Unsupervised Feature Selection for Biomarker Identification in Chromatography and Gene Expression Data

Marc Strickert[a], Nese Sreenivasulu[b], Silke Peterek[c], Winfriede Weschke[b], Hans-Peter Mock[c], and Udo Seiffert[a]

[a] Pattern Recognition Group, [b] Gene Expression Group, [c] Applied Biochemistry
Leibniz Institute of Plant Genetics and Crop Plant Research Gatersleben,
{stricker, srinivas, peterek, weschke, mock, seiffert}@ipk-gatersleben.de

Abstract. A novel approach to feature selection from unlabeled vector data is presented. It is based on the reconstruction of original data relationships in an auxiliary space with either weighted or omitted features. Feature weighting, on one hand, is related to the return forces of factors in a parametric data similarity measure as response to disturbance of their optimum values. Feature omission, on the other hand, inducing measurable loss of reconstruction quality, is realized in an iterative greedy way. The proposed framework allows to apply custom data similarity measures. Here, adaptive Euclidean distance and adaptive Pearson correlation are considered, the former serving as standard reference, the latter being usefully for intensity data. Results of the different strategies are given for chromatography and gene expression data.

Keywords: Feature selection, adaptive similarity measures.

1 Introduction

Recently developed metabolomic and genomic measuring technologies share the common property to yield in parallel thousands of metabolites and gene expression values from single probes of a given tissue/plant sample. Tools used for these purposes are mass spectrometry, chromatography, and micro- and macroarrays. In high-throughput approaches the number of probe attributes (metabolites, genes) is usually much higher than the number of probes, which is paradigmatic of the curse of dimensionality. Thus, it is desirable for analysis to consider as many experimental probes as data quality allows. Such desire for maximum information preservation for only few unlabeled data samples excludes the utilization of prototype-based data abstractions like supervised neural gas proposed for labeled data [2]. Principal component analysis PCA, the classical approach to factor analysis of unlabeled data, has got different limitations: the analytic focus is shifted away from the data matrix towards the attribute covariance matrix of which eigenvalues are computed to rate the importance of the axes of principal data directions. These axes, however, are linear combinations of the original data

F. Schwenker and S. Marinai (Eds.): ANNPR 2006, LNAI 4087, pp. 274–285, 2006.

attributes – this situation requires a complex interpretation of the eigenvector entries ('loadings') in order to rate the original data attributes. PCA finally results in the amount of feature contribution to the overall data variance. Both, implicit rotation of the data coordinate system and the restriction to variance, implying the Euclidean data metric for a reasonable interpretation, are circumvented in the following approach. In terms of feature subset selection (FSS) the proposed method will be a filter rather than a wrapper [3]. Custom data similarity measures can be integrated to the framework, and, furthermore, the new reconstruction-based feature selection does not require class labels, which complements other approaches such as proposed in [5]. For the lack of data samples, distribution-based separability criteria and expectation maximization methods for unlabeled data, like FSSEM-TR/ML [1], cannot be properly applied in the present case. In the proposed solution, no external clustering is required for evaluating the changes before and after masking (veiling) subset of features; instead, a built-in filter criterion is used which optimizes the reconstruction quality of the veiled data according to the strategy discussed in the following.

2 Unsupervised Feature Selection Based on Maximum Reconstruction Quality

Feature selection and weighting do both refer to the process of characterizing the relevance of components in fixed-dimensional data vectors. Unfortunately, many biological data sets do not possess an absolute reference coordinate system upon which a proper attribute analysis can be grounded: the organic material itself and many external influences affect the measurements, and the obtained data are thus, in a certain degree, situated in empirical domains. For example, in gene expression data, a theoretical lower bound of zero intensity exists, but due to background noise this value is never observed in practice. Subsequent standard operations like the logarithm might further amplify this uncertain domain, especially for near zero intensities. The *ad hoc* definition of absolute data domains can be avoided by dealing with relationships expressed by the data similarity matrix. This requires to choose an appropriate similarity measure. In case of the Euclidean metric, the resulting distance matrix is invariant to data (baseline) shifts and coordinate rotations. Invariance can be realized already at data level by using Pearson correlation which is invariant to vector shifting and scaling. This beneficial property is used as quality criterion for comparing data similarity matrices. Using the above ingredients, feature ranking for data from an observation-driven domain is realized by sensitivity analysis, i.e. by analyzing the effect of measure-specific feature veiling on the quality of reconstruction of the original data relationships. This general approach is sketched in Fig. 1. It is required that the data similarity measure d is chosen in advance, such as Euclidean distance or Pearson correlation in the following. If weighting is considered instead of feature dropping, also a parametric counterpart $d^{\boldsymbol{\lambda}}$ of d is necessary.

