Journal of Global Optimization **18:** 399–416, 2000.
© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

A Case Study of Stochastic Optimization in Health Policy: Problem Formulation and Preliminary Results

DAVID DRAPER and DIMITRIS FOUSKAKIS
Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, UK (e-mail: d.draper@maths.bath.ac.uk and df@maths.bath.ac.uk, web http://www.bath.ac.uk/∼masdd and ∼mapdf).

Abstract. We use Bayesian decision theory to address a variable selection problem arising in attempts to indirectly measure the quality of hospital care, by comparing observed mortality rates to expected values based on patient sickness at admission. Our method weighs data collection costs against predictive accuracy to find an optimal subset of the available admission sickness variables. The approach involves maximizing expected utility across possible subsets, using Monte Carlo methods based on random division of the available data into N modeling and validation splits to approximate the expectation. After exploring the geometry of the solution space, we compare a variety of stochastic optimization methods — including genetic algorithms (GA), simulated annealing (SA), tabu search (TS), threshold acceptance (TA), and messy simulated annealing (MSA) — on their performance in finding good subsets of variables, and we clarify the role of N in the optimization. Preliminary results indicate that TS is somewhat better than TA and SA in this problem, with MSA and GA well behind the other three methods. Sensitivity analysis reveals broad stability of our conclusions.

Key words: Bayesian decision theory, Cross-validation, Genetic algorithm, Input-output analysis, Logistic regression, Maximization of expected utility, Messy simulated annealing, Monte Carlo methods, Prediction, Quality of health care, Sickness at hospital admission, Simulated annealing, Tabu search, Threshold acceptance, Variable selection

1. Introduction

An important topic in health policy is the assessment of the quality of health care offered to hospitalized patients (Daley et al., 1988; Kahn et al., 1990b). Quality of care is usually thought to depend mainly on three ingredients (Donabedian, 1981): (i) *process*, which is what health care providers do on behalf of patients, (ii) *outcomes*, which are what happens to patients as a result of the care they receive, and (iii) patient *sickness at admission*, because the appropriateness of outcomes cannot be judged without taking account of the burden of illness brought to the hospital by its patients.

A direct audit of the processes of care is usually regarded as the single most informative component in an evaluation of quality, but process is much more ex-

pensive to measure than outcomes or admission sickness (Kahn et al., 1990a). Interest has therefore focused in recent years, in countries such as the United States and Britain, on an indirect method of assessment – which might be termed the *input-output* approach (Draper, 1995)1 – in which hospital outcomes (for instance, death within 30 days of admission) are compared after adjusting for differences in inputs (sickness at admission). The idea is to treat what goes on inside the hospital – process – as a black box, with the contents of the box inferred by examining its outputs after taking account of its inputs (Jencks et al., 1988, Kahn et al., 1988).

Quality of care measurement. In practice, to indirectly measure quality of care at any given moment in time, this strategy proceeds by (a) taking a sample of hospitals and a sample of patients in the chosen hospitals, (b) obtaining death outcomes for the sampled patients (for example, from central government data bases), (c) extracting information on admission sickness from the medical records of these patients, (d) forming an expected mortality rate for each hospital based on (c), and (e) comparing observed and expected mortality to identify unusual hospitals (on both the 'good' and 'bad' ends of the spectrum). This would involve abstracting data from the charts of many thousands of patients if it were attempted on a large scale; thus the *cost-effective* measurement of admission sickness is crucial to this approach (because, beyond a certain point, money spent on improving the sickness adjustment would be better spent by getting data on new patients and hospitals).

Quality of care assessment is a highly disease-specific activity: for instance, the right admission sickness variables to examine for pneumonia would be quite different from those for heart attack. With any given disease there will be on the order of 100 separate variables potentially available in the medical record that are directly or indirectly related to admission sickness (Keeler et al., 1990). In the case of pneumonia, for example, on which we focus exclusively in this paper, a list of the important variables from a clinical perspective would include such things as systolic blood pressure on day 1 of admission, the presence or absence of prior respiratory failure, and the blood urea nitrogen (BUN) level (a measure of kidney functioning; see Table 1 below).

Constructing a sickness scale. The standard method for creating an expected mortality rate from these admission sickness inputs is logistic regression (Hosmer and Lemeshow, 1989), with 30-day death as the outcome, and using a nationally-representative sample of patients to normalize the expectation to average care across the nation. Typically a frequentist variable-selection method – such as all-subsets regression (Weisberg, 1985) – is employed to find a parsimonious and clinically reasonable subset of the available sickness variables. In a major US study of quality of hospital care for elderly patients conducted by the Rand Corporation in the late 1980s, this approach was used (Keeler et al., 1990) to reduce the list of 83 available sickness indicators for pneumonia down to a core of 14 predictors.

As good as the resulting 14-variable scale may be on grounds of simplicity and ease of clinical communication, we take the view in this paper that – when the goal is the creation of a sickness scale that may be used prospectively to measure quality of care on a new set of patients not yet examined – the original Rand approach is sub-optimal, because it takes no account of differences in the *cost of data collection* among the available predictors (which varied for pneumonia from roughly 30 seconds to 10 minutes of abstraction time per variable). The Rand approach represents a kind of benefit-only analysis; we propose a cost-benefit analysis, in which variables are chosen for the final scale only when they predict mortality well enough *given how much they cost to collect*.

The need for optimization methods. Weighing data-collection costs against the accuracy of prediction creates a large optimization problem which cannot be solved by brute-force enumeration: for example, when $p = 83$ it is necessary to compare $2^p \doteq 9.7 \cdot 10^{24}$ subsets of sickness variables, and even at the rate of 100 subsets examined per second – which is far faster than present computational resources permit – it would take more than $3 \cdot 10^{15}$ years to find the optimal subset by looking at all of them. Our main methodological goal here is therefore to find an efficient global optimization technique by comparing a variety of such methods on their ability to find good subsets.

