From Modeler-free Individual Data Fitting to 3-D Parametric Prediction Landscapes: A Research Expedition

Sue E. Kase (skase@ist.psu.edu), Frank E. Ritter
College of Information Sciences and Technology, Pennsylvania State University
University Park, PA 16802 USA

Michael Schoelles
Cognitive Science Department, Rensselaer Polytechnic Institute
Troy, NY 12180 USA

Abstract

A parallel genetic algorithm on a high-performance cluster is used to fit individual differences found in subjects' performance of a stressful mental serial subtraction task. The approach leads to a succession of unexpected results, interesting questions, and atypical analyses including 3-D parametric visualizations of the cognitive architecture and the serial subtraction model.

Keywords: Cognitive modeling; Individual differences; Performance distributions; Parameter space visualizations

Introduction

A complex cognitive model of a serial subtraction task was developed as part of a project to study the effects of stress, task appraisal, and caffeine on cognition (Whetzel, Ritter & Klein, 2006). Generally, cognitive models attempt to explain average performance across subjects, but in this case the subjects exhibited a wide range of task performance. The pitfalls associated with averaging over subjects have been known for a long time (e.g., Siegler, 1987). The performance variability associated with serial subtraction suggested a high degree of individual differences. This performance variability made it difficult to fit the serial subtraction model to the human data (Ritter, Schoelles, Klein & Kase, 2007). Fitting the model to each individual subject's performance appeared both necessary (because the model's predictions did not match the range of subjects), and desirable (because it supported understanding individual differences, as suggested by authors such as Gobet & Ritter, 2000, and by Siegler, 1987).

The model's cognitive architecture, ACT-R, offers many parameters (over 60) for manipulating the model's performance. Considering the combinative parametric search space, and substantial computational and time resources, individual data fitting did not appear a viable approach if done by hand.

A research expedition commenced by considering a relatively unused approach to individual data fitting. We describe a parallel optimization algorithm's efficient, accurate, and non-biased fit of the serial subtraction model to individual data from 15 subjects. We put this option forward as an approach, not necessarily as the most efficient or best way to do model fitting. By plotting the search algorithm's progress, stochasticity was detected in the model and architecture making interpretation of the data difficult. The stochasticity directed analysis towards individual performance distributions. Upon closer examination of the search algorithm's results, different ACT-R parameter sets were producing nearly perfect fits for many subjects. This prompted a second distribution exploration, this time of ACT-R parameter values. Analysis of the ACT-R parameter distributions raised questions about the terrain of the parametric landscape. In the final phase of the expedition, snapshots record the parametric landscape from a 3-D visualization perspective. The paper concludes with thoughts about what the expedition may contribute to cognitive modeling.

Serial Subtraction Task

A cognitive model was developed to simulate a human subject performing a serial subtraction task. Serial subtraction is the mental arithmetic stressor portion of the Trier Social Stressor Test (TSST, Kirschbaum, Pirke & Hellhammer, 1993). The TSST has been used to provide an acute physiological stress response in human subjects in 100's of studies since the 1960's. The task consists of four 4-minute blocks of mentally subtracting by 7's and 13's from 4-digit starting numbers.

Subjects complete the task mentally without visual clues; speaking the solution to each subtraction problem. The task is performed in front of an experimenter and video camera. Subjects are timed and prodded to go faster during the course of each block.

Experimental Data

Data from 15 subjects in a control condition in a larger study were used for the individual differences fitting approach discussed here (the others ingested caffeine).

During the serial subtraction task, subjects' answers were scored by the experimenter against a list of correct answers from the starting number. For each subject the number of subtraction problem attempts were recorded and a percent correct score was calculated by dividing the total number of correct attempts by the total number of attempts for each block of the subtractions.

Table 1 shows the subtraction rates for the subjects' performance on the two 4-minute blocks of subtracting by 7s. The large standard deviations indicate that there is a wide range of performance on this task suggesting a high degree of individual differences. A previous study found that this variability made it difficult to fit the serial subtraction model to human data (Ritter, Schoelles, Klein & Kase, 2007).