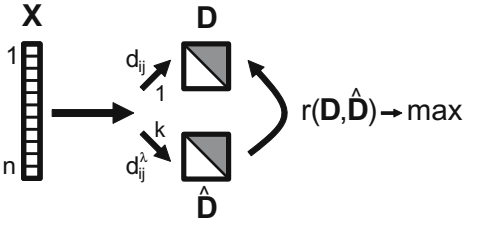

Fig. 1. Feature selection by reconstruction quality maximization. Due to symmetry, a number of $n \cdot (n-1)/2$ relationships of data vectors in $\mathbf{X}$ is once computed with static similarity measure d to yield triangular reference matrix $\mathbf{D}$ (upper path). Features are dropped or weighted by the $\boldsymbol{\lambda}$-parametrized measure $\mathrm{d}^{\boldsymbol{\lambda}}$ in a k-iterative manner (lower path): for greedy selection, those features providing highest correlation between feature-reduced similarity matrix $\mathbf{D}$ and reference matrix $\mathbf{D}^{\boldsymbol{\lambda}}$ are further considered important; for parallel selection, the average response to random feature perturbation is calculated.

The parametric Euclidean distance $\mathrm{d}_{ij}^{\boldsymbol{\lambda}} = {}^{E}\mathrm{d}_{ij}^{\boldsymbol{\lambda}} \in [0;\infty)$ is given by

$$ {}^{E}\mathrm{d}_{ij}^{\boldsymbol{\lambda}}(\mathbf{x}^i, \mathbf{x}^j) = \sqrt{\sum_{l=1}^{q} \lambda_l \cdot (x_l^i - x_l^j)^2}\,. \tag{1} $$

Canonic feature weighting is obtained by inserting weight factors to the squared differences – setting $\lambda_l = 1$ for $l = 1 \ldots q$ yields the original Euclidean distance. If just one parameter λ_l is zero, the others one, this expresses dropping of feature l.

The parametric Pearson correlation $\mathrm{d}_{ij}^{\boldsymbol{\lambda}} = \mathsf{r}_{ij}^{\boldsymbol{\lambda}} \in [-1;1]$ is given by

$$ \mathsf{r}_{ij}^{\boldsymbol{\lambda}} = \frac{\sum_{l=1}^{q} \lambda_l^2 \cdot (x_l^i - \mu_{\mathbf{x}^i}) \cdot (x_l^j - \mu_{\mathbf{x}^j})}{\sqrt{\sum_{l=1}^{q} \lambda_l^2 \cdot (x_l^i - \mu_{\mathbf{x}^i})^2} \cdot \sqrt{\sum_{l=1}^{q} \lambda_l^2 \cdot (x_l^j - \mu_{\mathbf{x}^j})^2}}\,. \tag{2} $$

Each of the mean-subtracted vector components $(x_l^m - \mu_{\mathbf{x}^m})$ has got its proper relevance factor λ_l – again, setting $\lambda_l = 1$ for $l = 1 \ldots q$ yields the original Pearson correlation. Note that, in contrast to Euclidean distance, setting $\lambda_l = 0$, $\lambda_m = 1, m \neq l, m = 1 \ldots q$ is not equivalent to dropping feature l, because it still contributes to the vector averages $\mu_{\mathbf{x}^i}$ and $\mu_{\mathbf{x}^j}$. Instead, the feature's induced mean deviation from average is measured.

For feature selection, parameters λ_l are searched that provide maximum correlation of parametrized data relationships and original data relationships. Trivial solutions $\lambda_l = C, C > 0, l = 1 \ldots q$ are avoided by construction.

- For dropping, correlation values $\mathsf{r}(\mathbf{D}, \mathbf{D}^{\boldsymbol{\lambda}})$ are computed for all attributes separately masked. Those with maximum correlation degradation are considered especially important. This attribute can be wiped out and the procedure can be repeated iteratively.
- For weighting, Monte-Carlo sampling around an optimum $\boldsymbol{\lambda}$-vector is performed and the average restoring forces are calculated by a gradient ascent

approach, by analyzing absolute values of gradients pointing into the parameter direction of high correlation values $\mathsf{r}(\mathbf{D}, \mathbf{D}^{\boldsymbol{\lambda}})$.

2.1 Feature Dropping

Feature relevance can be systematically probed by excluding single attributes from data similarity calculation and testing the impact of that operation on the correlation $\mathsf{r}(\mathbf{D}, \mathbf{D}^{\boldsymbol{\lambda}})$. By feature dropping, as a basic assumption, highly important features will induce a larger loss of r than less important ones. Thus, a first approach to relevance rating is the correlation loss resulting from feature dropping. Such a *top-level feature evaluation* can be recursively formulated in a greedy manner. This *iterative feature dropping* approach stores the index and then really excludes the currently most relevant feature from further calculations. It iteratively isolates those attributes that do maximum decorrelate the original similarity matrix $\mathbf{D}$ and the feature-reduced distance matrix $\mathbf{D}^{\boldsymbol{\lambda}}_S$:

$$S(k) = \arg\min_i \mathsf{r}^2(\mathbf{D}, \mathbf{D}^{\boldsymbol{\lambda}}_{S(k-1)\cup i}),\ i \in (1 \ldots T) \backslash S(k-1), k = 1 \ldots T-1.$$

$S(k)$ is the growing set of index pointers to features which have been isolated until iteration number k; by definition $S(0) := \{\}$, and by construction $|S(k)| = k$. $\mathbf{D}^{\boldsymbol{\lambda}}_{S(k-1)\cup i}$ is the similarity matrix that has been calculated by using the data vectors, thereby skipping the features indexed by the set $S(k-1) \cup i$.

