The Independent Sign Bias: Gaining Insight from Multiple Linear Regression

Michael J. Pazzani (pazzani@ics.uci.edu)
Stephen D. Bay (sbay@ics.uci.edu)
Department of Information and Computer Science
University of California, Irvine
Irvine, CA 92697

Abstract

As electronic data becomes widely available, the need for tools that help people gain insight from data has arisen. A variety of techniques from statistics, machine learning, and neural networks have been applied to databases in the hopes of mining knowledge from data. Multiple regression is one such method for modeling the relationship between a set of explanatory variables and a dependent variable by fitting a linear equation to observed data. Here, we investigate and discuss some factors that influence whether the resulting regression equation is a credible model of the data.

Introduction

Multiple linear regression (e.g., Draper and Smith, 1981) is a technique for finding a linear relationship between a set of explanatory variables (x_i) and a dependent variable (y): $y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n$. The coefficients, ($b_i$) provide some indication of the explanatory variables effect on the dependent variable. With the wide availability of personal computers and the inclusion of regression routines in commonly available statistics or spreadsheet software such as Microsoft Excel©, there is an increased recognition of the value of gaining insight from data. There are also free web servers (e.g., Autofit http://www.lava.net/~seekjc) for fitting data to linear models. As a consequence, multiple linear regression is being applied to a wide array of problems ranging from business to agriculture. The goal of this application is to 'convert data to information'. Such information might help guide future decision-making. For example, many lenders use a credit score to help determine whether to make a loan. This score is a combination of many factors such as income, debt, and past payment history which positively or negatively affect the credit risk of a borrower. In this paper, we show that multiple regression as used in practice can produce models that are unacceptable to experts and layman because factors that should positively affect a decision may have negative coefficients and vice versa. We introduce a constrained form of regression that produces regression models that are more acceptable.

To illustrate the problem we are addressing, consider forming a model of professional baseball players' salaries as a function of statistics describing the players' performance. An agent representing a player might use the model as part of an argument that the player is underpaid. A player could use the model to determine how to improve certain aspects of his game to increase his salary. We have created a model of baseball players' salaries in 1992 as a function of the players' performance in 1991 using data on 270 players collected by CNN/Sports Illustrated and The Society for American Baseball Research. The model is listed below:

```
s=-180+10r+5hit+0.9obp+15hr+14rbi-0.8ave-18db-39tr
```

where `s` is the salary (in thousands), `r` is the number of runs scored, `h` is the number of hits, `obp` is the on base percentage (between 0 and 1000), `rbi` is the number of runs batted in, `ave` is the batting average (between 0 and 1000), `db` is the number of doubles and `tr` is the number of triples. Most people knowledgeable about baseball are confused by the negative coefficients for `ave`, `db`, and `tr`. It is unlikely that someone familiar with the sport would consider this insightful or advise anyone to act upon this model. If a baseball player interested in maximizing his income were to follow this equation literally, he would always stop at first base when hitting, rather than trying for a double. In this paper, we discuss why incorrect signs occur in multiple linear regression and present some alternative means of inferring linear models from data that do not suffer from this problem.

In multiple linear regression, the best equation fitting the data is found by choosing the coefficients (b_i) so that the sum of the squared error for the training data points is minimized. The coefficients, b_i, can be found with matrix manipulations, also known as a least squares approach (Draper and Smith, 1981). Mullet (1976) discusses a variety of reasons that multiple linear regression produces the "wrong sign" for some coefficients:

- Computational Error. Some computational procedures for computing least squares have problems with precision when the magnitudes of variables differ drastically. To avoid this problem, we internally convert variables to standard form (i.e., 0 mean, and unit variance) for calculations and convert back to the original form for displaying the coefficients.
- Coefficients that don't significantly differ from zero. In this case, the sign of the coefficient does not matter because it is small enough so that it does not significantly affect the equation. One recommended way to avoid this problem is to eliminate these

irrelevant variables. Forward stepwise regression methods (Draper and Smith, 1981) do not include a variable in the model unless the variable significantly improves the fit of the model to the data. We used this method (with an alpha of 0.05) to model the baseball data and obtained the following equation:

```
s=-114+16r+17rbi-59tr
```

Although this equation reduces the number of violations, it still has the wrong sign for the variable `tr`.

- Multicollinearity. When two or more explanatory variables are not independent, the sign of the coefficient of one of the variables may differ from the sign of that coefficient if the variable were the only explanatory variable. One approach to deal with this problem is manually eliminating some of the variables from the analysis.

