RICE UNIVERSITY

Using Simulation to Assess Prediction Performance Change with Simulated Annealing on Probability Arrays

by

Jason Deines

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Master of Arts

Approved, Thesis Committee:

James R. Thompson, Chairman
Noah Harding Professor of Statistics

Katherine B. Ensor

Katherine B. Ensor
Professor of Statistics
Chair, Department of Statistics

Sharad W. Borle
Assistant Professor of Management
Jones Graduate School of Management

Houston, Texas

July, 2004

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Using Simulation to Assess Prediction Performance Change with Simulated Annealing on Probability Arrays

by

Jason Deines

Experimental results suggest that significant improvements in forecast performance can be obtained by applying the simulated annealing on probability arrays (SAPA) algorithm to grouped event probability forecasts. Such forecasts are frequently probabilistically incoherent, even when elicited expert subjects. The algorithm corrects any incoherence within the set of responses from each subject, while at the same time minimizing the sum of the absolute adjustments made to the original probability estimates. These adjusted coherent probability estimates appear to yield improved overall forecast performance, as measured by several different metrics. However, with the only published results consisting of several small experiments, definitive conclusions regarding potential forecast improvements in wider applications are difficult to justify. To address this lack of experimental data, a method for extending the exist-

ing published results using simulation is described, and the SAPA algorithm and its effects on forecast performance are examined.

Acknowledgements

I would first like to thank Dr. Daniel Osherson, now at Princeton University, for facilitating my participation on this project. During my time at Rice I have enjoyed the help and support of numerous faculty members. In particular, I would like to thank the members of my thesis committee: Dr. Katherine Ensor, Dr. Sharad Borle, and Dr. James Thompson, who served as my thesis advisor. Dr. Thompson's guidance and counsel were instrumental throughout my research and writing efforts.

I also wish to express my appreciation to Dr. David Scott, who generously provided me with the opportunity to work under him this academic year, and to Dr. Peter Olofsson for his help and advice. Finally, my efforts would not have been successful without the support of my friends and family. My deepest thanks to everyone.

Research supported by NSF grant IIS-9978135.

Contents

Abstract ii

Acknowledgements iv

List of Figures vii

List of Tables ix

1 Introduction 1

1.1 Description of the Problem 1

1.2 Methods to Correct for Incoherence 2

1.3 Simulated Annealing on Probability Arrays 4

2 Analysis Using Simulated Hybrid Experiments 7

2.1 Methodological Basis 7

2.2 Implementation Details 11

2.3 Pcrformancc Mctrics 13

3 Analysis 17

3.1 Published Results and Verification 17

3.2 Inferences from Published Results 19

3.3 Simulation Results from Original Response Data 21

3.4 Simulation Results from Perturbed Data 29

4 Conclusions 38

4.1 Summary of Findings 38

4.2 Practical Considerations 39

4.3 Possible Future Research 41

A Finance experiment variables 42

Bibliography 44

List of Figures

3.1 SAPA results after iteration steps on raw data 22

3.2 Paired-comparison t-test for fit, MSE, MAD, and slope, 100 trials . . 23

3.3 Correlation between fit, MAD, MSE and slope by iteration, 100 trials, raw data 24

3.4 SAPA fit on raw data 25

3.5 MAD after SAPA on raw data 25

3.6 MSE after SAPA on raw data 26

3.7 Slope after SAPA on raw data 26

3.8 Paired-difference t-test statistics for raw response data 27

3.9 Paired-difference t-test statistics for data perturbed towards correct . 31

3.10 SAPA fit, perturbed towards correct, level 0.5, rate 1.0 31

3.11 MAD after SAPA, perturbed towards correct, level 0.5, rate 1.0 . . . 32

3.12 MSE after SAPA, perturbed towards correct, level 0.5, rate 1.0 32

3.13 Slope after SAPA, perturbed towards correct, level 0.5, rate 1.0 . . . 33

3.14 Paired-difference t-test statistics for data perturbed towards incorrect 33

3.15 SAPA fit, perturbed towards incorrect, level 0.5, rate 1.0 34

3.16 MAD after SAPA, perturbed towards incorrect, level 0.5, rate 1.0 . . 34

3.17 MSE after SAPA, perturbed towards incorrect, level 0.5, rate 1.0 . . . 35

3.18 Slope after SAPA, perturbed towards incorrect, level 0.5, rate 1.0 . . 35

List of Tables

1.1 The four laws governing coherence 3

Chapter 1

Introduction

1.1 Description of the Problem

When asked to estimate probabilities of sets future events containing compound or conditional structures, human respondents frequently give estimates that are probabilistically incoherent. The following example, similar to an example from Deines, Osherson, Thompson, Tsavachidis and Vardi (2002) [3], illustrates the issues involved. An individual gives their probability estimates of the following events occurring:

a. Prob(Clinton is re-elected to the Senate in 2006) = 0.75

b. Prob(Giuliani runs for the Senate in 2006) = 0.50

c. Prob((Clinton is re-elected to the Senate in 2006) and (Giuliani runs for the Senate in 2006)) = 0.10

The probability of a conjunction of the form p $\wedge$ q is bounded above by the minimum of the individual probabilities of p and q, and bounded below by sum of the probabilities of p and q minus 1. In this example, if estimates a and b are maintained, the bounds dictate that the conjunction of the two estimates must lie on the interval $[0.25, 0.50]$. The estimate for c, 0.10, is therefore inconsistent, and the group of estimates (a, b, c) are probabilistically incoherent. It is important to note that it is not estimate c alone that is incoherent--the estimates must be considered as a group. Were one to maintain estimates a and c, for example, the lower bound on the conjunction of 0.10 would force the estimate for b to be no higher than 0.35 (since $0.75 + 0.35 - 1.0 = 0.10$).

Common sense dictates that as the number of interconnected judgments increases, maintaining probabilistic coherence becomes increasingly difficult. In practice, when grouped with their probability estimates of the fundamental events, even expert respondents almost always produce some incoherent estimates of the probabilities of compound or conditional events. This tendency is well-known and has been extensively studied for many years [3].

1.2 Methods to Correct for Incoherence

Several methods have been proposed to correct for this incoherence of probability estimates. These methods fall into two broad categories: on-line and off-line. In an

1	$\text{Prob}(p \wedge q)$	$\leq$	$\min[\text{Prob}(p), \text{Prob}(q)]$
	$\text{Prob}(p \wedge q)$	$\geq$	$\text{Prob}(p) + \text{Prob}(q)$ - 1
2	$\text{Prob}(p \wedge \neg q)$	$\leq$	$\min[\text{Prob}(p)$, 1 - $\text{Prob}(q)]$
	$\text{Prob}(p \wedge \neg q)$	$\geq$	$\text{Prob}(p)$ - $\text{Prob}(q)$
3	$\text{Prob}(p \vee q)$	$\geq$	$\max[\text{Prob}(p), \text{Prob}(q)]$
	$\text{Prob}(p \vee q)$	$\leq$	$\text{Prob}(p) + \text{Prob}(q)$
4	$\text{Prob}(p \vee \neg q)$	$\geq$	$\max[\text{Prob}(p)$, 1 - $\text{Prob}(q)]$
	$\text{Prob}(p \vee \neg q)$	$\leq$	$\text{Prob}(p)$ - $\text{Prob}(q) + 1$

Table 1.1: The four laws governing coherence (from [3], page 48)

on-line incoherence correction method, as the questions are asked of the respondent, the range of coherent estimates is presented at the same time. For the example given in 1.1, the subject might be asked to estimate the probability of events a and b. The respondent would be allowed to make an estimate from the full range of probability [0, 1], since events a and b are stochastically independent. Having given probability estimates for events a and b, when questioned for an estimate for event c, the subject would be limited to the range of coherent responses, from 0.25 to 0.50, in the example,. However, when estimating a large set of events, the order of questioning could influence the estimates given [3]. In the method just described, the initial, stochastically independent responses would remain unchanged, with the restrictions imposed only on those complex or conditional constructs where incoherence might present itself.

An off-line correction method does not impose such restrictions, and as such could be considered a superior solution to the problem. Off-line correction also affords greater flexibility in administering sets of event estimates to human participants. A paper questionnaire can suffice, with the responses later converted into a machine-readable format for adjustment by computer.

1.3 Simulated Annealing on Probability Arrays

The simulated annealing on probability arrays (SAPA) algorithm is an off-line method conceived by Daniel Osherson and first described in Batsell, Brenner, Osherson, Tsavachidis and Vardi (2002) [1]. Like other off-line methods, SAPA adjusts sets of estimates to make them probabilistically coherent. A summary of the four probability laws governing coherence are shown in Table 1.1.

The motivating principle behind the algorithm is straightforward. Adjustments are made to a given set of probability estimates to bring the set into a coherent probabilistic state, while simultaneously attempting to minimize the sum of the absolute changes made to the original subject estimates. In essence, SAPA combines an off-line incoherence correction method with an iterative optimization procedure. A variety of suitable optimization techniques exist, and before simulated annealing was adopted, precursors to SAPA used a optimization method based on genetic algorithms. Simulated annealing was ultimately chosen because it was significantly faster than the genetic algorithms implementation while achieving essentially equivalent results. Speed is frequently cited as one of the benefits of simulated annealing optimization, so the efficiency improvement is not unexpected. Many references exist comparing the relative advantages and disadvantages of these and other optimization methods [4, 6, 10, 13]. Technical details of the algorithm can be found in [8] and will not outlined in greater detail here.