The straightforward greedy algorithm does not require further parameters, however, two alternative design criteria need further attention. First, $\mathbf{D}^{\boldsymbol{\lambda}}_{S(k-1)\cup i}$ is correlated with $\mathbf{D}^{\boldsymbol{\lambda}}$, not with $\mathbf{D}^{\boldsymbol{\lambda}}_{S(k-1)}$. The reason is that a drift away from the original data set towards the subsequently reduced data features might occur otherwise, so $\mathbf{D}^{\boldsymbol{\lambda}}$ constitutes a fixed reference. Second, features are iteratively masked out from high relevances to low ones, not the other way round. This way, much of the relation-explaining attributes are already cleared off in the first steps, instead of realizing a culmination towards the crucial data attributes by least-attributes-first exclusion. This is beneficial in large scale applications with thousands of dimensions, because it allows early stopping when the remaining absolute correlation r^2 drops below a critical near-zero threshold, or in case of reaching a plateau. These two options – there are certainly many more – and the different results from the alternative greedy feature selection designs are circumvented by parallel feature selection as discussed in the next paragraph.

2.2 Feature Weighting

In the following approach, gradients are calculated for rating the data features. Decent perturbations are induced to the parameters λ_l of the adaptive similarity measure $\mathrm{d}^{\boldsymbol{\lambda}}$, close to the optimum values. The higher, on average, the return forces (gradients) of the disturbed parameters, the more important are the corresponding attributes for restoring maximum correlation $\mathsf{r}(\mathbf{D}, \mathbf{D}^{\boldsymbol{\lambda}})$. The proposed method uses several paradigms from artificial neural networks: the perturbation and pattern presentation processes are stochastic, a principle of correlation-maximization is pursued, and parametric similarity measures are optimized – or

are at least rated – using gradient dynamic. For the derivatives, an approach is chosen which has been proposed earlier for efficient multi-dimensional scaling [4]. In order to prevent saturation at boundaries of the correlation domain $[-1;1]$, the widely used Fisher z′-transform with its derivative is utilized:

$$\mathrm{z}'(\mathsf{r}) = \frac{1}{2} \cdot \log\left(\frac{a+\mathsf{r}}{a-\mathsf{r}}\right) \quad \Rightarrow \quad \frac{\partial \mathrm{z}'(\mathsf{r})}{\partial \mathsf{r}} = \frac{a}{a^2 - \mathsf{r}^2}\,.$$

In Fisher's original formulation a is set to 1, but here it is kept variable $a = 1+\epsilon$ in order to avoid infinitely large values in case of perfect correlation. For example, $a = (1+\sqrt{401})/20 \approx 1.05$ limits the transformed derivative domain to $[-10; 20/(1+\sqrt{401})]$. Desired gradients for λ_l with negative correlation transform result from application of the chain rule to the nested stress function formulation:

$$\mathrm{s} = -\mathrm{z}' \circ \mathsf{r} \circ \mathrm{d}^{\boldsymbol{\lambda}} \circ \boldsymbol{\lambda} \quad \Rightarrow \quad \frac{\partial \mathrm{s}}{\partial \lambda_l} = -\sum_{i=1}^{n} \sum_{j=1\ldots n}^{j\neq i} \frac{\partial \mathrm{z}'(\mathsf{r})}{\partial \mathsf{r}} \cdot \frac{\partial \mathsf{r}}{\partial \mathrm{d}^{\boldsymbol{\lambda}}_{ij}} \cdot \frac{\partial \mathrm{d}^{\boldsymbol{\lambda}}_{ij}}{\partial \lambda_l}\,. \tag{3}$$

