Clustering Optimization - Bob Agnew (raagnew1@gmail.com, www.raagnew.com)

Consider the binary integer program where the population associated with each of *m* observations is to be assigned to one and only one of *k* clusters to minimize total "cost."

Minimize: $\sum_{i=1}^{m}\sum_{h=1}^{k} p_i c_{ih} t_{ih}$ (total cost objective function, normally weighted squared distance)

Subject to: $\sum_{h=1}^{k} t_{ih} = 1$ for $i = 1,\ldots,m$ (each observation assigned to a single cluster)

$\sum_{i=1}^{m} p_i t_{ih} \geq b_h$ for $h = 1,\ldots,k$ (population lower bound imposed on each cluster)

$t_{ih} \in \{0,1\}$ (assignments must be binary)

If all $p_i \equiv 1$, then we have the Bradley-Bennett-Demiriz constrained k-means problem which can be solved with the linear programming transportation algorithm and an exact binary solution is guaranteed, e.g., using the lp.transport function in R package lpSolve. If the p_i vary, then the binary solution can be approximated in one of two ways and the resulting solution does not attain the lower bounds exactly, but in many instances the approximation will be acceptable, i.e., the bounds are treated as targets. This is the case in congressional districting where we aim for approximately equal cluster populations.

The first approximation is to solve the following transportation linear program where x_{ih} is the population assigned to cluster *h*.

Minimize: $\sum_{i=1}^{m} c_{ih} x_{ih}$ (total cost objective function)

Subject to: $\sum_{h=1}^{k} x_{ih} = p_i$ for $i = 1,\ldots,m$ (population for each observation assigned)

$\sum_{i=1}^{m} x_{ih} \geq b_h$ for $h = 1,\ldots,k$ (population lower bound imposed on each cluster)

$x_{ih} \geq 0$ (solution automatically in integers)

We then assign *all* the population for observation *i* to a single cluster with maximal allocation x_{ih}. This approximation works well because there are normally few split allocations in the transportation solution.

The second approximation is to relax the binary requirement in the integer program above to $t_{ih} \geq 0$, thus obtaining a linear program with rather simple dual formulation.

Maximize: $\sum_{i=1}^{m} u_i + \sum_{h=1}^{k} b_h v_h$

Subject to: $u_i + p_i v_h \leq p_i c_{ih}$ for $i = 1,\ldots,m$ and $h = 1,\ldots,k$

u_i unconstrained in sign and $v_h \geq 0$

Clearly $u_i = min_h (p_i c_{ih} - p_i v_h)$ in the optimal dual solution so the optimal v_h can be obtained by nonlinearly solving a collapsed version of the dual problem.

Maximize: $\sum_{i=1}^{m} min_h(p_i c_{ih} - p_i v_h) + \sum_{h=1}^{k} b_h v_h$

Subject to: $v_h \geq 0$

This collapsed objective function is continuous and concave, but not everywhere smooth. Nonetheless, it is smooth enough to be approximately solved using a standard nonlinear solver like *optim* in R which encompasses lower bounds. In other words, it's unnecessary to employ a specialized derivative-free solver. Once the quasi-optimal v_h are obtained, observation i is assigned to a cluster h with minimum $p_i c_{ih} - p_i v_h$. This dual approximation works well in some, but not all, applications. For us, it worked very well in a standard clustering problem with moderate bounding; it provided the same solution as the transportation approach in much less time. On the other hand, it performed worse in a congressional districting problem with widely varying populations and very tight cluster population bounds. The collapsed dual approach can in principle extend to applications with a very large number of observations since the objective function involves simple arithmetic operations. In past years, we have used the collapsed dual approach to solve very large marketing optimization problems where offers are assigned to prospects in order to maximize profit subject to offer constraints.