1 Tools for Learning about Computational Models

Mark A. Pitt and Daniel J. Navarro
Ohio State University

In the broad field of psycholinguistics, the modeling of language processing has evolved into a prominent subfield over the last couple of decades that now exerts substantial influence on the direction of the discipline (Christiansen & Chater, 2001). It has sparked new ways of thinking about how language is produced and perceived, most notably in the context of localist connectionist models. With these positive developments have come new challenges, such as devising tests to distinguish among competing models.

The experimental method has proven to be well-suited for testing theoretical assumptions from which computational models are built. Differences between models can lead to contrasting qualitative predictions across experimental conditions, such as two main effects for one model and an interaction for its competitor. When successful, this method of model testing can yield evidence that convincingly discriminates between models.

Because such definitive tests are not always possible, researchers must explore the intricacies and nuances of the models' in order to identify conditions in which the models could be discriminated. This can be very much a hit-and-miss undertaking because most language processing models are often sufficiently complex that it is difficult to understand, let alone anticipate their full range of behaviors. Two consequences of this are evident in the literature. One is the discovery of an emergent property of a model, whereby it exhibits a behavior that was not purposefully or knowingly built into it. The model always possessed the behavior, but the difficulty in understanding the full consequences of our design choices when building the model can leave us unaware of some of its capabilities. A related problem, which shows up too often in the broader cognitive science literature, is making what seems like a reasonable qualitative prediction about model performance that turns out to be wrong. For example, a researcher may collect what appear to be compelling data against a model (e.g., double dissociation), only to be shown afterwards through simulations or data fitting that the model in question can indeed produce the observed pattern of results. Because it is difficult to discern the full capabilities of one model, let alone assess the similarities and differences of two, experiments that clearly discriminate between two models are not as common as one would like.

An additional challenge has to do with the incremental approach to model development. Results that once discriminated between two models will no longer do so after the inferior model is modified to accommodate new data. Although this process should result in the models converging on the design of the language system, the similarity is functional, not necessarily structural. That is, the models will perform similarly across many testing situations (i.e., fit data or simulate phenomena), but be architecturally different. Performance differences to distinguish such models can be difficult to find.

In this chapter, we introduce two methods for comparing quantitative models that can assist in tackling the aforementioned problems. The first focuses on inspecting the properties of the model itself to learn about its built-in power to simulate results. In the second, we introduce a method for identifying an experimental design that has the potential to distinguish between pairs of models.

MINIMUM DESCRIPTION LENGTH: A METHOD FOR CHOOSING BETWEEN TWO MODELS

The primary criterion used to choose among a set of models is the ability to simulate an experimental result. Most often this is quantified as a model's goodness of fit (GOF) to data collected in an experiment. This is a necessary condition that all models must satisfy to be considered a possible description of the language process under study. The ability of a model to fit the data is determined not only by whether the model is a good approximation to the language process, but also by two properties of the model itself, collectively referred to as its *complexity* (Myung, 2000). The property most readers will be familiar with is the number of parameters a model possesses. The more parameters there are in a model, the better it will fit the data. Essentially, each parameter adds an additional degree of freedom to the model that allows it to absorb more variance in the data, thus improving fit.

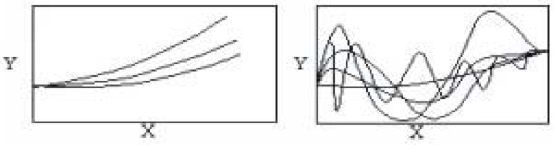

FIG 1. Simple models (left panel) produce only a few patterns, whereas complex models (right panel) can produce a diverse range of patterns.

Another dimension of a model that affects its ability to fit data is its functional form, which refers to the way in which the parameters, and possibly input, are combined in the model's equation. For example, Oden and Massaro's (1978) Fuzzy Logical Model of Perception has two parameters on a given trial. Anderson's (1981) Linear Integration Model also has two. As can be seem in the equations below, the parameters are combined differently. They are multiplied in FLMP, but added in LIM. It turns out that FLMP's multiplicative functional form makes it much more flexible in fitting data than LIM (Pitt, Kim, & Myung, 2003).

$$\text{FLMP: } p_{ij} = \frac{\theta_i \lambda_j}{\theta_i \lambda_j + (1-\theta_i)(1-\lambda_j)}$$

$$\text{LIM: } p_{ij} = \frac{\theta_i + \lambda_j}{2}$$

A model's complexity is directly related to its flexibility in fitting diverse data patterns. With its many parameters and powerful functional form, a complex model can produce many different data patterns, as depicted in the right-hand graph of Figure 1. A simpler model will have fewer parameters and a less powerful functional form. As shown in the left graph, it generates only one pattern, which changes little as the parameters of the model are varied across their ranges.