Using the abbreviations $\mathsf{r}(\mathbf{D}, \mathbf{D}^{\boldsymbol{\lambda}}) = \mathscr{H}/\sqrt{\mathscr{W} \cdot \mathscr{U}}$ with

$$\begin{aligned} \mathscr{H} &= \textstyle\sum_{l=1}^{n} \sum_{m=1}^{n} (\mathrm{d}_{lm} - \mu_{\mathbf{D}}) \cdot (\mathrm{d}^{\boldsymbol{\lambda}}_{lm} - \mu_{\mathbf{D}^{\boldsymbol{\lambda}}})\,, \\ \mathscr{W} &= \textstyle\sum_{l=1}^{n} \sum_{m=1}^{n} (\mathrm{d}_{lm} - \mu_{\mathbf{D}})^2\,, \\ \mathscr{U} &= \textstyle\sum_{l=1}^{n} \sum_{m=1}^{n} (\mathrm{d}^{\boldsymbol{\lambda}}_{lm} - \mu_{\mathbf{D}^{\boldsymbol{\lambda}}})^2\,, \end{aligned}$$

the derivative of the z′-transformed Pearson correlation is calculated by

$$\frac{\partial \mathrm{z}'(\mathsf{r})}{\partial \mathsf{r}} \cdot \frac{\partial \mathsf{r}}{\partial \mathrm{d}^{\boldsymbol{\lambda}}_{ij}} = \frac{a \cdot \left((\mathrm{d}^{\boldsymbol{\lambda}}_{ij} - \mu_{\mathbf{D}^{\boldsymbol{\lambda}}}) \cdot \mathscr{H} - (\mathrm{d}_{ij} - \mu_{\mathbf{D}}) \cdot \mathscr{U}\right) \cdot \sqrt{\mathscr{W}}}{(\mathscr{H} - a \cdot \sqrt{\mathscr{U} \cdot \mathscr{W}})^2 \cdot \sqrt{\mathscr{U}}}\,. \tag{4}$$

The term $\mathscr{W}$ needs to be calculated only once, even the mean of the static similarity matrix can be initially removed $\mathrm{d}_{lm} \leftarrow (\mathrm{d}_{lm} - \mu_{\mathbf{D}})$ in order to save computing operations. Eqn. 3 is evaluated for all features and the absolute values are averaged over a sufficient number of small random perturbations. For better comparison, these averaged gradient responses are rescaled to an upper limit of one representing the most sensitive feature.

Eqn. 4 is generic enough to plug in any differentiable parametric similarity measure. Two interesting choices are the parametric Euclidean distance for data comparisons and an adaptive version of the Pearson correlation that plays an important role in biopattern processing. These measures require derivatives $\partial {}^{E}\mathrm{d}^{\boldsymbol{\lambda}}{}_{ij}/\partial\lambda_l$ and $\partial \mathsf{r}^{\boldsymbol{\lambda}}_{ij}/\partial\lambda_l$ as rightmost factors in equation 3, respectively.

Parametric Euclidean. The derivative of the parametric Euclidean is easily obtained as:

$$\frac{\partial {}^{E}\mathrm{d}^{\boldsymbol{\lambda}}{}_{ij}}{\partial \lambda_l} = \frac{\partial}{\partial \lambda_l} \sqrt{\sum_{m=1}^{q} \lambda_m \cdot (x^i_m - x^j_m)^2} = \frac{(x^i_l - x^j_l)^2}{\sqrt{\sum_{m=1}^{q} \lambda_m \cdot (x^i_m - x^j_m)^2}} = (x^i_l - x^j_l)^2 \,/\, {}^{E}\mathrm{d}^{\boldsymbol{\lambda}}{}_{ij}\,.$$

Parametric Pearson Correlation. For deriving the $\boldsymbol{\lambda}$-weighted correlation $\mathsf{r}_{ij}^{\boldsymbol{\lambda}}$, a focus on component l will be a convenient abbreviation. Similar to the previous matrix correlations, the notation $\mathsf{r}_{ij}^{\boldsymbol{\lambda}} = \mathscr{H}_l / \sqrt{\mathscr{W}_l \cdot \mathscr{U}_l}$ of the correlation term is considered using

$$\begin{aligned}
\mathscr{H}_l &= \lambda_l^2 \cdot (x_l^i - \mu_{\mathbf{x}^i}) \cdot (x_l^j - \mu_{\mathbf{x}^j}) + \textstyle\sum_{\substack{u \neq l \\ u=1}}^{q} \lambda_u^2 \cdot (x_u^i - \mu_{\mathbf{x}^i}) \cdot (x_u^j - \mu_{\mathbf{x}^j}) \,, \\
\mathscr{W}_l &= \lambda_l^2 \cdot (x_l^i - \mu_{\mathbf{x}^i})^2 + \textstyle\sum_{\substack{u \neq l \\ u=1}}^{q} \lambda_u^2 \cdot (x_u^i - \mu_{\mathbf{x}^i})^2 \,, \\
\mathscr{U}_l &= \lambda_l^2 \cdot (x_l^j - \mu_{\mathbf{x}^j})^2 + \textstyle\sum_{\substack{u \neq l \\ u=1}}^{q} \lambda_u^2 \cdot (x_u^j - \mu_{\mathbf{x}^j})^2 \,.
\end{aligned}$$