In this paper, we consider an alternative approach to address the "wrong sign" problem in multiple regression. The goal is to produce linear models that are as accurate predictors of the dependent variable as the least-squares model but are more acceptable to people knowledgeable in the domain. Following the methodology commonly used in machine learning, we evaluate accuracy not by goodness of fit to a collection of data but by the ability to generalize to unseen data. Furthermore, we report on an experiment that evaluates what types of linear models are more acceptable to people in the baseball salary domain.

Constrained Regression

We hypothesize that a linear model is more acceptable to people knowledgeable in a domain when the effect of each variable in the regression equation in combination with the other variables is the same as the effect of each variable in isolation. That is, if in general, baseball salaries increase as the number of doubles increases, we would like the sign of the coefficient of this variable to be positive in the full linear model. Here, we propose and evaluate three methods to constrain regression to make this true. We call this constraint the *independent sign bias*.

Independent Sign Regression (ISR*)* treats the problem of fitting the linear model to the data as a constrained optimization problem: i.e., find the regression coefficients (b_i) that minimize the squared error on the training data subject to the constraint that all coefficients must have the same sign as they would in isolation (simple regression). There are many numerical algorithms for performing constrained optimization, and Lawson and Hanson (1974) present a comprehensive set for this case. The new contribution in ISR is that the constraints (i.e., the sign of the coefficient) are determined automatically by analysis of the data. Explanatory variables positively correlated with the dependent variable have a positive sign, while those negatively correlated have a negative sign.

We used ISR to create the following model of the baseball salary data:

```
s=-207+15r+0.8hit+11hr+11rbi+0.33ave+5db
```

In the next sections, we evaluate how well a constrained form of regression fits the data and whether people prefer regression equations with this constraint. Here, we note that the signs agree with our intuition. However, it does eliminate some variables such as the on base percentage (`obp`). This occurs because the best fit to the data subject to the constraint that `obp` $\geq$ `0` is that `obp = 0`. This occurs because `obp` is correlated with other variables such as `ave` (r = 0.81).

Next, we consider how to modify forward stepwise regression to constrain the sign of the variable. In forward stepwise regression (Draper and Smith, 1981), we start with an empty set of variables and then add the single variable that improves the model's fit to the training data the most. We continue this process of adding variables to those present until we have either included all variables or the remaining variables do not significantly improve the fit based on the partial F-test (Draper and Smith, 1981). *Independent sign forward regression* (ISFR) modifies this procedure by adding the constraint that the entering variable must also not result in sign violations (i.e., we add the variable that improves model fit the most subject to the constraint of no sign violations in the fitted equation). ISFR produces the following equation on the baseball data:

```
s=-148+15r+15rbi
```

The previous two constrained forms of regression both may eliminate variables. Here, we introduce a form of regression we call *Mean Coefficient Regression* (MCR) that does not eliminate variables but ensures that the signs agree with the sign in isolation. MCR finds the regression coefficients for each of the variables in isolation and then simply uses those values (dividing by the number of variables) for the multiple regression case. This is equivalent to treating each variable as a predictor and then averaging the results. The intercept is found automatically through the conversion of coefficients from standard form to the original scaling and minimizes the mean squared error of the linear equation with those coefficients. If all of the explanatory variables are uncorrected, MCR would produce the same equation as multiple linear regression. MCR produces the following equation on the baseball data:

```
s=-162+4r+2hit+1.1obp+10hr+3rbi+1.2ave+9db+16tr
```

Accuracy and the Independent Sign Bias

In this section, we evaluate the five regression algorithms on the several data sets. In each case, we report the squared multiple correlation coefficient (R^2), which is the percent of the total variance explained by the regression equation, and the descriptive mean squared error (MSE) of the regression routines on the entire data set. Both of these statistics measure how well the algorithms fit the data. Note that multiple linear regression always has the best fit to the training data, because it by definition minimizes squared training error. The more constrained forms of regression are limited in their ability to fit the data. We also report on the predictive mean squared error, which measures the ability of the regression algorithm to produce models that generalize

to unseen data. The predictive mean squared error is found by 5-fold cross-validation: The entire data set is randomly divided into five equal sized partitions. The data from four of the partitions is used to form a linear model that is evaluated on the fifth partition. This is repeated five times with each partition used exactly once for evaluation. The predictive MSE is almost always higher than the descriptive MSE. However, the algorithm with the lowest descriptive MSE does not necessarily have the lowest predictive MSE because the less constrained algorithms can overfit the data.