The increase in flexibility that comes with additional complexity means that GOF will also increase positively with complexity. This relationship is depicted schematically in the top graph of Figure 2, with complexity on the *x* axis and a measure of fit, such as percent variance accounted for (r^2) on the *y* axis. By virtue of its complexity alone, not its close approximation to the language process, a model can provide the best fit to the data. It is this problem that makes GOF a poor model

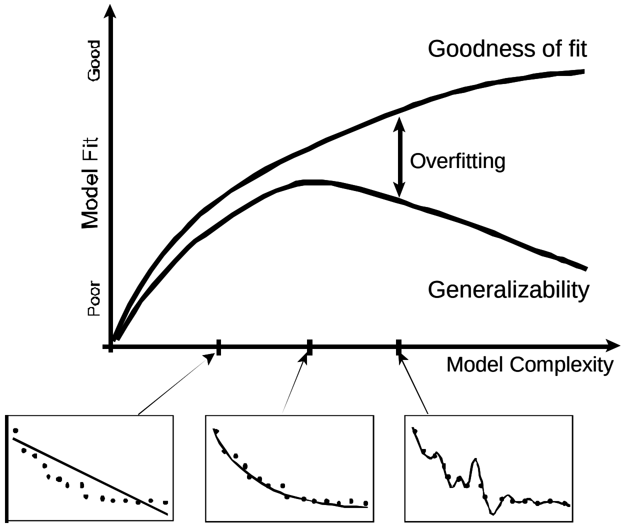

FIG 2. The relationship between goodness of fit, generalizability and model complexity.

selection criterion.

If GOF should be avoided, what should replace it? This question has been studied intensively in allied fields (Linhart & Zucchini, 1986). The consensus is that GOF should be replaced by *generalizability*, which seeks to choose the model whose performance (i.e., fit) generalizes best to data sets from replications of that same experiment. That is, do not choose the model that fits a single sample of data well. Rather, choose the model that fits all samples (i.e., replications) well. By doing so, the problem that befalls GOF is avoided - an inability to distinguish variation due to random error across samples from variation due to the language process itself.

The problem with GOF, and how generalizability overcomes it, is illustrated in Figure 2. The data points in the three bottom graphs are the same. The models (lines) increase in complexity from left to right. As they do, GOF increases as well. If GOF were the selection criterion, the model in the right-most graph would be chosen. It fits the data perfectly! The model in the middle graph fits the data less well, but notice that it captures the main downward trend and not the minor deviations of each point from this trend, which the right-hand model picks up. Which of these models best describes the data? Advocates of generalizability would pick the middle model because it captures the main trend well and is not side-tracked by the noise in the data (i.e., random error present in each data point). The sensitivity of the right-most model to the random noise is what makes it overly complex. Model A, in contrast, is overly simple. The straight line does not capture the decelerating trend in the data.

The lower line in the top graph depicts how fit and complexity are related when generalizability is used as a model selection criterion. It is an inverted U-shaped function that can be thought of as having two halves. In the first half, the complexity of the model must match the complexity of the pattern in the data. This is why generalizability increases as fit improves. If model complexity exceeds the peak of the function, generalizability will start to drop because the model will begin to fit random error, not just the regularity we attribute to the language process under study. Another way to think about generalizability is that it tries to strike a balance between the complexity of the model and the complexity needed to describe the regularity in the data.

Although the concept of generalizability is easy to describe, quantifying it has been a nontrivial undertaking. Short summaries of various measures can be found in Pitt, Myung, and Zhang (2002). The state of the art method today is the Minimum Description Length (MDL; Rissanen, 1996, 2001). It is elegant and conceptually quite simple to understand, although sometimes computationally challenging to implement. Given a set of data and two models, imagine that you varied the parameters of each model across their ranges and for each combination of parameter values fit the models to the data. You would end up with a very long list of fits, some being much better fits than others (MDL uses a lack-of-fit measure so smaller values are best). After summing these best fits, you would end up with a measure of each models flexibility. The smaller the value, the greater the model's flexibility.