The plan of the present paper is as follows. In Section 2 we formulate the basic problem more precisely; Section 3 presents results in a special case involving only $p = 14$ variables, where direct examination of all possible models suffices to identify the best subsets; and in Section 4 we explore the geometry of the solution space. The quantity we propose to optimize cannot be computed exactly in closed form, so we estimate it by Monte Carlo methods, and Section 5 clarifies the role of N, the number of simulation replications, in the optimization process. In Section 6 we present some preliminary findings comparing five optimization techniques in a version of the $p = 14$ case in which all of the methods are severely constrained on the total CPU time available for the search; Section 7 offers a variety of sensitivity analyses exploring the robustness of our formulation and results; and in Section 8 we conclude with discussion and comments on future work.

2. Problem Formulation

Suppose (a) the 30-day mortality outcome y_i and data on p sickness indicators $(x_{i1}, \ldots, x_{ip})$ have been collected on n individuals sampled randomly from a population $\mathcal{P}$ of patients with a given disease, and (b) the goal is to predict the death outcome for m new patients who will in the future be sampled randomly from $\mathcal{P}$, (c) on the basis of some or all of the predictors x_j, when (d) the marginal costs of data collection per patient $c_1, \ldots, c_p$ for the x_j vary considerably. What is the best subset of the x_j to choose, if a fixed amount of money is available for this task and you are rewarded (and penalized) for the quality of your predictions?

To solve this problem we use a Bayesian decision-theoretic approach based on maximization of expected utility (Bernardo and Smith, 1994). Any utility function given this setup would have two components, one to quantify data collection costs and one to keep track of predictive accuracy. (Lindley, 1968) recommended an approach similar to ours in a more general framework, using squared error loss to quantify predictive performance; we use a utility structure more closely tailored to the health policy context of our problem.

Data-collection utility. We follow traditional statistical usage and refer to a subset of the x_j as a *model*. One difficulty with the problem statement above is that by definition the future patients are unobserved, but – given that both the present and future samples are randomly drawn from $\mathcal{P}$ – a random subsample of the available data will be a good proxy for the future data. Thus to estimate the predictive success of a given model on future patients we use the cross-validation idea (Hadorn et al., 1992) of (1) dividing the available data at random into modeling and validation subsamples M and V, of size n_M and $n_V = n - n_M$ (respectively); (2) fitting the model to the data in M; and (3) evaluating its predictive accuracy on V. In Sections 3–6 we present results with the choice $\left(\frac{n_M}{n}, \frac{n_V}{n}\right) = \left(\frac{2}{3}, \frac{1}{3}\right)$; Section 7 contains some results on the sensitivity of our findings to this choice.

In our approach we quantify utility in monetary terms, so that the data collection utility is simply the negative of the total amount of money required to gather data on the specified predictor subset. Letting $I_j = 1$ if x_j is included in a given model (and 0 otherwise), the data-collection utility associated with subset $I = (I_1, \ldots, I_p)$ for patients in the validation subsample is

$$U_D(I) = -n_V \sum_{j=1}^{p} c_j I_j, \tag{1}$$

where c_j is the marginal cost per patient of data abstraction for variable j. In the Rand study described in Section 1, the data – on which we demonstrate our methods below – consisted of a representative sample of 16,792 elderly American patients hospitalized in the period 1980–86 with one of six high-prevalence diseases. As mentioned above, we focus here on pneumonia, for which the sample size was $n = 2{,}532$; the marginal costs per variable in this study were obtained by approximating the average amount of time needed by qualified nurses to abstract each variable from medical records and multiplying these times by the mean wage (about US$20 per hour in 1990) for the abstraction personnel. Table 1 shows the 14 variables in the Rand scale mentioned in Section 1 (APACHE II (Knaus et al., 1985) is a sickness scale developed for intensive care patients), together with their marginal costs and simple correlations with 30-day death (a measure of univariate predictive accuracy).

Predictive utility: one approach. To measure the accuracy of a model's pre-

Table 1. The 14 variables in the Rand pneumonia admission sickness scale, together with their approximate data collection costs per patient and correlation with 30-day death (CHF = congestive heart failure). The final column will be explained in Section 3

Variable	Cost c_j (US$)	Correlation with death	Good?
Total APACHE II score (36-point scale)	3.33	0.39	
Age	0.50	0.17	*
Systolic blood pressure score (2-point scale)	0.17	0.29	**
Chest X-ray CHF score (3-point scale)	0.83	0.10	
Blood urea nitrogen (BUN)	0.50	0.32	**
APACHE II coma score (3-point scale)	0.83	0.35	**
Serum albumin (3-point scale)	0.50	0.20	*
Shortness of breath (yes, no)	0.33	0.13	**
Respiratory distress (yes, no)	0.33	0.18	*
Septic complications (yes, no)	1.00	0.06	
Prior respiratory failure (yes, no)	0.67	0.08	
Recently hospitalized (yes, no)	0.67	0.14	
Ambulatory score (3-point scale)	0.83	0.22	
Temperature	0.17	−0.06	*

dictions, a metric is needed which quantifies the discrepancy between the actual and predicted values, and in our problem this metric must come out in monetary terms on a scale comparable to that employed with the data-collection utility. In the setting of this case study the actual values y_i are binary death indicators and the predicted values $\hat{p}_i$, based on statistical modeling, take the form of estimated death probabilities. We have chosen an approach to the comparison of actual and predicted values that involves dichotomizing the $\hat{p}_i$ with respect to a cutoff, to mimic the decision-making reality that actions taken on the basis of input-output quality assessment will have an all-or-nothing character at the hospital level (for example, regulators either may or may not subject a given hospital to a more detailed, more expensive quality audit based on process criteria). Other, continuous, approaches to the quantification of predictive utility are possible (e.g., a log scoring method (Bernardo and Smith, 1994)); we intend to explore this in future sensitivity analyses.