When evaluating the five regression algorithms, we also report on the number of sign violations where a sign violation occurs if the sign of the coefficient in the equation differs from the sign of the coefficient in the simple regression case. In this work, the principle goal is not to find regression routines that generalize better than multiple linear regression, but to find routines that generalize equally well and produce equations that people would be more willing to use.

We ran each of the 5 regression approaches on six data sets available from either the Statlib repository (http://www.stat.cmu.edu) or the UCI archive of databases (http://www.ics.uci.edu/~mlearn). The databases Autompg, Housing, and Pollution deal with automobile mileage, housing costs, and mortality rates respectively. CS Dept is available from the Computing Research Association (http://www.cra.org) and involves computer science department quality ratings. The Alzheimer's database was collected by UCI's Institute of Brain Aging and Dementia and involves predicting the level of dementia from the results of tests that screen for dementia.

In all of these domains, a sign violation could cause credibility problems. For example, in the CS Dept domain, linear regression indicated that the more publications per faculty member the lower the quality of the program, while in isolation this variable has the opposite effect.

Table 1. Summary of five approaches to creating linear models on six data sets

Database	Multiple Linear Regression	Independent Sign Regression	Mean Coefficient Regression	Stepwise Forward Regression	Independent Sign Forward Regression
Alzheimer					
R^2	0.750	0.743	0.420	0.727	0.727
Descriptive MSE	0.124	0.127	0.287	0.135	0.135
Predictive MSE	0.184	**0.166**	0.297	**0.166**	**0.166**
Violations	7	0	0	0	0
Autompg					
R^2	0.849	0.844	0.500	0.844	0.844
Descriptive MSE	9.3	9.47	30.5	9.4	9.55
Predictive MSE	10.6	**10.5**	30.8	10.7	10.6
Violations	4	0	0	1	0
Baseball					
R^2	0.478	0.472	0.375	0.476	0.470
Descriptive MSE	8e+5	8.09e+5	9.57e+5	8.02e+5	8.11e+5
Predictive MSE	8.74e+5	8.55e+5	9.66e+5	8.37e+5	**8.33e+5**
Violations	3	0	0	1	0
CS Dept					
R^2	0.859	0.858	0.414	0.844	0.844
Descriptive MSE	0.135	0.136	0.559	0.148	0.148
Predictive MSE	0.244	**0.213**	0.605	0.24	0.24
Violations	1	0	0	0	0
Housing					
R^2	0.740	0.698	0.328	0.740	0.696
Descriptive MSE	21.9	25.6	56.7	21.9	25.7
Predictive MSE	**23.7**	27.6	57.2	24.4	27.7
Violations	3	0	0	0	0
Pollution					
R^2	0.768	0.728	0.224	0.719	0.719
Descriptive MSE	895	1.04e+3	2.96e+3	1.08e+3	1.08e+3
Predictive MSE	3.53e+3	**1.6e+3**	3.3e+3	1.82e+3	1.78e+3
Violations	5	0	0	2	0

In Table 1, we show the results of the five regression approaches on the six data sets. The best (lowest) predictive MSE is shown in bold to allow simple comparison of the predictive ability. It is typical in such a simulation that no algorithm stands out as uniformly superior on all problems. However, the results indicate that independent sign regression is usually at least as accurate as multiple linear regression. Due to the sign violations, one might prefer to use independent sign regression. Mean coefficient regression does not fit the data nor generalize as well as the other regression algorithms. Independent sign forward regression is usually at least as accurate as stepwise forward regression.

Note that multiple linear regression has at least one sign violation on every data set. This shows that correlated variables frequently occur in naturally occurring databases and that the techniques designed to correct for sign violations may be applicable to a broad range of problems. Although stepwise forward regression mitigates the problem of sign violations, it does not eliminate it entirely.

In the next section, we report on the results of an experiment in which subjects indicate their willingness to use regression equations to make predictions. The goal of the study is to determine whether subjects have a preference for the independent sign bias: i.e., equations in which the sign of each coefficient is the same as the sign in isolation.

Baseball Salary Experiment

In this experiment, subjects are asked to imagine that they are an agent representing a baseball player. Subjects are shown various linear equations and told that they "might be used as a starting point to get a rough estimate of what a player should be paid." For each equation, they were asked to indicate on a [-3,+3] scale "How willing would you be to use this equation as a rough estimate of a baseball player's salary?" We are interested in exploring whether subjects have a preference for regression equations without sign violations.

