W.P. No. 89-90-11

Management Sciences Research Report No. 560

The Weighted Uni-Dimensional Similarities Problem with Least Absolute Value Metric Is *NP*-Hard

by

Nathan P. Ritchey*
and
Gerald L. Thompson**

April 1990

* Department of Mathematical and Computer Sciences
Youngstown State University
Youngstown, OH 44555

** Graduate School of Industrial Administration
Carnegie Mellon University
Pittsburgh, PA 15213

This report was prepared as part of the activities of the Management Science Research Group, Carnegie Mellon University, under Contract No. N00014-85-K-0198 NR 047-048 with the Office of Naval Research and as part of the Contract No. DE-FG02-85ER13396 - I.H. with the Department of Energy. Reproduction in whole or in part is permitted for any purpose of the U. S. Government.

Management Sciences Research Group
Graduate School of Industrial Administration
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

The Weighted Uni-Dimensional Similarities Problem with Least Absolute Value Metric Is *NP*-Hard

by

Nathan P. Ritchey and Gerald L. Thompson

ABSTRACT

The purpose of this paper is to prove that the weighted uni-dimensional similarities problem with least absolute value metric (USPAM) is, in general, *NP*-Hard. In the first four sections of the paper, the USPAM problem and four lemmas are presented which will be used in Section 6 to prove the main theorem of this paper. It is shown that the simple max cut problem can, in a polynomial number of steps, be converted into a special case of the USPAM problem, which shows that the USPAM problem is *NP*-Hard. Finally, some special cases of the USPAM problem are described for which polynomial solutions exist.

Keywords:

Weighted similarities, *NP*-Hard, Simple Max Cut, Integer Programming

The Weighted Uni-Dimensional Similarities Problem with Least Absolute Value Metric Is *NP*-Hard

by

Nathan P. Ritchey and Gerald L. Thompson

1. Problem Description

The USPAM problem can be defined as follows: Given m attributes and $m(m-1)/2$ measurements of the distances d_{ij} between attributes i and j for $i,j=1,\ldots,m$ and $i < j$, find model locations z_i, $i=1,\ldots,m$ on the real line so that W is minimized, where

$$W = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (w_{ij}\left|d_{ij}-\left|z_j-z_i\right|\right|) \qquad (1)$$

and $w_{ij} \geq 0$ is the weight attached to the deviation between the measured distance d_{ij} and the model distance $|z_j-z_i|$.

Geometrically, a positioning of the m points on the number line is sought such that the sum of the absolute values of the differences between the observed distance d_{ij} and the model distance $|z_j-z_i|$ between pairs i and j is minimized. Problems such as this, but using a least squares metric, arise often in economics and psychometrics, see Poole, 1984. As far as we know, this is the first paper to consider an absolute value metric for the problem.

2. A Preliminary Result

LEMMA 1. If the observed distances, d_{ij}, $i,j=1,\ldots,m$, $j > i$, are integral, then the optimal model location of each z_i, $i=1,\ldots,m$ is at an integer location on the real line.

PROOF. Consider an arbitrary ordering of the z_i's, renumbered, if necessary, so that $z_1 \leq z_2 \leq \ldots \leq z_m$. The locally optimal solution for this ordering can be found by solving the following linear program.

$$\text{Min } W = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (w_{ij}e_{ij}^{+} + w_{ij}e_{ij}^{-}) \qquad (P)$$

subject to

$$z_j - z_i + e_{ij}^{+} - e_{ij}^{-} = d_{ij}$$

$$z_i, e_{ij}^{+}, e_{ij}^{-} \geq 0 \quad \text{for all} \quad i,j=1,\ldots,m, \; i < j.$$

Notice that the coefficient matrix formed by this constraint set is the transpose of a node-arc incidence matrix for a complete graph together with an identity matrix and a negative identity matrix, which is well known to be totally unimodular. Hence, the locally optimal placement of each point, z_i, for $1,\ldots,m$ given by the solution to the linear program (P) for this ordering, and in fact, for every feasible ordering, is integral. Since every local solution occurs at an integer point, the global solution must occur at an integer point, which completes the proof.