The flexibility of a complex model will allow it to produce a few exceptionally good fits to the data, but this very same flexibility, which is due to excessive parameterization and its functional form, will cause it to generate a majority of fits that are poor, making the MDL value large. In essence, overly complex models are penalized for having more complexity than is needed to capture the regularity in the data. For a simpler model, the situation is very different. Although no one fit will be as good as the complex model, the reduced flexibility of the simpler model will mean that there will be fewer fits overall, the fits will not differ greatly from one another, and quite possibly all of the fits might not be too poor (compare the graphs in Figure 1) The MDL value of this simpler model could well be smaller. In short, a simpler model is penalized less severely because of its reduced flexibility, whatever the reason, be it fewer parameters or a simpler functional form.

Although this discussion has centered on model complexity, it is important to note that MDL does not favor the simpler of two models just because of the model's simplicity. Rather, a model's fit to the data is evaluated relative to its complexity to make the best inference as to which model most likely generated the data. MDL is a statistical inference tool that, at its most basic level, is not unlike statistical inference used in hypothesis testing. Given a small sample of data, we decide which conclusion to draw given its probability of being sampled by chance from the population. Similarly, MDL extracts as much information from the data sample *and the models* to make the best inference as to which model generated the data. We have found that it works quite well in choosing models in multiple areas of cognitive psychology (information integration, categorization, psychophysics; Pitt et al, 2002).

This short discussion of model complexity is meant to raise awareness of the difficulties of model selection. Although a model's good fit to data can, on the surface, seem like convincing evidence in support of a model, caution should be exercised in interpreting the fit until the reason for the good fit is known. Is it because the model is a good approximation of the language process being studied, or is it due to the model's complexity? Sensitivity to this issue will ensure a good fit is not misinterpreted.

LANDSCAPING: INVESTIGATING THE RELATIONSHIP BETWEEN MODELS AND DATA

Although neutralizing the effects of complexity is important to avoid selecting the wrong model, MDL only scratches the surface in informing us about model behavior. In addition to knowing that model A is more complex than model B, we would like to know more than this, such as how is it more complex and how and when does this extra complexity affect model performance. In short, we would like insight into the inner workings of the models, their similarities and differences, so that informative tests to distinguish them can be designed and carried out.

We have begun to develop tools for gaining this insight. *Landscaping* is the first of these. It has been successfully applied to statistical models (Navarro, Pitt, & Myung, in press; see also Pitt, Kim, & Myung, 2003), and as of this writing it is being adapted to localist connectionist models. The approach is the same in both modeling contexts. What differs is how it is implemented. We describe and demonstrate it in the context of statistical models because its application has been worked out more fully.

Two computational models that are functionally quite similar can be difficult to distinguish because, as mentioned in the introduction, an experimental setup that will lead to differing predictions can be elusive. One reason for this is that the experimental method is a rather course procedure for testing quantitative models. The choice of experimental design and the exact levels of the independent variables are decisions most often made from a consideration of the verbal model and intuitions about how best to manipulate the variables. The most well-thought-out experiment can yield data that are minimally informative because both models end up fitting the data well enough that neither can be rejected with confidence. This outcome could be avoided if, before conducting the experiment, we knew how the models would behave relative to one another. Landscaping provides this information. It does so by taking advantage of the precision of computational models to identify the circumstances in which model performance differs.

Landscaping relies on GOF to compare models, but does so in a way that is consistent with the spirit of generalizability. Instead of comparing models on their ability to fit a single data set, we compare their fits to a large number of data sets. When graphed, they yield a landscape of fits that inform us about model distinguishability. This is illustrated in Figure 3. Maximum Likelihood (ML) is used as the measure of fit. When the log ML value is taken, a negative value is obtained, with values closer to zero

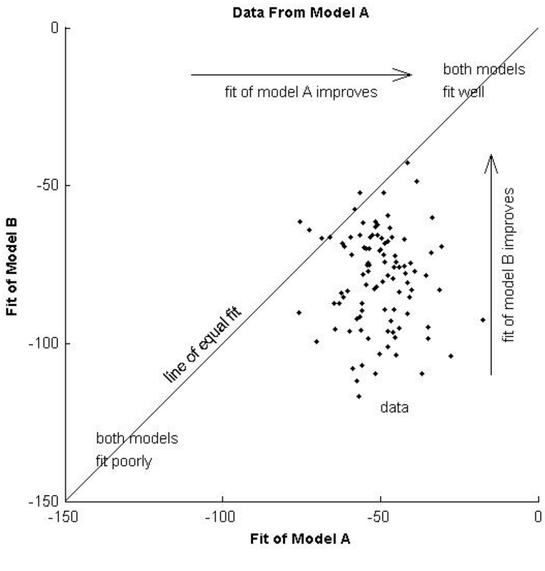

FIG 3. Schematic diagram of a landscape.

indicating a better fit.