Logic suggests that the minimization of total change is a desirable characteristic, under the assumption that experts, even if providing estimates that are incoherent when taken as a group, are likely to have insight on the level of individual event estimates [3]. It follows that an incoherence correction method should attempt to disrupt these initial probability estimates as little as possible, while still attaining the goal of coherence. Potentially more intriguing, however, are the findings from experimental results that the resulting coherent estimates generated by SAPA yield superior forecasting performance when compared to their incoherent counterparts [3, 1]. If true, the algorithm would be potentially useful in numerous applications involving groups of experts forecasting sets of of future events.

Unfortunately the published experimental results have several significant limitations, making justification for large-scale experimental evaluation or for applied use difficult on the basis of these results alone. The experimental results published were generally of a proof-of-concept nature—to illustrate the behavior and application of the algorithm to an easily definable problem. This is entirely reasonable approach given the time frames and goals for academic research. In contrast, rigorous, long-term experiments require far more effort, resources, and time, and may yield no marginal benefit over simpler experiments for research purposes. The lack of such rigorous experimental results, however, limits what inferences can be made about the forecast performance benefits of SAPA. Conversely, without compelling evidence that conducting large-scale experimentation is worth the investment involved, it is unlikely such experiments will be undertaken.

Of the six sets of experimental results published, only two, both involving economic events, could be considered typical to what one might encounter in a business or public-policy application. The remaining experiments focused on such things as predicting aspects of the outcome of sporting events [1]. Another limitation is that the experiments generally did not involve subjects likely to have above-average insight in the topic area of interest. Most participants were undergraduate students with no particular specialization or area of expertise, although in some cases at least a percentage of the experimental participants possessed expertise in the subject area of interest. Finally, most predictive time frames were short, with the longest experiment running for three months. This experiment, referred to as "Finance" in [3], was also the most comprehensive of those published. The structure and participant responses of the Finance experiment formed the basis of generating experimental results via simulation, which enabled more extensive examination of SAPA performance under a variety of controlled subject response conditions.

Chapter 2

Analysis Using Simulated Hybrid Experiments

2.1 Methodological Basis

Using computer simulation to generate and run experimental results to test SAPA has several advantages over human subject experimentation. Simulated experiments take only seconds to run, require little effort once the simulation code is written, and the subject responses can be modified in a controlled fashion. This speed and control allows examination of the algorithm using large numbers of repeated trials, under a variety of response conditions. Such a volume of results and control is not possible using experiments with human participants.

The following outlines a general framework for testing SAPA by utilizing repeated simulation. The simulation generates an arbitrary array of events, along with corresponding outcomes. Sets of questions are created from the base event array to form a questionnaire, and for each simulated subject, the questionnaire responses are generated according to algorithmic guidelines. The set of completed questionnaires is analyzed to determine the base performance of the simulated group. Finally, SAPA is applied to the batch of simulated data, and the SAPA-adjusted results are analyzed for comparison purposes. The process can be repeated an arbitrary number of times, changing parameter values as desired.

This approach, termed the "pure" simulation model, has two significant difficulties in practice. The first is that a thorough simulation implementation would be complex and have many free parameters. Important parameters to consider include the number of events, the number of questions, the structure of the questions and of the questionnaire, and the number of subjects. Beyond this there would be a number of additional parameters controlling the simulated subject responses, such as relative expertise (how likely the subject is to achieve a relatively low forecast error score by some performance measure), and the probability of incoherence on complex constructs. To ensure stable results and for confidence in any inferences made, the simulation would have to be run repeatedly with the same parameters, and there may be a wide range of parameters of interest. A complete simulated analysis, therefore, may still take a long time to complete. This is not a weakness of the simulation technique *per se*, but proper application of a pure simulation for analyzing the SAPA algorithm would still require considerable effort and time.

The second problem is best categorized as a philosophical issue. One could argue that results from repeated simulation, based on randomized values controlled by a set of parameter values, would not accurately demonstrate how human participants would respond in a real application situation. In essence this revisits the concern regarding the numerous parameters that must be identified and appropriately controlled as part of the simulation. The argument is extended in a philosophical sense, however, effectively stating that it is impossible to completely identify and control all the relevant parameters, and that the simulation would not capture all of the nuances in response that would be present in human subjects. While true, it is a criticism that can be leveled at almost any attempt at modeling system behavior, whether involving human participants or not. Generalizations and approximations of systems, often surprisingly crude ones, can still yield useful information.

There is, however, a variation on the pure simulation paradigm that has advantages in both problem areas, when compared with the pure simulation model as described. Rather than generate all the of the response data for the simulation using parameterized random variables, this method, termed the "hybrid" simulation method, starts instead from an existing experimental data set actually administered to human participants. The method uses the human subject response data as a basis for the randomized perturbations of responses that are used in the simulation trials.

With human responses as its underlying basis, the method is potentially more likely to incorporate subtle interrelationships between the individual estimates within a set of events than pure randomized data. This requires the structure of event sets used in the experiment to be utilized unchanged, however, and thus greatly simplifies the simulation problem. By constraining the structure of the experimental format to that administered to human subjects, the need to generate the experimental structure by simulation is eliminated.

The hybrid simulation method also affords an established, baseline reference to compare the performance of the SAPA algorithm before and after changes to the response data are made. Since a known performance standard has already been established from the human subject experiment, this known starting point can be used to make various controlled alterations in the data, from which performance measurements can be repeatedly taken. For example, if the characteristics of the actual human subject data indicate poor forecast performance, and hence likely no real expertise or insight, this can be corrected using several approaches. Since the true outcome of the events is known, *post-hoc* adjustments can be made to the human subject estimates, thereby yielding better forecast performance. Making these changes allows the performance of the simulated subjects to be adjusted as desired, allowing the performance of SAPA to be studied under various experimental conditions.

2.2 Implementation Details

As previously mentioned, the hybrid simulation technique used the Finance experiment data set as the basis for all simulated input and resulting SAPA results. The author was largely responsible for creating the set of 30 event questions that formed the experiment, which was designed with several goals in mind. The target audience was second-year MBA students at Rice University in Houston, Texas. It was reasoned that by participating in a full-time MBA program and by the corresponding self-selection and curriculum exposure, the second-year MBA students would likely demonstrate a higher-than-average level of knowledge and expertise regarding the economic and financial events under consideration. In addition, between their first and second years in the MBA program, almost all students participate in a summer internship, frequently with a Houston-based company. This factor was also incorporated in the event selection process, and questions involving specific company performance emphasized businesses based locally. The expectation was that at least a few participants would encounter questions involving the prospects of companies with which they had recent first-hand experience, and it was also more likely that the participants would have greater familiarity with local companies in general. A list of the 30 event variables that were used for the generated questionnaires can be found in Appendix A.

As a performance incentive, the subjects with the best forecast results were to be awarded prizes at the end of the forecast period. The criteria for forecast performance was explained to all participants beforehand, and was defined as the mean squared error between the forecast probabilities and actual outcomes. The exepected forecast period was to be six months long, from 1 October 2001 through 31 March 2002, meaning that the subjects were to give their estimates effective as of 1 October 2001, and the performance of the predictions would be measured as of 31 March 2002. Economic and financial data are generally considered "noisy," and evidence of true future insight into overall trends would be more apparent the longer the forecast horizon used. In addition, many economic performance measures are only reported on a monthly basis, so even after six months, an event based upon an economic indicator might be determined from the outcome of only six data observations. A longer time period would have been desirable, but logistically it was not practical to combine a longer period with the desired performance incentives. Even extending to a nine-month period, for example, would take the end date to 30 June 2002, by which time the second-year MBA students would have graduated, making prize distribution considerably more difficult.

Several issues arose to confound the initial design goals, and may have adversely affected the experimental results. The terrorist attacks of 11 September 2001 created tremendous uncertainty regarding short and medium-term impact to the economic and business climate, only weeks before the start of the prediction period on 1 October 2001. In a desire to produce results as quickly as possible, initially the time horizon for the forecasts was only one month in duration, despite the experimental design assumptions. Given the uncertainty from the terrorist attacks, however, the time horizon was extended to three months, with the period of interest running from 1 October 2001 to 31 December 2001. Nevertheless, this was likely still too short a prediction window. Finally, the scandal involving the most prominent Houston-based firm at the time, Enron, unfolded during the forecast period. Given Enron's stature in the local economy and the large numbers of Rice University MBA interns and graduates the company hired, Enron was an obvious corporate subject to include in the question pool. Since the vast majority of professional security analysts failed to predict the difficulties Enron would face in the latter months of 2001, in retrospect Enron was a poor choice to include as an event topic for the experiment.[1]

2.3 Performance Metrics

The primary forecast performance measurement reported for the published experimental results is the "quadratic penalty" ([11], cited in [3]). It is a measurement of squared error from the true outcome value, and is computed as follows. The probability forecast for an event is designated as *Prob*. For events that occur (the event is true), *Outcome* is equal to 1; for events that do not occur (the event is false), *Outcome* is correspondingly equal to 0. In all but conditional events, the quadratic score is simply $(\mathit{Outcome} - \mathit{Prob})^2$, or the squared error. For error scores of conditional events,

[1] Another Houston-based company, Compaq, was initially included as well. It was removed shortly before the experiment began because of an unexpected buyout offer from Hewlett-Packard.

however, if the conditioning event does not occur, the question has no inherent meaning, and the question is discarded from the set of responses. If the conditioning event does occur, the quadratic score for the conditional question is scored in the same manner given above. The sum of all the resulting quadratic score values is divided by the number of retained responses, giving the average quadratic score, equivalent to the mean squared error. For results generated by hybrid simulation, the more statistically common term "squared error" and "mean squared error" (MSE) will be used, but the value is computed in an identical fashion as the quadratic penalty, and the corresponding average value of "mean quadratic penalty" or "mean quadratic deviation" (MQD).