Table 1: Human subject (N=15) mean performance and standard deviation for serial subtraction on the two 4-minute blocks of subtracting by 7s.

	7s – 1st block	7s – 2nd block
Number of Attempts	47.3 (15.2)	47.8 (19.2)
Percent Correct	82.0 (10.0)	88.8 (7.0)

Cognitive Model and Architecture

Many instances of cognitive architectures exist, for example: ACT-R (Anderson, 2007), Soar (Newell, 1990), and Epic (Meyer & Kieras, 1997). Our research uses the ACT-R version 6.0 architecture. ACT-R has been used in modeling working memory tasks and arithmetic processing tasks by other cognitive science researchers. ACT-R is a two-layer modular cognitive architecture on a production system framework. In ACT-R cognition emerges through the interaction of a number of independent modules. Each of these modules is associated with specific brain regions and theories about the internal processes of these modules (Anderson, 2007). The modularity of ACT-R permits the parallel execution of the verbal system with the control and memory systems (specifically involved in the serial subtraction task).

Our model does subtraction in a right to left column order with borrow. It has 25 rules and hundreds of declarative memory elements. The goal buffer holds the task control, the imaginal buffer holds the problem representation, and the response is output through the speech buffer.

Parallel Genetic Algorithm

A new approach to fitting the serial subtraction cognitive model to the human data was attempted using a genetic algorithm. Genetic algorithms (GAs) are based on principles of natural selection and genetics, and have been applied successfully to numerous problems in business, engineering, and science (Goldberg, 1994). GAs are randomized, parallel search algorithms that search from a population of points. The points (often referred to as genotypes) represent individuals in a population. The genotypes are evaluated for fitness, then propagated to later generations by means of probabilistic selection, crossover, and mutation.

In a cognitive modeling context, the genotypes are sets of cognitive architecture parameters applied to the cognitive model. The population evolves to find better 'solutions' by selecting the most fit parameter sets (those that give the best match to the human data), and propagating these solutions to the next generation. The fitness evaluation consists of running the model, analyzing the model's performance output, and calculating a fitness value for the model's predictions compared to the data. A parallel version of the GA (PGA) distributes the computational load of the fitness evaluation among multiple processors reducing the time required to reach acceptable solutions.

The ACT-R architecture and cognitive model are written in the Lisp programming language. A message-passing interface (MPI) implemented the parallel processing portions of the PGA. ACT-R and the cognitive model were packaged into an executable Lisp image file. A prototype of the model-fitting PGA is running on a high-performance computing cluster at the National Center for Supercomputing Applications. Implementation details of the PGA, and MPI with ACT-R and the cognitive model are explained in Kase (forthcoming).

The Fitting Process

Fifteen PGAs were set to run 100 generations of 200 genotypes taking approximately two hours of runtime on 200 processors. Each PGA fit the serial subtraction model to an individual subject's performance data. A previous study (Kase, Ritter & Schoelles, 2007) reported preliminary tests of the PGA code fitting the serial subtraction model to post-task appraisal group means representing a challenge and a threat appraisal of the serial subtraction task. The model's predictions could match the mean but not the wide distribution of the data.

The PGA uses genotypes each representing the value of three ACT-R parameters. This set is offered only as a plausible and useful set to demonstrate this process. We investigated: activation noise representing variance in retrieving declarative information (ANS), the base level constant affecting declarative memory retrieval (BLC), and syllable rate, seconds per syllable (SYL)—because the model verbalizes the answers as the human subjects do. One processor was allocated for each genotype.

By running the model with the associated parameters and comparing the resulting predictions to the data, each genotype is associated with a fitness value. In this case, sum of the squared error was calculated on both number of attempts and the percent correct from a block of serially subtracting by 7s. The fitness is in terms of error (or cost) and is the discrepancy between the model's predictions and the actual human performance on the cognitive task. The genotypes offering the best fits were run an additional 200 times to ensure stable model predictions.

