The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 8–10, 2007
, pp. 302–308

Linkage Disequilibrium Map by Unidimensional Nonnegative Scaling

Michael K. Ng Eric S. Fung Hai-Yong Liao

Centre for Mathematical Imaging and Vision and Department of Mathematics
Hong Kong Baptist University, Kowloon Tong, Hong Kong

Abstract A unidimensional nonnegative scaling model is used to construct linkage disequilibrium (LD) maps for human genome. The proposed constrained scaling model can be solved efficiently by transforming it to an unconstrained model. A LD map based on Hapmap data of 50000 SNPs is constructed by the proposed method using PC Clusters.

Keywords Linkage equilibrium; undimensional scaling; nonnegativity; Hapmap.

1 Introduction

Genetic linkage maps have long been invaluable tools for gene localization. These maps have been useful and successful for positional cloning of many disease genes. Genetic maps are to provide locations for genetic polymorphisms on the recombination and the corresponding map distances reflect recombination intensity. However, linkage maps have not performed well for the common diseases because of poor reproducibility and low power (Collins et al., 2004). Many researchers have studied allelic association or linkage disequilibrium (LD) rather than linkage. Here we note that linkage disequilibrium is often termed "allelic association." When alleles at two distinctive loci occur in gametes more frequently than expected given the known allele frequencies and recombination fraction between the two loci, the alleles are said to be in linkage disequilibrium. Evidence for linkage disequilibrium can be helpful in mapping disease genes since it suggests that the two may be very close to one another.

Linkage disequilibrium (LD) analysis offers the prospect of fine scale localization of genetic polymorphisms of medical importance, particularly when single nucleotide polymorphisms (SNPs) are densely typed in a candidate region. The role of LD is to identify and then narrow a candidate region. Because of the complex observed patterns, the modeling of the relationship between SNP markers and disease phenotypes is required (Sham, 1998). Maniatis et al. (2002) developed a metric LD map with additive distances in LD units based on the Malecot model. The application of LD maps to association mapping and positional cloning was also studied (Maniatis et al., 2004).

Consider two diallelic SNPs, where the rarest allele has frequency p, and is positively associated with an allele at the other SNP, which has frequency q. The haplotype frequencies of the 2 SNPs can then be represented in a 2-by-2 table as follows:

	Allele B	Allele b	
Allele A	$pq+d$	$p(1-q)-d$	p
Allele a	$(1-p)q-d$	$(1-p)(1-q)+d$	$1-p$
	q	$1-q$	1

The parameter d is defined as the linkage disequilibrium (LD) between the two SNPs. Because of the above rare allele assignment for p and q, we have $p <= 1/2$, $p \leq q$, $p \leq 1-q$ and $d \geq 0$. The scaled measure of linkage disequilibrium between the two SNPs is given by

$$d' = \frac{d}{p(1-q)}.$$

It is obvious that $d' = 1$ if $d = p(1-q)$. Since d' decays by a factor of $1-\theta$ per generation where θ is the recombination fraction, the function $-\ln d'^2$ has the property that it is proportional to $-\ln(1-\theta)$. Note that for small values of θ, $-\ln d'^2$ is approximately proportional to θ, and therefore to genetic map distance measured in units of Morgan.

The LD distance between the ith SNP and the jth SNP is given by $-\ln d'^2_{ij}$. For a set of n SNPs, their inter-marker LD distances can be represented in an n-by-n matrix $[-\ln d'^2_{ij}]_{i,j=1,2,\cdots,n}$. We require a 1-dimensional representation of the SNPs, preserving the order of the SNPs on the chromosome, such that the distances between SNPs along this dimension are close to the distances in the n-by-n LD distance matrix.

The scaled distances is the basis for constructing linkage disequilibrium maps which illuminate differences and similarities in linkage disequilibrium patterns between populations and chromosome regions. The role of recombination in defining linkage disequilibrium patterns and the focus on association mapping prompts the development of a genetic map. This genetic map is derived from linkage disequilibrium data and is analogous to the linkage map but differs substantially by accommodating recombination events that have accumulated (Jeffreys et al., 2001). Here the scaled distances in the proposed method are defined a genetic map distance in each SNP intervals in terms of linkage disequilibrium units. For SNPs separated by large scaled distances there is no useful LD and so these pairs are uninformative. The location of the recombination intense regions correspond closely to the steeper segments on the linkage disequilibrium unit map, whereas recombination cool areas are represented as high fairly flat lands, see the figures in the next subsection. Combining the characterization of linkage disequilibrium patterns with inferences of recombination will facilitate the search of signatures of recent selective sweeps across the human genome, i.e., regions that show more extensive linkage disequilibrium than predicted by the underlying recombination rate and which exhibit unusually low nucleotide diversity (Przeworski, 2002).