Landscape creation begins by generating 1000 data sets from model A, and these data are fit by both models. The *x* axis in the figure denotes model A's fit, and the *y* axis denotes model B's fit. Each of the dots represents one data set. By drawing a diagonal line across the middle of the plot (at x=y), we observe that points above the line correspond to data sets that model B fits better than model A, whereas the opposite is true of points below the line. This line is referred to as a *criterion line*, or *decision threshold*. Data that both models fit very well will fall in the top right corner, whereas data that both models fit poorly will fall in the bottom left corner. By plotting the relative fits to the data, we obtain a landscape that enable us literally to see how closely model B can mimic model A. It would be nice if model A always provided better fits to its own data, but in practice this is not always true.

Construction of a landscape requires that data be generated from one of the models. In order to produce a data set, parameter values are needed. In the real world, it is rarely if ever known in advance which parameters values are most likely to be good ones (i.e., ones that yield model behavior that is similar to human performance). This is, after all, the very reason for the existence of free parameters. When comparing two models it is crucial to acknowledge this uncertainty. One way to do this is to specify a probability distribution over parameter values, and then sample the parameter values from this distribution. While we have used the "noninformative" Jeffreys' distribution (e.g., Robert 2001), a range of distributions (e.g., uniform) might be used for this purpose.

ILLUSTRATIVE APPLICATIONS

In this section, we present three concrete examples of how landscaping can be fruitfully employed to learn about model distinguishability. In the first example, we show how landscaping (and MDL) can be used to help design more discriminating experiments. In the second, we demonstrate how it can be used to assess the informativeness of past data in discriminating between models. In the final example, we briefly show how landscaping can be used to highlight the complex ways in which models can interact with one another, and the implications this has for model selection. More details on these examples can be found in Navarro, Myung, Pitt & Kim (2003) and Navarro et al. (in press).

Experimental Design

The first example we consider uses the information integration models LIM and FLMP presented earlier. Suppose that we want to discriminate between them using a two-choice phoneme categorization task (e.g., choose /ba/ or /da/) with a two by eight design, and 24 participants. This design involves two different levels of one information source (e.g., visual) and eight different levels of the other (e.g., auditory). Thus there are a total of 16 stimuli that may be produced by combining the two evidence sources. This is not an uncommon experimental setup, yet the landscaping plots shown in Figure 4 reveal that model distinguishability is asymmetric across data sets. When data are generated by FLMP (left graph), the FLMP model provides a superior fit to virtually all data sets. The long tail of the distribution indicates a sizeable majority of these are quite decisive. When LIM generated the data (right graph), the two models fit the data bout equally well, as indicated by the tightly packed distribution that hugs the criterion line.

What are the implications of this outcome? If the language process is truly FLMP-like, then there will be no problem validating this with the 2x8 design because FLMP will provide the best fit to the data. If the process is actually LIM-like, then it will be much more difficult, if not

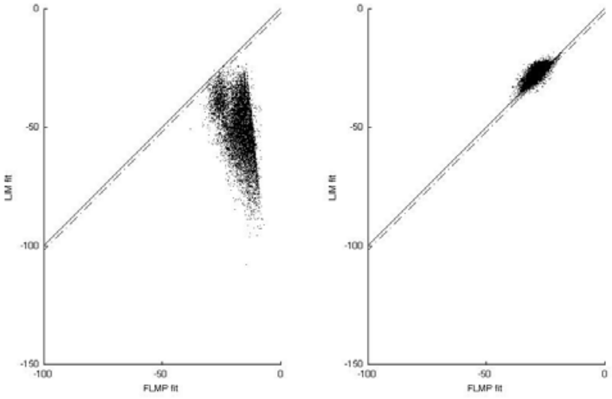

FIG 4. Landscapes for FLMP (left panel) and LIM (right panel), assuming a 2 by 8 design without unimodal conditions.

impossible, to distinguish between them. It would be preferable to conduct an experiment whose design does not suffer from this limitation so the models could be distinguished regardless of the form of the language process. Landscaping can assist in identifying such a design.