We hypothesized that subjects would give higher ratings to equations formed with Independent Sign Regression and Mean Coefficient Regression to equations formed with Multiple Linear Regression because such equations do not contain sign violations. Note that Independent Sign Regression does not necessarily use all of the variables, and on the baseball data it typically uses 4-6 of the 8 variables.

We also hypothesized that subjects would give higher ratings to equations found with Independent Sign Forward Regression than Stepwise Forward Regression because they also did not contain sign violations.

Subjects. The subjects were 47 male and female undergraduates attending the University of California, Irvine who indicated that they were somewhat or very familiar with baseball. The subjects participated in this experiment to receive extra credit in an artificial intelligence course. We did not enroll subjects with little or no familiarity with baseball in the study.

Stimuli. The stimuli consisted of 15 linear equations that were displayed to the user in a web browser. Figure 1 contains an example of the type of stimuli used. Three equations were generated by each of five different regression routines:

- Multiple Linear Regression
- Independent Sign Regression
- Mean Coefficient Regression
- Stepwise Forward Regression
- Independent Sign Forward Regression

Three different equations for each algorithm were formed on different random subsets of the baseball data resulting in different coefficients. The coefficients of the equations were rounded to two significant digits. The stimuli were presented in random order for each subject.

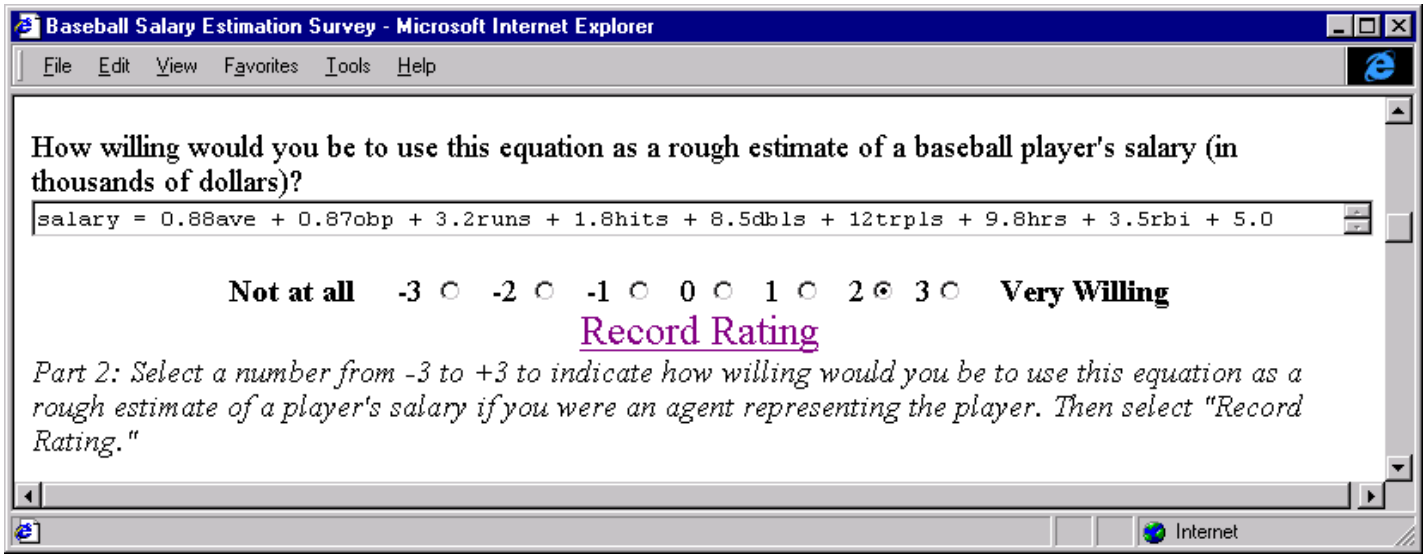

Figure 1. An example of the stimuli used in the experiment.

Procedures. Each subject was shown a single equation at a time, in random order, and asked to indicate on a scale from -3 to +3 how willing they would be to use the equation as a rough estimate of a player's salary by clicking on a radio button. Next they clicked on "Record rating" and were shown another equation. The radio button was reset to 0 before displaying the next equation. This continued until the subject rated all 15 equations.

