AN APPLICATION OF STOCHASTIC OPTIMIZATION TO COMBINE TWO FILES

Richard A. Griffin , Chilton Research Services

Introduction

Chilton Research Services has conducted an empirical study on methods to perform an operation we call FUSION. Two microdata files are available, the first includes a common set of variables, A, and the second includes the set A variables and a larger set of variables, B. This paper describes the creation of a synthetic sample for the first file that includes the actual A variables for each record along with the B variables. The technique used was developed by Gerhard Paass (Paass 1989). The procedure uses stochastic optimization to produce a file with best fit to available marginal distributions, followed by adding small deviations from the optimal solution to maximize entropy. The result is a synthetic file which may be considered as a sample from the distribution estimated from the EM-algorithm. For the empirical study, A and B variables are available for both files providing for a thorough evaluation.

Background (Paass 1989)

For simulation analyses in social science and economics comprehensive samples of persons, firms, or other simulation units are often required. Often, however, such microdata files are not available because a sample with the specific vector of all the required variables has never been collected or cannot be published because of privacy issues. To perform the desired analyses the investigators have to construct a file from available incomplete sources such as a number of microdata files each of which contain a subset of the total number of variables needed for analysis. If the number of variables is small the well known iterative proportional fitting algorithm (IPF) can be used to estimate a joint discrete distribution from available marginal tables. If there are more than 10-20 variables each of which has two or more possible responses the number of cells whose probabilities have to be updated by the IPF gets prohibitively large (combinatorial explosion) and the procedure can no longer be applied. The Stochastic Modification Algorithm (SMA) is a method which can construct a synthetic data file for which each record has all the desired variables. This synthetic data file has similar properties as a sample from the optimal distribution generated by the IPF. The algorithm can be carried out even if the number of variables is large since it operates on the synthetic sample instead of the set of all basic cells. The resulting synthetic data file can be used as input to microanalytic models. Unlike the IPF, where weights of fixed record are modified, the SMA changes the values of variables in the different records. This is controlled by a stochastic optimization procedure which improves the fit between the synthetic sample and the given marginals according to the maximum likelihood principle (or some other cost function such as the Chi-Square). The algorithm is able to generate approximate solutions with reasonable computational effort and eventually converges to a global optimum.

The input to the SMA is a set of J marginal tables each of which consists of a number of observed counts and a starting synthetic data file of n records that has all the variables for each record. In this paper we will deal with categorical variables only. Counts for the synthetic data file are compared to the observed counts using a cost function such as maximum likelihood, Chi-Square, or Kullbacks minimum discrimination information statistic. A stochastic optimization method called the simulated annealing algorithm is used to make random changes to the synthetic data file in an iterative fashion. This process is expected to converge to a unique equilibrium distribution that no longer changes as the algorithm proceeds. At each step of the iterative process one record is selected at random and one variable for that record is randomly selected to be randomly changed. The cost values for the synthetic data file before and after the change are compared. If the change results in lower cost, the change is accepted.

If the change results in higher cost it is accepted with a specified probability that is close to 1 at the start of the algorithm and gradually approaches 0 as the algorithm proceeds. Initially accepting changes that increase the cost is necessary so that the limit distribution of the process concentrates on samples with minimal cost. For most problems for which SMA is used there will be multiple solutions with minimal cost. According to the maximum entropy principle it is sensible to select from these multiple solutions one with maximum entropy. This is accomplished by modifying the cost function so that changes which produce a small increase in cost are accepted. The optimal counts are changed by 1 or 2 which permits a free fluctuation in the vicinity of the optimum without significant changes in the overall fit.

Application

Mediamarc Research Incorporated (MRI) uses the field interviewing department of Chilton Research Services to conduct the MRI Media Study. The MRI Media Study is a nationwide personal interviewing survey of 20,000 respondents annually. Information is collected about consumer interests and involvement in newspapers, magazines, television and radio. Through this information it is hoped that the consumer can be better served because publishers and broadcasters will have a better idea of the various kinds of publications and broadcast media that consumers find interesting. This study is an important one to many people in all lines of business throughout the U.S. It is a syndicated study and, therefore, all the data collected is put together and given to clients as a whole.