In the first step of our approach, given a particular predictor subset I, we fit a logistic regression model to the modeling subsample M and apply this model to the validation subsample V to create predicted death probabilities $\hat{p}_i^I$. In more detail, letting $y_i = 1$ if patient i dies and 0 otherwise, and taking $x_{i1}, \ldots, x_{ik}$ to be the k sickness predictors for this patient under model I, the statistical assumptions underlying logistic regression in this case are

$$\begin{aligned} (y_i \mid p_i^I) &\overset{\text{indep}}{\sim} \text{Bernoulli}(p_i^I), \\ \log\left(\tfrac{p_i^I}{1-p_i^I}\right) &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}. \end{aligned} \tag{2}$$

Table 2. Cross-tabulation of actual versus predicted death status. The left-hand table records the monetary rewards and penalties for correct and incorrect predictions; the right-hand table summarizes the frequencies in the 2×2 tabulation

		Rewards and penalties		Counts	
		Predicted		Predicted	
		Died	Lived	Died	Lived
Actual	Died	C_{11}	C_{12}	n_{11}	n_{12}
	Lived	C_{21}	C_{22}	n_{21}	n_{22}

We use maximum likelihood to fit this model, obtaining a vector $\hat{\beta}$ of estimated logistic regression coefficients, from which the predicted death probabilities for the patients in subsample V are given by

$$\hat{p}_i^I = \left[1 + \exp\left(-\sum_{j=0}^{k} \hat{\beta}_j x_{ij} \right) \right]^{-1}, \tag{3}$$

where $x_{i0} = 1$ ($\hat{p}_i^I$ may be thought of as the sickness score for patient i under model I).

In the second step of our approach, we classify patient i as predicted dead or alive according to whether $\hat{p}_i^I$ exceeds or falls short of a cutoff p^*, which is chosen – by searching on a discrete grid from 0.01 to 0.99 by steps of 0.01 – to maximize the predictive accuracy of model I. We then cross-tabulate actual versus predicted death status in a 2×2 contingency table, rewarding and penalizing model I according to the numbers of patients in the validation sample which fall into the cells2 of the right-hand part of Table 2. The left-hand part of this table records the rewards and penalties in US$. The predictive utility of model I is then

$$U_P(I) = \sum_{l=1}^{2} \sum_{m=1}^{2} C_{lm} n_{lm}. \tag{4}$$

Here C_{11} and C_{22} should be positive and C_{12} and C_{21} negative, and since it is easier to correctly predict that a person lives than dies with these data (the overall pneumonia 30-day death rate in our sample was 16%) it is natural to choose the C_{lm} so that $C_{11} > C_{22}$. It is also clear from the fact that it is worse to label a 'bad' hospital as 'good' than the other way around that one should take $|C_{12}| > |C_{21}|$, and furthermore that the magnitudes of the penalties should exceed those of the rewards. Based on discussions with health experts in the US and UK, the results below in Sections 3–6 use the values $(C_{11}, C_{12}, C_{21}, C_{22}) = 8.7 \cdot (4, -16, -8, 1)$; in Section 7 we present a sensitivity analysis on the choice of the C_{lm}.

Total expected utility. The overall expected utility function to be maximized over I is then simply

$$E[U(I)] = E[U_D(I) + U_P(I)]. \quad (5)$$

In practice we use Monte Carlo methods to evaluate this expectation, averaging over N random modeling and validation splits.

3. Full Enumeration Results in the Case $p = 14$

With p predictors to choose from, the expected utility maximization is over 2^p possible subsets of variables. With our data it takes about 0.4 second on a Sun `UltraSPARC` Enterprise 250 computer running Unix at 400 Mhz to evaluate Equation (5) for a single modeling/validation split with efficient code, so (as mentioned in Section 1) it is computationally infeasible given present computing resources – even with a moderate choice of N – to perform exhaustive enumeration for all $p = 83$ sickness indicators for pneumonia. Attention thus naturally focuses on stochastic optimization as a way to find 'good' (near-optimal) subsets for large p.

The most straightforward way to compare optimization methods in this situation is to create a test-case in which the truth about all 2^p models is known (up to small Monte Carlo uncertainty), so that the actual quality of subsets discovered by any given optimization method may be ascertained. We chose to do this by performing a full enumeration3 of the estimated expected utility of all $2^p = 16,384$ possible subsets of the $p = 14$ variables chosen in the Rand sickness scale for pneumonia (Table 1), in which each estimate of Equation (5) was based on $N = 500$ random splits (this choice of N was sufficient to yield a Monte Carlo standard error for each expected utility estimate of only about US$0.05).

Figure 1 presents parallel boxplots4 of the estimated expected utilities of the 16,384 models in the $p = 14$ case as a function of k, the number of predictors in each model. Several conclusions are evident from a detailed examination of this figure and the data on which it is based, as follows.

- The trace of the median expected utilities (the white lines in the middle of the boxes) as a function of k clearly shows the tradeoff between data collection cost and predictive accuracy: for small k the models don't cost very much but predict poorly, and for large k the predictions are excellent but the cost is too high, so that the best models are in the middle. In particular the full 14-variable Rand scale is highly inefficient (and slightly worse in monetary terms than using no sickness indicators at all, i.e., predicting death at random with probability 0.16).
- The 20 best models include the same 3 variables 19 or more times out of 20, and never include 5 of the other variables; the five best models are minor variations on each other, and include 4-6 variables. The eight variables which occur most frequently in the 20 best models are identified with asterisks in