It turns out that a minor alteration remedies the asymmetry. The preceding design does not ask how participants would respond when only one source of evidence is provided, even though the models make different predictions in these circumstances. LIM predicts $p_i = \theta_i$ whereas FLMP predicts that $p_i = \theta_i$ / (1- θ_i). By adding the 10 extra "unimodal" stimuli (two visual alone and eight auditory alone) to the design and then repeating the analysis, we obtain the landscapes in Figure 5. Clearly, the new design is far better able to discriminate between FLMP and LIM. Most notably, the data generated by LIM yield a distribution of relative fits has now moved above the criterion line.

The effects of differences in model complexity can also be evaluated in a landscape plot. Because within an experimental design complexity will be constant between models, the criterion line will shift toward the more complex model by the amount the two models differ in complexity. The dashed lines in Figures 4 and 5 incorporate this adjustment, and actually represent the MDL criterion instead of the ML criterion. Notice that in Figure 5 the relative-fit distributions are so far from both lines that it

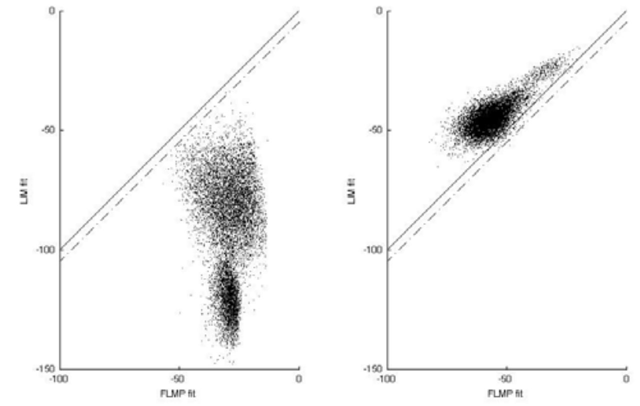

FIG 5. Landscapes for FLMP (left panel) and LIM (right panel), assuming a 2 by 8 design with unimodal conditions added. The solid line is the ML threshold and the dashed line is the MDL one.

really does not matter which model selection criterion one would use to choose the between the models. Both would perform about equally well. The right panel in Figure 4, in contrast, is an example of a case where good statistics can sometimes make up for the flaws in a design. With the ML criterion, 31% of the LIM distribution falls on the wrong (FLMP) side of the (solid) decision line. Although the (dashed) MDL criterion is very close to the ML criterion, it makes an enormous difference in model selection accuracy. Only 3.6% of the LIM data sets are now misclassified (on the wrong side of the criterion line).

In sum, the second design is far more likely to distinguish the models and has the attractive property of being able to collect data that clearly favor one model or the other because both relative-fit regions are distinct. Furthermore, it is much less sensitive to the choice of model-selection statistic.

Informativeness of Empirical Data

In addition to assisting with the design of future experiments, landscaping can be used to shed light on the informativeness of data collected in past experiments. A content area in which it has been fruitfully applied in this manner is modeling the time course of recognition, in particular because these models tend to mimic each other quite well. Furthermore, since commonly used MDL approximations such as the one discussed by Pitt et al. (2002) tend to fail in these cases (see Navarro, submitted), it is all the more important to have a quantitative methodology to guide model comparison.

Consider the functions $y = a \exp(-bt^c)$ and $y = a_1 \exp(-b_1t) + a_2 \exp(-b_2t) + a_3$. The first "power-exponent" (PE) function is from Wickelgren's (1972) strength-resistance theory of retention, while the second "sum of exponentials" (SE) function was suggested by Rubin, Hinton and Wenzel (1999). Both functions produce the decreasing, negatively accelerated curves that are highly typical of retention data, and provide good fits to the large number of data sets available (e.g., Rubin & Wenzel, 1996). Moreover, both satisfy Jost's law: If two traces have equal strength at time t, but are of different ages, then the older one should decay less rapidly from that point on.

Nevertheless, the two models represent different theoretical ideas: The PE function is based on the notion of a single memory trace whose decay is subject to two different factors. It is the action of these factors that produces Jost's law. However, in the SE function, there are three different memory stores, each decaying with a constant deceleration. In this function, Jost's law is produced by the multiplicity of stores.