Individual Data Fits

We would like the serial subtraction model to predict the same range and distribution of performance produced by the human subjects. Table 2 summarizes the results from fitting the model to performance data from the 15 subjects. The last row in Table 2 shows results of fitting the model to the average performance across all 15 subjects.

Considering the complexity of the serial subtraction task and the human performance variability, these are exceptional model to human data fits. For number of attempts and percent correct, all subjects were fit within a fractional part of a subtraction problem. Table 2 lists only one fit for each subject—the solution resulting in the lowest fitness value. The PGA actually produced a set of solutions for each subject. For example, when fitting to S16's performance, the PGA found 9 different genotypes with fitness values less than 1.0. Genotypes are sets of ACT-R parameter values (ANS, BLC, SYL); the PGA finding different genotypes yielding nearly perfect fits was unexpected.

Table 2: PGA fit results for the 15 subjects and an average across subjects (last row) comparing human performance and model predictions in number attempts and percent correct (both rounded), and fitness value associated with the genotype (ANS, BLC, SYL).

Subject	Human Performance	Model Prediction	Fitness Value	Genotype (ACT-R parameters)
1	28, 67.9	28.0, 67.8	0.0006	0.83, 2.76, 0.87
47	29, 62.1	29.3, 62.0	0.0866	0.66, 2.25, 0.83
25	31, 80.7	30.8, 80.8	0.0487	0.48, 2.25, 0.76
11	35, 65.7	34.5, 65.1	0.6836	0.82, 2.49, 0.69
14	37, 75.7	36.3, 75.8	0.5523	0.83, 2.75, 0.62
2	37, 78.4	36.2, 78.6	0.7682	0.81, 2.80, 0.63
46	45, 80.0	44.7, 80.4	0.2510	0.43, 1.90, 0.47
27	46, 87.0	46.1, 87.7	0.4917	0.76, 2.96, 0.46
16	50, 92.0	50.4, 92.3	0.2233	0.50, 2.46, 0.41
43	54, 89.0	53.9, 89.0	0.0214	0.72, 2.88, 0.38
41	55, 87.3	55.2, 86.8	0.2261	0.54, 2.32, 0.36
23	57, 84.2	56.8, 84.4	0.0744	0.79, 2.71, 0.35
9	57, 87.7	57.2, 87.1	0.4089	0.78, 2.92, 0.35
21	65, 90.8	64.8, 91.2	0.1997	0.53, 2.24, 0.29
26	83, 94.0	83.3, 94.2	0.1463	0.47, 2.14, 0.16
Avg	47, 82.0	47.0, 81.8	0.1652	0.76, 2.65, 0.45

Two notable regularities should be mentioned about the results in Table 2. The value of SYL (last ACT-R parameter in the genotype) represents seconds per syllable in speaking the answer. The model uses the ACT-R Vocal Module to speak the subtraction problem answers. Looking at Table 2 from the top down we see the value of SYL decreasing as performance increases. The results show top performers speaking a syllable more quickly than the poor performers.

In the first row, the ANS part of the genotype producing S1's fitness value is 0.83. ANS is ACT-R's activation noise parameter. The value 0.83 is higher than what is normally used within the ACT-R community. Modelers using a traditional manual fitting process would generally not assign a value for ANS over 0.5. 60% of the values for ANS in Table 2 are substantially above 0.5, and may reflect the results of the stress this task creates in some subjects.

The serial subtraction task was specifically designed to maximize a stress response in human subjects. The PGA's non-bias randomized search of the ACT-R parameter space yielded high ANS values when predicting performance on this task. A non-bias fitting approach may be a useful diagnostic tool for cognitive theory development.

Predicted Performance Distributions

Previously presented plots of the minimum fitness value as the PGA evolved generations of genotypes showed nonmonotonic patterns and non-convergence of the population on a solution (Kase, forthcoming). We hypothesized that stochastic effects embedded in the ACT-R architecture caused these effects. To investigate, the 15 genotypes producing the best fits listed in Table 2 were run in parallel on 200 processors. The results indicated that a static set of ACT-R parameters yields a distribution of performance predictions, not a single prediction, but that the distributions are quite narrow. While this is considered known, Figure 1 shows this difference between model and data is probably worse than most people believe. Absolute frequency histograms of the model's predictions for each individual subject's performance were plotted and overlaid in Figure 1.