With these isolated subterms, the derivative of interest is

$$\frac{\partial \mathsf{r}_{ij}^{\boldsymbol{\lambda}}}{\partial \lambda_l} = \frac{\lambda_l \cdot \left(2(x_l^i - \mu_{\mathbf{x}^i})(x_l^j - \mu_{\mathbf{x}^j}) \cdot \mathscr{W}_l \mathscr{U}_l - \mathscr{H}_l \cdot \left(\mathscr{U}_l \cdot (x_l^i - \mu_{\mathbf{x}^i})^2 + \mathscr{W}_l \cdot (x_l^j - \mu_{\mathbf{x}^j})^2 \right) \right)}{(\mathscr{W}_l \cdot \mathscr{U}_l)^{\frac{3}{2}}}.$$

Parameters of the feature weighting approach are the gradient delimiter which has been set to $a = 1.01$, the perturbation interval, and the number of iterations for calculating the average response gradients. The interval for random perturbations has been determined by studies on several data sets, including the ones presented in the application section. It has turned out that in case of both parametric Euclidean and Pearson similarity measures, parameters uniformly chosen $\lambda_l \in [0.75; 1.25]$ produce stable results. A number of $k = 1000$ iterations is chosen. Stability has been additionally tested by letting the parameters iteratively adapt according to stochastic descent on s using the calculated gradients: after noise induction, the parameters quickly return to constant values $\lambda_i \approx \lambda_j, \forall i, j$.

3 Applications

The presented methods are applied to three data sets of interest: to benchmark data related to absorbance spectra from Infratec Tecator food analyzer, publicly available from statlib data collection at http://lib.stat.cmu.edu/datasets/tecator, (215 samples, 100 dimensions [frequency channels]); to chromatography data from the in-house tomato germplasm database focusing on chemical compound detection at a wavelength of $280nm$ (19 samples, 3000 dimensions [retention time points]); and to gene expression data from macroarray hybridization experiments of developing endosperm barley tissue at 0–26 days after flowering sampled in steps of two days (two series, 14 samples each, 11786 dimensions [genes]).

Tecator Benchmark Spectral Data. The first data set has been included for illustration and reference purposes. It contains 215 food samples analyzed in a near infrared frequency range of 850–$1050nm$ measured with the Tecator Infratec Food and Feed Analyzer. The 100-dimensional spectra, originally used for predicting high and low fat content, are smoothly shaped, as shown for 10 examples in the top left panel of Fig. 2. Looking at the other panels of Fig. 2, several observations are made.