The main aim of this paper is to propose and develop a unidimensional nonnegative scaling model to construct linkage disequilibrium (LD) maps. The proposed constrained scaling model can be efficiently solved by transforming it to an unconstrained model. A LD map based on Hapmap data of 50000 SNPs is constructed by the proposed method using PC Clusters.

The outline of this paper is as follows. In Section 2, we develop the unidimensional nonnegative scaling model. In Section 3, an example is presented. Finally, some concluding remarks are given in Section 4.

2 The Unidimensional Nonnegative Scaling Model

The classical metric unidimensional scaling problem is to place n objects on the real line, so that the interpoint distances best approximate the observed dissimilarities between pairs of objects. Formally, the problem is to minimize the objective function:

$$F(x_1,x_2,\cdots,x_n)=\sum_{i>j}(d_{ij}-|x_i-x_j|)^2, \tag{1}$$

where x_i is the coordinate of the ith object and d_{ij} is the observed dissimilarity between the ith object and the jth object. We assume that the dissimilarity matrix (d_{ij}) is symmetric with nonnegative elements for all $i\neq j$ and $d_{ii}=0$ for $i=1,2,\cdots,n$. We note that if $x_i>x_j$ (or $x_i<x_j$), the error is $[d_{ij}-(x_i-x_j)]^2$ (or $[d_{ij}-(x_j-x_i)]^2$). Therefore the minimization of F is equivalent to minimizing the sum of the minimum between $[d_{ij}-(x_i-x_j)]^2$ and $[d_{ij}-(x_j-x_i)]^2$. It can be written as a nonlinear integer programming model and this problem is equivalent to an NP-hard combinatorial problem (Lau et al., 1998). Therefore the problem can only be solved for small n, and various heuristic algorithms have been proposed to solve this combinatorial problem, see for instance Hubert and Arabie (1986, 1988) and de Leeuw and Heiser (1977 and 1980). Recently, Lau, Leung and Tse (1998) formulated this unidimensional scaling problem as a nonlinear programming problem and solved it by optimization algorithms. Hubert, Arabie and Meulman (2002) further studied and compared different optimization algorithms for solving this problem.

In this paper, we consider constrained unidimensional scaling problems. The problem is to place n objects in a given order on the real line, so that the interpoint distances best approximate the observed dissimilarities between pairs of objects. In the literature, researchers have been interested in constrained multidimensional scaling problems. Bentler and Weeks (1978) used least squares scaling with the configuration in a Euclidean space and simply incorporated the required equality constraints in the least squares loss function. Lee (1984) used least squares scaling to allow not only for equality constraints but also inequality constraints.

In our unidimensional nonnegative scaling problem, the objects are required to place in a given order. For simplicity, we assume that the order of the objects is given as follows: 1st, 2nd, 3rd, $\cdots$, nth. This is the requirement for the objects in linkage disequilibrium maps. Therefore, the key issue is to determine the nonnegative interpoint distances among the ordered objects that best approximate the observed

dissimilarities between pairs of objects. Mathematically, the problem is to minimize the objective function:

$$J(z_1, z_2, \cdots, z_{n-1}) = \sum_{i>j} w_{ij} \left(d_{ij} - \sum_{k=j}^{i-1} z_k \right)^2, \tag{2}$$

subject to

$$z_k \geq 0, \quad k = 1, 2, \cdots, n-1,$$

where z_k is the scaled distance between the kth object and the $(k+1)$th object and w_{ij} is a positive weighting parameter that reflects the accuracy of the dissimilarity d_{ij}. We assume these parameters w_{ij} are fixed and known. In our application to genomics, these parameters w_{ij} are the inverses of the length of the confidence interval of d_{ij}. When the length is long, the weighting is small and therefore the importance is less.

In the model, we consider $\sum_{k=i}^{j-1} z_k$ to be the scaled distance between the ith object and the $(j-1)$th object $(i < j)$; this distance should be close to the dissimilarity between the ith object and the $(j-1)$th object. It is clear that the objective function value in (2) is equal to zero and $z_1 = d_{12}, z_2 = d_{23}, \cdots, z_{n-1} = d_{n-1n}$ iff

$$d_{ij} = \sum_{k=i}^{j-1} d_{kk+1},$$

for all $i < j$. We see that the objects preserve their original positions iff the objective function value is equal to zero.