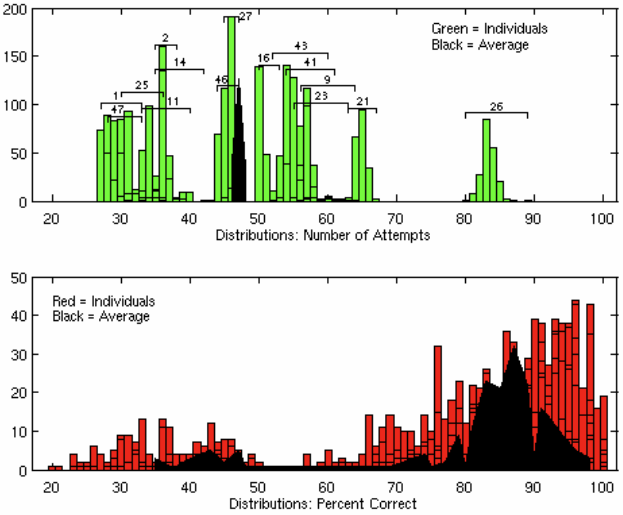

Figure 1: Model performance distributions (200 runs of each genotype in Table 2) for individual subject fits (labeled by subject number, top plot). Top plot, number of attempts (green); and bottom plot, percent correct (red). Black distributions represent predictions from fitting to an average across all subjects.

The top plot in Figure 1 shows the 15 performance distributions produced by the model for the 15 parameter sets in Table 2. The distributions are green (gray) and labeled by subject number. The black distribution represents fitting to the average number of attempts across all subjects. Similarly, the bottom plot in Figure 1 shows the 15 individual percent correct distributions overlaid in red (gray) with a black distribution representing the average percent correct across all subjects.

The large standard deviations in Table 1 hinted that fitting to individual subjects was the best approach. The top plot of

Figure 1 confirms an individual differences approach. When comparing the average attempts distribution to the 15 individual attempts distributions, the model's average attempts distribution covers only 7% of the full range of performance produced by the subjects. The average percent correct distribution is a better approximation of the range of performance produced by the individual subjects, but even here there remain substantial differences.

ACT-R Parameter Distributions

Upon closer examination of the PGA's results, the genotypes listed in Table 2 were only a subset of good fits produced by the PGA. To analyze possible similarities and differences across all the best-fitting solutions, genotypes from Table 2 were combined with other genotypes found by the PGA over the 200 generations with fits < 1. The results for the parameter values composing these genotypes are plotted in Figure 2. Here, subjects on the X-axis are ordered by number of attempts from low to high performance, and show the range of performance for ANS, BLC, and SYL. The range for each parameter across all subjects is represented by the vertical bar on the right-end of each plot. A line connects the best-fitting genotypes from Table 2.

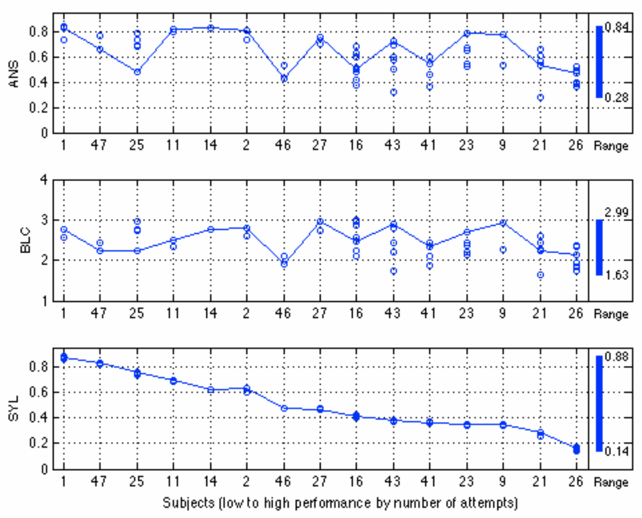

Figure 2: Distributions of ACT-R parameters values (ANS, BLC, SYL) for genotypes with fitness < 1.0 ordered by number of attempts. Small circles represent parameter values. A line connects parameter values associated with the fittest genotypes for each subject. Vertical bars at right show the range of parameter values across all subjects.

