Correlation-based Data Representation

Marc Strickert and Udo Seiffert

Leibniz Institute of Plant Genetics and Crop Plant Research Gatersleben
Bioinformatics Division

{stricker,seiffert}@ipk-gatersleben.de

Abstract. The Dagstuhl Seminar *Similarity-based Clustering and its Application to Medicine and Biology* (07131) held in March 25–30, 2007, provided an excellent atmosphere for in-depth discussions about the research frontier of computational methods for relevant applications of biomedical clustering and beyond. We address some highlighted issues about correlation-based data analysis in this seminar postribution. First, some prominent correlation measures are briefly revisited. Then, a focus is put on Pearson correlation, because of its widespread use in biomedical sciences and because of its analytic accessibility. A connection to Euclidean distance of z-score transformed data outlined. Cost function optimization of correlation-based data representation is discussed for which, finally, applications to visualization and clustering of gene expression data are given.

Keywords. correlation, data representation, gradient-based optimization, clustering, neural gas, visualization, multi-dimensional scaling

1 Introduction

Data comparison is one of the most fundamental operations in data analysis. Comparisons are used to induce an ordering of data, which can be regarded as precursor to clustering, i.e. grouping of similar data. Although ordering and clustering can be defined as two stand-alone problems, there are many efforts to combine both tasks: self-organizing maps realize such a combination by vector quantization (clustering) and mapping to a low dimensional grid (ordering); as another example, hierarchical clustering generates clustering trees, which are usually post-structured by means of leaf-ordering procedures. In addition, graphical data representations are frequently found in biomedical publications, such as principal component projections, for visually illustrating closeness (clustering) of reference data points and their arrangement (ordering) along axes of principal changes, such as induced by time, probe concentrations, stress application, and so forth. Ordering of multi-dimensional data items, and likewise their centroid representations, according to similarity is thus a non-trivial task, because the diversity of complex orthogonal relationships needs to be reduced to lists or other low-dimensional structures that can be intuitively called ordered.

Dagstuhl Seminar Proceedings 07131
Similarity-based Clustering and its Application to Medicine and Biology
http://drops.dagstuhl.de/opus/volltexte/2007/1134

Relative and absolute distance sources can be identified for data comparisons.

Relative distances are characterized by knowledge of an adjacency matrix, for example, gradients between concentrations of chemical agents. Corresponding network nodes need not have a proper physical representation to which an absolute measure is applicable; the network might be defined only by proximity relationships between the nodes. If needed, an auxiliary physical representation of such proximity network can be obtained, for example, by multi-dimensional scaling methods that embed proximities into a target vector space with fixed dimensionality, possibly Euclidean, for convenience. Such embedding is computationally costly, though, and a loss of information is inevitable in many cases. This motivates direct functional methods for dealing with proximity data [1] which also facilitate comparisons of non-vector objects like strings or graphs.

Absolute distances provide a comparison of data items, here, vectors of fixed dimensionality. Although metrics, distances, and similarity measures differ in their degree of mathematical strictness, there is some common sense about the basic intention in clustering: maximum data similarities are sought, or, on the contrary, minimum distances and dissimilarities. In such an intuitive manner, Minkowski metrics, the Malahanobis distance, and Kendall correlation are examples of absolute 'distances', although the mathematical definition of distance is more stringent.

Since clustering results are depending only on the data, the data measure, and the computing method, the the data measure has to be chosen carefully. In many biological applications, correlation measures are preferred because of their favorable invariance properties: adding a constant offset to components of a data sample, or applying a multiplicative factor does not affect correlation. Such invariance helps, up to a certain degree, to circumvent calibration issues connected to measuring devices. As will be discussed, normalization does not always allow to boil down data for treatment with the standard Euclidean distance.

In the following, correlation measures will be briefly revisited, Pearson correlation will be considered in detail, and, connected to this, cost function based optimization for clustering and visual data screening will be presented.

2 Correlation Measures

In general, correlation quantifies the strictness of dependence of two vectors: 'the more of one amount, the more of the other', corresponds to positive correlation, while negative correlation indicates 'the more of one, the less of the other'. High absolute correlation values, though, are no guarantee that two observations really influence one another. Correlations might be caused by spurious dependence, either mediated by a hidden factor controlling both, or simply by chance. This explains, why 'measure of association' is a misleading synonym of correlation.

Different types of correlation can be specified. For brevity, we refer to existing literature and focus on the three most frequent ones [2]. These are Kendall's Tau [τ], comparing occurrences of positive and negative signs of differences of components in two vectors; Spearman rank correlation, comparing rank order-

ings, i.e. monotonic relationships, of the entries in two vectors; and Pearson correlation which is a measure of linear correlation between real-value entries of two vectors. The three measures yield numerical values in a range between -1, meaning negative correlation, and +1, meaning positive correlation; values around zero indicate uncorrelatedness.

2.1 Kendall's Tau [τ]

The Kendall coefficient τ measures the strength of the common tendency of two d-dimensional vectors $\mathbf{x} = (x_k)_{k=1\ldots d}$ and $\mathbf{w} = (w_k)_{k=1\ldots d}$ in a very direct manner. Data pairs (x_i, w_i) and (x_j, w_j) are considered. If $x_i - x_j$ and $w_i - w_j$ have the same sign, the pair is concordant, else it is discordant. The number of concordant pairs is C, the number of discordant pairs is D, for $i < j$ in both cases. Then

$$\tau = \frac{C - D}{d \cdot (d-1)/2}$$

describes the amount of bias towards concordant or discordant occurrences, normalized by the effective number of component pairs. Versions for handling tied data are also available. For its fundamental counting statistics and its easy interpretation Kendall's τ might be considered as favorable characterization of correlation. However, the computing complexity is $\mathcal{O}(d^2)$, i.e. quadratic in the number d of data dimensions. Also, τ does change in discrete steps as data relationships change, which makes it difficult to use in optimization scenarios, such as the optimum data representation task described below.