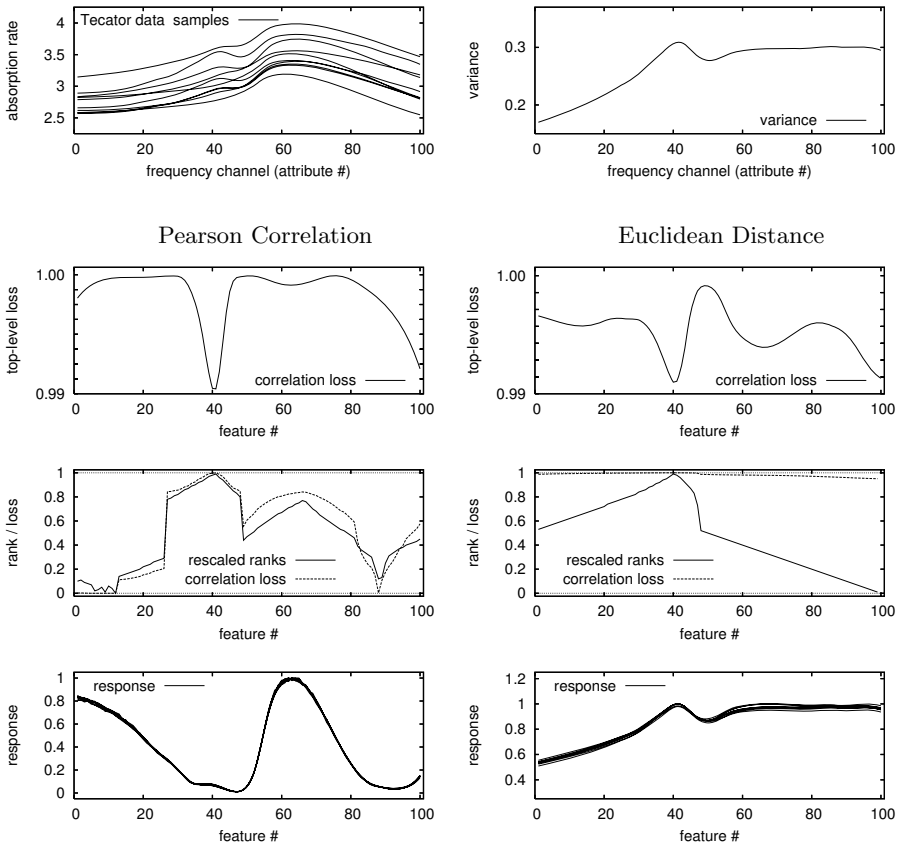

Fig. 2. Feature selection for Tecator data set. Top row: left panel displays 10 samples from 100-dimensional spectra; right panel channel variances for entire data set containing 215 samples. Subsequent rows: top-level feature sensitivity measuring the loss of squared correlation with original similarity matrix, $\mathsf{r}^2(\mathbf{D}, \mathbf{D}_l^{\boldsymbol{\lambda}})$, caused by dropping feature l (low correlation indicates high feature sensitivity); loss of squared correlation and corresponding feature rank caused by iterative dropping of the currently most sensitive feature (multiple application of top-level analysis with recursively reduced feature set); feature weighting based on gradients that point towards optimum state after random parameter perturbations of the adaptive similarity measure (graphs of ten independent runs are overlaid showing high reproducibility). Left column refers to adaptive Pearson correlation, right column to parametric Euclidean distance for the three investigated methods.

Most importantly, pairs of plots in the left column – corresponding to Pearson correlation similarity – and in the right column – displaying results for Euclidean distance – are rather different. Thus, as expected, the choice of data similarity measure has crucial influence on the highly rated features.

Row two for top level loss contains plots of the loss of correlation $\mathsf{r}^2(\mathbf{D}, \mathbf{D}_l^{\boldsymbol{\lambda}})$ after deletion of attribute l. Both plots exhibit a common minimum around

feature $l = 41$, pointing out these attributes as highly sensitive for both similarity measures. However, for other attributes just ratings are computed. It is pointed out that, due to data redundancy and the high number of 100 dimensions, the maximum correlation loss for many dropped features is still very close to one.

Row three, showing plots of rank and loss for iterative feature dropping, gives support to the top-ranked features around index 41 of high importance for correlation reconstruction. The loss described by the dotted lines has much higher variability in case of Pearson similarity than in case of Euclidean distance. Subsequent feature ranking is obtained by assigning their ascending sorting indices. These ranks have been divided by the number of dimensions in order to obtain a mapping into the value range of squared correlation. As a consequence of greedy feature selection, large-scale discontinuities appear in the resulting graphs.

Row four with response plots contains smooth non-ranked attribute weights obtained by gradient calculations. Three important properties are observed. First, indices around 41 for Pearson similarity are remarkably insensitive in contrast to the results from the other two approaches. Second, the results of ten independent runs of the gradient method display very high reproducibility. Third, the graphs for Euclidean distance in the bottom right panel are strikingly similar to the simple variance plot in the top right panel – the average squared correlation between the ten response graphs with the variance is $\mathsf{r}^2(\text{variance}, \text{Euclidean response}) = 0.991$. This is a key observation. On one hand, this meets the expectation of the role of variance for Euclidean distance as a natural measure of data variability – although the presented approach measures, inversely, the sensitivity of the parametric Euclidean distance. On the other hand, this essential solution for the Euclidean distance induces high confidence in analog results for non-Euclidean case, like those given for the adaptive Pearson similarity. This approach can thus be regarded as generalization of the concept of variance to other types of parametric data similarity measures.

To conclude, quite different feature evaluations are obtained for the different approaches. This points out that feature dropping is structurally different from parametric measure perturbation. The case of correlation measure shows insensitivity to attribute scaling where entire feature dropping produces the highest loss, around index 41. However, in case of masked or weighted Euclidean distance, the special importance of that feature set around index 41 is common sense for all methods.

Tomato Peel Chromatograms for Chemical Compound Analysis. High performance liquid chromatography (HPLC) allows recording of high resolution spectra related to compound-specific absorbance rates. Especially the group of health protective flavonoids is of great interest for the evaluation of food crops. Here, a collection of tomato plants is studied at a wavelength of $280nm$ to capture the chemical constituents within the fruit peel. A measuring duration of $50min$ considered with a sampling of $1Hz$, producing values for 3000 retention times per fruit. Biological attention is put on 19 of these chromatograms to find intervals of retention times with characteristic variability in absorption. The integrated values in those intervals are proportional to the abundance of the corresponding chemical compounds. For precise further calculations, the chromatogram have