At the subject-level each plot indicates a *distribution* of ACT-R parameter values when a nearly perfect fit is achieved. The number of circles per subject equates to the number of nearly perfect fits produced by the PGA. For example, the PGA produced 9 nearly perfect fits for S16's data (listed in column 3 of Table 3). Therefore, each plot shows 9 circles for each parameter value for S16. Columns 4, 5, and 6 of Table 3 list the minimum and maximum values for each parameter by subject.

If the circles appear stacked, such as the case for SYL, this means the values for SYL were stable across genotypes producing nearly perfect fits. In addition to stability, the SYL parameter shows the expected downward trend previously discovered in Table 2.

Table 3: Range of parameter values (ANS, BLC, SYL) for genotypes with fitness < 1.0 by subject number.

Subject	Human Data	Fittest	ANS Range	BLC Range	SYL Range
1	28, 67.9	3	0.73–0.84	2.57–2.76	0.86–0.88
47	29, 62.1	2	0.66–0.77	2.25–2.43	0.82–0.83
25	31, 80.7	5	0.48–0.79	2.25–2.95	0.74–0.76
11	35, 65.7	2	0.80–0.82	2.35–2.49	0.68–0.69
14	37, 75.7	1	0.83	2.75	0.62
2	37, 78.4	3	0.73–0.81	2.59–2.80	0.60–0.63
46	45, 80.0	2	0.43–0.53	1.90–2.12	0.47
27	46, 87.0	2	0.70–0.76	2.74–2.96	0.46–0.47
16	50, 92.0	9	0.37–0.68	2.09–2.99	0.40–0.42
43	54, 89.0	6	0.32–0.72	1.74–2.88	0.37–0.38
41	55, 87.3	4	0.36–0.60	1.86–2.44	0.36–0.37
23	57, 84.2	6	0.52–0.79	2.15–2.71	0.34–0.35
9	57, 87.7	2	0.53–0.78	2.28–2.92	0.34–0.35
21	65, 90.8	5	0.28–0.66	1.63–2.59	0.26–0.29
26	83, 94.0	6	0.36–0.52	1.75–2.37	0.14–0.17
Avg	47, 82.0	4	0.25–0.76	1.59–2.65	0.44–0.45

The two other parameters in the genotype, ANS and BLC, show a distribution of values as non-stacking circles aligned vertically for each subject. These distributions appear to widen as performance increases. Noted in Table 3, for example, when fitting the model to S43's data, ANS values ranged from 0.32 to 0.72 (difference 0.4), BLC values ranged from 1.74 to 2.88 (difference 1.14), and SYL varied by only 0.01. This subject showed the greatest variability in ANS and BLC values overall.

When considered across subjects, ANS and BLC have pattern similarities. The increases and decreases for the best-fitting genotypes connected by a line mirror one another. For example, both ANS and BLC drop for S46, then both spike for S27, then both drop again for S16. Evidence of a pattern is not that surprising—ANS and BLC are both parameters in ACT-R's Declarative Module used to calculate chunk activation. Their relationship would be important for the modeler to understand, for example, if a specific range of ANS was validated for a task, then the corresponding range for BLC could be estimated.

Parameter Space Visualizations

The distributions of parameter values for ANS and BLC producing nearly perfect fits were an unexpected finding A 3-D visualization method called slice planes was used to investigate the situation further (Figure 3).