The other measure of a subject's forecast performance utilized in the published results is called the "slope." (In [1], the same measure is called the "discrimination index" and [3] cites a discussion of the measure found in Chapter 3 of Yates [12].) For a given subject's set of probability estimates, the slope is calculated as follows:

(mean estimate for "true" events) - (mean estimate for "false" events)

The rationale for the measure is that if a subject, on average, assigns higher probability estimates to outcomes that occur than to those that do not, this indicates some discrimination ability between the two types of outcomes. Positive values are therefore considered an indication of insight. A typical example might have a mean estimate of true outcomes of 0.6, and a mean estimate of false outcomes of 0.4, resulting in a slope of 0.2.

There are some potential difficulties with the use of slope as a measure of performance, however. First, taken in isolation, it is difficult to interpret the potential significance of a given slope value. A small positive value averaged across all subjects might be shown to be statistically significant greater than 0, and thus arguably demonstrate insight, but it is difficult to ascertain how high this value must get before insight of practical significance is shown. Second, although not a problem in the experiment examined here, when applied to a relatively small set of events, it is possible that only very few will fall into a particular outcome. This could happen either by chance, poor questionnaire design, or a combination of the two, and the forecast performance regarding these few events could affect the resulting slope value disproportionately.

The following hypothetical example underscores the caution that must be exercised when considering the slope value alone as a measure of forecast performance. In this example a subject has responded with a mean estimate of true events of 0.9, and a mean estimate of false events of 0.7. As defined, the slope would therefore equal 0.2, which for the purposes of the example is stipulated as statistically significantly greater than 0, therefore indicating insight. Yet the subject is clearly heavily biased towards positive forecasts. The argument in support of the slope as a performance measure is that despite this bias, there is still discrimination between true and false events, so insight is evident. An overall bias in the estimates is nevertheless undesirable, however, and examination of the slope measure alone would not indicate that such a bias existed.

Another common error measure is mean absolute deviation (MAD), computed as the absolute value of the forecast from the outcome. The MAD measure is not cited in the published performance results, and Osherson indicated that MAD has certain deficiencies as an accuracy measure in this experimental context, and therefore performance analysis utilizing it was not performed [7]. Nevertheless, since MAD provides an additional benchmark with a straightforward definition, for the simulated results MAD error measures were also analyzed and are given for comparison purposes.

Chapter 3

Analysis

3.1 Published Results and Verification

Details of the experimental results from the Finance experiment are discussed in three published articles [3, 8, 9]. The results from the Finance experiment were first published in [3] in 2002. As might be expected given the variety of issues confounding the experiment, the forecasting performance of the participants was poor. The MSE across all 47 subjects, before application of SAPA, is 0.314, with a standard deviation of 0.077. After optimization with SAPA, the MSE was reduced to 0.285, with a standard deviation of 0.076 [3]. This improvement after SAPA is stated as significant by correlated t-test, giving $[t(46) = 7.6, p < 0.001]$. Applying the simulation implementation to the unadjusted subject response values gives similar results, with an MSE after SAPA of 0.287, and a standard deviation of 0.072. A paired-sample t-test performed on the original subject MSE values and those after applying the simulation SAPA implementation confirms the published test values, with $t(46) = 6.91, p < 0.001$.

Results are also given for the "slope" measure of forecast accuracy. The average slope for the raw estimates across the 47 subjects is 0.053, with a standard deviation of 0.130. These values are stated as reliably greater than zero $[t(46) = 2.81, p < 0.01]$. As with MSE, application of SAPA improves the slope error measure, increasing the value to 0.115, with a standard deviation of 0.129. This difference is stated as reliable by a correlated t-test $[t(46) = 8.46, p < 0.001]$. Applying the simulation implementation again confirms the published values, yielding a slope value of 0.107, with a standard deviation of 0.1294, and by a paired-sample t-test, $t(46) = 7.24, p < 0.001$.

The most detailed description of the SAPA algorithm's construction and parameter space can be found in [9], and the parameter values used for application of SAPA in the simulation studies were identical to the recommendations in the article. SAPA results obtained using both the original code and the simulation implementation always gave similar results for a given response set, and examination of internal results as the algorithm was in progress were also verified as effectively identical. Finally, several complete code comparisons were made by hand between the original code and the simulation code to confirm algorithmic consistency. Given the non-deterministic nature of the optimization algorithm used in SAPA, identical results are extremely unlikely, but the results from the two implementations were consistently statistically equivalent.

3.2 Inferences from Published Results

Based upon the results just discussed and the tests of statistical significance performed, it is claimed that application of SAPA improves forecast accuracy on the Finance experimental data [3, 8, 9]. In a statistical sense the claim may be valid, but it is hard to justify on several pragmatic levels.

The chain of reasoning for the improved results with SAPA rests completely on its primary distinctive attribute compared with other off-line incoherence correction schemes—that is, the iterative optimization process that minimizes the absolute changes made to the initial subject responses [1]. This is a logical goal if there is insight or expertise to preserve on an individual-estimate level. However, in the absence of any conclusive evidence of insight, it is unclear why applying SAPA would result in improved forecast performance.

In the Finance experiment, the average forecast error measures for the subject group demonstrate a low level of insight at best. By the MSE measure, the group of subjects indicates no evidence of insight or expertise. A constant response of 0.50 for every question would result in a MSE of 0.25, since for the two possible *Outcome* states of 0 or 1, $(\mathit{Outcome} - 0.5)^2 = 0.25$. 0.25 can therefore be considered a "zero-insight" boundary for the MSE measure, since this score is achievable regardless of insight or knowledge. As mentioned earlier, MSE for the initial forecast estimates for all subjects is 0.314, with a standard deviation of 0.077, and is significantly higher than the 0.25 boundary level by t-test. Even after application of SAPA, the improved group average MSE equals 0.285, with a standard deviation of 0.076. The improved values remain significantly greater than 0.25 by a t-test, with $t(46) = 3.476, p < 0.005$.

In contrast to the MSE values, initial subject insight can be claimed using slope as a measure, since the zero-insight boundary with the slope measure lies at zero. The average raw slope scores were reliably higher than zero, as discussed in the previous section, and were significantly improved after adjustment by SAPA. Nevertheless, the small magnitude of the slope score, even after SAPA adjustment, does not make a strong argument in support of subject insight.

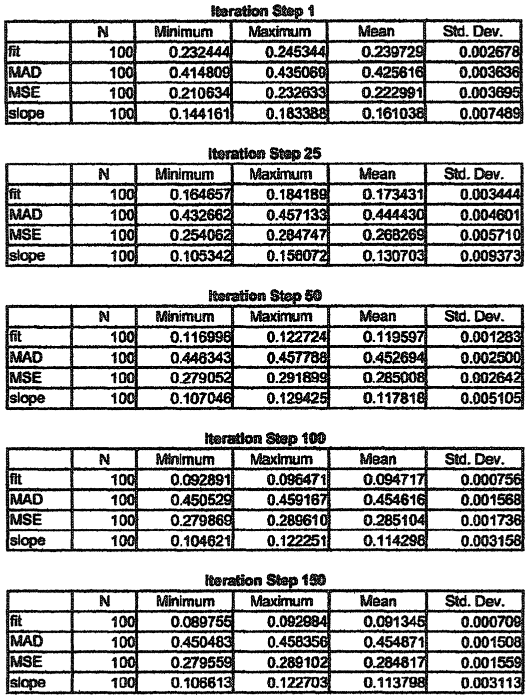
Based upon the published results of the MSE and slope performance measures of the subjects in the Finance experiment, there is conflicting evidence in support of subject insight or expertise, and whatever insight demonstrated by the subject group must be considered low, at best. Logically, there is no reason to expect forecast performance improvements via an incoherence correction mechanism that attempts to approximate the original subject probability estimates as closely as possible, as SAPA does, since the subject estimates are so poor to begin with. Yet the forecast performance measures do show a statistically significant improvement after applying SAPA to the incoherent estimates. Closer examination of forecast performance using the simulation code implementation provides a likely explanation for this behavior.