Results. The average rating of all subjects for each type of equation is shown in Table 2.

Table 2. Average Subjects Ratings for Linear Equations.

Regression Algorithm	Mean Rating
Multiple Linear Regression	-0.816
Independent Sign Regression	0.603
Mean Coefficient Regression	0.851
Stepwise Forward Regression	-1.09
Independent Sign Forward Regression	-0.113

We analyzed the results of the experiment as follows. For each subject, we found the mean rating of the three equations generated by each of the five algorithms. An analysis of variance showed that the algorithm had a significant effect on the rating $F(4,184) = 22.11$, $p < .0001$. A Tukey-Kramer test at the .05 level was used to evaluate the three hypotheses. The critical difference is 0.706 so all three differences are significant:

- Subjects gave significantly higher ratings to equations found by independent sign regression than equations found by multiple linear regrssion.
- Subjects gave significantly higher ratings to equations found by mean coefficient regression than equations found by multiple linear regression.
- Subjects gave significantly higher ratings to equations found by independent sign forward regression than equations found by stepwise forward regression.

Discussion. The results support the notions that subjects have a preference for linear models that conform to the independent sign bias, i.e., those in which the sign of the coefficients of each explanatory variable agrees with the effect that the explanatory variable has in isolation.

The independent sign regression routine introduced in this paper automatically determines the sign of the coefficient of explanatory variables, produces models that have similar predictive accuracy as those produced by multiple linear regression, and produces linear models that subjects would be more willing to use. A possible disadvantage of the independent sign regression routine is that it may eliminate some variables from the linear equation. This may be a benefit in some cases (e.g., if it was expensive to collect some variables) or if simplicity is a consideration. However, the results of the experiment in this paper suggest that ignoring many explanatory variables may reduce the willingness of subjects to use a linear model. Although this is not a focus of this study, it appears that both independent sign forward regression and stepwise forward regression received relatively low rankings by subjects.

If one is interested in eliminating variables, the simulations showed that independent sign forward regression produces equations with similar predictive accuracy to stepwise forward regression and the experiment showed that users preferred equations created by independent sign forward regression.

We proposed mean coefficient regression as a means of eliminating sign violations while using all explanatory variables. Although it received the highest average ranking by subjects in our experiments, it does not fit the data nor generalize as well as the other regression routines. We suspect that our subjects are sensitive to the sign and perhaps order of magnitude of the coefficients, but there's no reason to believe they'd be able to determine whether two similar equations with slightly different coefficients are a better fit to the data. The inferior accuracy of mean coefficient regression is a result of correlations between the explanatory variables. These same correlations result in multiple linear regression getting the "wrong sign" on the coefficients. It remains an open question whether a linear model can be found that has similar accuracy to multiple linear regression, gets the signs right, and uses all of the variables when there are correlations among the explanatory variables. Because there is a relationship between averaging multiple linear models and mean coefficient regression, it is possible that some of the methods for correcting for correlations in linear models (e.g., in Leblanc and Tibshirani, 1993; Merz and Pazzani, in press) may be useful in this case.

Related Work

The purpose of the independent sign bias is to produce linear models that are as accurate as those produced by multiple linear regression, yet are more acceptable to users because they do not violate the users' understanding of the effect that each explanatory variable has on the dependent variable. If knowledge-discovery in database systems is to produce insightful models that are deployed in practice, it is important that users be willing to accept the models. One implication of this bias is that as additional explanatory variables are added to a model, the magnitude of the effect of the other variables may be changed but not the direction of the effect (cf. Kelley, 1971).

Credit scoring is one important application that may benefit from the global sign bias. In this application, the risk that a borrower may not pay back a loan is assessed as a function of a number of factors such as income, debt, payment history, etc. If a potential borrower is turned down for a loan, it is necessary to explain why. It is important to get the signs of the coefficients right on the models so that the explanation makes sense to the lender and the borrower.

Here, we have introduced constrained regression routines that produced linear models conforming to the independent sign bias. Monotone regression (Lawson and Hanson, 1974) is a related type of constrained regression in which the user indicates the sign constraint on the variables. In contrast, in independent sign regression, the sign is determined automatically.

Having the wrong sign in the regression equation results from having correlated explanatory variables. One way to deal with this is to introduce additional variables to represent

the interaction between two explanatory variables. The focus of such work has been to produce models that improve the fit of the data to the model and not to improve the comprehensibility or acceptance of the learned models.