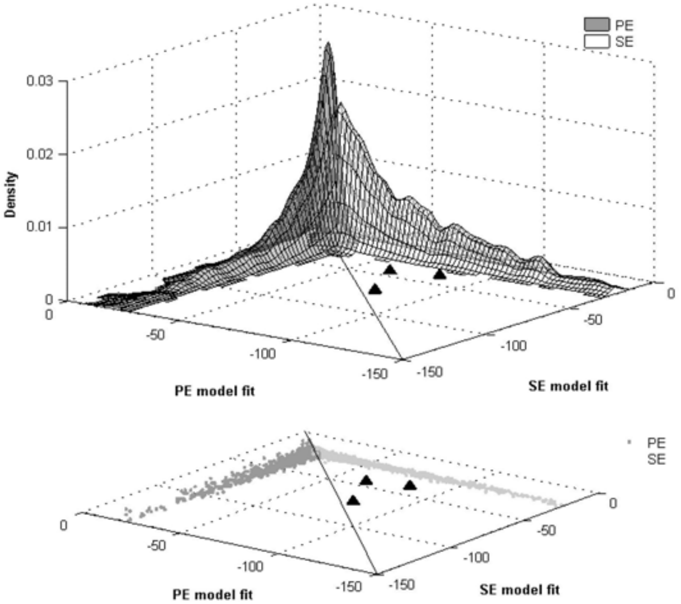

FIG 6. Landscapes for the retention functions. The lower panel shows raw landscapes, while the upper panel shows estimated densities. The locations of empirical data are shown by triangles.

Landscapes for the two models are shown in the lower panel of Figure 6. The experimental design that was used to generate the data comes from a recognition memory experiment of Rubin et al (1999), chosen because the design was large (in number of time steps and participants) and because it was replicated three times. Two landscapes are shown on the same graph. The lightly shaded one was generated by the data from SE, while the darker one came from the data of PE. The upper panel gives the estimated probabilities of a given relative fit (see Navarro et al, in press for details). The highest density regions are concentrated near the criterion line. The models also have quite pronounced tails, indicating that they can be distinguished. Furthermore, when we overplot the fits of the models to the empirical data from the three recognition experiments of Rubin et al (triangles), clear evidence for SE over PE is visible. Two of the three points are not just below the criterion line, but quite close to the SE distribution, and just as importantly, far from the PE distribution.

In hindsight, this outcome makes a good deal of sense. The experimental designs used in the Rubin et al. paper did not employ a distractor task, so the empirical data may represent a mixture of traces from short-term and long-term stores (Wixted, personal communication). Since the PE function incorporates only a single trace and is not designed to accommodate short-term memory, these findings are highly interpretable.

Note that the landscapes make such a conclusion much easier to support. Imagine, for instance, that we had presented Figure 6 with only the triangles (which is equivalent to a Table of ML fits). It would seem a little rash to draw such strong conclusions, particularly since SE has more parameters than PE, so its superior fit is a little suspect. However, the landscapes provide information about the representativeness of a relative fit, which assists in interpreting the empirical data. The landscapes allow us to conclude with more confidence that SE really does perform better on these data, but there are also some aspects of the data that it clearly does not capture.

Model Selection

In this final example, we demonstrate how landscaping can reveal some of the complexities of model selection. Consider Nosofsky's (1986) Generalized Context Model (GCM), and an extension of this model, GCM-γ (Shin & Nosofsky, 1992). In the GCM, the probability that an observed stimulus is judged to belong to a particular category is proportional to the stimulus' similarity to a set of stored exemplars from that category. In the GCM-γ model, the probability of category membership is assumed to be proportional to some power γ of this similarity. Obviously, the GCM is a special case of the GCM-γ when $\gamma = 1$. While these notions are quite simple, the models gain considerable complexity from the underlying similarity measure. When we landscape these models using the similarity representation reported by Shin and Nosofsky (1992), the results are rather surprising. As is immediately apparent in Figure 7, the landscapes are remarkably different from each other. This is true despite the fact that GCM is nested within GCM-γ. This outcome arises because the γ parameter adds a large set of new data patterns that GCM-γ can produce and GCM cannot. This set is so large that GCM-like patterns are very atypical of GCM-γ.

Comparison of the solid decision threshold (ML) to the broken one (MDL) reveals that the latter is far superior. Since the models are nested, ML classifies all patterns as belonging to the more complex model, GCM-γ. To compensate for complexity differences between the models, the criterion line should be shifted downward by 5.2 units in both landscapes. Although this minimally affects model selection when fitting GCM-γ data (misclassification errors are still close to 0), selection improves for the GCM data, but errors are still quite high at 67%.