3.3 Simulation Results from Original Response Data

In addition to the MSE and slope performance measures, the MAD error measure of forecast error was also recorded in the simulation implementation of SAPA. The MAD error measure, which is not reported with the published results, has a zero-knowledge boundary of 0.50. Assuming an equal probability of true and false outcomes and responses and random responses between 0 and 1, the MAD score is equal to the expectation of a uniform[0, 1] random variable, 1/2.

The average MAD for all subjects raw estimates is 0.483, with a standard deviation of 0.0697. This is not reliably different from 0.50 by t-test $[t(46) = -1.174, p = 0.088]$, and consequently does not demonstrate significant subject insight. Nevertheless, after applying SAPA, the adjusted estimates again improve, to the point where they are reliably lower than 0.50. The adjusted average MAD is 0.458, with a standard deviation of 0.0654, and by t-test, $t(46) = -4.381, p < 0.001$. Once again, by the MAD measure there is little evidence to support overall group insight, yet application of SAPA to the estimates yields a small but significant forecast performance improvement.

That SAPA adjustment was able to make small but statistically significant improvements in the average error measures, despite the poor quality of subject forecasts in the Finance experiment, might cause one to conclude that SAPA can extract performance improvements under even the most challenging of circumstances. However, the additional measurement capabilities in the simulation implementation of SAPA,



Iteration Step 1

	N	Minimum	Maximum	Mean	Std. Dev.
fit	100	0.232444	0.245344	0.239729	0.002678
MAD	100	0.414809	0.435069	0.425616	0.003636
MSE	100	0.210634	0.232633	0.222991	0.003695
slope	100	0.144161	0.183388	0.161038	0.007489

Iteration Step 25

	N	Minimum	Maximum	Mean	Std. Dev.
fit	100	0.164657	0.184189	0.173431	0.003444
MAD	100	0.432662	0.457133	0.444430	0.004601
MSE	100	0.254062	0.284747	0.268269	0.005710
slope	100	0.105342	0.156072	0.130703	0.009373

Iteration Step 50

	N	Minimum	Maximum	Mean	Std. Dev.
fit	100	0.116998	0.122724	0.119597	0.001283
MAD	100	0.446343	0.457788	0.452694	0.002500
MSE	100	0.279052	0.291899	0.285008	0.002642
slope	100	0.107046	0.129425	0.117818	0.005105

Iteration Step 100

	N	Minimum	Maximum	Mean	Std. Dev.
fit	100	0.092891	0.096471	0.094717	0.000756
MAD	100	0.450529	0.459167	0.454616	0.001568
MSE	100	0.279869	0.289610	0.285104	0.001736
slope	100	0.104621	0.122251	0.114298	0.003158

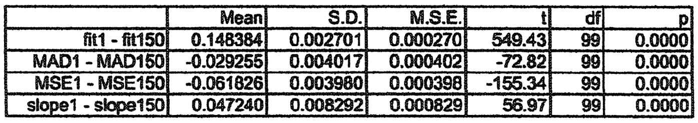
Iteration Step 150

	N	Minimum	Maximum	Mean	Std. Dev.
fit	100	0.089755	0.092984	0.091345	0.000709
MAD	100	0.450483	0.458356	0.454871	0.001508
MSE	100	0.279559	0.289102	0.284817	0.001559
slope	100	0.106613	0.122703	0.113798	0.003113

Figure 3.1: SAPA results after iteration steps on raw data.

along with the capability to make controlled changes in initial subject estimates, provide evidence that suggest this is not the case. The performance gains noted in publication are likely the result of the incoherence correction alone, and not becuase of the unique approximation optimization aspect of SAPA.

Evidence for this conclusion is found by examination of the measures of fit and forecast accuracy at each optimization iteration in SAPA, rather than simply comparing measures of accuracy before SAPA and after the final optimization iteration. (150 iteration steps were used for all results, the same value used in the original SAPA code implementation.) 100 simulation trials were performed with different randomized starting values for the SAPA algorithm, with each trial applying SAPA to the



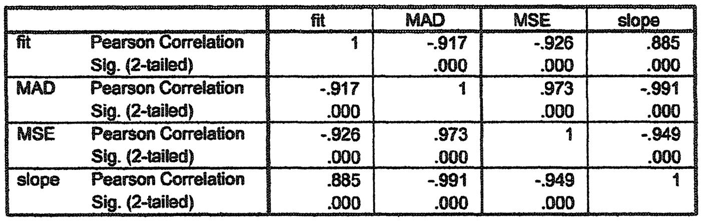
	Mean	S.D.	M.S.E.	t	df	p
fit1 - fit150	0.148384	0.002701	0.000270	549.43	99	0.0000
MAD1 - MAD150	-0.029255	0.004017	0.000402	-72.82	99	0.0000
MSE1 - MSE150	-0.061826	0.003980	0.000398	-155.34	99	0.0000
slope1 - slope150	0.047240	0.008292	0.000829	56.97	99	0.0000

Figure 3.2: **Paired-comparison t-test for fit, MSE, MAD, and slope, 100 trials.**

original subject responses. As can be seen in Figure 3.1, although the mean fit measure declines as expected as the optimization progresses starting at at first iteration, as the fit improves and more closely approximates the original subject estimates, the forecast error measures move in the direction opposite from that desired. The MAD and MSE measures increase, and the slope measure decreases. The measure of coherent fit to the original responses is at its worst after the first iteration—over 100 trials, the mean fit is 0.2397 after the first iteration. The mean value of MAD, MSE and slope, however, all attain their best respective values after the first iteration.

Since every iteration results in coherent adjusted estimates, as this is an algorithmic constraint, the coherent estimate approximations resulting from the first iteration serve as a starting point for assessing the level of fit for subsequent coherent approximations. The simulated annealing optimization plays no role in the initial coherent estimate approximations—the coherent approximations are computed using a separate method, based upon either randomized starting values and the initial incoherent estimates, or upon the previous iteration state, as appropriate. Additional details on the method can be found in [8].

It appears that the improvements in forecast performance evident when applying SAPA to the original subject estimates arise from the initial incoherence correction,



		fit	MAD	MSE	slope
fit	Pearson Correlation	1	-.917	-.926	.885
	Sig. (2-tailed)		.000	.000	.000
MAD	Pearson Correlation	-.917	1	.973	-.991
	Sig. (2-tailed)	.000		.000	.000
MSE	Pearson Correlation	-.926	.973	1	-.949
	Sig. (2-tailed)	.000	.000		.000
slope	Pearson Correlation	.885	-.991	-.949	1
	Sig. (2-tailed)	.000	.000	.000	

Figure 3.3: Correlation between fit, MAD, MSE and slope by iteration, 100 trials, raw data.

and that the SAPA algorithm running to completion only provides an improvement in that the coherent estimate fit optimization does not completely eliminate the initial forecast performance improvements realized after the first iteration. For the original subject responses, the mean performance measures across all subjects are MAD = 0.48229, MSE = 0.31418, and slope = 0.05324. Over 100 simulation trials, after the initial iteration of SAPA, the mean MAD declines to 0.42562, the mean MSE declines to 0.22299, and the mean slope increases to 0.16104. Upon completion of the final iteration, however, all the values have worsened: the mean MAD is 0.45487, the mean MSE is 0.28482, and the mean slope is 0.11380. Performing a pairwise comparison between the respective measures at the first and last iteration shows a statistically significant difference in fit and all three forecast error measures ($p < 0.001$ in every case), as shown in Figure 3.2.

Significant correlations of the opposite sign to that desired are also apparent between the fit measure and the respective forecast performance measures. Figure 3.3 shows a bivariate correlation analysis on 15,000 observations for the four measures, recorded from each of the 150 iterations per trial, with 100 simulation trials.

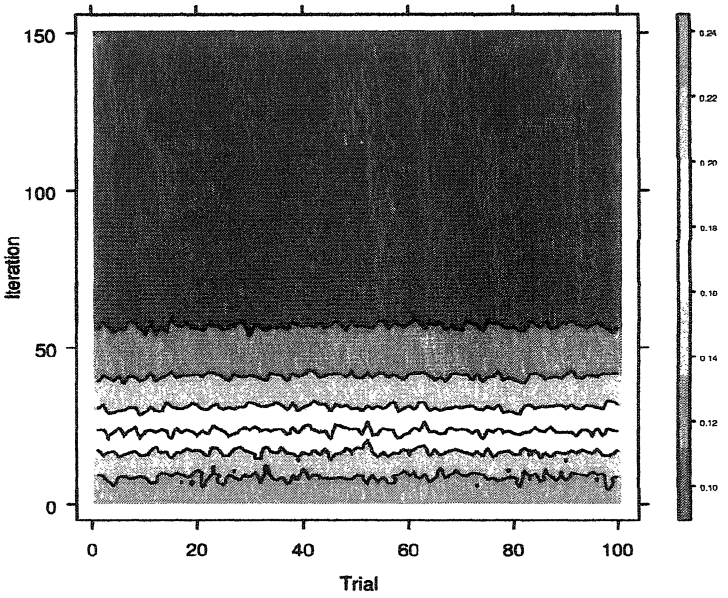


Figure 3.4: SAPA fit on raw data.

