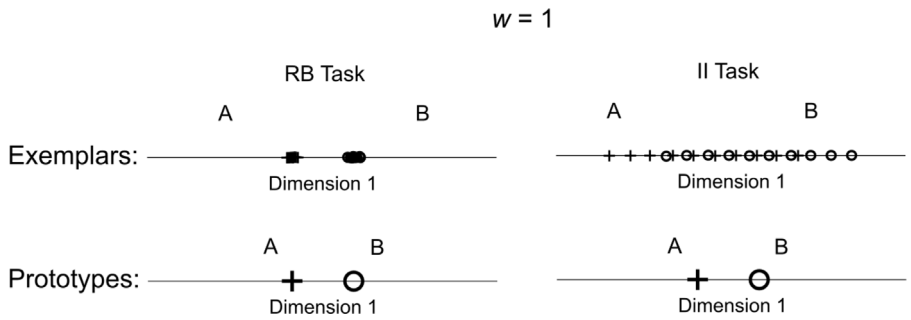


**Fig. 7.** Representations of the exemplars and prototypes of the RB and II categories shown in Fig. 3 when all attention is allocated to dimension 1.

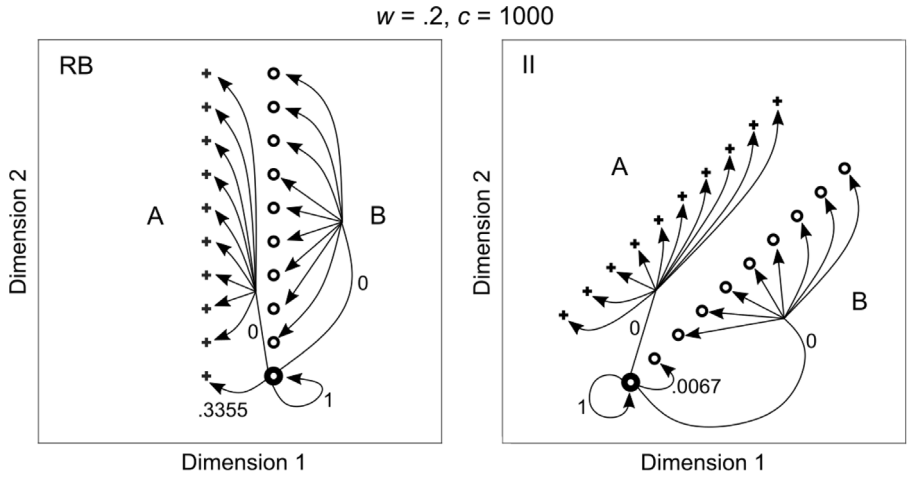


**Fig. 8.** The RB and II categories shown in Fig. 3 along with the similarities of the lowermost stimulus in category B to all exemplars from both categories. All similarities were computed from the GCM with $w = .2$ and $c = 1000$.

classifier. Categorization decisions depend almost exclusively on the category membership of the stored exemplar that is most similar to the presented stimulus. This property of the GCM was first noted by Smith and Minda (2001), who reported that best-fitting versions of the GCM to real data frequently mimic this type of nearest-neighbor classifier. Note that in this region of data space, the GCM could be interpreted as being more consistent with prototype theory than with exemplar theory, in the sense that, like prototype theory, the GCM assumes that categorization decisions are based on a single stored representation. The GCM and prototype models only disagree about the nature of this single stored representation. Prototype models assume the single critical stored representation is of the category prototype, whereas the GCM assumes the single critical representation is of the nearest neighbor.

When $c$ is small, all similarities are greater than 0 and the GCM is a true exemplar model, in the sense that all exemplars contribute to the categorization decision. As $c$ increases, the similarities of the most distant exemplars drop to 0 (or effectively to 0) and these distant exemplars no longer contribute to the decision. Finally, for large values of $c$, categorization decisions are based exclusively on the category membership of the single nearest neighbor. So by varying $c$, the GCM can toggle among a variety of qualitatively different decision strategies. In contrast, in the prototype model, categorization decisions always depend on the similarity of the presented stimulus to the two stored prototypes, regardless of the value of $c$. This observation suggests that the GCM's greater flexibility in selecting a decision strategy is the primary reason it is so much more complex than the prototype model.

