ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Sanger-driven MDSLocalize – A comparative study for Genomic Data

Marc Strickert[1], Nese Sreenivasulu[2], Udo Seiffert[1]

1 - Pattern Recognition Group, 2 - Gene Expression Group
Institute of Plant Genetics and Crop Plant Research Gatersleben, Germany
{stricker,srinivas,seiffert}@ipk-gatersleben.de

Abstract. Multidimensional scaling (MDS) methods are designed to establish a one-to-one correspondence of input-output relationships. While the input may be given as high-dimensional data items or as adjacency matrix characterizing data relations, the output space is usually chosen as low-dimensional Euclidean, ready for visualization. MDSLocalize, an existing method, is reformulated in terms of Sanger's rule that replaces the original foundations of computationally costly singular value decomposition. The derived method is compared to the recently proposed high-throughput multi-dimensional scaling (HiT-MDS) and to the well-established XGvis system. For comparison, real-value gene expression data and corresponding DNA sequences, given as proximity data, are considered.

Keywords. MDS, dimension reduction, proximity data, visualization.

1 Introduction

Dimension reduction and visualization are ever-challenging topics in data processing. Large data bases from high-throughput measuring devices in bio-, geo- and other sciences are ready for analysis. Before detailed analyzes and model creation, data inspection is very important for the identification of relevant system parameters and interdependencies. Faithful data displays help to get a feeling about data densities and class distributions. Kohonen's self-organizing map (SOM) is one of the most widely used methods for dimension reduction and visual data inspection, and it has been successfully applied to virtually any kind of data [7]. SOM mappings project data onto nodes ('clusters'), which yields desired complexity reduction at the undesired expense of loss of specific data characteristics: boundary SOM-nodes capture data outliers and might produce a misleadingly homogeneous data representation. For this reason, a sensitive one-to-one correspondence of input item and output counterpart might be preferred. Multidimensional scaling (MDS) methods seek to establish such correspondence between input and output data with a minimum of model parameters. An additional advantage of MDS approaches is their ability to directly deal with proximity data, which for SOM is an ongoing issue [2, 5, 6]. In earlier studies, the SOM has been compared with Sammon's mapping, a straight-forward realization of MDS. Usually, the computational requirements of MDS are quite high in contrast to SOM, but recent developments allow to handle even large datasets accurately [10]. Another efficient MDS version, called MDSLocalize, is pretty much related to principal component analysis (PCA), but it is ready for directly dealing with proximity data. In this paper, a reformulated version of this method is studied in detail.

2 Multidimensional Scaling (MDS) revisited

Two main applications of multidimensional scaling are the reduction of the input space dimension and the – mostly visual – reconstruction of interrelationships between the input data. Dimension reduction is attained by calculating pairwise data vector distances according to a definable similarity measure; these are used for reconstructing replacements given by adaptive points in low-dimensional Euclidean space. Thereby, metric conversion is realized if input vectors are compared by metrics different from the Euclidean, such as general Minkowski metrics or just measures of dissimilarity, such as 1-correlation. This way, abstract data relations are embedded in an intuitive visual space. The crucial ingredient to MDS is a cost function that maximizes the quality of reconstruction, or equivalently, an expression that minimizes the stress. Several stress functions with specific convergence properties and optimization goals exist that lead to different target point configurations, i.e. to distinct data views. Here, three iterative MDS approaches are considered, a new formulation of MDSLocalize, the recently proposed HiT-MDS, and the XGVis system.

2.1 Sanger-driven MDSLocalize

If n data vectors $\mathbf{x}^i$ were available, and not just n^2 mutual similarities d_{ij}, principal component analysis (PCA) would be a standard for dimension reduction. It can be shown that the projection of data to the most prominent eigenvectors of the data correlation matrix corresponds to the configuration of target points that, by their distances $\hat{\mathrm{d}}_{ij}$, minimize the classical stress function $\mathrm{s} = \sum_{i<j}^{n} (\mathrm{d}_{ij} - \hat{\mathrm{d}}_{ij})^2 \rightarrow \min$ [4]. Thereby and in the following symmetric proximity data are assumed. For mid-size data sets the required eigendecomposition can be efficiently computed by using linear algebra. The original MDSLocalize algorithm proposed by Drineas et al. [3] is a closed form approach to MDS, but for directly processing a given data distance matrix, not a correlation matrix. MDSLocalize is based on singular value decomposition (SVD):

1. **Centering:** $\tau(\mathbf{D}) = -\frac{1}{2}\mathbf{L}\,\mathbf{D}\,\mathbf{L}$, with $\mathbf{L} = \mathbf{I} - \frac{1}{n}\mathbf{1}$.
2. **SVD:** Compute rank-d approximation to $\tau(\mathbf{D})$ by $\tau_d(\mathbf{D}) = \mathbf{U}_d\,\mathbf{S}_d\,\mathbf{U}_d^{\mathsf{T}}$.
3. **Return** $\hat{\mathbf{X}} = \mathbf{U}_d\,\mathbf{S}_d$.