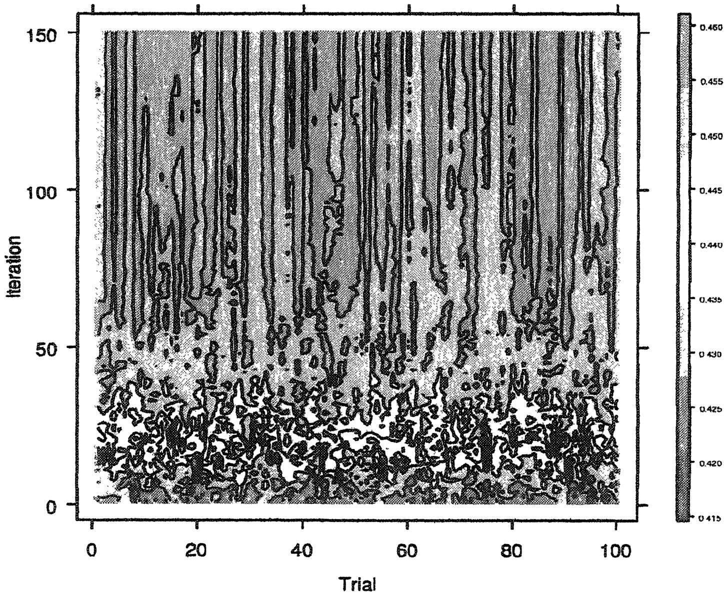


Figure 3.5: MAD after SAPA on raw data.

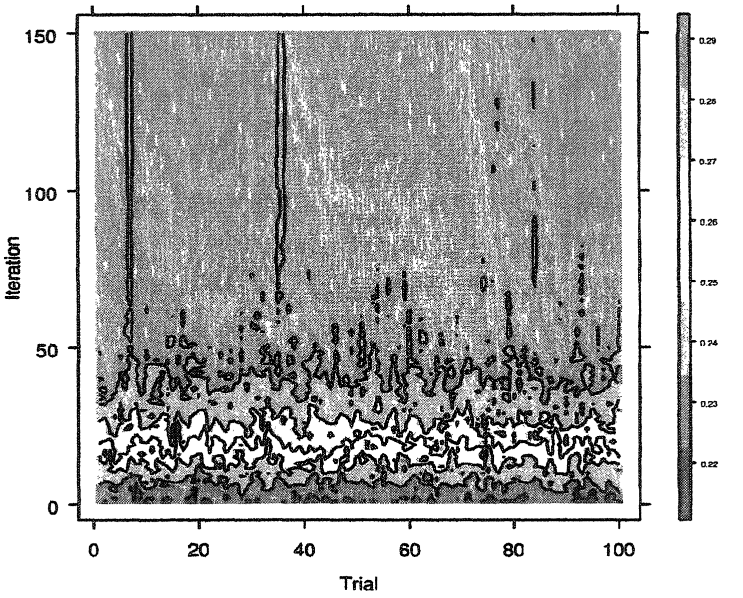


Figure 3.6: MSE after SAPA on raw data.

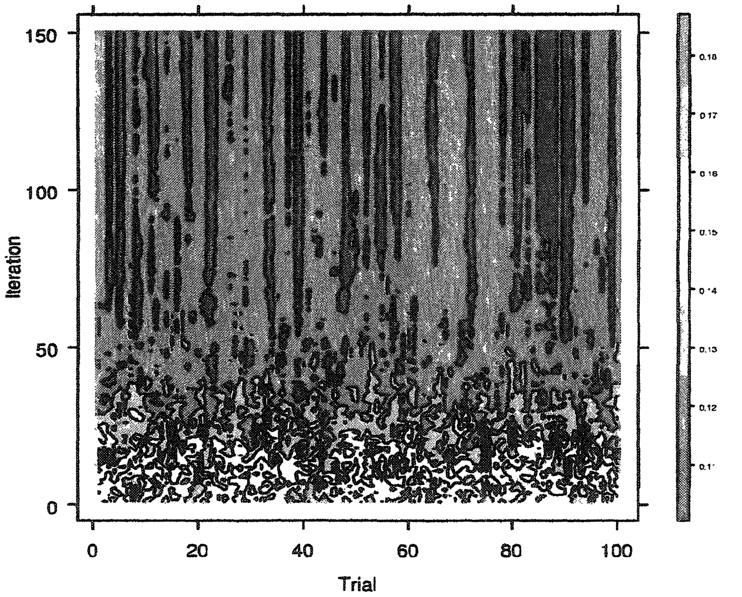
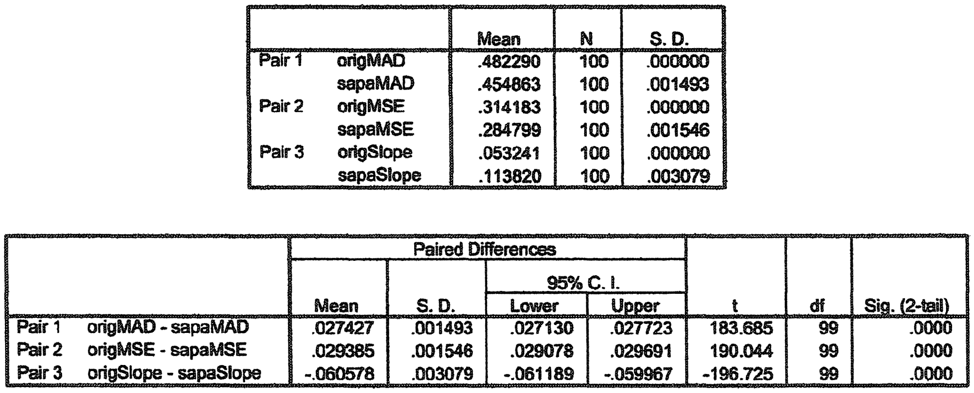


Figure 3.7: Slope after SAPA on raw data.



		Mean	N	S. D.
Pair 1	origMAD	.482290	100	.000000
	sapaMAD	.454863	100	.001493
Pair 2	origMSE	.314183	100	.000000
	sapaMSE	.284799	100	.001546
Pair 3	origSlope	.053241	100	.000000
	sapaSlope	.113820	100	.003079

		Paired Differences						
				95% C. I.				
		Mean	S. D.	Lower	Upper	t	df	Sig. (2-tail)
Pair 1	origMAD - sapaMAD	.027427	.001493	.027130	.027723	183.685	99	.0000
Pair 2	origMSE - sapaMSE	.029385	.001546	.029078	.029691	190.044	99	.0000
Pair 3	origSlope - sapaSlope	-.060578	.003079	-.061189	-.059967	-196.725	99	.0000

Figure 3.8: Paired-difference t-test statistics for raw response data.

Finally, contour plots of the measures afford a visualization of the overall trends. In Figure 3.4, the fit of the coherent estimates is plotted against the iteration step and simulation trial. Across all simulation trials, the fit consistently improves (the value of fit declines) as the optimization progresses, as expected. Looking at the trends of the MAD, MSE, and slope measures (Figures 3.5, 3.6, and 3.7), however, show the accuracy measures worsening as the optimization proceeds. Although the MAD, MSE and slope measures have considerably more trial-to-trial variation than the fit measure, the overall trend is readily apparent. The highest forecast accuracy is achieved in the early iterations of SAPA. Once the fit measure has stabilized close to the minimum value, usually around iterations 60 to 70, the accuracy measures have all deteriorated from their earlier values by a significant degree.

These findings are not surprising, and are in fact consistent with the rationale behind the SAPA design. It makes sense that if there is little to no information content to preserve in the incoherent responses, there is no reason to expect any benefit to be realized from attempting to preserve those estimates as closely as possible when correcting for incoherence.

There is, however, some apparent benefit afforded from the initial coherence correction step, and this behavior has been noted in the existing literature. In [3], work by de Finetti from 1974 is cited, including a theorem that states that a set of incoherent estimates over a set of absolute events (excluding conditional events) can always be replaced by a set of coherent estimates that have lower forecast errors, regardless of the outcome states. In the same work, de Finetti outlines a method to generate coherent approximations to a set incoherent estimates that will result in a lower quadratic penalty (MSE) in all states. However, this method, as with the theory, is not applicable to sets containing conditional events [5, 3]. Since the Finance experiment involved estimates of 12 conditional events out of the 46 total, the theorem is not directly applicable, nor can di Finetti's method be applied. Later work by Bernardo and Smith from 1994 [2] (and also cited in [3]) extended de Finetti's theorem to certain sets of estimates involving conditional events, although this extension is not possible for every possible set of estimates [3].

Despite this limitation, the theory suggests that in many cases, corrected coherent estimates can attain superior forecast error measures. This supports the conclusion that it is the correction for incoherence, produced by the initial coherent approximations to the subject estimates after the first SAPA iteration, that is the source of the improved forecast performance measures in the published experimental results.

3.4 Simulation Results from Perturbed Data

Since the mean subject estimates from the Finance experiment demonstrate little to no insight regarding the set of events forecast, the simulation results showing that the SAPA optimization works contrary to improving forecast performance is not surprising. Such findings are consistent with the theoretical basis of the algorithm, and this consistency could be further confirmed by applying SAPA to subject estimates with significantly higher, or lower, levels of insight. This is the central difficulty the simulation technique was designed to overcome. If the subject estimates from the Finance experiment could be modified in some way to change the resulting insight or expertise level of the participants, applying SAPA to those revised subject estimates might lend additional support for the algorithm's design rationale.