2.2 Spearman Rank Correlation

Another non-parametric correlation measure is obtained by calculating the normalized squared Euclidean distance of the ranks of x_k and w_k according to

$$\rho(\mathbf{x}, \mathbf{w}) = 1 - \frac{6}{d \cdot (d^2 - 1)} \cdot \sum_{k=1}^{d} \big(\mathrm{rnk}(x_k) - \mathrm{rnk}(w_k)\big)^2 .$$

Ranks of real values c_k are defined by $\mathrm{rnk}(c_k) = |\{c_i < c_k \,, i = 1 \ldots d\}|$, which can be easily derived from the ordering index (minus one) after an ascending sorting operation. This induces a common computing complexity of $\mathcal{O}(d \cdot \log(d))$. Again, tie handling strategies are available.

Spearman correlation has got the interesting property that a conversion of a non-linear data space into a special Euclidean one takes place. Replacing vector entries by their ranks leads to a compression of outliers and to a magnification, of close values, which, in the absence of ties, results in a uniform distribution with unit spacing and invariant statistical moments of the data vectors. In case of a low noise ratio, this simple conversion is a robust preprocessing step for getting standardized value discriminations, not only in correlation analysis. Unfortunately, as for Kendall's τ, these favorable features cannot be easily transfered from their discrete ranking basis into a desirably continuous optimization framework.

2.3 Pearson Correlation

In the majority of publications, data dependencies are described by values of Pearson correlation. The reason is that this measure is closely connected to linear regression analysis via the residual sum of squares to the fitted line. Pearson correlation describes the degree of linear dependence of vectors $\mathbf{x}$ and $\mathbf{w}$ by

$$\mathsf{r}(\mathbf{x},\mathbf{w}) = \frac{\sum_{i=1}^{d}(x_i-\mu_{\mathbf{x}})\cdot(w_i-\mu_{\mathbf{w}})}{\sqrt{\left(\sum_{i=1}^{d}(x_i-\mu_{\mathbf{x}})^2\right)\cdot\left(\sum_{i=1}^{d}(w_i-\mu_{\mathbf{w}})^2\right)}} =: \frac{\mathscr{B}}{\sqrt{\mathscr{C}\cdot\mathscr{D}}}\,. \tag{1}$$

This equation has got advantageous properties, because r requires only linear computing complexity $\mathcal{O}(d)$, and it is continuous in $\mathbb{R}$, except for non-interesting constant vectors with zero standard deviations of $\mathbf{x}$ or $\mathbf{w}$, i.e. zero denominators.

In principle, the covariance $\mathscr{B}$ of $\mathbf{x}$ and $\mathbf{w}$ in Equation 1 gets standardized by the product of the standard deviations $\sqrt{\mathscr{C}}$ and $\sqrt{\mathscr{D}}$ of $\mathbf{x}$ and $\mathbf{w}$, respectively, after mean subtraction of $\mu_{\mathbf{x}}$ and $\mu_{\mathbf{w}}$ from $\mathbf{x}$ and $\mathbf{w}$. As with any calculation that involves mean or variance, these first two statistical moments have highest reliability in case of well-behaved data distributions, possibly uni-modal and symmetric, such as the normal distribution; this condition, though, is hardly ever considered in practical calculations of Pearson correlation. Data standardization makes Pearson correlation invariant to rescalings of whole data vectors by common multiplication factors and to additive component offsets, such as induced by the gain of measuring devices and homogeneous background signals. In other words, the favorable invariance feature of Pearson correlation results from implicit data normalization realized by Equation 1. This raises the question, if Pearson correlation can be and should be replaced by simple covariance analysis of preprocessed data.

Relationship between Pearson correlation and Euclidean distance. The z-score transform $\mathbf{x}^z = (\mathbf{x}-\mu_{\mathbf{x}}\cdot\mathbf{1})/\sqrt{\mathrm{var}(\mathbf{x})}$ discards the mean value of $\mathbf{x}$ and yields unit variance. For z-score transformed vectors $\mathbf{x}^z$ and $\mathbf{w}^z$ the correlation of $\mathbf{x}$ and $\mathbf{w}$ can be expressed in terms of covariance using the scalar product $\langle\cdot,\cdot\rangle$:

$$\mathsf{r}(\mathbf{x},\mathbf{w}) = \langle\mathbf{x}^z,\mathbf{w}^z\rangle/(d-1)\ , \quad \langle\mathbf{x}^z,\mathbf{w}^z\rangle = \sum_{k=1}^{d} x_k^z\cdot w_k^z\ .$$

Because of invariance $\mathsf{r}(\mathbf{x},\mathbf{w}) = \mathsf{r}(\mathbf{x}^z,\mathbf{w}^z)$. When this notation is applied to the squared Euclidean distance of z-score transformed data this yields

$$\begin{aligned}\mathrm{d}^2(\mathbf{x}^z,\mathbf{w}^z) = \sum_{i=1}^{d}(x_i^z-w_i^z)^2 &= \langle\mathbf{x}^z,\mathbf{x}^z\rangle - 2\cdot\langle\mathbf{x}^z,\mathbf{w}^z\rangle + \langle\mathbf{w}^z,\mathbf{w}^z\rangle \\ &= 2\cdot(d-1)\cdot\big(1-\mathsf{r}(\mathbf{x},\mathbf{w})\big)\,.\end{aligned}$$

Thus, correlation r can be easily expressed as distance d^2. However, one must not forget about the crucial step of z-score normalization. In optimizations operating

on dynamic data, static pre-computation by the z-score transform is not available for computational improvements over Equation 1. Furthermore, for analytic considerations, such as the derivative computation discussed below, it is much more natural to think of the 'correlation' rather than of the 'negative rescaled and shifted squared Euclidean distance'.