This algorithm has been shown to yield good reconstruction results also in the presence of noisy distance information, and even missing distances can be inferred [3]. In closed-form SVD realizations memory requirements and runtime might become a bottleneck for large scale applications with dense proximity data. For this reason an alternative formulation with striking simplicity is offered here. It is based on PCA, realized by Sanger's rule, a well-known cascading of Oja's neural approach to eigenvector approximation. Sanger's one-layer feed-forward networks yield eigenvectors ordered descendingly according to their importance for explaining data variances [9]. Both the centering step and the PCA iterations only require few lines of code. The total memory requirement is less than $\mathcal{O}(2 \cdot n^2)$, determined by the distance matrix which, due to symmetry, can be used to store intermediate centering results.

SVD can be easily reformulated by PCA: in the SVD step $\tau(\mathbf{D}) = \mathbf{U}\,\mathbf{S}\,\mathbf{U}^{\mathsf{T}}$ the columns of $\mathbf{U}$ contain the eigenvectors of $\tau(\mathbf{D})\,\tau(\mathbf{D})^{\mathsf{T}}$. Since symmetry of $\mathbf{D}$ and,

consequently, of $\tau(\mathbf{D})$ is a precondition, the required eigenvectors $\mathbf{U}$ of $\tau(\mathbf{D})\,\tau(\mathbf{D})^\mathsf{T}$ and the eigenvectors $\mathbf{V}$ of $\tau(\mathbf{D})$ are the same, because eigendecomposition into eigenvectors $\mathbf{V}$ and eigenvalue matrix Λ is:

$$\tau(\mathbf{D}) = \mathbf{V}\,\Lambda\,\mathbf{V}^\mathsf{T} \quad \Rightarrow \quad \tau(\mathbf{D})\,\tau(\mathbf{D})^\mathsf{T} = (\mathbf{V}\,\Lambda\,\mathbf{V}^\mathsf{T})\,(\mathbf{V}\,\Lambda\,\mathbf{V}^\mathsf{T})^\mathsf{T} = \mathbf{V}\,\Lambda^2\,\mathbf{V}^\mathsf{T}\,.$$

After subtracting the average row from the centered distance matrix $\tau(\mathbf{D})$, the first d eigenvectors of $\mathbf{V} = (\boldsymbol{v}_i)$ are obtained per component by Sanger's rule:

$$\Delta\boldsymbol{v}_{ij} = \gamma \cdot < \boldsymbol{x}, \boldsymbol{v}_i > \cdot\,(\boldsymbol{x} - \sum_{k=1}^{i} < \boldsymbol{x}, \boldsymbol{v}_k > \cdot\, \boldsymbol{v}_{kj})\,, \quad i = 1 \ldots d,\ j = 1 \ldots n\,.$$

This iterative scheme cycles through shuffled rows $\boldsymbol{x}$ of $\tau(\mathbf{D})$, starting with randomly initialized orthogonal eigenvectors of unit length or the standard basis vector system. After convergence, eigenvector normalization to unit length is forced for subsequent calculations. The corresponding eigenvalues λ_i are extracted from the eigensystem equations $\tau(\mathbf{D})\boldsymbol{v}_i = \lambda_i \cdot \boldsymbol{v}_i$: the first row, for example, of $\tau(\mathbf{D})$ is projected to the i-th eigenvector and divided by the first component of that eigenvector, which yields λ_i. For redundancy, any other than the first row and component could be used. The singular values s_i required on the diagonal of $\mathbf{S}$ in the return step are just the square roots $s_i = \sqrt{\lambda_i}$ of the calculated eigenvalues.