Remark 1. Since there are $m!/2$ different orderings, the global solution to (1) can be found by solving that many linear programs (P). However, this is a very inefficient solution procedure.

Remark 2. It is always possible to shift any local solution up or down the number line and not change the value of the objective function. This fact is obviously true since the objective function involves only distances between pairs of points.

Remark 3. From Lemma 1 and Remark 2 it follows that, after shifting, any feasible ordering has clusters of model locations at $r+1$ integer points $0,1,\ldots,r$ on the real line, with k_o positions located at 0, k_1 located at 1, etc., and $k_0+k_1+\ldots+k_r = m$.

3. Binary Data

Suppose that all of the observed weights and distances are required to be binary, that is, $w_{ij} = 0$ or 1 and $d_{ij} = 0$ or 1 for $i,j=1,\ldots,m$ and $j > i$.

The uni-dimensional similarities problem with absolute value metric and binary data is called the *binary USPAM problem.*

LEMMA 2. For any binary USPAM problem, there exists at least one optimal solution such that $z_i = 0$ or 1, for all $i=1,\ldots,m$.

PROOF. By Remark 2 above, we can shift any solution so that there is at least one i such that $z_i = 0$ and that there is no other index, k such that $z_k < 0$. From Lemma 1, we know that a global solution to this problem has integral values. Assume that an optimal solution has been found and that there exists at least one h such that $z_h \neq 0$ or 1. Therefore, $z_h \geq 2$.

By Remark 3 we know there are $r+1$ model locations n the real line with k_0 at 0, k_1 at 1, etc., with $k_0 + k_1 +\ldots+ k_r = m$. If $r \geq 2$ we will show that it is possible to move all the k_r model positions from location r to $r-2$ without increasing the objective function (1). Let z_j be located at integer r and z_i be located at integer $k < r$. There are two cases: $k=r-1$ and $k \leq r-2$.

For the first case, $k=r-1$ so that when z_j is located at r

$$\left||z_j-z_i|-d_{ij}\right| = \left||r-r+1|-d_{ij}\right| = |1-d_{ij}|$$

and when z_j is located at $r-2$ we have

$$\left||z_j-z_i|-d_{ij}\right| = \left||r-2 - r+1|-d_{ij}\right| = |1-d_{ij}|$$

which is the same.

For the second case, $r - k \geq 2$, so that when z_j is located at r we have

$$\left||z_j-z_i|-d_{ij}\right| = \left||r-k|-d_{ij}\right| \geq |2-d_{ij}|$$

and when z_j is located at $r-2$

$$\left||z_j-z_i|-d_{ij}\right| = \left||r-2 - k|-d_{ij}\right| > \left||2-2|-d_{ij}\right| = d_{ij}.$$

Because d_{ij} is a binary variable

$$|2-d_{ij}| \geq d_{ij}$$

which shows that moving z_j from r to r-2 does not increase the objective function in this case either.

All the other terms in (1) which do not involve the model location r, stay the same so the objective function is not changed by moving z_j from r to r-2.

By repeatedly applying this argument we can find an optimal solution which uses only the model locations 0 and 1, completing the proof.

4. Graphical Interpretation

The binary USPAM problem can be interpreted as a problem on the graph G(N,E), whose node set is $N = \{1,\ldots,m\}$ and whose edge set is $E = \{(i,j) | i \neq j$ and $d_{ij} = 1\}$. A feasible solution to the binary USPAM problem is a partition of N into two disjoint subsets S_0 and S_1 with S_0 containing the indices i such that $z_i=0$ and S_1 contains those with $z_i = 1$. In order to construct a partition that minimizes (1) try to use the following rules:

(a) if $d_{ij}=1$ for two nodes i and j, place i in one set and j in in the other

(b) If $d_{ij}=0$, place i and j in the same set.

An optimal partition is one that violates these two rules a minimum number of times, since violating either rule causes a penalty of 1 in the objective function (1).