Why was the complexity adjustment not better? Comparison of the landscapes reveals that complexity only partly accounts for the differences between the models. Complexity measures like MDL consider the relationship between a model and data, but do not consider the interrelationship between models as well. This limitation of scope results in a complexity measure that can suggest only a constant correction to the ML criterion. In the Shin and Nosofky experiment, however, GCM and GCM-γ have a nonlinear relationship with each other as well as the data. Because the GCM landscape is so sharply defined, almost any pattern inside that region (which is basically a semi-circular area) is more representative of GCM. Anything outside of this area is more representative of GCM-γ. Therefore, the best way to discriminate between these models would be to define a nonlinear decision threshold along the borders of this semi-circular region. Measures of model complexity like those provided by MDL cannot achieve this.

CONCLUSION

Computational modeling has advanced the field of psycholinguistics by sharpening our understanding of theoretical ideas and their potential. To build a model, assumptions about process and representation must be formulated, which adds precision to one's description of the language system.

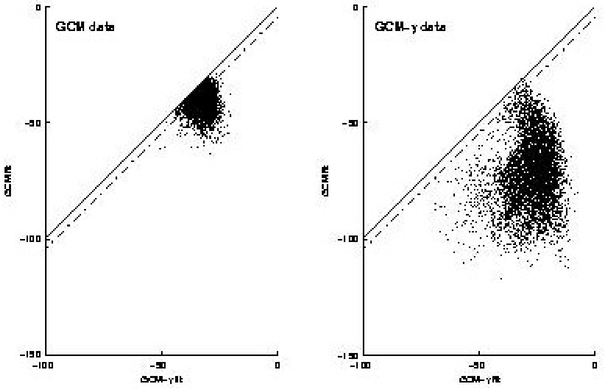

FIG 7. Landscapes for GCM (left panel) and GCM-γ (right panel). The solid line denotes the ML decision criterion, while the broken line is the MDL criterion.

The very success of modeling raises new and difficult issues. Two of these is how to compare and select between competing models. The consequences of the design choices made in model construction must be understood to succeed in either. Otherwise one runs the risk of being misled in the same way a garden-path sentence misleads a reader. The two quantitative tools introduced in this chapter are intended to assist in this enterprise. Landscaping is a simple yet powerful tool for assessing model distinguishability. MDL is useful for selecting among quantitative models, where the goal is to maximize generalizability, not goodness of fit. The three examples presented here (assessing the distinguishability of models within an experimental design, evaluating the informative of data in distinguishing models, and discovering complex relationships between models and data) are meant to demonstrate the usefulness and versatility of these complementary tools. We hope they are of practical use.

ACKNOWLEDGEMENTS

The authors were supported by NIH grant R01-MH57472. DJN was also supported by a grant from the Office of Research at OSU.

References

Anderson, N. H. (1981). *Foundations of Information Integration Theory*. Academic Press.

Christiansen, M.H., & Chater, N. (2001). *Connectionist Psycholinguistic*s. Westport, CT: Ablex.

Linhart, H. & Zucchini, W. (1986). *Model Selection*. New York: Wiley.

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

Navarro, D. J. (submitted). On the misbehavior of the Fisher information approximation to Minimum Description Length. Submitted to *Neural Computation.*

Navarro, D. J., Myung, I. J., Pitt, M. A. & Kim, W. (2003). Global model analysis by landscaping. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*.

Navarro, D. J., Pitt, M. A. & Myung, I. J. (in press). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology.*

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

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review, 85*, 172-191.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Science, 6*, 421-425.

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

Pitt, M. A., Kim, W., & Myung, I. J. (2003). Flexibility versus generalizability in model selection. *Psychonomic Bulletin & Review, 10*, 29-44.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transaction on Information Theory, 42*, 40-47.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory 47,* 1712-1717.

Robert, C. P. (2001). *The Bayesian Choice* (2nd Ed). New York: Springer.

Rubin, D. C., Hinton, S. & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory & Cognition, 25,* 1161-1176.

Rubin, D. C. & Wenzel, A. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103*, 734-760.

Shin, H. J. & Nosofsky, R. M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General, 121*, 278-304.

Wickelgren, W. A. (1972). Trace resistance and decay of long-term memory. *Journal of Mathematical Psychology, 9*, 418-455.