If the initial incoherent estimates are significantly accurate when evaluated individually, it makes sense to preserve that accuracy as much as possible when correcting for incoherence. Under such circumstances, one would expect that as the SAPA coherent approximations fit more closely to the original estimates, the forecast performance would be better when compared with a coherence correction mechanism that does not attempt such a fit. The initial coherent approximation forecast performance measures, calculated after the first iteration of the optimization procedure, can serve for comparison purposes. These initial estimates should be comparatively poor, improving as the coherent approximation estimates more closely fit the original responses. Conversely, if the initial subject estimates are significantly inaccurate, in a manifestation of "reverse" insight, the opposite should be true. As the coherent approximations fit more closely to the lower-insight estimates, the performance measures should worsen.

Several methods to perform adjustments to the original subject responses were implemented and examined, and the results from repeated simulation trials for each of the methods were consistent and similar to the simple adjustment method outlined here. Two parameter values are chosen, *level* and *rate*, each with a range between 0 and 1 inclusive. For each subject response, adjustments are made at random in proportion to the *rate* parameter; if *rate* is 0.5, approximately 50% of the responses will be adjusted (or *perturbed*). Perturbed responses are then adjusted by generating a unform random number between 0 and *level*, and adding or subtracting that value from the existing estimate to bring the result closer to the true, correct outcome. If the response is not perturbed, the original subject estimate is left unchanged. For example, if an event comes true (an outcome of 1), and the initial subject estimate is 0.3, a random value (for example, 0.443) would be added, resulting in a new estimate of 0.743, significantly closer to the correct value. Conversely, false outcomes would have the random value subtracted from the existing estimate. When the sum or difference of the existing estimate and the random value exceed 0 or 1, the new estimate is set at 0 or 1, respectively. A new set of adjusted estimates is generated for each simulation trial, with SAPA applied to each set.

For the results examined here, *rate* was set to 1, and *level* was set to 0.5, meaning

		Mean	N	S. D.
Pair 1	origMAD	.277187	100	.002373
	sapaMAD	.278053	100	.003811
Pair 2	origMSE	.145158	100	.002047
	sapaMSE	.131347	100	.003146
Pair 3	origSlope	.460686	100	.005010
	sapaSlope	.460855	100	.007685

		Paired Differences						
				95% C. I.				
		Mean	S. D.	Lower	Upper	t	df	Sig. (2-tail)
Pair 1	origMAD - sapaMAD	-.000866	.002871	-.001436	-.000297	-3.017	99	.0032
Pair 2	origMSE - sapaMSE	.013811	.002346	.013345	.014276	58.869	99	.0000
Pair 3	origSlope - sapaSlope	-.000170	.005876	-.001335	.000996	-.288	99	.7736

Figure 3.9: Paired-difference t-test statistics for data perturbed towards correct.

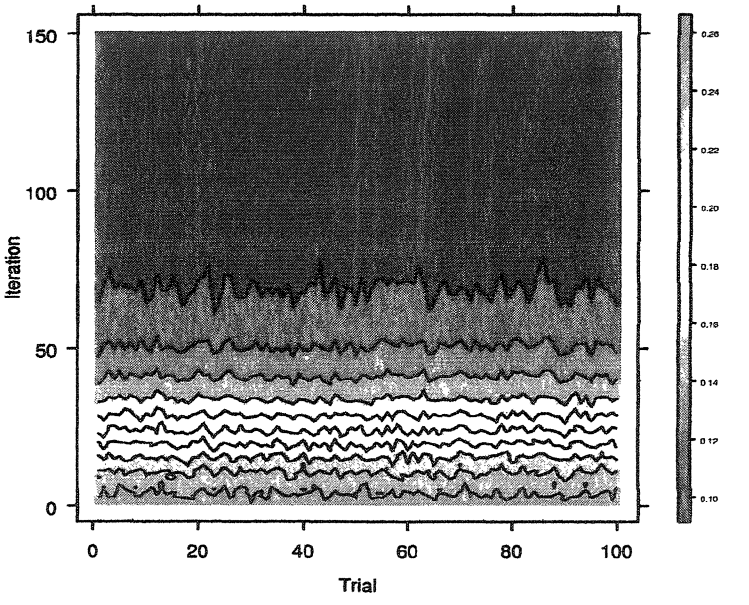


Figure 3.10: SAPA fit, perturbed towards correct, level 0.5, rate 1.0.

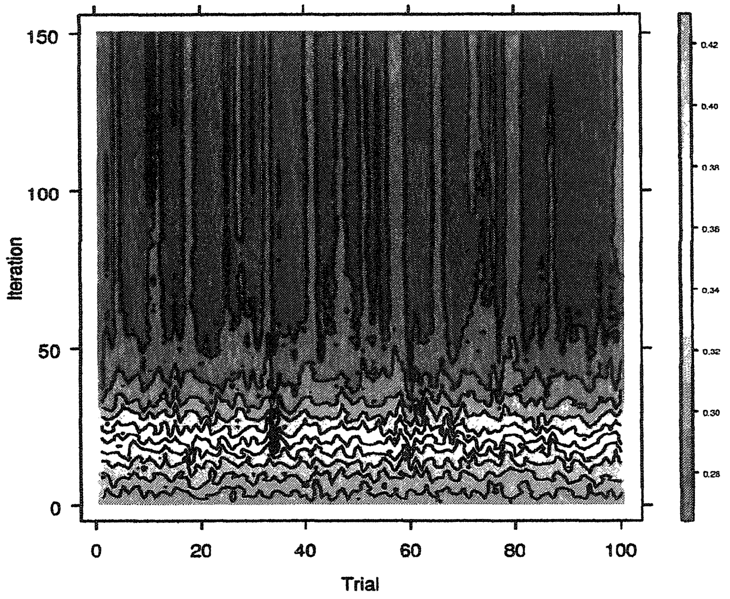


Figure 3.11: MAD after SAPA, perturbed towards correct, level 0.5, rate 1.0.

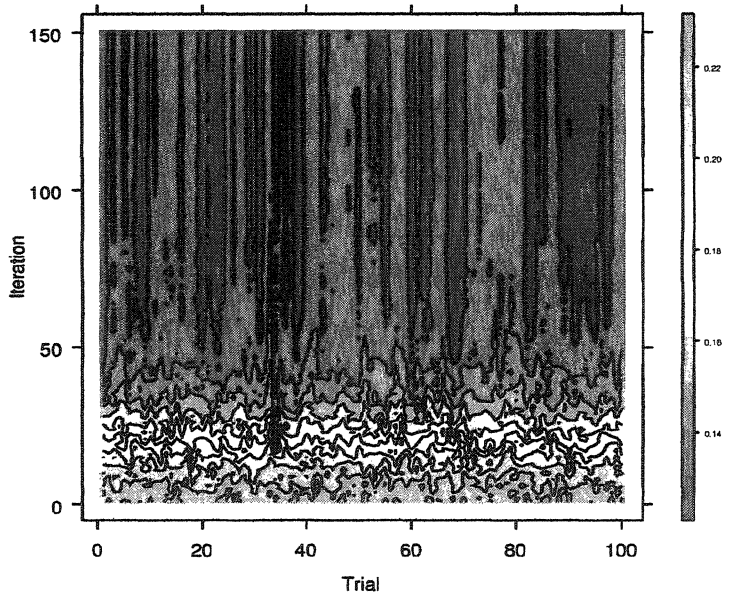


Figure 3.12: MSE after SAPA, perturbed towards correct, level 0.5, rate 1.0.

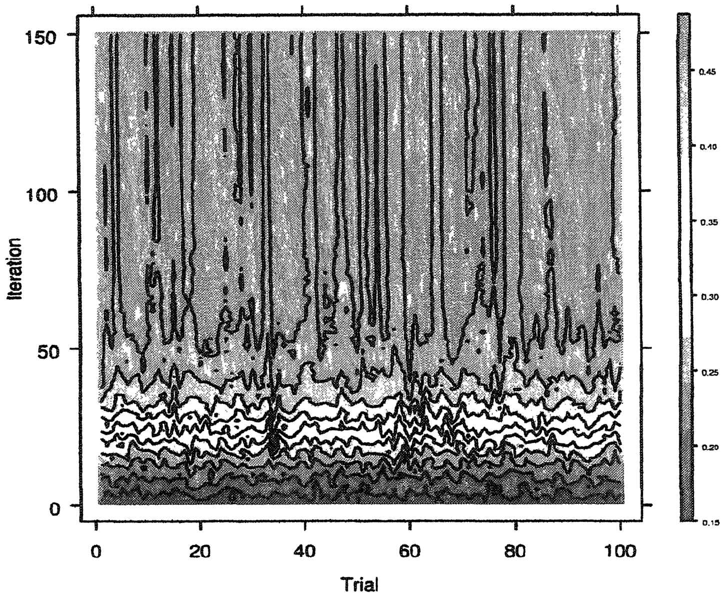
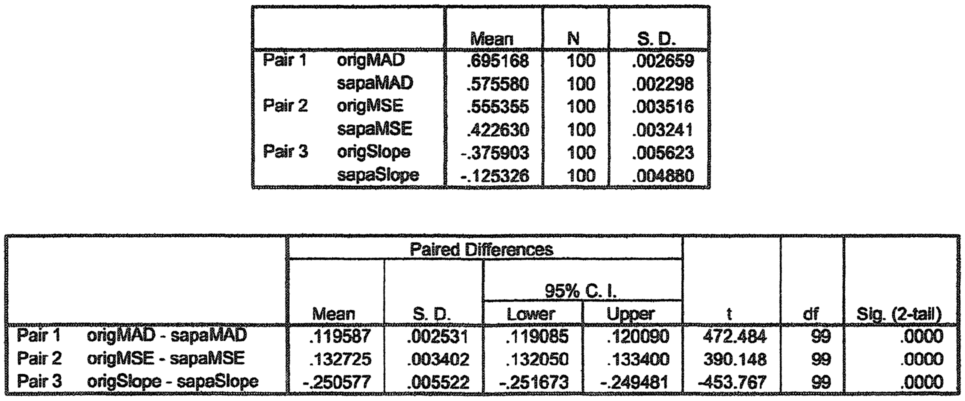