Compactification of z-score transformed data.
The z-score transform is often used for visualizing correlated data, when different data ranges and offsets complicate a common plotting display or coloring scheme. It is the aim to transform highly correlated data into data items that are compact in Euclidean sense. However, the z-score vector transformation, producing zero mean and unit variance, is not optimum in terms of Euclidean compactness. Further compactification of n z-score transformed data vectors $\mathbf{x}^j$ can be obtained by minimizing the sum-of-squares quantization term

$$E_Q = \sum_{j=2}^{n} \sum_{i=1}^{n} \sum_{k=1}^{d} \left(\upsilon_i \cdot x_k^i + \nu_i - (\upsilon_j \cdot x_k^j + \nu_j) \right)^2 \rightarrow \min \ . \tag{2}$$

Notice that the correlation $\mathsf{r}(\upsilon_i \cdot \mathbf{x}^i + \nu_i \cdot \mathbf{1}, \upsilon_j \cdot \mathbf{x}^j + \nu_j \cdot \mathbf{1})$ is not influenced by different choices of the free parameters $\upsilon_l \in \mathbb{R}_0, \nu_l \in \mathbb{R}$. The cost function can be minimized, for example, by gradient descent, using $\partial E_Q / \partial \upsilon_j$, $\partial E_Q / \partial \nu_j$, initializing $\upsilon_i = 1, \nu_i = 0, i = 1 \ldots n$. By the heuristic trick of starting at $j = 2$ in Equation 2, the trivial solution $\upsilon_i = 0, \nu_i = 0, i = 1 \ldots n$ is effectively prevented, because the first pattern remains fixed, inducing an anchoring constraint on the other parameters.

An example of optimized data alignment is given in Figure 1. A number of 48 temporal gene expression profiles are aligned by standard z-score, followed by optimization. Since optimization reduces overall variance, the dimensionless coefficient of variation (cv) is computed for a standardized comparison. As desired, it turns out that this measure of dispersion is especially low for attributes in the set of optimized expression profiles.

3 Approaches to Correlation-based Analysis

In a plain view, data analysis is essentially data modeling, followed by model analysis and interpretation. This view allows to consider, for example, even a simple averaging operation as a modeling task, namely as solution of k-means with one ($k = 1$) centroid. In general, model selection and the modeling process itself are crucial ingredients to proper analysis. This motivates our focus on the derivation of optimum correlation-based data models.

In the following, optimality is always defined in terms of mathematically rigorous cost functions. These cost functions are continuous in almost every practical case, which allows their optimization by means of gradient techniques. Therefore, the partial gradient of Pearson correlation with respect to a target vector component is revisited, which can be used for attribute characterization [3], clustering, classification [4], and visualization [5]. Here, gradients will be used in order to optimize cost functions related to clustering and visualization.

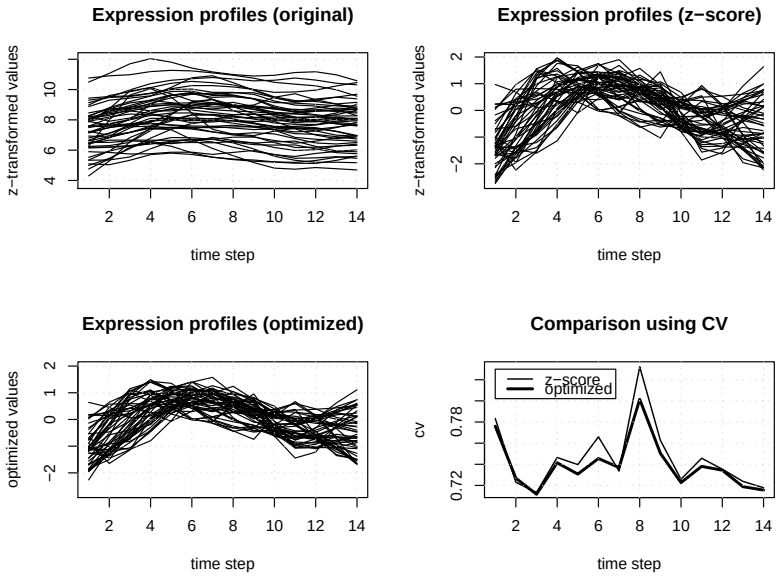


Fig. 1. Data alignment of temporal gene expression profiles. Original data, consisting of 48 14-dimensional expression profiles (top left), are transformed by z-score (top right), which is refined by optimization of Equation 2 (bottom left). For scale-free comparison of both alignments, coefficients of variation ($cv = \sigma/\mu$) have been calculated as measures of dispersion, separately between all pairs of 42^2 distances available at each time step (bottom right). Optimized data exhibit smaller cv-levels, i.e. higher compactness, than data only transformed by z-score.

3.1 Correlation-based Representation

Alternative data representations help to reduce complexity if these data models require only a low number of parameters, such as simple models of data distribution. Then the model can be analyzed instead of the possibly large and/or high-dimensional data sets. Very intuitive and characteristic data representations can be obtained by means of centroids, also known as prototypes, or, in classification setups, as codebook vectors. It is commonly considered a useful or even undoubted strategy to apply vector averaging in order to obtain representative centroid vectors for faithful data coverage. The well-known k-means algorithm is one such example using a center-of-gravity approach. Other methods, like learning vector quantization (LVQ) and self-organizing maps (SOM), rely on a similar reasoning, implementing incremental prototype adaptation based on the plain Hebbian learning term $(\mathbf{x} - \mathbf{w})$. This term expresses the movement of a centroid $\mathbf{w}$ on a straight line in Euclidean space towards the currently processed pattern $\mathbf{x}$. If data are completely considered in Euclidean space and is not only processed there, everything works fine. However, there are many computer programs avail-

able, offering k-means, LVQ, SOM, and some more methods in combination with a bunch of data measures, ranging from uncentered Minkowski metrics to Kendall correlation; yet, they do only change the comparison and not the essential step of centroid update. Since Pearson correlation is a widely used measure with advantageous analytic properties, this measure will be considered in more detail in the following, for realizing appropriate model updates.