At each sample household, an adult is selected at random and a questionnaire providing key demographic data is completed. A Product Booklet is left at each sample household and the interviewer makes arrangements to pick it up after it is completed. The product booklet is about 100 pages long and asks questions about thousands of products. For most products the respondent indicates if they have used the product in the last specified time period (i.e., 6 or 12 months) and how often in another time period (such as in a day for toothpaste or last 12 months for a brand of oil filter).

MRI is interested in having information about product usage for persons for whom only the demographic data is available. Referring to the discussion in the introduction, the A variables would be the demographic variables obtained from the questionnaire filled out by the interviewer at the first visit. The B variables are the questions asked in the Product Booklet. For the empirical study, we have 2000 records for which all the demographic variables (A) and all the product variables (B) are available. These records are randomly divided into one group of 1000 for which we assume both A and B are available and the remaining 1000 for which just A is available. We want to run SMA on this remaining 1000 using distributions from the first 1000 as the observed counts to which we are trying to get a good fit. For this work we reduced the number of B variables to 196 magazine variables each of which was recoded as a binary variable where 1 indicated purchasing the magazine and 0 indicated not purchasing the magazine and 30 product usage variables such as car phone, type of car, fast food, and camping equipment. Seventeen of the product variables were binary, six had three possible responses, 3 had 4 responses, and the remaining 4 had 5,6,8 and 9 responses respectively. There were six A demographic variables.

Since the 1000 records with no magazine or product data are a random sample from the same population as the 1000 records for which we have all the data, we can use the observed joint distributions of the variables from the 1000 records with A and B variables as target distributions for the SMA synthetic data file. In addition both groups have 1000 records so we can deal with counts instead of percentage distributions. Joint distributions are of primary importance since we want to capture any correlations that may exist. With 232 magazine, product and demographic variables we have many more possible sets of joint distributions we could consider than are practical. We decided to only look at the joint distributions of pairs of variables within magazine items or within product items and at pairs of items involving a demographic variable and either a magazine variable or a product variable. However, the combinations of 196 magazine variables taken two at a time is 19,110 and this is still many more controls than is practical.

Thus for magazines we formed groups of similar items and considered all possible pairs of items within each group. There were 32 magazine groups formed. For the magazine groups we used SAS to compute the odds ratio and 95% confidence interval for the odds ratio for each 2X2 table formed by each pair of magazines within each group (all magazine variables were binary). A odds ratio close to 1 indicates little relationship between the magazines in a pair. Out of 945 pairs of variables we selected 182 for which the odds ratio confidence interval indicated the strongest relationship between the two magazine variables.

There were six demographic items common to both files (set A variables). These were age of household head, number of persons in household, presence of children, sex of respondent, head of household income, and race. Each of these six demographic variables was crossed with each of the 196 magazine variables for a total of 1176 cross-classifications. Exactly one-half of these were 2X2 tables and the other half were 2XN with N>2. For the 2X2 tables, 74 were selected for control using the odds ratio. For the others, the p value from a Chi-Square test of independence was calculated. Pairs with small p values were selected for use as target control distributions. Thus, 147 additional tables were selected for control for a total of 74+147+182 = 403 magazine variable control tables.

A similar procedure was done for the product variables. There were 91 product variable pairs selected for control for which either the odds ratio indicated a strong relationship from a 2X2 table or a small p value from a Chi-Square test for independence indicated a strong relationship for a 2XN table with N>2. Similarly 100 tables, indicating a strong relationship between a demographic variable and a product variable, were selected for control for a total of 191 control tables involving product variables.

Finally, any magazine or product variable that was not included in any pair of items involved in a control table cross-classification was designated for control as a table defined by the frequency distribution of that variable. There were 28 such variables resulting in a grand total of 622 control tables.

To produce the starting synthetic file the set B variables for each of the 1000 records with both set A and B variables available were randomly assigned to the 1000 records with the set B variables missing. Thus each record in the starting synthetic file consisted of the actual set A demographic variables for one of the 1000 records for which set B variables are not available and the set B variables from one of the records for which all variables are available.

The simulated annealing algorithm was applied as follows.

Let Xi denote the present synthetic sample (this is the starting synthetic data file).