It is easy to see that for some problems, this procedure cannot satisfy both of these rules for all pairs of indices. Consider the following simple example: Let $d_{ij} = 1$ for i=1,2,3 for $j > i$. Graph G is a triangle because by rule (b) edges (1,2), (1,3) and (2,3) are in E. The objective function for this problem is

$$W = \left|1-|z_1-z_2|\right| + \left|1-|z_1-z_3|\right| + \left|1-|z_2-z_3|\right|.$$

The minimum value of W is 1 and can be obtained (for instance) by choosing $z_1=0$, $z_2=1$, and $z_3=1$. It is impossible to satisfy both rules (a) and (b) for this example.

Let $\tilde{E} = \{(i,j) | i \neq j \text{ and } (i,j) \notin E\}$ be the complement of E; then it follows that

$$|E| + |\tilde{E}| = m(m-1)/2$$

where $|X|$ is the number of elements in set X.

Given a partition of N into two subsets S_0 and S_1 define the following:

(a) (i,j) is *external* if $i \varepsilon S_0$ and $j \varepsilon S_1$ or $j \varepsilon S_0$ and $i \varepsilon S_1$.

(b) (i,j) is *internal* if i and j belong to S_0 or i and j belong to S_1.

For the same partition of N we define

E_x the set of external edges of E

E_i the set of internal edges of E

$\tilde{E}_x$ the set of external edges of $\tilde{E}$

$\tilde{E}_i$ the set of internal edges of $\tilde{E}$.

From these definitions it follows that

$$|E_x| + |E_i| + |\tilde{E}_x| + |\tilde{E}_i| = m(m-1)/2. \qquad (2)$$

Using these definitions it is obvious that the binary USPAM problem can be restated as follows.

LEMMA 3. The binary USPAM problem is to choose S_0 and S_1 so as to optimize either of the two objective functions

(a) Minimize $W = |\tilde{E}_x| + |E_i|$

(b) Maximize $Z = |E_x| + |\tilde{E}_i|$.

PROOF. Statement (a) follows from the graphical interpretation of the problem. The equivalence of the two objective functions follows by rewriting (2) as

$$|E_x| + |\tilde{E}_n| = m(m-1)/2 - |\tilde{E}_x| - |E_i|$$

completing the proof.

LEMMA 4. An optimal solution to the binary USPAM problem gives $W=0$ if and only if G is a complete bipartite graph and S_0 and S_1 are chosen accordingly.

PROOF. A complete bipartite graph $G^* = (N^*, E^*)$ has the property that N^* can be partitioned into two subsets N_0 and N_1 such that E^* consists exactly of all of the edges (i,j) with $i \epsilon N_0$ and $j \epsilon N_1$. If we choose $S_0=N_0$ and $S_1=N_1$, then $W=0$. Conversely if $G = (N,E)$ and there exist subsets S_0 and S_1 such that $W=0$ then $|\tilde{E}_x| = 0$ and $|E_i| = 0$ so that $\tilde{E}_x=\phi$ and $E_i=\phi$ which implies that $E=E_x$ and $E_i=\phi$ so that G is a complete bipartite graph.

Remark 4. From the objective function in Lemma 3(a) another interpretation of the binary USPAM problem can be stated as follows. Given a graph G, choose a partition S_0, S_1 of N so that if a complete bipartite graph is constructed by adding the edges in $\tilde{E}_x$ and deleting those in E_i, the smallest total number of changes must be made.

5. The Simple Maximum Cut Problem

Closely related to the binary USPAM problem is the simple maximum cut problem which can be stated using the same notation as follows. Given a graph $G = (N,E)$ find a partition S_0 and S_1 of N so that the number of external arcs, E_x, connecting the two sets is maximized.

Although the maximum cut problem is known to be, in general, NP complete (Garey and Johnson, 1979), there is a considerable literature about the problem. Grotschel, et al., 1988, provide a good literature review of the problem. See also Barahona, 1983; Barahona, et al., 1985; Barahona, et al., 1986; and Fonlupt, et al., 1984.

6. Conversion of the Simple Maximum Cut Problem into a Binary USPAM Problem

We now show that an simple max cut problem can be converted into a binary USPAM problem in a polynomial number of steps, which will show that the binary USPAM problem is NP-hard.

THEOREM. The uni-dimensional similarities problem is, in general, NP-hard.