In fact, this analysis also suggests that a strong argument can be made that calling the GCM an exemplar model is a misnomer. Within the machine-learning literature, the GCM would more accurately be labeled as a $k$-nearest-neighbor classifier, in which categorization decisions are based on the category membership of the $k$ nearest neighbors of the presented stimulus. If we denote the total number of exemplars in categories A and B by $N_A$ and $N_B$, respectively, then the GCM is a complex model because it allows $k$ to take any value from 1 to $N_A + N_B$. The specific value that $k$ takes in any application is controlled almost completely by the numerical value assigned to the parameter $c$ (i.e., $k$ is inversely related to $c$).

Ashby and Rosedahl (2017) showed that the GCM also has another, completely different interpretation. Specifically, they showed that the GCM is mathematically equivalent to a model in which categorization training does not create any memory representations, but instead alters the synaptic strengths between input and output units in a feedforward neural network. In this account, categorization decisions are made without ever activating memory representations of any category exemplars. Instead, during training, the summed similarities of Eq. (3) are encoded as synaptic strengths.

So the GCM has three very different psychological interpretations: (1) as an exemplar model in which categorization decisions depend on all stored exemplar representations; (2) as a $k$-nearest neighbor classifier in which categorization decisions depend only on the $k$ nearest stored exemplar representations, where $k$ can take any value from 1 to $N_A + N_B$; and (3) as a feedforward neural network in which summed similarities are encoded as synaptic strengths and no memory representations of any exemplars are ever activated. Of these three interpretations, the first dominates the literature, but is the least accurate since the GCM mimics a nearest-neighbor classifier over much of its parameter space.

## 6. Discussion

Information geometry defines the complexity of a mathematical model as the volume of its statistical manifold. The greater this volume, the more different data patterns the model predicts, and so the greater the model complexity. Despite its intuitive appeal, this measure of model complexity is rarely used in the psychological literature because (1) computing the volume of a model's statistical manifold for anything but the simplest possible models is a difficult computational challenge, (2) statistical manifolds are usually impossible to visualize because of their high dimensionality and because their points are probability distributions, and (3) the volume of a model's statistical manifold is a property of the whole model and does not allow one to examine the contribution of individual parameters to this volume. All of these problems disappear when the data space is a regular two-task state-trace plot. First, because the state-trace is a plot of one summary statistic against another, for any set of parameter values, a model predicts one point in this data space, rather than a probability distribution. Second, when the state traces are generated from a model by varying a single parameter, then the volume of the model's topological manifold in data space equals the LOST, which is easy to compute, the resulting state traces are simple to visualize, and the LOST can easily be computed separately for each of the model's parameters.

For signal-detection models, the obvious data space for a LOST analysis is the ROC curve. In the case of the normal, equal-variance model, the average LOST when $X_C$ is varied is much greater than the average LOST when $d'$ is varied (i.e., see Figs. 1 and 2). Therefore, the $X_C$ parameter grants the model much more complexity than the $d'$ parameter. Simulations show that this difference has profound effects on the results of fitting the model to random data. Applying the signal-detection model to random data almost always leads to the conclusion that all the points share the same value of $d'$ but were generated under different values of $X_C$.

The LOSTs associated with a parameter will tend to be greater the more curved its state traces. For example, the state traces generated by varying the signal-detection parameter $X_C$ (i.e., isosensitivity curves) are more curved than the traces generated by varying $d'$ (i.e., isobias curves). A logical inference that follows from this principle is that parameters that have non-monotonic effects on performance will tend to have large LOSTs, and therefore contribute to model complexity more than parameters that have monotonic effects on performance. As examples of this effect, we saw that in both the GCM and prototype models, the LOSTs for the attention-weight parameter $w$, which predicts non-monotonic effects on II accuracy, are much greater than the LOSTs for the parameters $c$ and $\beta$, which predict monotonic effects on accuracy.

Computing LOSTs for the popular normal, equal-variance, signal-detection model, for the GCM, and for the prototype model shows that a large discrepancy in LOSTs (e.g., of more than 2-to-1) among parameters of the same model should not be considered unusual. In hindsight, this should not be too surprising. If all parameters contributed equally to model complexity, then model-selection statistics that define the complexity of a model simply by the number of its free parameters (such as AIC and BIC) would always agree with statistics that depend on more sophisticated complexity measures, such as normalized maximum likelihood and free energy.