1. Set B = 1000
2. for t = 1 to 1000
3. a record from the present synthetic sample and a variable on that record are randomly selected. The selected variable is randomly changed giving a modified synthetic sample Xj.
4. The Chi-Square cost values C(Xi) and C(Xj) are computed.
   (Basically the Chi-Square cost for a synthetic sample is the sum over all the cells in the control tables of (C-S)^2/S, where C is the count in the cell from the control table and S is the count in the same cell using the synthetic sample).
5. the probability, Pacc(j/i), of accepting the modification is given by

$$\text{Pacc}(j/i) = \begin{cases} 1 & \text{if } C(Xj) < C(Xi) \\ \exp((C(Xi)-C(Xj))/B) & \text{otherwise} \end{cases}$$

6. Draw a uniform random number between 0 and 1. If it is smaller than Pacc(i/j) then Xj becomes the present sample. Otherwise keep the old sample Xi.
7. next t
8. B = m*B with m=.9 at the start
9. go to 2
10. continue until B = 100.

The initial B = 1000 is selected so that at the start almost all modifications are accepted. At the end of the process if convergence has not been reached, the process is continued from where it left off until B = 1; m can also be changed. Appropriate values are determined experimentally to reach convergence.

The simulated annealing algorithm gives multiple solutions with minimal cost. This means that a different synthetic sample and/or different parameter selections (starting and stopping B values, and m) can result in different final synthetic samples. It is sensible to select from these multiple solutions one with maximum entropy (or "disorder"). Maximum entropy is likely in the "real world".

In order to understand entropy consider the following example:

The entropy, E, of a discrete distribution which takes on the value 1 with probability p and the value 0 with probability 1-p is given by -E = plogp + (1-p)log(1-p) so d(-E)/dp = log(p/(1-p)). Setting equal to 0 gives p = .5. Since the second derivative is greater then 0, p = .5 is a relative minimum for -E or a relative maximum for E.

We want a synthetic sample with a cost value near the minimum which has maximum entropy. Allowing a small insignificant increase in cost this is accomplished by stochastic modifications.

Denote Xopt as the synthetic sample that comes out of the simulated annealing algorithm. Then define

$$C't(X) = \begin{cases} Ct(Xopt) \text{ if } Ct(X) <= Ct(Xopt) + z \\ Ct(X) \text{ otherwise} \end{cases}$$

where Ct(X) is the Chi-Square cost for control table t for synthetic sample X and C(X) is the sum over all control tables t of Ct(X) as a modified cost function. This means if an alternative synthetic sample X has cost greater for a control table t than the cost of Xopt for table t by an amount less than or equal to z then we assign table t the optimal cost value (i.e., Ct(Xopt)). If not, we assign table t Ct(X). If the cost of X is less than the cost of Xopt for table t, it is also assigned the optimal cost value. Note : This is possible since C(Xopt) is minimum for the sum of the cost values over all the control tables, not necessarily for each table.

Assuming the cell counts are all larger than three, the optimal count may be changed by 1 or 2 without significantly increasing the total cost. Changes of 1 or 2 are required for a free fluctuation in the vicinity of the optimum.

Starting with Xopt the simulated annealing algorithm is performed with cost function C'(X) equal to the sum over all control tables of C't(X) and B = 0 (actually the probability of acceptance is set equal to 0 if the modified cost is greater than the present cost).

Hence modifications are accepted only if C'(X) = C'(Xopt).

The cut off threshold is to be large enough so that there are small changes to many of the control cells.

For the simulated annealing algorithm with B = 0 there is just one loop of the first cycle so that 1000 new synthetic samples were tried. The z value is set experimentally by monitoring the amount of changes in the control counts. There must be small changes in many of the cells.

The suggestion by Paass is to assume that for a table with k cells 2Ct(X) is distributed as a Chi-Square random variable with k-1 degrees of freedom. Pick a z value from this Chi-square distribution that is small enough so that only small increases in cost are accepted but large enough so that after one pass through the loop, there is a slight change in many control cells. For any given control table with k cells the z value used was the 5 percentile point of a Chi-square random variable with k-1 degrees of freedom.