In summary, the steps of the Sanger-driven MDSLocalize are: $\langle 1 \rangle$ matrix centering, $\langle 2 \rangle$ mean subtraction moving the average row to origin, $\langle 3 \rangle$ Sanger-driven eigenvector determination, $\langle 4 \rangle$ calculation of corresponding eigenvalues/singular values, $\langle 5 \rangle$ scaling d eigenvectors by their singular values to yield $\hat{\mathbf{X}}$.

Parameters of the iterative MDSLocalize are: the number of eigenvector iterations, typically about 1000 epochs of the rows constituting the distance matrix, the eigenvector adaptation rate γ, usually a value in $[0.1/n; 0.0001/n]$, and the desired reconstruction dimension d, e.g. 1,2, or 3 for visualization.

2.2 HiT-MDS and XGVis

Two other approaches to MDS-based proximity data reconstruction are considered. One is the recently proposed high-throughput method HiT-MDS which implements a straight forward stress function, the maximization of Pearson correlation between the input and the reconstructed distance matrix [10]. The other one is the a well-established multiple purpose software package XGvis [1] which allows power transformations of the input distances prior and even during the embedding phase [1]. Its optimization goal is exact distance reconstruction, i.e. a unit slope in the corresponding Shepard plot. In contrast to that, the maximum correlation approach of HiT-MDS, corresponding to lines of arbitrary slope, imposes less obstructive conditions on the optimization, which leads to much faster convergence. For both methods the program defaults are used for the experiments, and a plateau in the stress function is the termination criterion.

2.3 Evaluation Criteria

Quality measures are required to evaluate the reliability of calculated MDS embeddings. Two complementing ways to look at the input-output relationships are, first, to

[1] http://public.research.att.com/~stat/xgobi/

which extent output space neighbors are also neighbors in the input space (backward-criterion: trustworthiness), second, to which extent neighbors in the input space do also have corresponding neighbors in the output space (forward-criterion: continuity). The formal definition of trustworthiness and continuity with comparative studies has been first given by Venna and Kaski [11]. Both measures are derived from the *ranks* of neighborhood point sets in input and output space, usually for a neighborhood size interval between 1 and $n/2$. Both quality measures have their theoretic maximum at one for perfect embedding and a minimum of zero for worst reconstruction. Another structurally different quality evaluation criterion is the squared *correlation* r^2 of the original input matrix and its reconstruction, calculated for the upper triangular matrix without diagonal elements. Zero means bad embedding, one perfect embedding.

3 Visualization of Barley Genomic Data

Two data sets of interest are taken from gene expression of developing barley grains. One set of 1660 differentially expressed genes obtained by macroarray hybridization is considered that characterize 14 developmental time points of barley grain tissue. For these genes, a second set of expressed sequence tags, given as DNA strings between 345 and 690 nucleotides, is available.

Expression data. For clustering of similar and co-expressed genes, different definitions of similarity are used to obtain different groupings. Two similarity measures are considered, the standard Euclidean distance of the log-normalized data, and the Euclidean distance of the ranks of the 14-dimensional expression vectors, which is closely related to Spearman rank correlation. Both distance matrices are computed once and used for the three MDS methods.

Results. The left column of Fig. 1 corresponds to results for Euclidean distance. All three MDS techniques yield very similar high-quality results, which is confirmed in the quality plots and by the first column of Tab. 1. Rank-based distances, however, as shown in the center column, lead to different results. Interestingly, the relatively high degree of dispersion of MDSLocalize (Fig. 1 top center) produces the worst quality curves (bottom center), but the distance correlations of original and reconstruction are still at intermediate level (Tab. 1 third column). Both Euclidean and rank-based sources yield two major clusters of higher data density. These are related to sequences of genetic temporal up- and down-regulation and revealed by all three methods.

Sequence proximity data. The DNA sequences have been aligned by the commonly used ClustalW package for multiple alignment. Alignment scores sc_{ij} are translated according to Oja et al. [8] into $\mathrm{d}_{ij} = -\log(sc_{ij}/(200 - sc_{ij}))$ ($sc_{kk} = 100$).

Results. The right column of Fig. 1 shows reconstructed DNA sequence distances. Embedding is difficult, because all pairs have got very similar large distances. This is confirmed by the low embedding qualities. A visual approximation by uniform spherical point distribution is obtained by XGVis in the third row. Since there is more structure in the data, other embeddings are obtained by MDSLocalize, which amplifies the asymmetry, and HiT-MDS, which seeks to find a compromise. MDSLocalize yields intermediate results for trustworthiness and continuity, but, at the same time, the correlation value r^2 in Tab. 1 is very low.