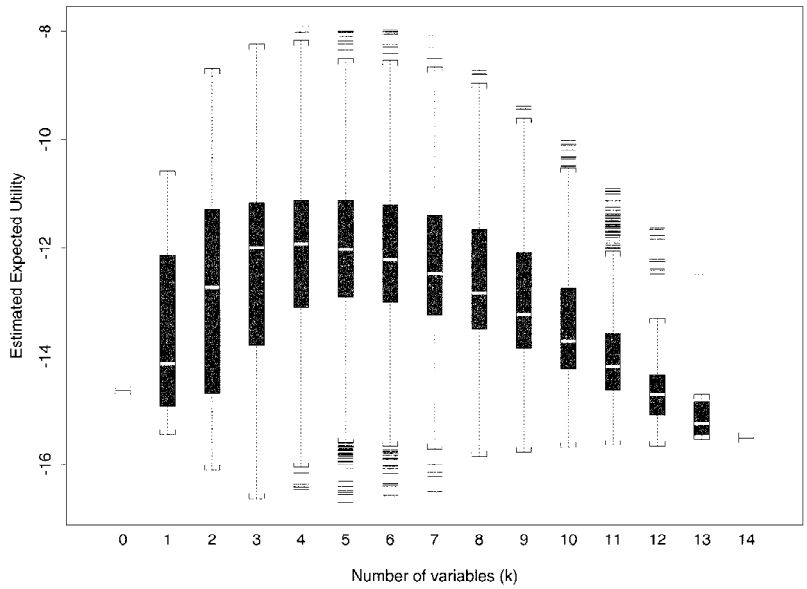

Figure 1. Parallel boxplots in the $p = 14$ case, showing estimated expected utility as a function of number of predictors (k) retained, based on $N = 500$ data splits.

Table 1; the global optimum subset of predictors has double asterisks. The single best univariate predictor, the total APACHE II score, does not appear in any of the good models because it is so costly to collect – BUN and the APACHE II coma score predict death (univariately) almost as well and are much cheaper to obtain.

– The best models cost almost US$8 per patient less than the full 14-variable model, which would yield significant savings annually if the input-output approach were to be implemented on a widespread basis.

4. Geometry of the Solution Space

Optimization methods such as simulated annealing (Kirkpatrick et al., 1983) and tabu search (Glover, 1989) require the specification of a neighborhood structure across models, so that – having evaluated the quality of a given model – one can judge where best to move next in the search for the global optimum. In our problem a model is a vector of p 1s and 0s specifying the presence or absence of each predictor in the subset of available variables, and a natural first choice for neighborhood structure is based on moves which select a single bit in the binary string and flip it from 0 to 1 or vice versa (call such moves *1-bit flips*).

Whatever the neighborhood structure, the space of all possible models can be visualized as a tree (Knuth, 1968). Figure 2 shows the $2^4 = 16$ models for one particular choice of $k = 4$ variables chosen from among the 14 predictors in

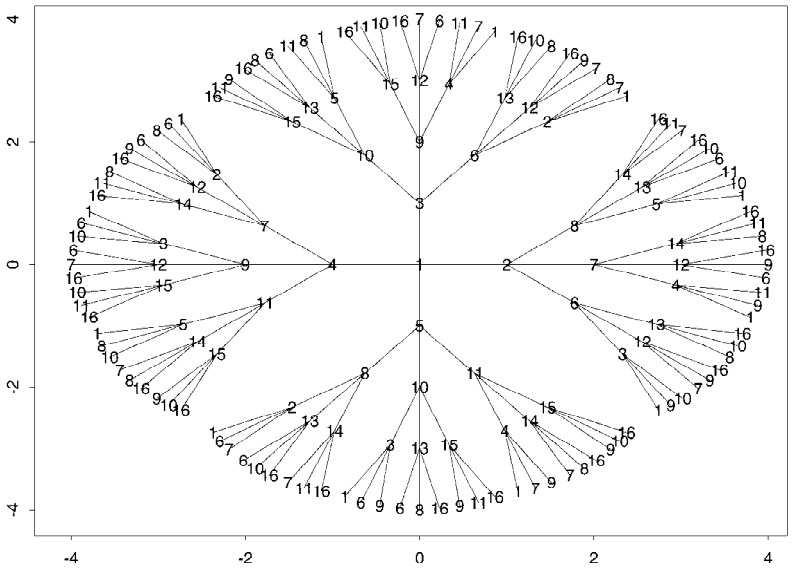

Figure 2. Tree of adjacent models ($k = 4$) expanded out to four levels, with the neighborhood structure induced by moves based on one-bit flips. The horizontal and vertical scales are arbitrary.

the Rand scale, with the tree expanded out to four levels (in the notation of this figure, models 1-16 are {0000, 1000, 0100, . . . , 1111}, respectively). Figure 3 is a perspective plot of the expected utility 'surface' corresponding to Figure 2, with all points not actually part of the tree set to zero for contrast. While it is true that the quality of a given model's neighbors is sometimes similar to that of the model itself, it is also evident that adjacent models can have sharply different expected utilities, demonstrating the discontinuity of the solution space in our problem: good models do not necessarily have good models as neighbors. This has implications for the optimal search strategy – methods that spend considerable time exploring local alternatives to good models may not perform as well as methods that frequently make large jumps around the model space, but too much jumping around in an unguided way will yield poor performance as well.

5. Optimal Choice of N

To make a fair comparison among optimization methods it is natural to ask how well each method performs when permitted no more than a fixed amount of CPU time. In our problem, under such a constraint, the role of N – the number of random modeling and validation splits of the data on which the Monte Carlo estimate of Equation (5) is based – requires consideration: if N is small many models can be evaluated but the estimates of their quality will be noisy, often leading to incorrect decisions about which models are good, whereas if N is large the quality of each

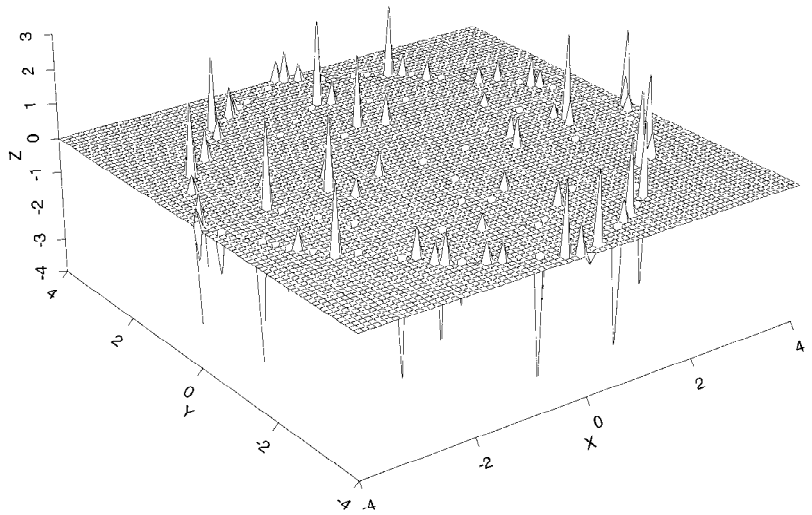

Figure 3. Perspective plot of the expected utility 'surface' for the 4-variable tree expanded out to four levels. The X and Y axes correspond to the horizontal and vertical axes in Figure 2; the Z axis plots the estimated expected utilities of the 16 models (from a full-enumeration exercise like that described in Section 3), shifted so that the median utility is zero.

model visited will be known accurately but there will not be time to evaluate many models.