Correlation-optimized centroid representation – A toy example. A small three-dimensional data set with three items is given in Table 1, for pointing out exemplary differences between the center-of-gravity $\bar{\mathbf{x}}$ and a correlation-optimized centroid location $\boldsymbol{s}$. Optimization is carried out on the cost function $\sum_{i=1}^{3} \mathsf{r}(\mathbf{x}^i, \boldsymbol{s}) \rightarrow \max$ via adaptation of centroid components in $\boldsymbol{s}$, initialized at the center-of-gravity. Gradient ascent on the optimization target yields vector $\boldsymbol{s}$ in Table 1, which is only one of infinitely many equivalent solutions $\hat{\boldsymbol{s}} = \upsilon \cdot \boldsymbol{s} + \nu \cdot \mathbf{1},\ \upsilon, \nu \in \mathbb{R}, \mathbf{1} \in \mathbb{R}^3$.

The quality of data representation should certainly depend on the similarity measure, i.e. analytic properties of the chosen measure should be considered for optimizing the representation, as sketched for the cost function above. Table 2 contains the obtained 'quantization' results in terms of individual and average Pearson correlations between the data vectors and the two centroids $\bar{\mathbf{x}}$ and $\boldsymbol{s}$.

It is not too much surprising that the measure-specific optimization outperforms the simple vector averaging. Still, it is surprising that widely accepted software tools most often do not realize such integrative modeling when they separate similarity computation and model update. If the pragmatism of Euclidean update is accepted, then why shouldn't it be acceptable, the other way round, to compare by Euclidean distance, but update in a correlation-optimum manner? Sometimes there are, of course, good reasons to stick to Euclidean updates. These are cases when analytic properties cannot be derived from the measure, such as the discrete counting statistics in Kendall's τ. Then Euclidean-driven optimization might be the only available choice. Still, one must keep in mind that Euclidean updates towards component-wise identity of centroids and data are not always compatible with more relaxed similarity measures. At least, the strict Euclidean dynamic does not distribute the centroids generously, which might induce the usage of more prototypes than would be actually required by a more relaxed similarity measure.

$\mathbf{x}^1$	$\mathbf{x}^2$	$\mathbf{x}^3$	avg.: $\bar{\mathbf{x}}$	alt.: $\boldsymbol{s}$
0	0	1	0.3333	0.1744
0	2	2	1.3333	0.1333
4	2	3	3	2.4923

Table 1. Toy example with three data vectors $\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3$ and their average centroid $\bar{\mathbf{x}}$ (center of gravity). The last column contains an alternative reference vector $\boldsymbol{s}$ derived from cost function optimization.

corr.	$\mathbf{x}^1$	$\mathbf{x}^2$	$\mathbf{x}^3$	avg.
$\bar{\mathbf{x}}$	0.92857	0.78571	0.98974	0.90134
$\boldsymbol{s}$	0.86605	0.866	1	**0.91068**

Table 2. Correlation table of $\bar{\mathbf{x}}$ and $\boldsymbol{s}$ vs. pattern vectors $\mathbf{x}^i$, including their average correlations (rightmost column). High values indicate good representations.

3.2 Gradient of Pearson Correlation

The definition of the gradient of a similarity measure is generally a very valuable tool for assessing the influence of vector components on the measured value, characterizing the relationship between two vectors. In unsupervised attribute selection tasks, this allows to identify attributes contributing most to the measure [3]. This is trivial for Euclidean distance, for which the derivative can be decomposed into independent components, except for a common scaling factor:

$$\mathrm{d}(\mathbf{x},\mathbf{w}) = \sqrt{\sum_{i=1}^{d}(x_i - w_i)^2} \qquad \rightarrow \qquad \frac{\partial \mathrm{d}(\mathbf{x},\mathbf{w})}{\partial w_k} = \frac{w_k - x_k}{\mathrm{d}(\mathbf{x},\mathbf{w})}\,.$$

In this case, the component k with maximum absolute difference (or highest variance, for simplicity) contributes most.

The situation is much more interesting when gradients of Pearson correlation corresponding to Equation 1 are considered:

$$\frac{\partial \mathsf{r}(\mathbf{x},\mathbf{w})}{\partial w_k} = \frac{(x_k - \mu_{\mathbf{x}}) - \frac{\mathscr{B}}{\mathscr{D}} \cdot (w_k - \mu_{\mathbf{w}})}{\sqrt{\mathscr{C} \cdot \mathscr{D}}}\,. \tag{3}$$

Independence of the components is not realized, because the components $\mu_{\mathbf{x}}$, $\mu_{\mathbf{w}}$, $\mathscr{B}$, $\mathscr{C}$ and $\mathscr{D}$, contribute knowledge of all other vector components in a non-trivial manner. Additionally, the Euclidean rule of opposite direction for argument flipping, does not hold, because $\partial \mathsf{r}(\mathbf{x},\mathbf{w})/\partial w_k \neq -\partial \mathsf{r}(\mathbf{w},\mathbf{x})/\partial w_k$ in the usual case of $\mathscr{B} \neq \mathscr{D}$.

