ADAPTIVE MATRIX METRICS FOR ATTRIBUTE DEPENDENCE ANALYSIS IN DIFFERENTIAL HIGH-THROUGHPUT DATA

M. Strickert[1], K. Witzel[1], J. Keilwagen[1], H.-P. Mock[1], P. Schneider[2], M. Biehl[2], and T. Villmann[3]*

[1]Leibniz Institute of Crop Plant Research Gatersleben, Germany,
[2]Institute for Math. and Computer Science, University of Groningen, NL,
[3]Research group Computational Intelligence, University of Leipzig, Germany.
[*]Corresponding author email: stricker@ipk-gatersleben.de

ABSTRACT

Data-driven metric adaptation is proposed for proteome analysis of 2D-gel electrophoretic plots aiming at identification of stress related proteins in two barley cultivars with different response towards different salt stress conditions. Gradient descent is applied to the ratio of intra- and inter-class distance sums to optimize the matrix parameters of generalized Mahalanobis distances in order to separate the several hundred dimensional data of protein intensities in the transformation space. The resulting matrix contains mutual dependence of spots, explaining differential stress reactions and putative protein interactions. We present interesting results obtained by the new metric learning method that possesses general applicability in biomedical data analysis.

Keywords: Supervised feature characterization, adaptive matrix metric, attribute dependence modeling.

1. INTRODUCTION

The identification of gene and protein dependences is an essential step for the inference of interaction networks from experimental data. Both network inference and the exploration of the obtained connectivity structure are hot topics in systems biology [5]. Typical approaches for the automatic reconstruction of network topologies make use of correlation measures [4], Bayesian inference [9], or information theoretic statistics [8] in order to model the mutual dependence of network nodes. Among different beneficial properties these models also possess some unwanted properties, ranging from being rather simplistic or suffering from high computational complexity to requiring additional assumptions like density estimates. The assessment of the quality of the inferred networks is usually problematic. One general reason is that test statistics might be inappropriate for reflecting biological experience [3]. A more specific problem is the biological probing and confirmation of the huge number of potential interaction partners. Alternatively, the promising concept of learning metrics from the area of machine learning research [6, 13] can be utilized for network construction by modeling attribute pairs. Data-driven metric adaptation also helps to reduce the curse of dimensionality occurring during the analysis of high-throughput data. In our case, protein data of 2D electrophoretic gels are considered providing intensities of many protein spots measured in a relatively low number of available experiments. A minimalistic attribute characterization method is used for rating the influence of attribute pairs on the spatial arrangement of class-specific data clouds in vector space, expressed by an adaptive matrix metric, as recently utilized in matrix learning vector quantization [11]. The method presented here aims at minimizing within-class differences while maximizing inter-class distances by rescaling the data space based on a trained transformation matrix without building an explicit classification model [12]. Although this aim resembles the one of linear discriminant analysis (LDA) [2], the transform maintains the original data dimensionality and is thus not reduced to an a priori low-dimensional LDA subspace. The new method yields an estimate of a label-specific inverse covariance matrix and might be considered as supervised whitening operation. Some concepts can be related to the threshold gradient descent method [7].

2. METHOD – MATRIX LEARNING

As input q-dimensional row vectors $\mathbf{x} \in \mathbb{R}^{1\times q}$ are assumed to be taken from a set containing n data vectors $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$. The proposed metric adaptation requires a class-specific label $\mathrm{c}(k)$ for each data vector $\mathbf{x}^k$. We define the main building block of the method, the matrix-based metric $\mathrm{d}^{ij}_{\mathbf{\Omega}} \in [0;\infty)$ for data vectors $\mathbf{x}^i$ and $\mathbf{x}^j$:

$$\mathrm{d}^{ij}_{\mathbf{\Omega}} = \mathrm{d}_{\mathbf{\Omega}}(\mathbf{x}^i, \mathbf{x}^j) = (\mathbf{x}^i - \mathbf{x}^j) \cdot \mathbf{\Lambda} \cdot (\mathbf{x}^i - \mathbf{x}^j)^{\mathsf{T}}\,, \quad (\mathbf{\Lambda} = \mathbf{\Omega} \cdot \mathbf{\Omega}^{\mathsf{T}}) \in \mathbb{R}^{q\times q}\,. \qquad (1)$$