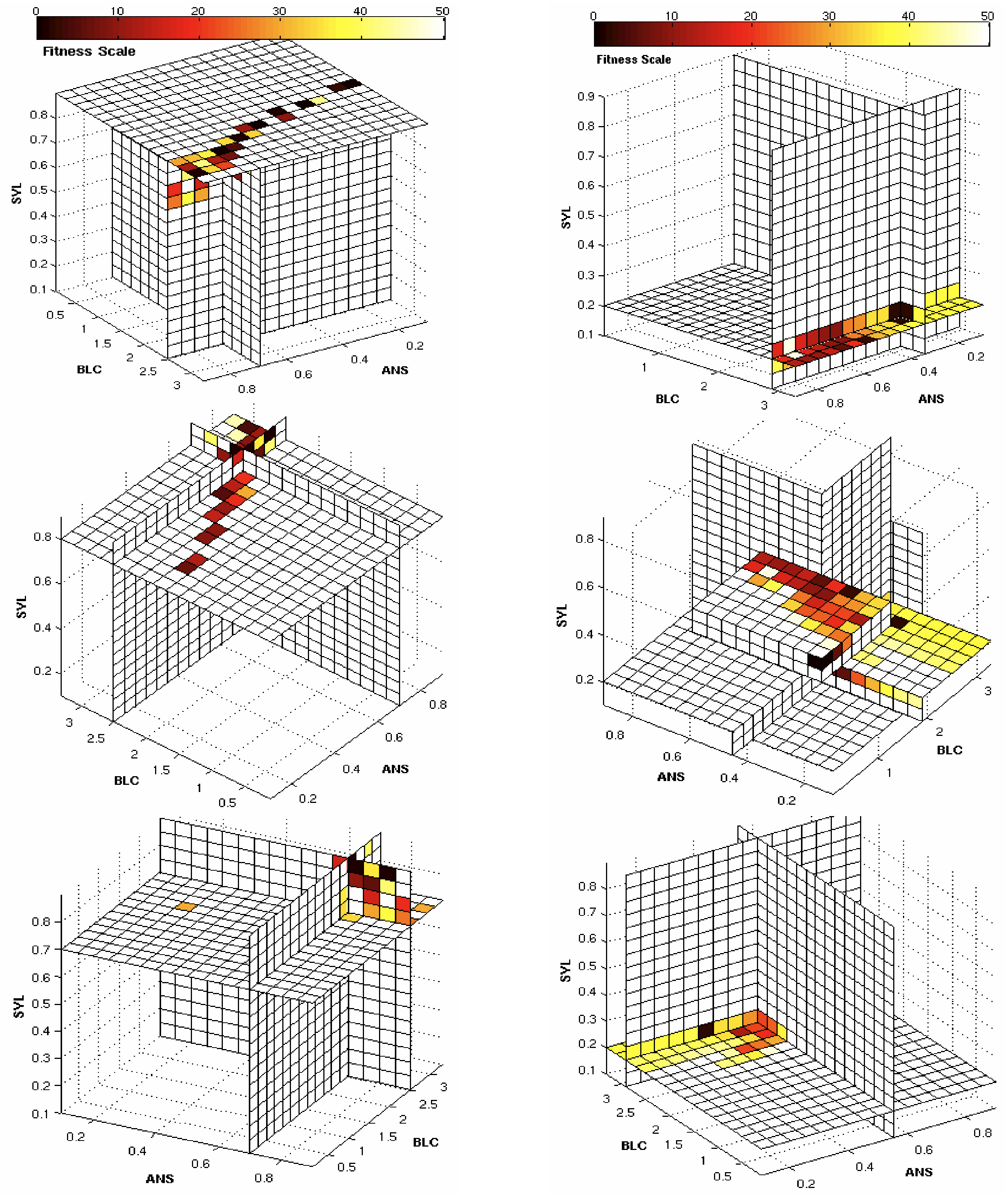

Figure 3: 3-D plots using slice planes to visualize the parametric space when fitting the model to S1's performance (left-side plots) and S26's performance (right-side plots). Fitness scale is shown at the top.

Two preliminary data volumes were constructed from coarse-grained grids of 4913 data points per volume. A data point is a parameter set and associated average fitness value calculated across 20 model runs—totaling 98,260 model runs consuming approximately 30 CPU hours per volume. The data is scalar with XYZ representing ANS, BLC and SYL, respectively. Fitness is the 4th dimension or V represented by a color bar scale from black (fitness = 0, perfect fit) to white (fitness > 50).