PROOF. Consider the simple maximum cut problem defined on a graph $G = (N,E)$. We show it is a special type of binary USPAM problem on the same graph with $d_{ij}=w_{ij}=1$ if $(i,j)\varepsilon E$ and $d_{ij}=w_{ij}=0$ if $(i,j)\varepsilon\tilde{E}$. The objective function for the binary USPAM problem is that given in Lemma 3(b). Using the weights defined above it is

$$\text{Maximize} \quad z=1|E_x| + 0|\tilde{E}_1| = |E_x| \tag{3}$$

because the weights are zero on edges $(i,j)\varepsilon\tilde{E}_1$. Since the objective function in (3) is exactly that of the simple max cut problem, the proof is complete.

There are certain instances where the USPAM problem can be solved in polynomial time. For example, if the data is perfect (i.e. the optimal objective function value, W=0), the positions of any pair of points, i and j, at a distance of d_{ij} away from each other will uniquely determine the positions of the rest of the points. Also, for the simple max cut problem, Barahona and Mahjoub, 1986, have presented a polynomial time algorithm for solving it on any graph, G, not contractible to K_5. This algorithm, called the separation algorithm, uses the ellipsoid method to solve LP problems. Earlier, Grotschell and Pulleyblank, 1981, presented an algorithm for weakly bipartite graphs which include graphs not contractible to K_5 and having nonnegative weights. Also, Orlova and Dorfman, 1972, and Hadlock, 1975, have used matching techniques and planar duality to solve planar max cut problems in polynomial time.

7. Conclusions

We have shown that the uni-dimensional similarities problem is, in general, NP-hard and have given several instances for which a polynomial solution method exists. In order to obtain a solution for a specific problem having general data, we could develop heuristic procedures or convert the problem into a 0-1 mixed integer program, see Ritchey, 1989. Many of the heuristics developed for the traveling salesman problem can be used in a solution process for this problem. Regardless of the chosen method of solution, at each iteration of a method, an L_1 estimation problem must be solved. Therefore, it is likely that good solutions to a larger problem, ($m > 20$), will be computationally expensive to obtain.

Bibliography

[1] Barahona, F. (1983), The Max Cut Problem in Graphs Not Contractible to K_5, *Oper. Res. Let.* 2, 107-111.

[2] Barahona, F., Grotschel, M. Junger, M. and Reinelt, G. (1986), An Applicaton of Combinatorial Optimization to Statistical Physics and Circuit Layout Design, Preprint No. 5, Institute fur Mathematik, Universitat Augsburg, Augsburg.

[3] Barahona, F., Grotschel, M., Mahjoub, A. (1985), Facets of the Bipartite Subgraph Polytope, *Math. of Oper. Res.* 10, 340-358.

[4] Barahona, F. and Mahjoub, A. (1986), On the Cut Polytope, *Math. Prog.* 36, 157-173.

[5] Fonlupt, J., Mahjoub, A., and Uhry, J. (1984), Composition of Graphs and the Bipartite Subgraph Polytope, Research Report No. 459, Laboratoire ARTEMIS (IMAG), Universite de Grenoble, Grenoble.

[6] Garey, M. R. and Johnson, D. S. (1979), *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman and Co., New York.

[7] Grotschel, M. Lovasz, L., and Schrijver, A. (1988), *Geometric Algorithms and Combinatorial Optimization*, Srpinger-Verlag, New York.

[8] Grotschel, M. and Pulleyblank, W. (1981), Weakly Bipartite Graphs and The Max-Cut Problem, *Oper. Res. Let.* 1, 23-27.

[9] Hadlock, F. (1975), Finding a Maximum Cut of a Planar Graph in Polynomial Time, *SIAM Journal on Comp.* 4, 221-225.

[10] Orlova, G. and Dorfman, Y. (1972), Finding a Maximum Cut in a Graph, *Engineering Cybernetics*, 10(3), 502-506.

[11] Poole, K. T. (1984), Least Squares Metric, Unidimensional Unfolding, *Psychometrica* 49(3), 311-312.

[12] Ritchey, N. P. (1989), Semi-linear Programming and the Uni-Dimensional Similarities Problem, Ph.D. Dissertation, Carnegie Mellon University, Pittsburgh, PA.