The identity matrix $\mathbf{\Lambda} = \mathbf{\Omega} = \boldsymbol{I}$ induces the special case of the squared Euclidean distance; other diagonal matrices yield weighted squared Euclidean distances. Arbitrary positive-definite matrices $\mathbf{\Lambda}$ lead to very general metrics that can express rotation and translation which do not affect distances between points, and scaling and shearing which do affect them. A triangular or symmetric matrix $\mathbf{\Omega}$ would be sufficient to express any such configuration by Eqn. 1. Faster convergence can be observed, though, if the full matrix is adapted in the matrix optimization scheme

for minimizing the label-specific metric stress criterion:

$$\mathrm{s}(\mathbf{\Omega}) \;:=\; \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{d}_{\mathbf{\Omega}}(\mathbf{x}^i,\mathbf{x}^j)\cdot\delta_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{d}_{\mathbf{\Omega}}(\mathbf{x}^i,\mathbf{x}^j)\cdot(1-\delta_{ij})} = \frac{\mathrm{d}_C}{\mathrm{d}_D}$$
$$\text{with} \qquad \delta_{ij} = \begin{cases} 0 : \mathrm{c}(i)\neq\mathrm{c}(j) \\ 1 : \mathrm{c}(i)=\mathrm{c}(j) \end{cases} . \qquad (2)$$

Distances $\mathrm{d}_{\mathbf{\Omega}}^{ij}$ between all n data vectors $\mathbf{x}^i$ and $\mathbf{x}^j$ depend on the adaptive matrix parameters $\mathbf{\Omega} = (\Omega_{kl})_{\substack{k=1\ldots q \\ l=1\ldots m}}$ of interest. The numerator represents within-class data variability, which should be small. The denominator is related to inter-class distances, which should be large. Thus, optimization of $\mathrm{s}(\mathbf{\Omega})$ handles both parts of the fraction simultaneously. Compromise solutions must be found in cases when within-class variation, potentially caused by outliers, needs compression, while inter-class separability would require inflation.

Although similar at first glance, the proposed approach is structurally different to LDA, because the inverse LDA-like ratio in Eqn. 2 is optimized in the original data space, not in the projection to the most prominent class separating LDA direction [12]. In contrast to LDA where covariance matrices and class centers can be initially computed and then reused, this is not possible in the proposed method, because the metric adaptation affects both class centers and data covariances. Full matrix adaptation, though, creates higher computational demands of the optimization method described in the following.

The cost function $\mathrm{s}(\mathbf{\Omega})$ gets iteratively minimized by gradient descent. This requires adaptation of the matrix $\mathbf{\Omega}$ in small steps γ into the direction of steepest gradient

$$\mathbf{\Omega} \leftarrow \mathbf{\Omega} - \gamma \cdot \frac{\partial \mathrm{s}(\mathbf{\Omega})}{\partial \mathbf{\Omega}} \qquad (3)$$

obtained by the chain rule

$$\frac{\partial \mathrm{s}(\mathbf{\Omega})}{\partial \mathbf{\Omega}} = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial \mathrm{s}(\mathbf{\Omega})}{\partial \mathrm{d}_{\mathbf{\Omega}}^{ij}} \cdot \frac{\partial \mathrm{d}_{\mathbf{\Omega}}^{ij}}{\partial \mathbf{\Omega}} . \qquad (4)$$

The derivative of the fraction $\mathrm{s}(\mathbf{\Omega}) = \mathrm{d}_C/\mathrm{d}_D$ in Eqn. 2 is

$$\begin{aligned} \frac{\partial \mathrm{s}(\mathbf{\Omega})}{\partial \mathrm{d}_{\mathbf{\Omega}}^{ij}} &= \frac{\delta_{ij}\cdot \mathrm{d}_D}{\mathrm{d}_D^2} + \frac{(\delta_{ij}-1)\cdot \mathrm{d}_C}{\mathrm{d}_D^2} \\ &= \begin{cases} 1/\mathrm{d}_D \;:\; \mathrm{c}(i)=\mathrm{c}(j) \\ -\mathrm{d}_C/\mathrm{d}_D^2 \;:\; \mathrm{c}(i)\neq\mathrm{c}(j) \end{cases} . \end{aligned} \qquad (5)$$

The right factor in Eqn. 4 is the matrix derivative of Eqn. 1:

$$\frac{\partial \mathrm{d}_{\mathbf{\Omega}}^{ij}}{\partial \mathbf{\Omega}} = 2\cdot(\mathbf{x}^i-\mathbf{x}^j)^{\mathsf{T}}\cdot(\mathbf{x}^i-\mathbf{x}^j)\cdot\mathbf{\Omega} . \qquad (6)$$

In practice, the gradient from Eqn. 4, is computed and reused as long the cost function decreases. Increase of $\mathrm{s}(\mathbf{\Omega})$ triggers a recomputation of the gradient. The step size γ is dynamically determined as the initial size γ_0, being exponentially cooled down by rate η, divided by the maximum absolute element in the matrix $\partial \mathrm{s}(\mathbf{\Omega})/\partial\mathbf{\Omega}$.