Furthermore, as Table 1 shows, a significant disparity in LOSTs can have profound effects on conclusions derived from fitting that model to noisy data. In fact, Table 1 shows that in extreme cases, almost nothing can be learned from such a model-fitting exercise because the outcome is preordained – the version of the model in which the single parameter that varies across points is the one with the largest LOSTs is virtually guaranteed to provide the best fit. This lesson reinforces cautions that model selection should not be based solely on goodness-of-fit (e.g., Myung, 2000). Especially in cases when the best-fitting version of a model is the one in which the free parameter has the greatest LOSTs, some other independent evidence should be sought before concluding that this model provides insight into the underlying psychological processes. On the other hand, if the best-fitting version of a model happens to be one in which the single varying parameter has small LOSTs, then knowledge that the LOSTs associated with that parameter are smaller than the LOSTs associated with other model parameters should raise confidence in the validity of the model-fitting outcome.

Finally, our examination of the complexity contribution of the parameters of the GCM and prototype models revealed another, somewhat unexpected benefit of a LOST analysis. While computing the various LOSTs associated with the two models, we noticed that the state traces for the GCM enclosed a larger region of data space than the corresponding state traces for the prototype model. The resulting follow-up analysis showed that the GCM is in fact, more complex than the prototype model for the two tasks we studied, even though both models are characterized by the same three free parameters and make predictions using the same base equations. A more detailed study then showed that the GCM is more complex because it can mimic a variety of different decision strategies (i.e., by varying the numerical value of the parameter $c$). In fact, this analysis indicated that calling the GCM an exemplar model could be interpreted as a misnomer. Rather, it might be more accurate to label the GCM as a $k$-nearest neighbor classifier, where $k$ can be set to any value. Thus, in addition to showing that the attention-weight parameter $w$ contributes to the complexity of both models far more than either of the other two parameters, our LOST analysis also led to some deep new insights into the structure of the GCM.

One possible limitation of the results described here is that our analyses ignored the variability inherent in any real data. For example, any empirical ROC curve must estimate $P(\text{H})$ and $P(\text{FA})$ with the proportion of observed hits and false alarms, respectively, and the standard error of these estimates will be inversely related to the sample size. However, note that the variability introduced by this sampling error will only exacerbate the problems identified here. Variability in the summary statistics adds noise to data space, which as we have seen, favors parameters with greater LOSTs. For example, consider an ROC curve in which the true underlying model is the standard normal, equal-variance, signal-detection model in which $d'$ varies but $X_C$ is constant. As a result, the true ROC curve is an isobias curve of the type shown in the top right panel of Fig. 1. However, note that any variability in the proportions that estimate $P(\text{H})$ and $P(\text{FA})$ will produce a scatter plot of points that no longer fall on a vertical line. The smaller the sample size that was used to estimate these probabilities, the greater the deviations we should expect from points that are aligned vertically. Table 1 shows that one result of this variability is to make it more likely that the best-fitting version of the signal-detection model will be the one that varies $X_C$ and holds $d'$ constant, even though this is not the model that generated the data. In other words, sampling variability of the type inherent to all empirical data will increase the odds that the best-fitting version of the model will be the one in which the single varying parameter is the one with the greatest LOSTs, and this will be true regardless of the psychological structure of the data.

There are also many questions that would be interesting to pursue in future research. For example, the present article focused exclusively on the case where the data space is two dimensional (i.e., $N_d = 2$). Although three-dimensional state traces

are somewhat more difficult to visualize, computing LOSTs when $N_d = 3$ is not really more difficult than when $N_d = 2$. As a result, it would be interesting to examine how a parameter's contribution to model complexity changes as the number of dependent variables is increased.

Another future research direction might be to generalize the LOST analysis described here to include the Bayesian notion of prior distributions on each of the model's parameters. For example, it was mentioned earlier that large values of the signal-detection parameter $d'$ are often viewed skeptically because experimental conditions are commonly arranged to expect small $d'$'s. A Bayesian approach would incorporate this belief into the data analysis by placing a prior distribution on $d'$ that assigns low likelihoods to large values. It seems straightforward to incorporate prior beliefs of this type into a LOST analysis. Let $f_i(\theta_i)$ denote the prior probability density function on parameter $\theta_i$. Then the Eq. (1) LOST measure can be generalized to include this prior distribution via

$$LOST(\theta_i) = \frac{\sum_{j=1}^{M-1} f_i(\theta_{i,j}) \sqrt{\left[DV_1(\theta_{i,j+1}) - DV_1(\theta_{i,j})\right]^2 + \left[DV_2(\theta_{i,j+1}) - DV_2(\theta_{i,j})\right]^2}}{\sum_{k=1}^{M-1} f_i(\theta_{i,k})}. \tag{6}$$