Figure 3 shows the parametric space when fitting to S1 (worst performer, left-column plots) and S26 (best performer, right-column plots). The slice planes are positioned and perspectives rotated to illuminate locations of nearly perfect fit within the parameter space.

In S1's plots, the ANS and BLC slice planes remain at a constant position, while the slice plane for SYL is initially positioned at 0.9 in the top plot and then incrementally lowered by 0.1 in the second and third plots. For S26's plots, the SYL slice plane remains at a constant position of 0.2; ANS is incremented by 0.1 from 0.35 to 0.55; and BLC is positioned from 2.9 to 1.9 then back to 2.9.

Figure 3 shows the small and scattered range of useful values—multiple data points resulting in nearly perfect model-to-data fits for both subjects (black squares). S1's visualizations show a band-like pattern of fair to good fits perforated by several nearly perfect fits. In contrast, the nearly perfect fits for S26 are massed in an area near the upper constraint of BLC and lower constraint of SYL.

As noted in Table 3, the PGA found twice as many nearly perfect fits for S26 then for S1. The limitations of this research, one data set and one model, restrict interpretative insights gained by the visualizations. Different patterns of nearly perfect fits might indicate types of strategies attempted by subjects performing the task while under different levels of stress.

Conclusion

Why the 'expedition' characterization of this research? An alternative approach to cognitive model fitting leads to a succession of unexpected results, interesting questions, and possible atypical analyses. What we have learned: (1) When variability in the data makes fitting to mean performance difficult and mean performance does not resemble any particular subject's performance, individual data fitting is a viable approach, (2) Non-bias fits produced by optimization algorithms mitigate the bias present in human (and modeler) problem solving, (3) Stochasticity detected in the architecture and model is an example of how multiple model runs help in understanding the distribution of model predictions, (4) Finding sets of good fits instead of only one may infer an architectural issue, flawed model or architecture, or alternative strategies for doing the task, (5) Visualizations of the prediction landscape are useful in understanding a cognitive architecture and how different combinations of parameters effect a model's performance.

Acknowledgments

This project is partially supported by ONR grant N000140310248. Computational resources were provided by TeraGrid grant DAC TG-IRI070000T. The authors would like to thank Laura Klein and her lab at the Department of Biobehavioral Health, Penn State University, for collection of the human performance data. David Davis and Suzanne Shontz have helped our thinking in this area.

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.

Gobet, F., & Ritter, F. E. (2000). Individual Data Analysis and Unified Theories of Cognition: A methodological proposal. In *Proceedings of the 3rd International Conference on Cognitive Modelling*, 150-157. Universal Press: Veenendaal, The Netherlands.

Goldberg, D. (1994). Genetic and evolutionary algorithms come of age. *Communications of the ACM , 37*, 3, 113-119.

Kase, S. E., Ritter, F. E., & Schoelles, M. (2007). Using HPC and PGAs to optimize noisy computational models of cognition. *International Joint Conferences on Computer, Information, and System Sciences, and Engineering*.

Kase, S. E. (forthcoming). A parallel genetic algorithm approach for individual data modeling optimization of cognitive tasks. Unpublished PhD thesis, College of IST, The Pennsylvania State University.

Kirschbaum, C., Pirke, K. M., & Hellhammer, D. H. (1993). The Trier Social Stress Test—A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28, 76-81.

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part 1: Basic mechanisms. *Psychological Review, 104*, 1, 3-65.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Ritter, F. E., Schoelles, M., Klein, L. C., & Kase, S. E. (2007). Modeling the range of performance on the serial subtraction task. *Proceedings of the 8th International Conference on Cognitive Modeling*. (pp. 299-304). Oxford, UK: Taylor & Francis/Psychology Press.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General*, *116*, 250-264.

Whetzel, C. A., Ritter, F. E., & Klein, L. C. (2006). DHEA-S and cortisol responses to stress and caffeine in healthy young men: Is DHEA-S a reliable marker for stress? *Psychosomatic Medicine, 68*, A-77.