For running the iterative optimization, the initial step size γ_0 can be chosen as a value below one, such as 0.01 used here. In general, between 50 and 2500 iterations are necessary, depending on the saturation characteristics of the logged cost function value. It was set to 50 in this study. The exponential cooling rate was set to $\eta = 0.995$. For initialization of matrix $\mathbf{\Omega}$ random matrix element sampling from uniform noise in the interval $[-0.5; 0.5]$ is proposed as first step. This noise matrix $\boldsymbol{A} \in \mathbb{R}^{q\times q}$ is then broken by QR-decomposition into $\boldsymbol{A} = \boldsymbol{Q}\cdot\boldsymbol{R}$, of which the $\boldsymbol{Q}$-part is known to form an orthonormal basis with $\boldsymbol{Q}\cdot\boldsymbol{Q}^{\mathsf{T}} = \boldsymbol{I}$. Thus, although $\mathbf{\Omega} = \boldsymbol{Q}$ contains random configurations, its self-product leads to the intuitive squared Euclidean distance in the beginning of optimization.

3. RESULTS – PROTEOME DATA ANALYSIS

Abiotic stress factors have severe effects on the growth as well as on the yield of crop plants, and proteome analysis of stress responses is widely used for unraveling tolerance mechanisms for crop improvement [1, 10]. Our data has been created in a proteomic study concerning metabolic reactions of two barley cultivars, Steptoe and Morex, to different salt stress conditions, ranging from zero NaCl concentration via 100mM to 150mM. The main task is the identification of protein pairs in root parts affected by salt stress, but with different regulation dynamics between the salt-sensitive Steptoe line and the salt-tolerant Morex line. Using 2D-gels separating along pH and mass gradient, images with protein-specific spot distributions were obtained. After image processing, a number of 997 common spots in all gel images was obtained for further analysis. Since three technical replicates per experimental condition were taken, a total number of 18 images was available for differential analysis of the spot combinations characteristic of the three salt treatments.

Matrix learning has been done independently for the Morex and Steptoe lines. In order to increase the reliability of the results, 100 repetitions with random matrix initializations have been created, leading to a total number of 200 trained 997x997 matrices $\mathbf{\Lambda}_i = \mathbf{\Omega}_i\cdot\mathbf{\Omega}_i^{\mathsf{T}}$. Within each such symmetric matrix the ranks of its lower triangular elements, including the diagonal, were calculated. Especially high and low ranks are linked to protein pairs separating between the three salt stress conditions. Since metabolic differences of Steptoe and Morex regarding salt treatments are looked for, only those pairs with very different ranks between both lines are of interest. Thus, the absolute differences of average ranks of the 100 Steptoe and 100 Morex results were taken as ordering criterion of all protein pairs. For illustration, the top 100 protein pairs are considered in more detail. In that list all standard deviations of ranks are below 12.8, which indicates a high reproducibility of the found protein pairs; for comparison, the expectation of randomly drawn rank differences would be $1/3\cdot 997\cdot(997+1)/2 = 165834.3$.

The connectivity structure of the strongly associated top 100 protein pairs is shown in Fig. 1. Two protein spots, 543 and 94, can be identified as network hubs. These are linked to many other spots of interest. Spots within bold ellipses were identified as candidate proteins in an