As an instance of this phenomenon, Figure 4 presents an example of the performance of simulated annealing (SA; see Section 6) in our problem, in a run with $p = 14$ in which SA found the global optimum solution described in Section 3. We used a geometric cooling schedule (Stander and Silverman, 1994) from a starting temperature of 1 to a final temperature of 0.001, and moves from one model to another were based on 1-bit flips. The run consisted of 4,989 iterations, with N beginning at 1 and increasing geometrically to 50, and the null model (with no predictors) was used as the starting value. Four aspects of the run are plotted: (apparent) estimated expected utility and N (the left- and right-hand vertical scales in the upper panel), and dimension k of the current model and temperature (the left- and right-hand scales in the lower panel). It is evident that from about iteration 3,000 to the end SA primarily visited good models with 3-7 predictors (the optimal range in Figure 1), but the method spent much of its time before that point looking at models known from the results in Section 3 to be inferior. This may well be (in part) because values of N that were too small were used early in the run: note, for example, that in the first 1,000 iterations (when N was at most 2) SA found several models with apparent estimated expected utility of about -6, which is much larger than the actual utility of the best models.

To explore the optimal choice of N in a simple setting, as a way of informing its choice in our main results, we compared two search strategies: random-walk in model space (a) with $N = 1$, and (b) with $N > 1$. Each strategy was given a budget of M utility evaluations (which is equivalent to a CPU constraint); strategy

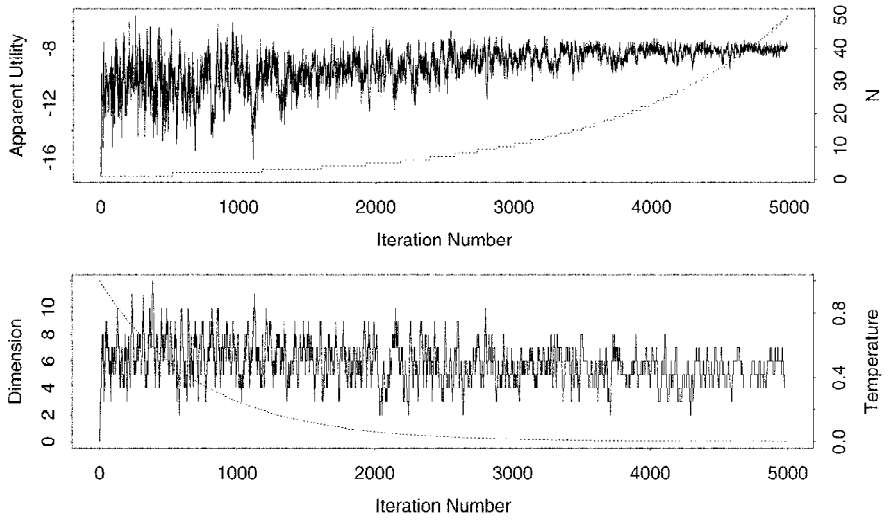

Figure 4. Performance of simulated annealing on a run that found the global optimum in the $p = 14$ case, allowing the method 24 hours of CPU time at 400 MHz. The top panel plots the (apparent) estimated expected utility (solid line, left-hand vertical axis) and N (dotted line, right-hand axis) against iteration number, and the bottom panel does the same for model dimension (solid line, left-hand vertical axis) and temperature (dotted line, right-hand axis).

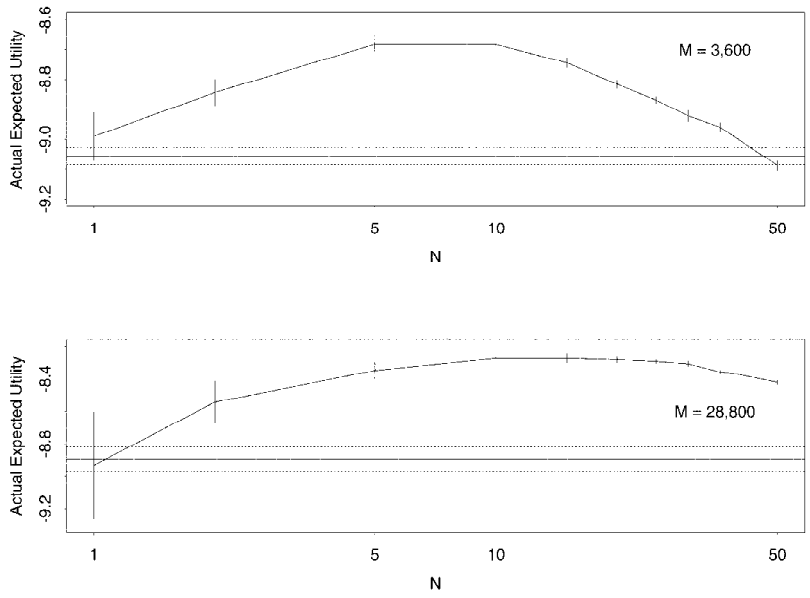

Figure 5. Actual expected utility as a function of N for a random-walk search strategy (the horizontal scale is logarithmic).

(a) visited M models chosen at random from the 16,384 possible models in the $p = 14$ case, with each model having its expected utility evaluated only once, whereas strategy (b) visited $\frac{M}{N}$ models at random and estimated the expected utility as an average of N evaluations across random model/validation splits, for N varying from 2 to 50. In each simulation replication the *actual* expected utility (from the full-enumeration results of Section 3) of the model with the maximum *apparent* expected utility across all models visited was recorded, and we repeated this exercise 3,000 times for each value of N and for M varying from 3,600 to 28,800.