In training artificial neural networks, weight decay (Krogh,. & Hertz, 1995) Sill, & Abu-Mostafa,, 1997) have been proposed as techniques for constraining models. However, the focus has been on improving generalization ability and not improving the user acceptance of the learned models.

Causal Models (Spirtes, Glymour, and Scheines, 1993) and Belief Networks (Pearl, 1988) also explicitly represent the dependencies among variables. The resulting models are more complex than linear models. In this work, we have adopted a fixed representation (linear equations) and addressed what constraints can be imposed upon this representation to improve user acceptance.

In previous work (Pazzani, Mani & Shankle, 1997), we addressed a related problem of rule learning algorithms including counterintuitive tests in rules by having an expert provide "monotonicity constraints." For nominal variables, a monotonicity constraint is expert knowledge that indicates that a particular value makes class membership more likely. For numeric variables, a monotonicity constraint indicates whether increasing or decreasing the value of the variable makes class membership more likely. By showing neurologists rules learned with and without these constraints, we showed that monotonicity constraints biased the rule learning system to produce rules that were more acceptable to experiments.

Pazzani (1998) extended the work on monotonicity constraints by introducing the globally predictive test bias. In this bias, every test in a rule must be independently predictive of the predicted outcome of the rule. Such a bias eliminated the need for a user to specify monotonicity constraints but provided the same advantages. The globally predictive test bias in rule learners is analogous to the independent sign bias in linear models in that the effect of a variable in combination with other variables is constrained to be the same as the effect of that variable in isolation. Here we have shown that such a constraint improves the willingness of people to use linear models without harming the predictive power of the models.

Conclusions

People are not computers and cannot easily find a linear equation that best fits a data set with 10 variables and 500 examples. However, we argue that people have certain constraints on the qualitative properties of the linear equations. We have shown that one factor that influences the willingness of subjects to use linear models is the independent sign bias. By creating regression routines that conform to this bias, we constrain the computer to produce results that are more acceptable to people.

New regression routines were produced that implement the independent sign bias. Experiments and simulations showed that independent sign regression produces linear equations that are approximately as accurate as multiple linear regression and that are more acceptable to users. Independent sign forward regression is similarly preferable to forward stepwise regression.

Acknowledgements

This research was funded in part by the National Science Foundation grant IRI-9713990. Comments by Dorrit Billman and Susan Craw on an earlier draft of this paper help to clarify some issues and their presentation.

References

Draper, N. and Smith, H. (1981). *Applied Regression Analysis.* John Wiley & Sons.

Kelley, H. (1971). Causal schemata and the attribution process. In E. Jones, D. Kanouse, H. Kelley, N. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp 151-174). Morristown, NJ: General Learning Press.

Krogh, A. & Hertz, J. (1995). A Simple Weight Decay Can Improve Generalization Advances in Neural Information Processing Systems 4, Morgan Kauffmann Publishers, San Mateo CA, 950-957.

Lawson, C. L. & Hanson, R. J. (1974). *Solving least squares problems*. Prentice-Hall.

Leblanc, M. & Tibshirani, R. (1993). Combining estimates in regression and classification. Dept. of Statistics, University of Toronto, Technical Report.

Merz, C. & Pazzani, M. (in press). *A Principal Components Approach to Combining Regression Estimates*. Machine Learning.

Mullet, G. (1976). *Why Regression Coefficients Have the Wrong Sign.* Journal of Quality Technology. 8(3), 121-126.

Pazzani, M., Mani, S., & Shankle, W. R. (1997). *Comprehensible Knowledge-Discovery in Databases.* In M. G. Shafto and P. Langley (Ed.), Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society (pp. 596-601). Mahwah, NJ: Lawrence Erlbaum.

Pazzani, M. (1998). Learning with Globally Predictive Tests. The First International Conference on Discovery Science Fukuoka, Japan.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Palo Alto.

Sill, J. & Abu-Mostafa, Y. (1997). Monotonicity Hints. Advances in Neural Information Processing Systems 9, Morgan Kauffmann Publishers, San Mateo CA, 634-640

Spiegelhalter, D., Dawid, P., Lauritzen, S. and Cowell, R. (1993). *Bayesian Analysis in Expert Systems*. Statistical Science, 8, 219-283.

Spirtes, P., Glymour, C. and Scheines, R. (1993). Causation, Prediction, and Search, New York, N.Y.: Springer-Verlag.