Figure 3.13: Slope after SAPA, perturbed towards correct, level 0.5, rate 1.0.



		Mean	N	S. D.
Pair 1	origMAD	.695168	100	.002659
	sapaMAD	.575580	100	.002298
Pair 2	origMSE	.555355	100	.003516
	sapaMSE	.422630	100	.003241
Pair 3	origSlope	-.375903	100	.005623
	sapaSlope	-.125326	100	.004880

	Paired Differences						
			95% C. I.				
	Mean	S. D.	Lower	Upper	t	df	Sig. (2-tail)
Pair 1 origMAD - sapaMAD	.119587	.002531	.119085	.120090	472.484	99	.0000
Pair 2 origMSE - sapaMSE	.132725	.003402	.132050	.133400	390.148	99	.0000
Pair 3 origSlope - sapaSlope	-.250577	.005522	-.251673	-.249481	-453.767	99	.0000

Figure 3.14: Paired-difference t-test statistics for data perturbed towards incorrect.

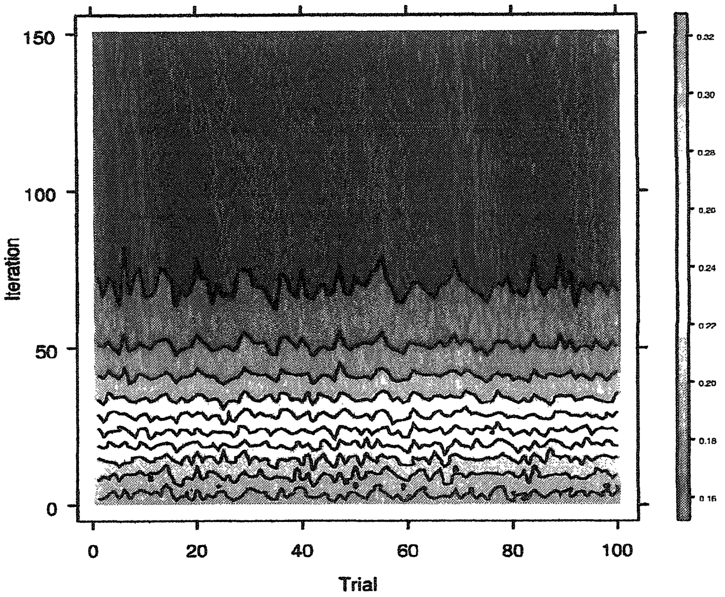


Figure 3.15: SAPA fit, perturbed towards incorrect, level 0.5, rate 1.0.

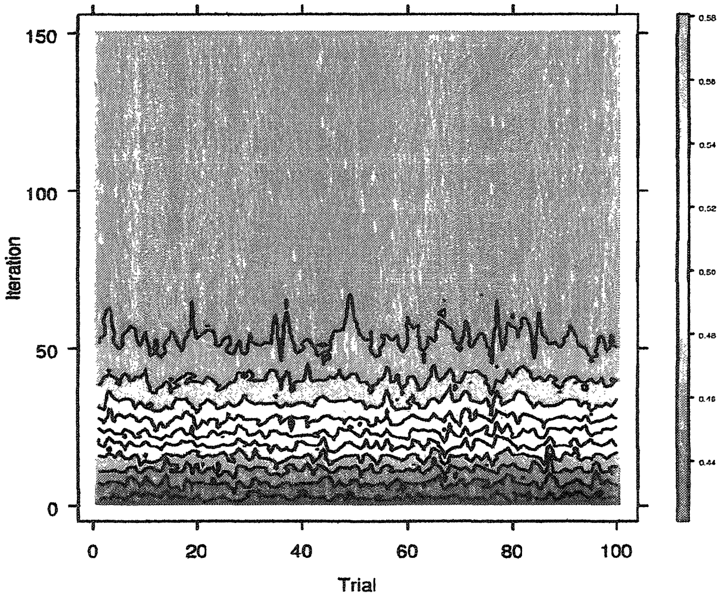


Figure 3.16: MAD after SAPA, perturbed towards incorrect, level 0.5, rate 1.0.

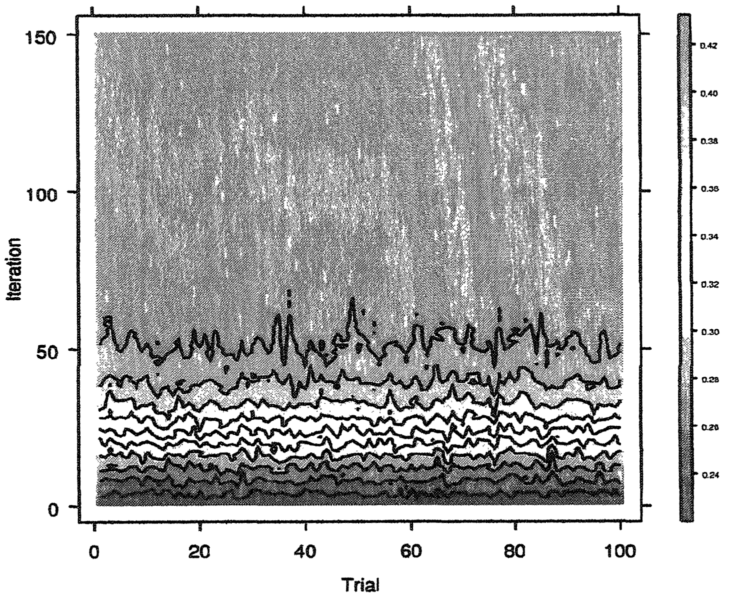


Figure 3.17: MSE after SAPA, perturbed towards incorrect, level 0.5, rate 1.0.

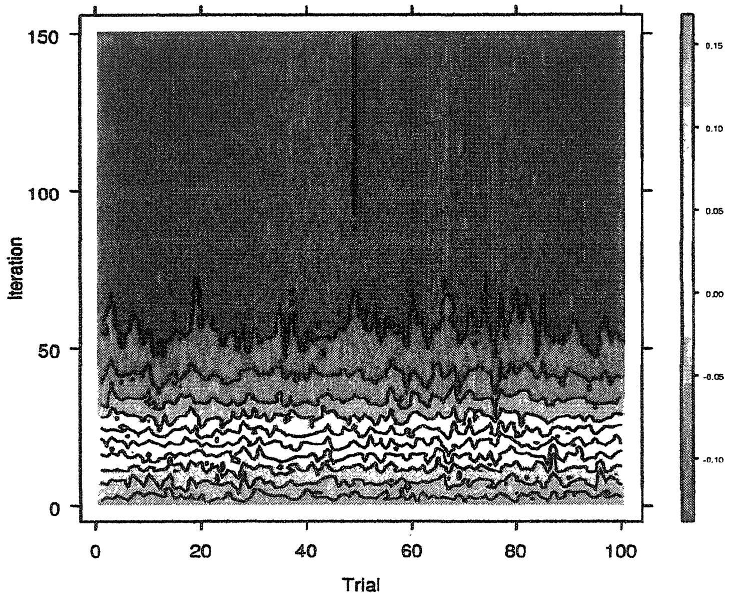


Figure 3.18: Slope after SAPA, perturbed towards incorrect, level 0.5, rate 1.0.

that for each simulation trial, every estimate (since *rate* = 1) was perturbed by a uniform random variable between 0 and 0.5, and the results are presented in the form of plots similar to those used with the unadjusted estimates previously. In Figures 3.10, 3.11, 3.12, and 3.13, the perturbations were made in the direction of the correct outcome. The plots show the fit, MAD, MSE, and slope measures, respectively.

With the perturbations made towards the correct outcome, the initial incoherent forecast performance measures will improve. As shown in the upper table of Figure 3.9, the measures for the set of improved estimates are: MAD, mean 0.2772, standard deviation 0.0024; MSE, mean 0.1452, standard deviation 0.0020; and slope, mean 0.4607, standard deviation 0.0050. All are significant improvements over the unadjusted values. If the optimization aspect of the SAPA approximation is providing any benefit, one should see evidence of improving error measures as the fit becomes better. MAD and MSE should both decline, and the slope should increase.