Figure 5 summarizes the results for the extremes of M we examined. The concave, roughly quadratic curves (with 95% uncertainty bands plotted as solid vertical lines) trace out the mean *actual* expected utility as a function of N across the simulation replications; the horizontal line in each plot (with 95% uncertainty bands as dotted lines) gives the results from a separate set of runs with $N = 1$ for comparison. For $M = 3,600$ the optimum N is attained between 5 and 10, with the results for $N = 50$ about as bad as those with $N = 1$. When the number of utility evaluations is increased by a factor of 8, both strategies naturally find better models and the optimal N increases to about 10 (although values of N between 10 and 30 do almost as well as $N = 10$). Even though strategies (a) and (b) are much less sophisticated than those examined in Section 6, we have found that the results described here are a good guide to the sensible choice of N when more intelligent global optimization methods are used instead.

6. Preliminary Comparison of Optimization Methods

We have completed a preliminary comparison, on the $p = 14$ case, of five5 stochastic optimization methods:

- a genetic algorithm (GA): a classic approach (Holland, 1975) whose crossover and mutation moves are motivated by evolutionary ideas;
- tabu search (TS): a promising heuristic method based on three phases: preliminary search (to find promising regions of model space), intensification (to explore these regions in more detail), and diversification (to strike out into completely new regions);
- simulated annealing (SA): another classic approach in which moves to models with lower values of the criterion function are accepted with decreasing probability (governed by the diminishing *temperature* of the process) as the iterative search progresses, to avoid getting stuck in local optima;
- messy simulated annealing (MSA): a variation (Kvasnička and Pospíchal, 1995) of SA in which moves from the current model are motivated by ideas similar to those in GA; and
- threshold acceptance (TA): a different variation of SA (Dueck and Scheuer, 1990) in which a move to a model worse than the current state is only accepted

if its criterion function value differs from that of the current model by no more than a threshold.

To give all of the methods a realistically small amount of CPU time with $p = 14$ (to simulate the situation with larger p), we made a number of runs forcing each method to only use 10 or 20 minutes of CPU time at 400 (Unix) MHz. With TA and SA we alternated 1-bit flips with a second move type: *2-bit swaps*, in which a random subset of two variables is chosen and their inclusion indicators, if different, are interchanged.

In these runs we also implemented an improvement involving adaptive choice of N: previously in a fixed–N run with (say) $N = 10$ all models were evaluated with $N = 10$. In the adaptive method (i) 20 models are chosen at random to initialize the search and evaluated with $N^* = 10$, creating a *league table* of the current 20 best models; and (ii) a new model is chosen and evaluated once. If its apparent utility would seem to place it somewhere in the current league table, the utility is evaluated for $(N^* - 1) = 9$ more random splits and the average over the N^* values is computed – if it still belongs in the league table it is added at the appropriate place; if not it is discarded. We found that this *adaptive–N^** approach was significantly better than the fixed-N approach for all optimization methods we examined.

Table 3 presents the results of our preliminary comparison with 10 and 20 minutes of CPU time. The adaptive-N^* method was used throughout; see the Appendix for implementation details on all five optimization techniques. Each row in the table represents the best of eight runs, corresponding to $N^* = \{1, 2, 3, 4, 5, 10, 15, 20\}$. In keeping with the likely use of our method in health policy, in which a list of the b best models would be presented to decision-makers for a check on clinical face-validity, we examined three summaries of how well each method recovered the $b = 20$ best models from the full-enumeration exercise: (1) how many of the actual 20 best models were in each method's announced list of 20 best, and the actual ranks of the (2) best and (3) worst models in the apparent 20 best. (In column 4 of Table 3 we also report the mean and standard deviation (SD) of the actual utility of the 20 best models found by each method.) In these preliminary comparisons6 (and others not shown here for reasons of space) we found that TS was the overall best method in our problem (routinely able to locate about 75% of the 20 best models in only 20 minutes of CPU time), with TA and SA not far behind; MSA came in an unimpressive fourth, and GA decisively brought up the rear. Given that differences of 0.20 or more on the utility scale are large in practical terms in our problem (because of the financial implications of such differences), the utility results in Table 3 convey a similar message.

By looking at the geometry of the solution space we showed in Section 4 that good methods in our problem need to strike a compromise between respecting the local neighborhood structure and making bold jumps around the model space. Careful examination of GA and TS results indicates that the crossover operation

Table 3. Preliminary results comparing simulated annealing (SA), messy simulated annealing (MSA), tabu search (TS), threshold acceptance (TA), and genetic algorithms (GA). The adaptive-N^* method was used in all cases. Boldface indicates the best result in each column for each CPU time constraint

10 minutes CPU time					
		20 Best Models Found			
		Number of	Mean (SD)	Actual Rank of	
Method	N^*	20 Actual Best	Actual Utility	Best	Worst
TS	5	**11**	**–8.36** (0.29)	4	**103**
TA	3	10	−8.43 (0.37)	2	119
SA	1	9	−8.65 (0.64)	2	1321
MSA	3	4	−8.84 (0.51)	**1**	499
GA	2	1	−9.58 (0.64)	9	2379

20 minutes CPU time					
		20 Best Models Found			
		Number of	Mean (SD)	Actual Rank of	
Method	N^*	20 Actual Best	Actual Utility	Best	Worst
TS	15	**15**	**–8.26** (0.32)	**1**	153
TA	10	12	−8.30 (0.29)	**1**	**67**
SA	4	13	−8.32 (0.30)	**1**	102
MSA	3	8	−8.57 (0.43)	**1**	190
GA	1	3	−9.22 (0.77)	6	5562

inherent in GA makes insufficient use of the modest amount of continuity present in our problem, while TS appears to achieve a happy balance between local exploration of good models and occasional leaps into fruitful new territory. It is the diversification stage of TS that appears to give it the edge over TA and SA in this problem.