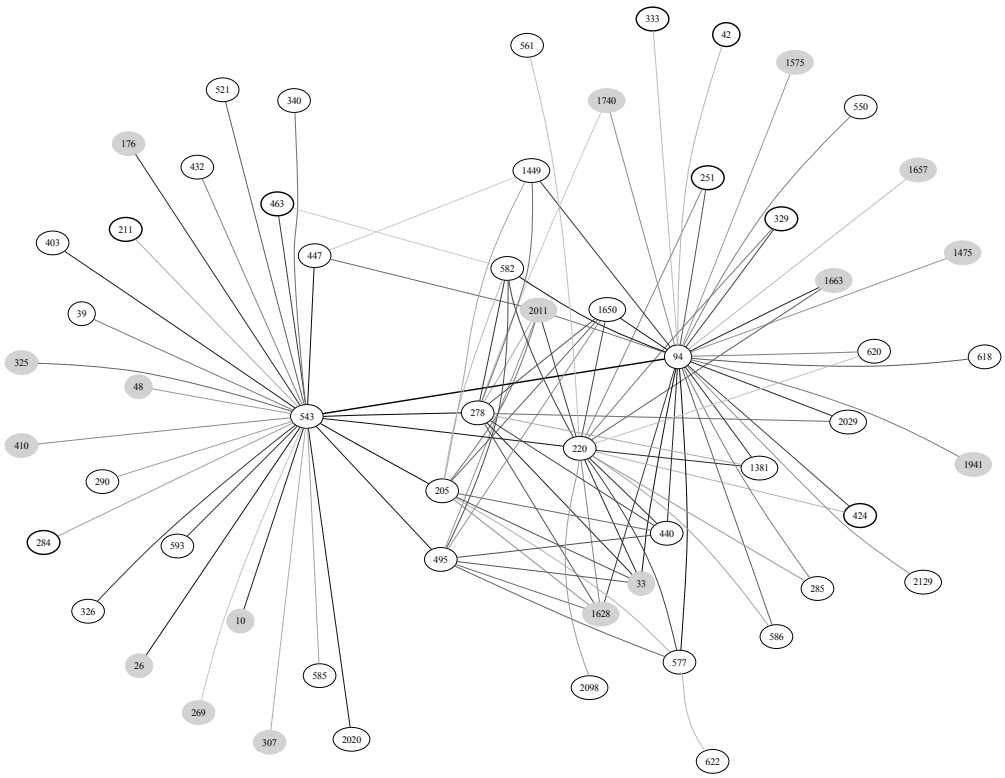

Figure 1. Protein-protein network derived from 2D gels containing patterns of differential protein abundance induced by salt-stress. The top 100 dependent pairs of protein spots, indexed by numbers, are shown. Edge gray levels indicate the ranking, the darker the stronger. The connection from 543 to 94 is the strongest. Bold face ellipses denote spots identified as interesting in previous studies, gray filling indicates spots close to the background intensity.

independent previous study. Plain ellipses are new candidates that have not been detected previously by single spot analysis. It must be stated, though, that also spots close to the background intensity have been found. These, shaded in gray, cannot be considered as biologically relevant. Yet, scale-free inspection indicates that magnitudes alone are not the only consistent class-separation criteria.

Model compression. Since the matrix model of 997x997 is huge in contrast to the (2x9) x 997 experimental protein spots, compression is an important issue. Eigen decomposition of $\mathbf{\Lambda} = \boldsymbol{S} \cdot \mathbf{W} \cdot \mathbf{W}^{-1}$ into the diagonal eigenvalue matrix $\boldsymbol{S}$ and the eigenvector matrix $\mathbf{W}$ helps to reach substantial reduction. For the protein-specific matrices the largest eigenvalues are about 7-fold greater than their predecessors which themselves are twice larger than their predecessors. These first two eigenvectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ therefore define outstanding directions in the scaling matrix $\mathbf{\Lambda}$. This matrix can be approximately reconstructed by $[\boldsymbol{w}_1 \boldsymbol{w}_2] \cdot [\boldsymbol{w}_1 \boldsymbol{w}_2]^\mathsf{T}$. Each experiment $\mathbf{x}$ is projected into a class-separating subspace by $\mathbf{x} \cdot [\boldsymbol{w}_1 \boldsymbol{w}_2]^\mathsf{T}$. This is shown in the right panel of Fig. 2, where the within-class variation of the technical repetitions is virtually completely suppressed in contrast to the scatter plot obtained by ordinary PCA projection, displayed in the left panel of Fig. 2. This result indicates that relevant directions for noise cancellation have been found by matrix learning.

4. CONCLUSIONS

The presented matrix metric learning approach offers a new way to extracting biomarkers, advancing the traditional assessment of individual data attributes to attribute pairs. As illustrated for protein data, dependent treatment-specific substances can be identified. This allows the construction of undirected network structures with weighted edges, a first step towards the inspection of possible protein interactions. Multi-parallel data sources like the considered protein gels create big challenges, because the number of experiments are usually substantially lower than the number of attributes. Therefore, metric adaptation is generally considered as beneficial to counter-act the curse of dimensionality. Confidence in the proposed method is derived from the observation that training showed very stable results despite random initializations of $\mathbf{\Omega}$. However, additional data for the validation of the trained metric are needed, and attention must be put to the role of pairs with low-intensity partners. In order to force further model regularization and for a significant speedup of adaptation, the direct training of only the first k eigenvectors of $\mathbf{\Lambda}$ are currently considered.