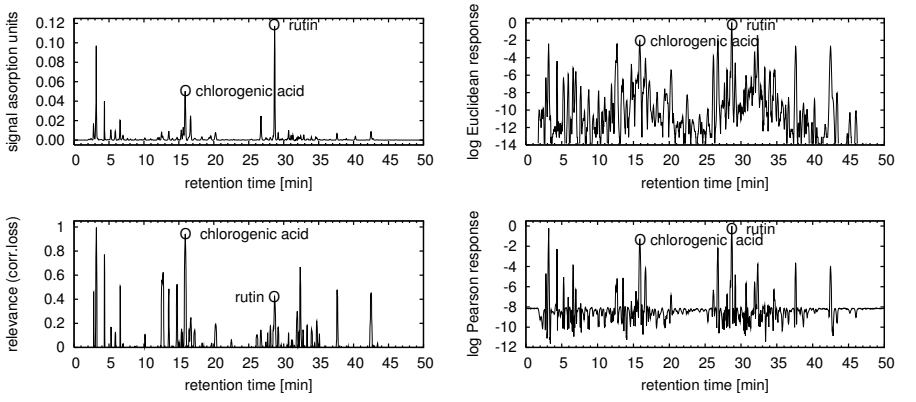

Fig. 3. Feature selection for tomato data set. Top left: one exemplary chromatogram from the data set used for feature rating. Bottom left: iterative correlation loss for feature dropping with Pearson similarity. Right: gradient responses for adaptive Euclidean distance (top) and parametric Pearson correlation (bottom). For comparison, two important substances representative for all chromatograms are encircled: chlorogenic acid and rutin.

been baseline-corrected and their peaks have been aligned by the correlation optimized warping method.

Fig. 3 contains results for different feature rating methods applied to the tomato peel chromatograms. The original chromatograms look very peaky if plotted with a condense time axis like in the top right panel; by zooming onto the time axis, however, smoother details become visible. A reason for not using adaptive Euclidean response is shown in the top right panel: two time intervals can be identified, $[10; 20]$ and $[25; 35]$, which intermediately drift to higher relevance values just because of a higher overall variability in these domains. A very good correspondence of Euclidean response and variance is supported by a high squared correlation value of $r^2(\text{log variance, log Euclidean response}) = 0.976$.

The strongly oscillating plot for feature variance is complemented by the Pearson correlation response, as shown in the lower left panel. A straight baseline is identified at a value of about -8, and the peaks provide much clear candidates of interesting retention times. For the present high-dimensional data set with few data points, iterative feature dropping for Pearson similarity yields very similar results, as given in the bottom left panel of Fig. 3. As a matter of fact, correlation-based chromatogram comparison usually has more biological impact than in Euclidean manner. This is already supported at the level of chromatogram peak alignment where correlation-optimized warping yields most accepted curve alignments. The Pearson correlation response in the lower left panel points out retention times that are in high agreement with biological knowledge. Moreover, the clear baseline can be used to define a threshold above which time intervals might be automatically integrated for further analysis.

Barley Endosperm Gene Expression Data. Discovery of sequential processes involved in tissue differentiation are available from gene expression

data. The search for key identity genes specific for observed tissue differentiation is a valuable desire. Micro- and macroarray technology allows parallel recording of the abundance of thousands of gene expression intensities. The identification of key regulators from such a usually long list of expression values is a particularly challenging task. Here, log-normalized expression values for 11786 genes from in-house macroarray hybridization experiments are analyzed. Two independent series of experiments are available concerning the development of endosperm barley tissue at 0–26 days after flowering, sampled in steps of two days.

Analytic focus has been put on correlation-based feature identification. This overcomes limitations of Euclidean distance approaches that emphasize genes which are mainly related to high variance. Two lists of top-rated 25 genes out of 11786 are computed, one by feature dropping and the other by response