7. Sensitivity Analysis

We have conducted three types of sensitivity analyses to explore the robustness of our problem formulation and preliminary findings, as follows.

– How sensitive are the optimality results to our specific choice of C_{lm}, the penalties and rewards for prediction accuracy, and c_j, the data collection marginal costs per variable? Starting with the values noted in Section 2, we multiplied all the C_{lm} by $\kappa = 2, 3, \ldots, 8$ and $\frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{8}$ (holding the data collection

costs constant at their Section 2 values throughout) and recomputed the 20 best models in each instance in the $p = 14$ case. Results were highly stable: for instance, with $\kappa = 2$, 14 of the original 20 best models were still among the 20 best, and for 11 of the 14 variables, the frequencies of occurrence in the 20 best models differed by 10% or less. We then multiplied all the c_j by the same κ values (this time holding the penalties and rewards constant at their Section 2 values) and again recomputed the 20 best models. Here the findings were even more robust: for example, with $\kappa = 2$, 18 of the original 20 best were still among the new 20 best, and for all 14 variables, the frequencies of occurrence in the 20 best models differed by 10% or less uniformly.

- When constructing an admission sickness scale from available predictors x_j, it is possible to include not only main effects (the x_j themselves) but also interactions and quadratic terms (of the form $\left(x_j - \bar{x}_j\right)\left(x_k - \bar{x}_k\right)$ and $\left(x_j - \bar{x}_j\right)^2$, respectively). How sensitive are the results presented here to omission of interaction and quadratic terms among the predictors? As an approximate answer to this question, we used the entire pneumonia data set ($n = 2{,}532$) to find all 2-way interactions (including quadratic terms) among the 14 variables which had $z = |\hat{\beta}/\widehat{SE}(\hat{\beta})| \geqslant 2$ when added one by one to the 14-variable model (here $\widehat{SE}(\hat{\beta})$ is the maximum likelihood standard error of $\hat{\beta}$) – there were 5 such interactions out of a possible 105. We then used TS with 20, 40, 80, 160, and 320 min of CPU time at 400MHz on the 19-variable model formed by adding the 5 new interaction terms. With a CPU time limit of 20 min, only one interaction appeared among the 20 best models, and for CPU time constraints in excess of 40 min none of the interactions appeared among the 20 best. We conclude that interactions play only a minor role in this problem and their omission has little effect on our findings.
- All of our main runs were with $\frac{n_M}{n} = \frac{2}{3}$. How sensitive are the optimality results to the relative choice of n_M and n_V? To address this question we repeated the brute-force full enumeration summarized in Figure 1 but with $\frac{n_M}{n} = \frac{1}{3}$ (and additionally we repeated the full-enumeration run with $\frac{n_M}{n} = \frac{2}{3}$ using different random number seeds). The median rank correlation among the 20 best models across these three sets of full enumeration results was $+0.81$ (minimum $+0.69$, maximum $+0.86$), indicating strong agreement no matter which value of $\frac{n_M}{n}$ (or random seed) was used, and the graph corresponding to Figure 1 with $\frac{n_M}{n} = \frac{1}{3}$ was almost identical.

8. Discussion

In this paper we have presented preliminary results on the use of stochastic optimization to solve an important problem in health policy, focusing on problem formulation and some implementation details along the way. Ongoing work includes the following:

- Conducting a large simulation experiment to investigate the quality of the solutions from each optimization method as a function of the method's inputs (for example, what are the best settings for TS with respect to number of intensification versus diversification searches?);
- Creating hybrid methods by combining several approaches (for example, an initial round of TS to identify promising models followed by a series of short SA runs involving rapid cooling from each of these candidate neighborhoods), and parallelizing the stochastic optimization to reduce computational time;
- Increasing the number of predictors to $p = 83$ (the full pneumonia model) and making a comparative study of the same stochastic optimization techniques in this much larger solution space; and
- Drawing conclusions about which optimization methods perform best in our problem for a given amount of CPU time as a function of p.

We expect that the final results from this project will have both a practical payoff for health policy and broader implications for stochastic optimization.

Acknowledgements

We are grateful to Katherine Kahn for making the data available to us, and to Chris Jennison and Andrew Wood for references and helpful comments on earlier versions of this material. The work described here represents part of the Ph.D. dissertation of the second author, under the supervision of the first author.

Appendix

In the version of tabu search used to produce the results in Table 3 we chose the following user-defined settings: a tabu list size of 7, r repetitions of the whole search process (where r varied from 1 to 11 as a function of N and the amount of CPU time allowed), 6 preliminary searches, 9 intensification searches, a maximum of 4 random restarts within each intensification search (a restart occurred whenever the globally best solution found so far was located), and 2 diversification searches. Our version of SA employed a geometric cooling schedule from a maximum temperature of 1.0 to a minimum of 0.1. The version of MSA used in Table 3 was identical to SA except with gene and allele mutation probabilities both set to 0.5. In our version of TA we varied the threshold, on the utility scale, geometrically throughout the run from an initial value of 1.0 to a minimum of 0.1. Finally, in the version of GA whose results are given in Table 3 the population size was 40, we used one-bit crossover with probability 0.7, and the mutation probability was 0.001.

Notes

1 In the UK this approach is also referred to as *league-table quality assessment* (Goldstein and Spiegelhalter, 1996), by analogy with the process of ranking football (soccer) teams.

2 To clarify the role of the probability cutoff, for each of the 99 values of p^* from 0.01 to 0.99 we calculated the entries in Table 2 and the resulting predictive utilities in Equation (4), and we chose the cutoff p^* which maximizes this utility. In practice the optimal cutoff was typically around 0.4.

