On convex formulation of 1D scaling problem

Michael Zibulevsky
Department of Computer Science
Technion–Israel Institute of Technology
Email: mzib@cs.technion.ac.il

October 12, 2008

Abstract

The 1D-scaling problem is to locate a set of points on an axis using [noisy] matrix of pair-wise distances between them. Straight-forward least-squares formulation of this problem is non-convex, and there is a danger of being trapped in a local minimum. We propose a convex variational formulation, which provides reliable solution to the problem. Noise statistics can be incorporated into the objective in quasi-optimal way. We also introduce a parametric family of objective functions, which gradually transforms from convex to "statistically optimal" non-convex. Using sequential optimizations with gradually updated "non-convexity" parameter, further improvement of solution can be achieved. A priori knowledge about the order of point locations can be incorporated as a set of linear constraints.

1 Convex variational formulation

Suppose that we need to recover coordinates x_i of n points on X-axis, using a set of [noisy] distances $\{d_{ij}\}$ between them. This problem appears in many areas of science and engineering; genetics is one of them [1]. The standard least-squares formulation of this problem

$$\min_x \sum_{i,j} (|x_i - x_j| - d_{ij})^2 \tag{1}$$

is non-convex and there exists a danger of being trapped in a local minimum.

1.1 Truncated least-squares

We introduce the following convex formulation of our problem

$$\min_x \sum_{i,j} \varphi(|x_i - x_j| - d_{ij}), \tag{2}$$

where $\varphi(t)$ is a convex monotone non-decreasing *one-sided* penalty function. Particularly, we can set

$$\varphi(t) = \begin{cases} 0 & t \le 0 \\ t^2 & t > 0 \end{cases} \tag{3}$$

In order remove indefinite shift and avoid the trivial solution $x = 0$, we treat two extremal points in a special way. Suppose that we know that x_k is the left and x_l is the right extremal point. Without loss of generality, we can set $x_k = 0$ and exclude it from the optimization. If the distance d_{kl} is given, we also add a term $\mu_l \psi(x_l - d_{kl})$, where ψ can be a *double-sided* penalty function, for example, $\psi(t) = t^2$. If d_{kl} is unknown, we can use any convex penalty, which will encourage x_l to increase. Even a linear term $-\mu_l x_l$ with small enough scalar μ_l can serve this mission.

1.2 Statistical approach

The functions φ and ψ can be also obtained by statistical considerations. Suppose that measurement errors are statistically independent and minus-log-likelihood $L_{ij}(d_{ij}; r_{ij})$ of a measurement d_{ij} as a function of the true value r_{ij} is convex in its increasing branch, for $r_{ij} > d_{ij}$. In this case we can set

$$\varphi_{ij}(t) = \begin{cases} 0 & t \le d_{ij} \\ L_{ij}(d_{ij}; t) & \text{otherwise} \end{cases} \tag{4}$$

and solve the convex problem

$$\min_x \sum_{i,j} \varphi_{ij}(|x_i - x_j|), \tag{5}$$

In order to further improve the solution, we can use several sequential optimizations, gradually coming back to the original non-convex log-likelihood $L(x) = \sum_{i,j} L_{ij}(d_{ij}; |x_i - x_j|)$. For this purpose we use a family of functions parameterized with a scalar μ

$$\varphi^{\mu}_{ij}(t) = \begin{cases} \mu L_{ij}(d_{ij}; t) & t \le d_{ij} \\ L_{ij}(d_{ij}; t) & \text{otherwise} \end{cases} \tag{6}$$

We first optimize the convex problem with $\mu = 0$; then we gradually increase μ towards 1, approaching the actual log-likelihood.

1.3 Order constraints

If we know a priori order of points $x_k \ge x_l$ for some given set M of pairs $\{k, l\}$, we can formulate an optimization problem with linear constraints:

$$\min_x \quad \sum_{i,j} \varphi_{ij}(|x_i - x_j|) \tag{7}$$

$$\text{subject to} \quad x_k \ge x_l, \quad \{k, l\} \in M \tag{8}$$

Note that for the indexes $\{i, j\} \in M$, the modulus can be removed from the corresponding terms in the objective function, so that one can use original double-sided log-likelihood terms $\varphi_{ij}(x_i - x_j)$ without damaging convexity.

2 Preliminary simulations

Presented approach was implemented in Matlab, using Quasi-Newton BFGS optimization function FMINUNC. At first stage we generated random set of 200 points uniformly distributed over interval [0 10]. Fully connected noiseless distance matrix was fed into the optimization algorithm. Accurate distances were recovered reliably in about 40 iterations of BFGS using about one second of CPU time of 2GHz Pentium-4 processor. Further experiments, verifying reconstruction from noisy distance matrix, are in process.

Acknowledgements

I am grateful to Avraham Korol for discussions, which motivated this work.

References

[1] D. Mester, Y. Ronin, E. Nevo, and A. Korol, *Fast and high precision algorithms for optimization in large-scale genomic problems*, Comp. Biol. & Chem, 28 (2004), pp. 281–290.