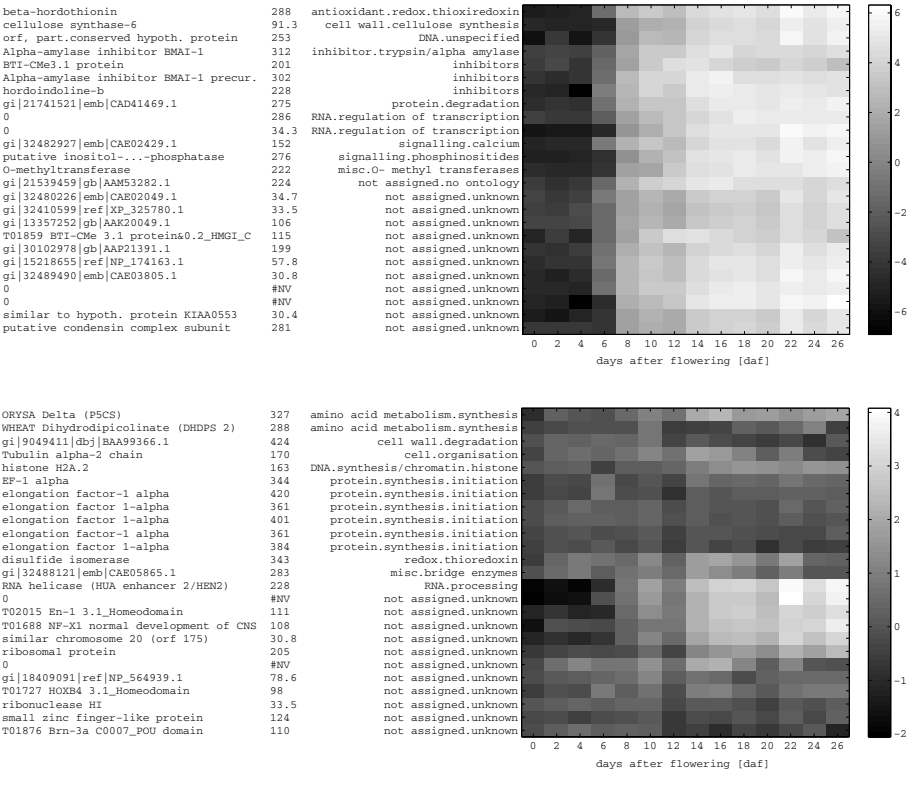

Fig. 4. Temporal gene regulation of top 25 genes in endosperm barley tissue. Upper panel: results from feature dropping. Bottom panel: results from gradient response analysis. Both panels are related to the Pearson similarity measure. Gene profiles have been ordered by a combination of one-dimensional self-organizing map and functional annotation. Shades of gray denote normalized gene expression intensities according to the reference bar. Text columns contain Blast description, Blast score, and functional category of genes.

gradients. Thereby, the two independent series of gene expression experiments are processed separately, their gene ranks are summed up, and the highest 25 sums of ranks are considered top candidate genes. Rank summation has been regarded as a valid operation after confirming that the squared correlation of the ranks of all genes is greater than 0.9. In order to compare results of feature dropping and response gradients, the temporal gene regulation profiles associated with the top-rated genes are plotted in Fig. 4.

In summary, feature evaluation based on Pearson correlation yields quite different results for feature dropping and gradient response analysis. The group of genes obtained by feature dropping (upper panel) exhibits common patterns of strong up-regulation. Among the top-rated 25 genes detected by feature dropping, most of them are found to be endosperm-specific and exclusively detected in *triticeae* species. In other words, the feature of up-regulation is considered important for characterizing the data set, which is a reasonable finding for the temporally related experiments. Nonetheless, this very prominent regulation characteristic could be captured by standard clustering techniques. Qualitatively new patterns are revealed by gradient response analysis (lower panel). Intermediate regulations are shown in addition to the group of moderately up-regulated genes (Fig. 4, lower panel). Thus, again, variability alone does not take too much influence on gene selection. Interestingly, most detected genes that are expressed during late endosperm development are connected to protein synthesis initiation processes. These are considered to have an important functional role during the peak of product accumulation.

4 Conclusions

A new approach to unsupervised feature selection has been proposed. Its basic principle is the detection of features that maximum decorrelate original and feature-masked data relationships. These features are supposed to be most critical for faithful relationship reconstruction. The considered data relationships are defined by appropriate data similarity measures, such as the presented Euclidean distance and Pearson correlation. Sensitivity is obtained as response to feature dropping or as gradient-based reaction to small perturbations in parametric formulations of the utilized similarity measure. Both ways measure correlation loss, but, by construction, they are structurally different. As has been demonstrated in the experiments, gradient-based response analysis can be regarded as for measure-specific counterpart of variance. This canonic interpretation, its computational advantage over greedy feature dropping, and the parallel feature probing makes response analysis the preferred feature selection method. Whichever technology is chosen, reasonable automatic feature rating essentially helps to pre-structure the data, to get different views and to formulate hypotheses about the data sets, like for the three examined high-dimensional data sets. In unsupervised scenarios the data-driven, method-intrinsic dynamic fully determines the outcome; therefore, since unsupervised methods optimize different goals, objective quality criteria are missing in comparisons, and the results must

be judged by a combination of subjectiveness and additional knowledge. In future studies, further potential of the proposed methods will be assessed in close cooperation with biological experts and their additional background knowledge.

Acknowledgement. The work is supported by BMBF grant FKZ 0313115, GABI-SEED-II.

References

1. J. Dy and C. Brodley. Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5:845–889, 2004.
2. B. Hammer, M. Strickert, and T. Villmann. Supervised neural gas with general similarity measure. *Neural Processing Letters*, 21(1):21–44, 2005.
3. N. Søndberg-Madsen, C. Thomsen, and J. Pena. Unsupervised feature subset selection. In *Proceedings on the Workshop on Probabilistic Graphical Models for Classification*, pages 71–82, 2003.
4. M. Strickert, S. Teichmann, N. Sreenivasulu, and U. Seiffert. High-Throughput Multi-Dimensional Scaling (HiT-MDS) for cDNA-array expression data. In W. Duch et al., editor, *Artificial Neural Networks: Biological Inspirations, Part I, LNCS 3696*, pages 625–634. Springer, 2005.
5. L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th International Conference on Machine Leaning (ICML-03)*, pages 856–863, 2003.