3 This exercise took 38 days of CPU time to complete.

4 Column k in this figure, as k runs from 0 to p, is a summary of the estimated expected utilities of all $\left(\begin{array}{c} p \\ k \end{array} \right)$ subsets consisting of k predictor variables (for example, while there is only one model with no predictors at all and only one with all 14 predictors, there are 3,432 ways to choose 7 predictors from among the available 14; the column marked 7 at the bottom of the plot is a summary of these 3,432 models.) The central rectangular part of each column (the 'box' portion of the boxplot (Tukey, 1977)) runs from the 25th to the 75th percentiles of the distribution being summarized, and the white band in the middle of the box is the median. The region between the lower and upper square brackets contains observations that would not be unusual if drawn from a Gaussian distribution; observations beyond this point (possible outliers) are plotted as thin horizontal bars.

5 As a referee noted, with values of p smaller than 14 it is possible that simpler algorithms than the ones we study here – such as methods that make completely random proposed moves from the current model and only accept such moves if the estimated expected utility increases – would be highly competitive, but with even modest values of p there are so many local optima and they are so highly dispersed in model space that random search would be dominated by more sophisticated methods.

6 As an example of the results in Table 3, the TS row in the part of the table based on runs with 10 minutes of CPU time is interpreted as follows: of the 20 best models TS found in the run with $N^* = 5$, 11 of these were among the list of 20 globally best models summarized in Figure 1; the mean and SD of the actual estimated expected utilities of the 20 best models found by TS were -8.36 and 0.29, respectively; and the best and worst models among the 20 best models found by TS ranked 4th and 103rd best in the overall enumeration of all 16,384 models.

References

1. Bernardo, J.M. and Smith, A.F.M. (1994), *Bayesian Theory*. New York: Wiley.
2. Daley, J., Jencks, S., Draper, D., Lenhart, G., Thomas, N. and Walker, J. (1988), Predicting hospital-associated mortality for Medicare patients with stroke, pneumonia, acute myocardial infarction, and congestive heart failure. *Journal of the American Medical Association* 260: 3617–3624.
3. Donabedian, A. (1981), Advantages and limitations of explicit criteria for assessing the quality of health care. *Milbank Memorial Fund Quarterly – Health and Society* 59: 99–106.
4. Draper, D. (1995), Inference and hierarchical modeling in the social sciences (with discussion). *Journal of Educational and Behavioral Statistics* 20: 115–147, 233–239.
5. Draper, D., Kahn, K., Reinisch, E., Sherwood, M., Carney, M., Kosecoff, J., Keeler, E., Rogers, W., Savitt, H., Allen, H., Wells, K., Reboussin, D. and Brook, R. (1990). Studying the effects of the DRG-based Prospective Payment System on Quality of Care: Design, sampling, and fieldwork. *Journal of the American Medical Association* 264: 1956–1961.

6. Dueck, G. and Scheuer, T. (1990). Threshold acceptance: A general purpose optimization algorithm appearing superior to simulated annealing. *Journal of Computational Physics* 90: 161–175.
7. Glover, F. (1989), Tabu search – Part I. *ORSA Journal on Computing* 1: 190–206.
8. Goldstein, H. and Spiegelhalter, D.J. (1996), League tables and their limitations: Statistical issues in comparisons of institutional performance (with discussion). *Journal of the Royal Statistical Society, Series A* 159: 385–444.
9. Hadorn, D., Draper, D., Rogers, W., Keeler, E. and Brook, R. (1992), Cross-validation performance of patient mortality prediction models. *Statistics in Medicine* 11: 475–489.
10. Holland, J.H. (1975), *Adaptation in natural and artificial systems.* Ann Arbor: University of Michigan Press.
11. Hosmer, D.W. and Lemeshow, S. (1989), *Applied Logistic Regression*. New York: Wiley.
12. Jencks, S., Daley, J., Draper, D., Thomas, N., Lenhart, G. and Walker, J. (1988), Interpreting hospital mortality data: The role of clinical risk adjustment. *Journal of the American Medical Association* 260: 3611–3616.
13. Kahn, K., Brook, R., Draper, D., Keeler, E., Rubenstein, L., Rogers, W. and Kosecoff, J. (1988), Interpreting hospital mortality data: How can we proceed? *Journal of the American Medical Association* 260: 3625–3628.
14. Kahn, K., Rogers, W., Rubenstein, L., Sherwood, M., Reinisch, E., Keeler, E., Draper, D., Kosecoff, J. and Brook, R. (1990), Measuring quality of care with explicit process criteria before and after implementation of the DRG-based Prospective Payment System. *Journal of the American Medical Association* 264: 1969–1973.
15. Kahn, K., Rubenstein, L., Draper, D., Kosecoff, J., Rogers, W., Keeler, E. and Brook, R. (1990), The effects of the DRG-based Prospective Payment System on quality of care for hospitalized Medicare patients: An introduction to the series. *Journal of the American Medical Association* 264: 1953–1955 (with editorial comment, 1995–1997).
16. Keeler, E., Kahn, K., Draper, D., Rogers, W., Sherwood, M., Rubenstein, L., Reinisch, E., Kosecoff, J. and Brook, R. (1990), Changes in sickness at admission following the introduction of the Prospective Payment System. *Journal of the American Medical Association* 264: 1962–1968.
17. Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983). Optimization by simulated annealing. *Science* 220: 671–680.
18. Knaus, W.A., Draper, E.A., Wagner, D.P. and Zimmerman, J.E. (1985). APACHE II: A severity of disease classification system for severely ill patients. *Critical Care Medicine*, 13: 818–829.
19. Knuth, D.E. (1968), *The Art of Computer Programming. Volume 1: Fundamental Algorithms.* Reading, MA: Addison-Wesley.
20. Kvasnička, V. and Pospíchal, J. (1995), Messy simulated annealing. *Journal of Chemometrics*, 9: 309–322.
21. Lindley, D.V. (1968), The choice of variables in multiple regression (with discussion). *Journal of the Royal Statistical Society, Series B*, 30: 31–66.
22. Stander, J. and Silverman, B.W. (1994), Temperature schedules for simulated annealing. *Statistics and Computing* 4: 21–32.
23. Tukey, J.W. (1977), *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.
24. Weisberg, S. (1985), *Applied Linear Regression*, second edition. New York: Wiley.