The solution of (2) can be formulated as the solution of a least squares problem with nonnegativity:

$$\min_{\mathbf{z} \geq 0} \|\mathbf{WAz} - \mathbf{Wd}\|_2^2, \tag{3}$$

where

$$\mathbf{W} = \text{ diag}(w_{21}, w_{31}, \cdots, w_{n1}, w_{32}, w_{42}, \cdots, w_{nn-1}), \quad \mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_{n-1} \end{pmatrix},$$

$\mathbf{A}_i$ is an $(n-i)$-by-$(n-1)$ matrix given by $[\mathbf{0}_i|\mathbf{T}_i]$, $\mathbf{0}_i$ is an $(n-i)$-by-$(i-1)$ zero matrix, $\mathbf{T}_i$ is an $(n-i)$-by-$(n-i)$ Toeplitz matrix with its first column $[1, 1, \cdots, 1]^T$ and its first row $[1, 0, \cdots, 0]$,

$$\mathbf{z} = (z_1, z_2, \cdots, z_{n-1})^T \quad \text{and} \quad \mathbf{d} = (d_{21}, d_{31}, \cdots, d_{n1}, d_{32}, d_{42}, \cdots, d_{nn-1})^T.$$

We consider the parameterization $z = e^y$, i.e., $z_i = e^{y_i}$. With this parameterization, we can transform the constrained minimization problems into unconstrained problems, which are convex in y. The transformed problems are then minimized by using efficient optimization techniques. For instance, we note in (3) that the matrix A is

structured and W is a diagonal matrix. We can use an approximate Hessian and solve each linear subproblem with a variant of the conjugate gradient methods. In each iteration we need to evaluate the matrices A and A^T, which can be done efficiently, and hence A need never be formed explicitly.

3 An Example

We have developed a parallel MATLAB program to solve the unconstrained optimization problem in the last section. The parallel program is run in a PC Cluster located in the High Performance Cluster Computing Centre at Hong Kong Baptist University. The program takes a file of d' values produced, for example, by HAPLOVIEW (Barrett et al., 2004). The weighting parameter w_{ij} in (2) or (3) is defined as: $w_{ij} = 1/(-\ln CIL_{ij}^2 + \ln CIH_{ij}^2)$. The variables CIL_{ij} and CIH_{ij} are from the outputs of genetics program HAPLOVIEW, and represent the lower 95% confidence interval and upper 95% confidence interval of d'_{ij}. We note that if the length is large, the weighting parameter is small and therefore the importance of such d' contributes to the scaled distance is small.

A LD map based on Hapmap data of chromosome 22 is constructed by the proposed method. There are 54254 SNPs, but some of them are filtered out by HAPLOVIEW. The number of SNPs for the construction of LD map is 34556. We know that when a physical distance between two SNPs is larger than 500kb, the corresponding LD can be ignored (see Figure 1). Therefore, we apply this strategy to reduce the computer memory requirement in a PC Cluster for the LD distance matrix. We note that this is the default setting in HAPLOVIEW. We use steepest descent method to compute the solution of the unconstrained optimization problem for (3). The results are shown in Figure 2. Figure 2 (right) is the scaled LD map for chromosome 22. It takes about 10 hours using 20 CPUs in a PC Cluster in order to obtain the results.

4 Concluding Remarks

As a summary, we have formulated and studied constrained unidimensional scaling models, where the objects are required to place in a given order on the real line. Numerical results are presented to demonstrate the model for the application in linkage disequilibrium maps. In the future work, we plan to generate a genome-wide LD map for the whole human genome. It is expected that a more efficient parallel program should be designed and developed since the number of SNPs to be handled would be more than a million.

References

[1] Bentler, P. and Weeks, D. (1978), "Restricted Multidimensional Scaling Models," *Journal of Mathematical Psychology*, 17, 138-151.

[2] Collins, A., Lau, W. and De La Vega, F. (2004) "Mapping Genes for Common Diseases: The Case for Genetic (LD) Maps," *Human Heredity*, 58, 2-9.

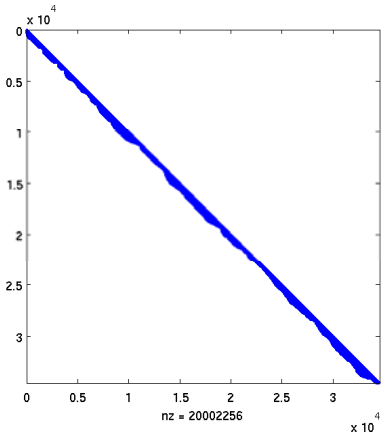

Figure 1: The pairwise LD of SNPs in chromosome 22.