In other words, each distance segment is weighted by the prior likelihood associated with the value of $\theta_i$ used to generate the state-trace points that define that segment. The denominator is just a normalizing term that ensures the sum of all weights equals 1. Therefore, if a prior distribution on $d'$ assumes that values of $d'$ above 4 (for example) are impossible, then Eq. (6) would assign a weight of 0 to all distance segments associated with values of $d' > 4$, and so they would not contribute to the LOST. Note that Eq. (1), which was used to compute all LOSTs reported in this article, is equivalent to assuming uniform (or noninformative) prior distributions on all parameters. As with all Bayesian approaches, the challenge with Eq. (6) would be to define nonuniform prior distributions that would be accepted as noncontroversial by the broad research community.

Despite all of these benefits of a LOST analysis, it is important to note that computing the volume of a model's statistical manifold should always be viewed as a superior measure of model complexity. For example, the volume of a model's statistical manifold computes complexity relative to a saturated data space, rather than to a space defined by summary statistics. In general, collapsing trial-by-trial data into a few summary statistics should be expected to provide only a limited view of a model's complexity landscape. Despite this limitation, the analyses presented here revealed striking differences in LOSTs for parameters of some of the most popular mathematical models in psychology, and the results described in Table 1 show that these differences can have profound consequences on results of routine model fitting. Coupling these facts with the enormous computational advantage that computing LOSTs has relative to computing the volume of a model's statistical manifold, suggests that there could be significant benefit to adding a LOST analysis to one's standard modeling toolbox.

## References


Amari, S.-i. (2016). *Information geometry and its applications, vol. 194*. Springer.

Ashby, F. G. (2014). Is state-trace analysis an appropriate tool for assessing the number of cognitive systems? *Psychonomic Bulletin & Review*, *21*, 935–946.

Ashby, F. G. (2019). State-trace analysis misinterpreted and misapplied: Reply to Stephens, Matzke, and Hayes (2019). *Journal of Mathematical Psychology*, *91*, 195–200.

Ashby, F. G., & Bamber, D. (2022). State trace analysis: What it can and cannot do. *Journal of Mathematical Psychology*, *108*, Article 102655.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*(3), 372–400.

Ashby, F. G., Paul, E. J., & Maddox, W. T. (2011). COVIS. In E. M. Pothos, & A. Wills (Eds.), *Formal approaches in categorization* (pp. 65–87). New York: Cambridge University Press.

Ashby, F. G., & Rosedahl, L. (2017). A neural interpretation of exemplar theory. *Psychological Review*, *124*(4), 472–482.

Ashby, F. G., & Wenger, M. J. (2023). Statistical decision theory. In F. G. Ashby, H. Colonius, & E. Dzhafarov (Eds.), *The new handbook of mathematical psychology, vol. 3*. Cambridge University Press.

Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, *19*(2), 137–181.

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, *95*(1), 91–101.

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 840–859.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley New York.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(6), 418–439.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology, vol. 1* (pp. 103–189). Wiley.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*(1), 190–204.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, *97*(21), 11170–11175.

Myung, J. I., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, *50*(2), 167–179.

Newell, B. R., Dunn, J. C., & Kalish, M. (2010). The dimensionality of perceptual category learning: A state-trace analysis. *Memory & Cognition*, *38*(5), 563–581.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*(3), 472–491.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345.

Smith, J. D. (2002). Exemplar theory's predicted typicality gradient can be tested and disconfirmed. *Psychological Science*, *13*(5), 437–442.

Smith, J. D., & Minda, J. P. (2001). Journey to the center of the category: The dissociation in amnesia between categorization and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(4), 984–1002.

Stephens, R. G., Matzke, D., & Hayes, B. K. (2019). Disappearing dissociations in experimental psychology: Using state-trace analysis to test for multiple processes. *Journal of Mathematical Psychology*, *90*, 3–22.

Veksler, V. D., Myers, C. W., & Gluck, K. A. (2015). Model flexibility analysis. *Psychological Review*, *122*(4), 755–769.