Clustering framework. In coexpression analysis, the ultimate goal is to find clusters of data containing highly correlated data vectors [6]. Centroids are a very natural schematic representation of such clusters. Faithful data representation requires robust centroid locations within the data. Self-organizing maps (SOM) realize a cooperative centroid placement strategy by iterative presentation of data points that trigger further improvements of previously placed centroids. A general formulation of this simple procedure is given in Algorithm 1. The SOM mode of algorithm 1 is not of interest here, because the visualization capabilities of SOM are required; yet, we are interested in high quantization accuracy. Neural gas (NG) [7] is our method of choice for finding high-quality centroids

Algorithm 1 SOM / NG centroid update

repeat
  chose randomly a data vector $\mathbf{x}$
  $k \leftarrow \arg\min_i \{ \mathrm{d}(\mathbf{w}^i, \mathbf{x}) \}$
    { $\mathbf{w}^k$ is closest centroid to data vector $\mathbf{x}$ }
  for all m centroids j **do**
    $\mathbf{w}^j \leftarrow \mathbf{w}^j + \gamma \cdot \mathrm{h}_\sigma \left(D(\mathbf{w}^k, \mathbf{w}^j) \right) \cdot U(\mathbf{x}, \mathbf{w}^j)$
      { $\gamma, \mathrm{h}, \sigma, D, U$: see text }
  end for
until no more major changes

in the original data space. The authors of NG showed that the NG algorithm asymptotically realizes a stochastic gradient descent on the cost function:

$$\mathrm{E}(\mathbf{W}, \sigma) = \frac{1}{C(\sigma)} \cdot \sum_{j=1}^{m} \sum_{i=1}^{n} \mathrm{h}_\sigma \left(\mathrm{rnk}(\mathbf{x}^i, \mathbf{w}^j) \right) \cdot \mathrm{d}(\mathbf{x}^i, \mathbf{w}^j) \,. \tag{4}$$

The scaling factor $C(\sigma) = \sum_{i=0}^{m-1} \mathrm{h}_\sigma(i)$ is used for normalization. In the limit $\sigma \to 0$, the NG mode of Algorithm 1 leads to a centroid placement that minimizes the total quantization error between the m centroids and n data vectors.

The benefits of neural gas are: mathematical understanding of centroid specialization, high reproducibility of results, neighborhood cooperation for robustness against initialization, and easy implementation. A fast batch version of neural gas with quadratic convergence has been proposed recently [8], complementing the iterative online approach discussed here.

Correlation described by Equation 1 can be plugged into the cost function Equation 4 being optimized by gradient tracking along partial derivatives of E with respect to the components of all centroids $\mathbf{w}^j$. Since the cost function should be minimized, correlation r is turned by negative sign into a dissimilarity measure. Therefore, the term $U(x_k, w_k) = -\partial \mathsf{r}(\mathbf{x}, \mathbf{w}) / \partial w_k$ is inserted into Algorithm 1, which constitutes the alternative version of neural gas for correlation-based centroid placement, NG-C for short.

It can be shown that this correlation-based update rule yields a valid gradient descent also at the boundaries of the receptive fields. A proof, originally for the Euclidean case, is provided by [7], where a vanishing contribution of the ranks was presented. Since the proof does not rely on specific properties of the Euclidean metric, a direct transfer to Pearson correlation is possible. Thus, Equation 4 is a valid cost function that gets optimized by the neural gas algorithm. If visualization is desired and the cost function criterion be relaxed, the correlation derivative can be used, of course, for an improved update of self-organizing maps.

Visualization framework. For visualization of individual data points one of the most widely used methods is principal component projection. However, PCA is restricted to linear mappings of high-dimensional data, thereby focusing on directions of maximum Euclidean variance. A more natural alternative goal is to obtain low-dimensional displays that reflect most faithfully the inter-vector similarities of the source data.

In principle, this goal can be reached by using multi-dimensional scaling (MDS) techniques to make distances of reconstructed low-dimensional points similar to distances between the input vectors. This optimization task can be very hard, because of ambiguous compromise solutions in the low-dimensional space. Most MDS methods define quite stringent cost functions, such as searching – by least squares approaches – strict identity of distances between the reconstructed point locations and the distances of corresponding input data.

Alternatively, Pearson correlation between the distance matrices of input data and reconstructed points allow, due to scale and shift invariance, infinitely many more solutions than for strict identity optimization. The following method, called high-throughput multidimensional scaling (HiT-MDS), describes how correlation is used to help alleviate the optimization task of finding proper low-dimensional point locations.

Let matrix $\mathbf{D} = (\mathrm{d}_{ij})_{i,j=1\ldots n}$ contain the pattern distances, and the matrix $\hat{\mathbf{D}} = (\hat{\mathrm{d}}_{ij})_{i,j=1\ldots n}$ those of the reconstructions. Then the correlation $\mathsf{r}(\mathbf{D}, \hat{\mathbf{D}})$ between entries of the source distance matrix $\mathbf{D}$ and the reconstructed distances $\hat{\mathbf{D}}$ is maximized by minimizing the following embedding cost function:

$$\mathrm{s} = -\mathsf{r} \circ \hat{\mathbf{D}} \circ \hat{\mathbf{X}} \quad \Rightarrow \quad \frac{\partial \mathrm{s}}{\partial \hat{x}^i_k} = -\sum_{j=1\ldots n}^{j\neq i} \frac{\partial \mathsf{r}}{\partial \hat{\mathrm{d}}_{ij}} \cdot \frac{\partial \hat{\mathrm{d}}_{ij}}{\partial \hat{x}^i_k} \to 0,\ i = 1 \ldots n \qquad (5)$$