[3] De Leeuw, J. and Heiser, W. (1977), "Convergence of Correction-Matrix Algorithms for Multidimensional Scaling," in *Geometric Representations of Relational Data: Readings in Multidimensional Scaling*, Ed., J. C. Lingoes, Ann Arbor, MI: Mathesis Press, 735-752.

[4] De Leeuw and Heiser, W. (1980), "Multidimensional Scaling with Restrictions on the Configuration of Points," in *Multivariate Analysis*, Ed., P. R. Krishnaiah, Amsterdam: North Hollandm Vol. V, 501-522.

[5] Hubert, L. and Arabie, P. (1986), "Unidimensional Scaling and Combinatorial Optimization," in *Multidimensional Data Analysis*, Ed., J. de Leeuw, W. Heiser, J. Meulman and F. Critchley, Leiden, The Netherlands: DSWO Press, 181-196.

[6] Hubert, L. and Arabie, P. (1988), "Replying on Necessary Conditions for Optimization: Unidimensional Scaling and some Extensions," in *Classification and Related Methods for Data Analysis*, Ed., H. H. Bock, Amsterdam: North-Holland, 463–472.

[7] Hubert, L. Arabie, P. and Meulman, J. (2002), "Linear Unidimensional Scaling in the L_2-Norm: Basic Optimization Methods Using MATLAB," *Journal of Classification*, 19, 303-328.

[8] Jeffreys, A. Kauppi, L. and Neumann, R. (2001), "Intensely Punctate Meiotic Recombination in the Class II Region of the Major Histocompatibility Complex," *Nat. Genet.*, 29, 217-222.

[9] Lau, K., Leung, P. and Tse, K. (1998), "A Nonlinear Programming Approach to Metric Unidimensional Scaling," *Journal of Classification*, 15, 3-14.

[10] Lee, S. (1984), "Multidimensional Scaling Models with Inequality and Equality Constraints", *Commun. Statist. Simula. Computa.*, 13, 127-140.

[11] Maniatis, N., Collins, A., Xu, C., McCarthy, L., Hewett, D., Tapper, W., Ennis, S., Ke., X. and Morton, N. (2002), "The First Linkage Disequilibrium (LD) Maps:

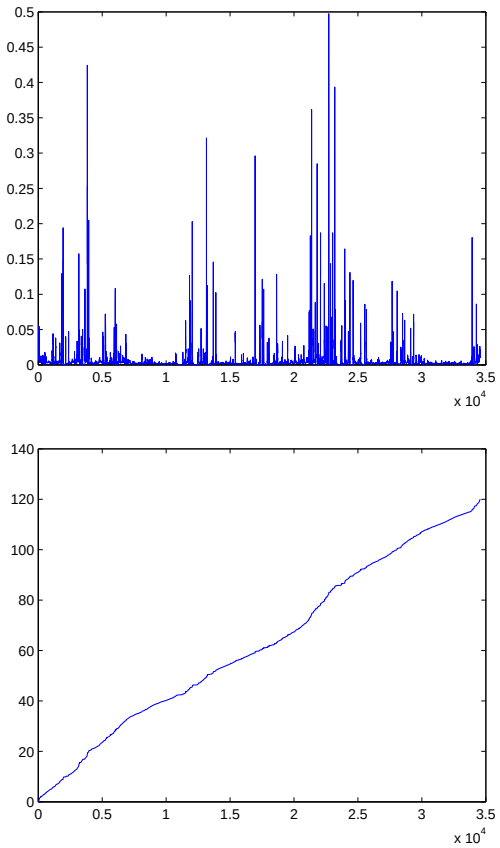

Figure 2: The solution of z (left) and the cumulative of the entries in z (right)

DeLineation of Hot and Cold Blocks of Diplotype Analysis", *Proc. Natl. Acad. Sci.*, U.S.A., 99, 2228-2233.

[12] Maniatis, N., Collins, A., Gibson, J., Zhang, W. Tapper, W. and Morton, N. (2004), "Positional Cloning by Linkage Disequilibrium", *Am. J. Hum. Genet.*, 74, 846-855.

[13] Przwworski, M. (2002), "The Signature of Positive Selection at Randomly Chosen Loci", *Genetics*, 162, 2053.

[14] Sham, P. (1998), "Statistics in Human Genetics", Edward Amold.