Thanks to the anonymous reviewer for the valuable comments. The work is supported by grant XP3624HP/0606T, Ministry of Culture Saxony-Anhalt, Germany.

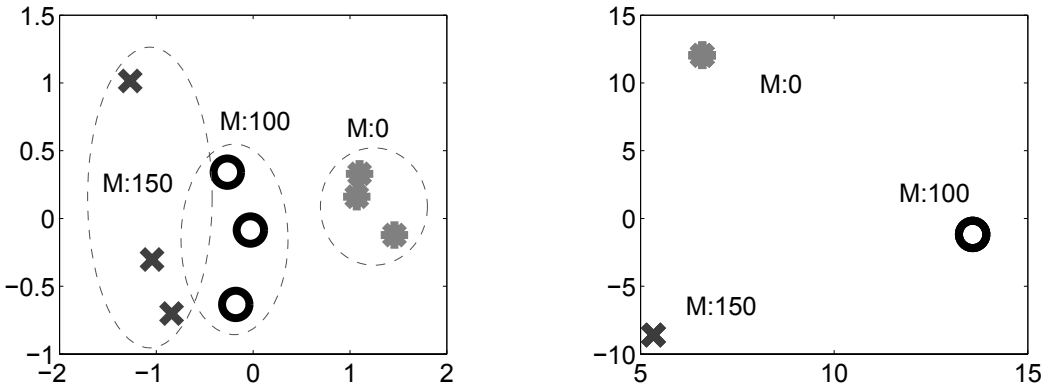

Figure 2. Scatter plots of 2D gels of Morex roots under salt stress. Left: PCA projection of original data to second vs. first eigenvector of data covariance matrix. Right: projection to second vs. first eigenvector of the trained metric matrix $\mathbf{\Omega} \cdot \mathbf{\Omega}^{\mathrm{T}}$. Labels M:0–M:150 denote salt stress concentrations in mM NaCl. The three technical replicates, belonging to specific salt levels, constitute a class.

5. REFERENCES

[1] S. Amme, A. Matros, B. Schlesier, and H.-P. Mock. Proteome analysis of cold stress response in arabidopsis thaliana using dige-technology. *Journal of Experimental Botany*, (57):1537–1546, 2006.

[2] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] D. Johnson. The insignificance of statistical significance testing. *Journal of Wildlife Management*, 63(3):763–772, 1999.

[4] F. Jourdan, R. Breitling, M. P. Barrett, and D. Gilbert. MetaNetter: inference and visualization of high-resolution metabolomic networks. *Bioinformatics*, 24(1):143–145, 2008.

[5] B. Junker and F. Schreiber. *Analysis of Biological Networks*. John Wiley and Sons, 2008.

[6] S. Kaski. From learning metrics towards dependency exploration. In M. Cottrell, editor, *Proceedings of the 5th International Workshop on Self-Organizing Maps (WSOM)*, pages 307–314, 2005.

[7] H. Li and J. Gui. Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, 7(2):302–317, 2006.

[8] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J. Bioinformatics Syst. Biol.*, 2007(1):8–8, 2007.

[9] B.-E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. d'Alche Buc. Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, 19:138–148, 2003.

[10] M. Rossignol, J.-B. Peltier, H.-P. Mock, A. Matros, A. Maldonado, and J. Jorrn. Plant proteome analysis: A 2004-2006 update. *Proteomics*, (6):5529–5548, 2006.

[11] P. Schneider, M. Biehl, and B. Hammer. Relevance Matrices in LVQ. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks (ESANN)*, pages 37–42, Bruges, Belgium, 2007.

[12] M. Strickert, P. Schneider, J. Keilwagen, T. Villmann, M. Biehl, and B. Hammer. Discriminatory data mapping by matrix-based supervised learning metrics. In L. Prevost, S. Marinai, and F. Schwenker, editors, *Lecture Notes in Computer Science, LNCS 5065*, to appear. Springer, 2008.

[13] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1473–1480. MIT Press, Cambridge, MA, 2006.