Locations of points in target space are obtained by gradient descent on the stress function s using the depicted chain rule. The derivatives in Equation 5 are [5]

$$\begin{aligned} \frac{\partial \mathsf{r}}{\partial \hat{\mathrm{d}}_{ij}} &= \frac{(\mathrm{d}_{ij} - \mu_{\mathbf{D}}) - \frac{\mathscr{B}}{\mathscr{D}} \cdot (\hat{\mathrm{d}}_{ij} - \mu_{\hat{\mathbf{D}}})}{\sqrt{\mathscr{C} \cdot \mathscr{D}}} \quad \text{(cf. Eqn. 3)} \\ \frac{\partial \hat{\mathrm{d}}_{ij}}{\partial \hat{x}^i_k} &= (\hat{x}^i_k - \hat{x}^j_k) \,/\, \hat{\mathrm{d}}_{ij} \quad \text{for Euclidean} \quad \hat{\mathrm{d}}_{ij} = \sqrt{\textstyle\sum_{l=1}^{d} (\hat{x}^i_l - \hat{x}^j_l)^2}\,. \end{aligned}$$

While, for planar and intuitive plotting purposes, target distances $\hat{\mathrm{d}}_{ij}$ are usually Euclidean, input distances can be mere dissimilarities, like flipped Pearson correlation $\mathrm{d}_{ij} = (1 - \mathsf{r}(\mathbf{x}^i, \mathbf{x}^j))$ or powers of which. Note that these data vector correlations are completely different from the target r in the correlation-based cost function optimization in Equation 5 of HiT-MDS.

In contrast to previous versions of HiT-MDS, a slightly modified, but efficient update step is proposed. Randomly drawn points $\hat{\mathbf{x}}^i$ are updated into the direction of the sign $\mathrm{sgn}(x) = x/|x|$ of the steepest gradient of s, scaled by the decreasing learning rate γ_t:

$$\Delta \hat{x}^i_k = -\gamma_t \cdot \mathrm{sgn}\left(\frac{\partial \mathrm{s}}{\partial \hat{x}^i_k}\right) \quad , \quad \gamma_{t \to t_{t-\max}} \to 0.$$

Convergence is forced by driving the learning rate monotonously to zero, in the limit of maximum cycles $t_{\max}+1$. In practice, the learning rate starts at $\gamma_0 = 1$, where it is kept until iteration number $t_{\max}/2$; it is then linearly decreased to zero. This update scheme is very robust against the choice of the learning rate and usually yields excellent results.

Maximum calculation efficiency, justifying the name high-throughput MDS, can be obtained by optimized procedures described in [9]: once matrices $\mathbf{D}$ and $\hat{\mathbf{D}}$ are computed in $\mathcal{O}(n^2)$, updates of the similarity matrix and incremental changes in correlations $\mathsf{r}(\mathbf{D}, \hat{\mathbf{D}})$ can be computed in $\mathcal{O}(n)$ instead of $\mathcal{O}(n^2)$.

4 Correlation-based Methods for Gene Expression Data

Using the two methods for clustering (NG-C) and visualization (HiT-MDS) described above, a number of interesting features can be derived from the data. Visualization is certainly helpful for initial data screening and for accompanying later steps of analysis. Clustering is one of the central tools for extracting common regulatory patterns, such as temporal up- and down-regulation, intermediate events and other characteristic processes. The particular advantage of the proposed methods is their perfect interplay, as they optimize correlation-based data representation.

The aim of the presented analysis is the identification of tissue-specific key regulatory genes that trigger critical pathways during the temporal development in growing barley grains [10]. This study has practical values for the improvement of seed quality which is of high interest to breeding companies. A total of 330 008 gene expression values were collected from 28 hybridization experiments with 12k macroarrays, covering 14 temporal developmental points of developing barley endosperm tissue from two independent series. Gene expression levels had to exceed twice the background to be considered as signal. Background subtraction and quantile normalization of $\log_2$-transformed data was carried out for the remaining genes. This processing was done separately for both experimental series to allow the comparison of signal intensities across time series. A filter based on Pearson correlation was then applied to select gene profiles time series that correlate at a conservative level of $r > 0.5$ between the two independent series. With this criterion, a qualified subset of 4824 out of 11 786 genes was created for analysis. For simplicity, data from only the second series are considered here.

4.1 HiT-MDS Visualization

For the 4824 genes of interest, HiT-MDS embedding requires only 100 data cycles, processed within a few minutes, to get the high-quality display shown in Figure 2. The characteristic sandglass shape results from using eighth power of the correlation measure, $(1 - \mathsf{r}(\mathbf{x}^i, \mathbf{x}^j))^8$. This power magnifies subtle dissimilarities in highly correlated genes which leads to focus on a good reconstruction – and thus a fair differentiation – of highly correlated, i.e. with near zero dissimilarities, rather than of obviously discorrelated genes. The exponent of eight

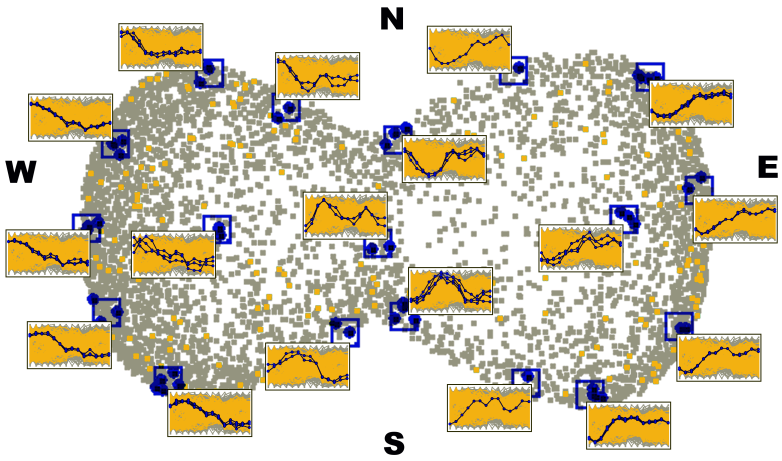


Fig. 2. HiT-MDS scatter plot of embedded temporal gene expression data. Correlation similarity $(1 - \mathsf{r}(\mathbf{x}^i, \mathbf{x}^j))^p$ is considered at $p = 8$ for magnification of high-correlation subsets, which explains the characteristic sandglass shape.

has turned out to be a good compromise for spreading highly correlated genes and for giving, at the same time, space also to intermediate regulations. Similar findings for higher order powers of correlation are reported in [11].