Results

The final synthetic data file set B variables were compared to the actual set B variables to evaluate SMA. This was done for both the before and after entropy files. Chi-square goodness of fit tests and the Jaccard

Statistic were used. The Jaccard Statistic used was for 2X2 tables. Consider a magazine variable taking the value 1 if the magazine is purchased and 0 if the magazine is not purchased. The 2X2 table is for the true response crossed with the response on the synthetic data file. a is the count in the (1,1) cell; b is the count in the (1,0) cell; c is the count in the (0,1) cell; and d is the count in the (0,0) cell. The Jaccard Statistic is defined as a/(a+b+c) or the proportion of matches ignoring cases with a match of 0 (did not purchase the magazine). A Chi-square goodness of fit test was done comparing each of the 622 SMA control tables computed from the synthetic data file and from the actual data file. In addition 31 magazine group binary variables were formed. Each magazine variable was placed in one group and the group variable was 1 if the respondent purchased any one of the magazines in the group and 0 otherwise. Each of these 31 magazine variables was crossed with each of the 6 demographic variables to produce 192 more tables for which Chi-square goodness of fit tests were also done. Results were as shown in Table 1.

Table 1

	Before Entropy	After Entropy
Jaccard by respondent	.251	.249
Jaccard by variable	.077	.077
Total Chi-square for 622 SMA control tables	7,655	7,622
# tables with a good fit (5%)	447	456
Total Chi-square for 192 magazine groupXdemo tables	11,346	11,771
# tables with a good fit (5%)	89	83

The Jaccard statistics are almost exactly the same before and after entropy. They are quite small particularly by variable. The respondent statistic is calculated by computing the Jaccard value for each of the 1000 respondents summing over all the variables and then averaging over the number of respondents. The variable statistic is calculated by computing the Jaccard value for each variable summing over all the respondents and then averaging over the number of variables. Clearly there are many variables with a very low match rate. However the purpose of the operation is not really to predict the actual response of each individual. It is more important to produce a file which can be used to produce data at an aggregate level or do analysis. Thus the Chi-square tests are of greater importance.

For the 622 control tables, SMA does a adequate job. The after entropy file is slightly better producing a good fit at the 5% level of significance in 456 of the 622 tables (73.3%) as compared with 71.9% before entropy. However neither the before or after entropy files performed well for the 192 cross-tabulations between magazine groups and demographic variables. Less than 47% of the tables produced a good fit.

A similar test was performed for seven other techniques also using Chi-square goodness of fit tests and Jaccard statistics for analysis. The file that has all the data (set A and B variables) is called the donor file and the file that has just the common demographic variables (set A) is called the recipient file. These techniques all use the common variables (set A) as the basis for deciding which respondents in the donor file will pass data to respondents in the recipient file. Some statistical test of the similarity between donors and recipients is used to find the best pairings. All the missing data from the matching donor is transferred to the matching recipient. Thus these methods are all different than SMA in that all the imputed data comes from one respondent while SMA randomly changes individual variables to conform to available marginal distributions. Included were a random donor selection both over all possible donors and within classes matching on some demographic variables, selection based on linear programming optimization, selection based on FUSION techniques

(Antoine and Santini 1986 and 1987), and selection using multivariate techniques such as cluster analysis and discriminant analysis. These techniques produced similar results to SMA using the Jaccard statistics. As would be expected they did not do as well as SMA for the 622 SMA control tables; however, the Two Fusion techniques did almost as well as SMA. For the 192 magazine group by demographic variables tables, one did much worse than SMA, three did about the same, and the two FUSION methods did quite a bit better. About 53% of the 192 tables had a good fit at the 5% significance level.

Conclusion

SMA did a fairly good job on producing a synthetic data file that could be used to produce the same cross-tabulations as those used for control in the SMA process. These were selected as those pairs of variables with the strongest relationship hoping that those with less of a relationship would fall out in an acceptable way via the random process. If these had small correlation this was expected to occur. However, for the 192 tables not controlled on SMA did not do as well as we had hoped. The FUSION techniques did better for these tables but they only produced a good fit in slightly over 50% of the tables. For this data set it appears that well designed imputation methods that imputed all the data from one donor could do better. However, the limited amount of variables in common between the two files probably is the major reason why a better imputation method was not discovered.

References

Paass, Gerhard (1989), "Stochastic Generation of a Synthetic Sample from Marginal Information", Proceedings of the Fifth Annual Research Conference, Bureau of the Census, 1989, pp. 431-445.

Antoine, Jacques and Santini, Gilles, "Fusion Techniques: Alternative to Single-Source Methods", European Research, August 1987.

Antoine, Jacques and Santini, Gilles, "An Experiment to Validate Fusioned Files Obtained by the Referential Factorial Method", ESOMAR, Helsinki, April 1986.