This is indeed what happens. Figure 3.10 shows the fit improving through iterations 70 to 80, and stabilizing thereafter. Correspondingly, by the same point in the iteration cycle Figure 3.11 shows the MAD declining to approximately 0.28, Figure 3.12 shows the MSE declining to approximately 0.14, and Figure 3.13 shows the slope increasing to approximately 0.45. In each case the behavior is consistent with the goal of minimizing coherence adjustments to the initial insightful estimates.

The perturbation is then applied in reverse with the results shown in Figures 3.15, 3.16, 3.17, and 3.18, with the perturbations made away from the correct outcome.

The worsened estimates should have significant poorer forecast performance measures compared with the unadjusted estimates, and this is confirmed by the upper table in Figure 3.14. It follows that as the coherent estimates improve and achieve a closer fit to these initially poor estimates, there should be a corresponding worsening of forecast performance.

Again the results are consistent with expectations. Figure 3.15 shows the fit improving and stabilizing in much the same manner as the with the improved estimators. Figures 3.16 and 3.17 show MAD and MSE increasing to approximately 0.58 and 0.42, respectively, and Figure 3.18 likewise shows the slope declining to a value of approximately -0.10.

Despite the confirmation of behavior consistent with the algorithm's design rationale, the results indicate that the the forecast performance improvement seen with the unadjusted values is significantly reduced or eliminated entirely when applied to subject estimates demonstrating conclusive levels of insight. Comparing the results for the improved estimates in the lower table of Figure 3.9 with the corresponding values for the unadjusted estimates in the lower table of Figure 3.8, only the improvement in MSE remains significant, with the mean improvement considerably lower in value. In contrast, the results for the worsened estimates in Figure 3.14 show even higher levels of improvement. Although not conclusive, the results suggest that when applied to sets of estimates from subjects with true expertise, the coherent estimates generated by SAPA are unlikely to yield significant performance improvements.

Chapter 4

Conclusions

4.1 Summary of Findings

Overall, the simulation results indicate that the design logic for SAPA is sound. Minimization the total adjustment made to the original incoherent responses does appear to effectively preserve insight or expertise, compared with the initial coherent approximations before the iterative optimization is performed. It is unlikely, however, that applying SAPA to responses from true experts will result in any significant improvement in forecast performance. When applied to estimates with lower levels of insight, significant forecast performance improvements are realized, but not to a sufficient degree that the resulting improved estimates would have any practical value.

4.2 Practical Considerations

Even if SAPA did yield forecast performance improvements with expert data, it is unclear whether or not SAPA could be used in an applied setting to improve forecast performance and afford better decision-making. There are several issues to consider that complicate the possible use of a SAPA-type mechanism in an applied capacity in business or government.

First, the questionnaire format used in the SAPA experiments is an extremely artificial construct. The construction of the event sets which the subjects are asked to estimate essentially guarantees probabilistic incoherence, and is hardly typical of how opinions or judgments are elicited in practice. The insight of the simulated subjects can be numerically manipulated for the purposes of analysis. It is possible, however, that even the best human experts may not be able to express their insight effectively using a SAPA questionnaire format. Potential evidence that this may be a problem is that the MBA students who participated in the Finance experiment were indistinguishable as a group in their forecast performance from the other, "non-expert" subjects [7]. This could be explained by the length of the forecast time horizon used, which may have been too short for insight to manifest itself, but the concern remains.

Bias in the event structures used to elicit estimates is another potential problem. With event constructs that involve qualification against an arbitrary fixed metric, as opposed to a relative one, setting the metric unrealistically can result in making the estimate too "easy." This results in high levels of apparent insight into an event that is almost certain to occur or not occur. For example, if the stock price performance of Microsoft is of interest, asking for an estimate regarding the probability of "Microsoft outperforms the S&P 500 over the next 12 months" is free of bias—the event is a comparison between two relative measures, the performance of Microsoft's stock, and of the market in general. However, asking for an estimate of the probability of "Microsoft's stock price will increase by at least 100% over the next 12 months" is effectively meaningless. The likelihood of a very large capitalization stock like Microsoft doubling in price in one year is extremely low. Most subjects familiar with the financial markets will realize this, and respond accordingly. The unlikely event (as defined by the question) does not occur, and the majority of subjects will get a low error score as a result, but this apparently high level of insight is of no practical value.

Finally, assuming insightful and meaningful estimates about future events can be obtained for use with SAPA, problems may arise integrating the results with the overall decision-making framework. The problem manifests itself when attempting to decide on an action based upon probabilistic information, since decisions and actions are frequently binary in nature and cannot be made proportionately. Using a "0 - 1" binary scoring mechanism for error measurement might be a useful extension to explore performance in an applied context.

4.3 Possible Future Research

There are several potential extensions to the simulation model. Adding the capability for finer control over the type and degree of change made to the initial subject estimates may provide additional information beyond the results obtained from the simple adjustment methods used here. The addition of other measures of performance may yield further insight regarding the practical applications of SAPA. In addition to the binary measure, another possibility would be to track "high confidence" estimates separately and examining how their performance characteristics are altered by SAPA. High confidence would be suggested by an estimate sufficiently close to 0 or 1, and could be regarded as comparatively more persuasive and easier to utilize in an applied context.

Finally, examining the performance of SAPA variants, such as forming an aggregate judge for generating forecast estimates [3], would likely be worthwhile, as the published results suggest that additional performance improvements may be possible using such approaches. Utilization of such extensions may yield additional performance gains, but additional examination is required.

Appendix A

Finance experiment variables

1. The Standard & Poor's 500 Index increases.
2. The Standard & Poor's 500 Index outperforms the NASDAQ Composite Index.
3. The NASDAQ Composite Index increases.
4. General Electric's stock price increases.
5. Reliant Energy's stock price increases.
6. Exxon Mobil's stock price increases.
7. Enron's stock price outperforms Reliant Energy's stock price.
8. El Paso Corp's stock price increases.
9. Enron's stock price increases by greater than 10%.
10. Wal-Mart's stock price increases.
11. Amazon.com's stock price increases.
12. Sears Roebuck's stock price increases.
13. Wal-Mart's stock price outperforms Amazon.com's stock price.
14. The U.S. prime lending rate increases.
15. The price of crude oil decreases by more than 10%.

16. U.S. 30-year fixed mortgage rates decrease.
17. The U.S. retail sales rate increases.
18. The U.S. Consumer Confidence Index increases.
19. The annualized U.S. Consumer Price Index inflation rate increases.
20. The U.S. unemployment rate increases.
21. Continental Airlines' stock price increases.
22. United Parcel Service's stock price increases.
23. Exxon Mobil's stock price outperforms United Parcel Service's stock price.
24. General Motors' stock price increases.
25. IBM's stock price increases.
26. Dell's stock price outperforms Sun Microsystems' stock price.
27. Intel's stock price increases.
28. Microsoft's stock price increases by more than 10%.
29. Dell's stock price outperforms IBM's stock price.
30. Dell's stock price outperforms Apple Computer's stock price.

Bibliography

[1] Randy Batsell, Lyle Brenner, Daniel Osherson, Moshe Y. Vardi, and Spyros Tsavachidis. Eliminating incoherence from subjective estimates of chance. *Proceedings of the 8th International Conference on the Principles of Knowledge Representation and Reasoning*, pages 353–364, 2002.

[2] J. Bernardo and A. Smith. *Bayesian Theory.* John Wiley and Sons, New York, NY, 1994.

[3] Jason Deines, Daniel Osherson, James Thompson, Spyros Tsavachidis, and Moshe Y. Vardi. Removing incoherence from subjective estimates of chance. Technical report, Rice University, Houston, TX, 2002.

[4] Urmila M. Diwekar. *Introduction to Applied Optimization.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.

[5] B. De Finetti. *Theory of Probability, Volume 1.* John Wiley and Sons, New York, NY, 1974.

[6] L. Ingber and B. Rosen. Genetic algorithms and very fast simulated reannealing: A comparison. *Mathematical and Computer Modelling*, 16(11):87–100, 1992.

[7] Daniel Osherson. Personal communication.

[8] Daniel Osherson, Spyros Tsavachidis, and Moshe Vardi. Eliminating incoherence from subjective estimates of chance. Technical report, Rice University, Houston, TX, 2002.

[9] Daniel Osherson and Moshe Vardi. Aggregating disparate estimates of chance. Technical report, Princeton University, Princeton, NJ, 2003.

[10] Panos M. Pardalos and Mauricio G. C. Resende, editors. *Handbook of Applied Optimization.* Oxford University Press, Oxford, UK, 2002.

[11] Detlov von Winterfeldt and Ward Edwards. *Decision Analysis and Behavioral Research.* Cambridge University Press, New York, NY, 1986.

[12] Jacques Frank Yates. *Judgment and Decision Making.* Prentice Hall, Englewood Cliffs, NJ, 1990.

[13] Zelda B. Zabinski. *Stochastic Adaptive Search for Global Optimization.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.