By posterior labeling with known gene annotations, the group of hormone and signaling related genes are highlighted in orange colors before other functional categories, marked in gray. Additionally, data boxes, brushed in blue, and their corresponding plots of temporal patterns have been manually picked in order to demonstrate the high spatial connectivity of similar regulatory profiles and their embedded two-dimensional counterparts. A smooth transition can be found from the western side (W) with patterns of down-regulation, via south (S) corresponding to patterns of intermediate up-regulation and up-regulation located in the east (E) to north (N) with intermediate down-regulation, back to west. Rare and unique regulation patterns are found in the interior of the sandglass structure.

The prominent temporal expression patterns are easily revealed by browsing the scatter plot in the way described above. The plot shows that the correlation space is very homogeneous, dominated by patterns of up- and down-regulation, according to the experimental design. Overall, the HiT-MDS embedding procedure applied to transcriptome data of endosperm development yields a faithful arrangement of genes with their typical temporal expressions. Using the freely available GGobi visualization software [12], data can be interactively browsed for picking candidate genes. This combination of embedding and visualization tools turned out to be very assisting in the derivation of potential regulatory pathways [5].

4.2 Clustering

Beyond the described visual sub-grouping of embedded gene expression data, dedicated clustering methods provide more reliable clusters. Since these methods do usually operate in the original data space and not in the somewhat lossy reconstruction of embedded data, higher quantization accuracy can be obtained.

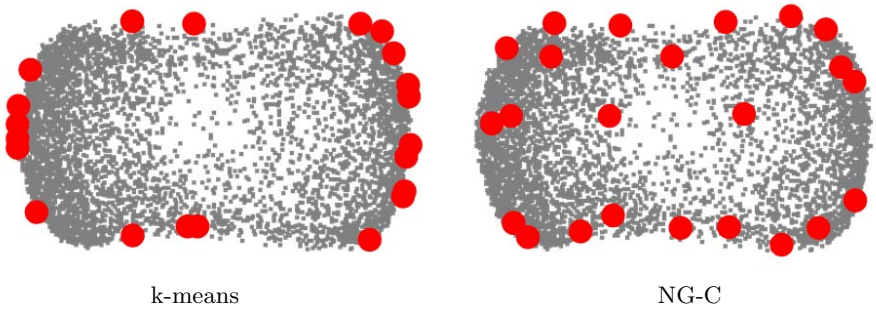
In a previous study, neural gas has been successfully applied to subdividing the set of 4824 genes into smaller sets of characteristic regulation patterns [5]. In principle, such a partitioning helps to get an abstract view upon the data set, because the data space gets faithfully covered by a pre-defined number of typical centroids. Such a relatively small number of centroids helps to identify potential cascades of regulatory events by looking at temporal delays of transcription activity. The benefits over hierarchical clustering are two-fold: firstly, the hierarchical clusters would need to be merged appropriately for data abstraction, requiring an extra step; secondly, with NG-C no decision must be taken about the linkage method – one of complete, single, or average linkage – needed for tree creation in hierarchical clustering. Hierarchical clustering is of interest rather for the identification of outliers.

Certainly the number of centroids is an interesting choice for NG-C, because it defines the level of abstraction. For k-means, which is used for the same purpose as NG-C, a number of heuristic methods exist to give rough estimates about the number k of centroids [13]. The pragmatic approach suggested here is to compute the centroids and embed them together with the data. The obtained display will then give reasonable hints, if all major modes, i.e. high density regions, are covered by the centroids. If not so, their number has to be further adjusted until a good correspondence of centroids and data is obtained.

Visualization of embedded centroids. Clustering and visualization of the 4824 barley endosperm genes introduced above yields scatter plots shown in Figure 3. For comparison with NG-C, Eisen's implementation of k-means has been taken as reference model [14]. Both methods make use of Pearson correlation for creating sets of similar patterns for centroid calculation, but according to the standards, k-means calculates averages in Euclidean space, whereas NG-C uses correlation-optimized updates. The exponential NG-C neighborhood influence is realized as exponential decay from $\sigma = 23$ to $\sigma = 0.001$, the update rate is set to $\gamma = 0.001$. Both methods were trained with 100 data cycles for 23 centroid positions.

As most fundamental difference, the final states in k-means, corresponding to the right panel of Figure 3, are quite close and dense at the boundaries of the embedded data manifold, while a more homogeneous spreading is observed for NG-C centroids.

Quality of representation. Beyond visualization, another quality criterion has been derived from 10 independent repetitions of k-means and NG-C clustering, starting from random initialization. Analogous to quantization accuracy, we determine the average correlation of centroids to their represented data. For k-means we obtained, over 10 runs, an average correlation of $\mathsf{r} = 0.9329 \pm 0.0017$



k-means NG-C

Fig. 3. Correlation-based clustering and visualization using HiT-MDS. Left: k-means; right: NG-C. Both methods use Pearson correlation similarity for computing locations of 23 centroids. NG-C centroids are more faithfully distributed among the data.

per centroid with an average standard deviation 0.0881 ± 0.0038. For NG-C the results are $\mathsf{r} = 0.9516 \pm 0.0001$ with standard deviation 0.0573 ± 0.0004. Thus, NG-C provides higher average correlation with much less standard deviation. The low standard deviation underlines a very important feature of NG-C: the high reproducibility of final centroids, independent of their initialization. This is one major advantage over k-means for which a poor reproducibility is known. Moreover, and contrary to k-means, unused prototypes do not occur in NG-C, because of its built-in neighborhood cooperation.

5 Conclusions

Gradients of Pearson correlation have been introduced for cost function optimization frameworks aiming at reliable data representation.

The quality of models can be assessed already during data processing by looking at the current cost value. Rigorous comparisons to existing methods have not been carried out in this work. There is one simple reason: it does not make much sense to compare an unsupervised model, optimized for a certain purpose to a model not optimized for it. From a different perspective, why should an optimization model be judged by another than the optimization criterion? For example, if PCA optimizes directions of maximum variance and HiT-MDS optimizes maximum correlation between two data spaces, the only reason for choosing either approach is its practical use.

A very central statement derived from the claim above, apply methods to targets they are designed for, is to make model update consistent with the data similarity. Here, derivative properties of Pearson correlation are used for optimization which provides a model space that is in good agreement with the data space. An introduced toy example has demonstrated the practical value of this consideration. The presented NG-C clustering method realizes an update con-

sistent with the Pearson correlation similarity measure which allows to generate highly reproducible partitions of the data.

Adequate visualization is certainly a very important tool to accompany most steps of data analysis. Initial screening yields hints if pronounced clustering can be expected, and it allows a reasonable choice of the number of centroids by embedding them together with the data. In the presented gene expression study it turned out that the correlation-based data space is rather homogeneous; it might thus be worth to look for interesting outliers by interactive browsing or by detection of special nodes in trees from hierarchical clustering.

Finally, the presented cost function frameworks focussing on the Pearson correlation measure are general enough to replace correlation by any measure for which mathematical derivatives are available. This opens a very wide perspective on future approaches for reliable data-driven biomedical analysis.

Availability

C, MATLAB (GNU Octave), and R source code of high-throughput multi-dimensional scaling (HiT-MDS) and supplemental data is online available at http://hitmds.webhop.net/.
C code of neural gas for Pearson correlation (NG-C) is online available at http://pgrc-16.ipk-gatersleben.de/~stricker/ng/.

Acknowledgements

This work is supported by BMBF grant FKZ 0313115 (GABI-SEED-II) and by the Ministry of Culture of Saxony-Anhalt, grant XP3624HP/0606T.

References

1. Hammer, B., Hasenfuss, A.: Relational clustering. In Biehl, M., Hammer, B., Verleysen, M., Villmann, T., eds.: Similarity-based Clustering and its Application to Medicine and Biology. Number 07131 in Dagstuhl Seminar Proceedings, Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany (2007) <http://drops.dagstuhl.de/opus/volltexte/2007/1118>.
2. Bolboaca, S., Jäntschi, L.: Pearson versus Spearman, Kendall's Tau correlation analysis on structure-activity relationships of biologic active compounds. Leonardo Journal of Sciences **5** (2006) 179–200
3. Strickert, M., Schleif, F.M., Seiffert, U.: Gradients of pearson correlation for analysis of biomedical data. In: Proceedings of the Argentine Symposium on Artificial Intelligence (ASAI 2007). (To appear.)
4. Strickert, M., Seiffert, U., Sreenivasulu, N., Weschke, W., Villmann, T., Hammer, B.: Generalized relevance LVQ (GRLVQ) with correlation measures for gene expression data. Neurocomputing **69** (2006) 651–659
5. Strickert, M., Sreenivasulu, N., Usadel, B., Seiffert, U.: Correlation-maximizing surrogate gene space for visual mining of gene expression patterns in developing barley endosperm tissue. BMC Bioinformatics **8** (2007)

6. Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J., Pavlidis, P.: Coexpression Analysis of Human Genes Across Many Microarray Data Sets. Genome Res. **14** (2004) 1085–1094
7. Martinetz, T., Berkovich, S., Schulten, K.: "Neural-gas" network for vector quantization and its application to time-series prediction. IEEE Transactions on Neural Networks **4** (1993) 558–569
8. Cottrell, M., Hammer, B., Hasenfuss, A., Villmann, T.: Batch and median neural gas. Neural Networks **19** (2006) 762–771
9. Strickert, M., Teichmann, S., Sreenivasulu, N., Seiffert, U.: High-Throughput Multi-Dimensional Scaling (HiT-MDS) for cDNA-array expression data. In Duch et al., W., ed.: Artificial Neural Networks: Biological Inspirations, Part I, LNCS 3696, Springer (2005) 625–634
10. Sreenivasulu, N., Radchuk, V., Strickert, M., Miersch, O., Weschke, W., Wobus, U.: Gene expression patterns reveal tissue-specific signaling networks controlling programmed cell death and ABA-regulated maturation in developing barley seeds. The Plant Journal **47** (2006) 310–327
11. Zhou, X., Kao, M.C., Wong, W.: Transitive functional annotation by shortest-path analysis of gene expression data. PNAS **99** (2002) 12783–12788
12. Buja, A., Swayne, D., Littman, M., Dean, N., Hofmann, H.: Interactive Data Visualization with Multidimensional Scaling. Report, University of Pennsylvania, URL: http://www-stat.wharton.upenn.edu/~buja/ (2004)
13. Dudoit, S., Fridlyand, J.: A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biology **3** (2002) 36.1–36.21
14. de Hoon, M., Imoto, S., Nolan, J., Miyano, S.: Open source clustering software. Bioinformatics **20** (2004) 1453–1454