Proc. IJCNN'01, International Joint Conference on Neural Networks, in press

Learning Metrics for Self-Organizing Maps

Samuel Kaski Janne Sinkkonen Jaakko Peltonen
Helsinki University of Technology
Neural Networks Research Centre
P.O. Box 5400, FIN-02015 HUT, Finland
{samuel.kaski, janne.sinkkonen, jaakko.peltonen}@hut.fi

Abstract

We introduce methods that adapt the metric of the data space to reflect relevance, as indicated by auxiliary data associated with the primary data samples. The derived metric is especially useful in descriptive data analysis by unsupervised methods such as the Self-Organizing Maps. In this work we use the new metric to refine SOM-based analyses of the factors affecting the bankruptcy risk of companies.

1 Introduction

The goal of this work is to develop methods for data-driven search of statistical dependencies in data. The so-called unsupervised learning methods such as clustering, density estimation, and visualization methods are useful for exploring data sets without explicit prior hypotheses. It is hoped that the resulting summaries and descriptions of the properties of the data help make new discoveries in an interactive, iteratively refined process.

The problem with unsupervised learning is that not all statistical properties in the data set are interesting. There is noise, and not even all the "true" dependencies are relevant or interesting to the analyst. In pattern discovery it is well known that many discovered patterns are trivial or not interesting. In clustering the distinctions between clusters may be made over irrelevant features. Things get worse when the sample size and dimensionality increase—indeed, the current challenge for the exploration methods lies in the massive amounts of electronically available data.

By contrast, in supervised methods the well-defined goal, be it the minimization of classification error or prediction error or something else, implicitly determines which aspects of the data are interesting. Many supervised methods are universal approximators: with increasing model complexity, they can asymptotically approximate any function from the inputs to the desired outputs. Hence the methods are invariant to a general class of data transformations and hence robust with respect to the representation of data.

In this work the novel insight is to utilize, in *unsupervised* learning, the knowledge that implicitly exists in signals that are traditionally used in *supervised* learning. In our case study we use the future bankruptcy state of companies as a guiding signal to explore financial statements. Unlike in supervised learning, however, the goal is to make data-driven discoveries from the statistical properties of the data *given the supervision*, i.e., after the supervising signal has been utilized. Technically, the goal is to automatically *learn* metrics which measure distances along important or relevant directions pointed out by the supervisory signal, and to use the metrics in unsupervised learning.

Metrics have been derived from probabilistic models but without the supervisory signal (see e.g. [1, 2]). In addition, numerous methods exist for transforming the data for improved classification or prediction accuracy. To our knowledge the principle and our solution [3] to carry out unsupervised learning in a "supervised" metric is new. Below we demonstrate its use with the Self-Organizing Maps, in the analysis of the bankruptcy risk of companies.

2 The Learning Metric

We aim at finding interesting or relevant features of the *primary data* $\mathbf{x} \in \mathbb{X} \subset \mathbb{R}^n$, samples of a vector-valued random variable X. Samples c, or *auxiliary data*, of an associated random variable are also available, and it is assumed that a change in the conditional distribution $p(c|\mathbf{x})$ signifies an interesting change in $\mathbf{x}$.

We will represent the relevance of a variation in $\mathbf{x}$ as *a distance*. Together such distances constitute a *metric*. A metric has the advantage of being a rather general description of the relationships of $\mathbf{x}$, and even a non-Euclidean metric can be readily incorporated into many unsupervised methods. Below we will first introduce a suitable metric as a mathematical, differential-geometric (see e.g. [4]) construction. The motivation for why precisely that kind of a metric is particularly useful for data analysis, and its relationship to alternative approaches will be presented at the end of the section.

Assume tentatively that the original Euclidean metric of the space $\mathbb{X}$ is arbitrarily chosen and hence irrelevant. The proximity relations of $\mathbf{x}$ or, disregarding possible singularities, the topology are important, however. Then all metrics obtained by local rescalings of the original Euclidean metric retain the important structure of $\mathbb{X}$. Such metrics are of the form

$$d_L^2(\mathbf{x}, \mathbf{x} + d\mathbf{x}) = d\mathbf{x}^T \mathbf{J}(\mathbf{x}) d\mathbf{x} \ , \tag{1}$$

where $\mathbf{J}(\mathbf{x})$ is a positive semidefinite matrix depending on $\mathbf{x}$. Although not necessary for the application in this paper, global distances are defined as minimal path integrals; this gives a Riemannian metric [4]. Note that the usual Euclidean metric can be expressed locally by $d^2(\mathbf{x}, \mathbf{x}+d\mathbf{x}) = d\mathbf{x}^T \mathbf{I} d\mathbf{x}$, where $\mathbf{I}$ is the identity matrix. Thus the two metrics span the same topology if $\mathbf{J}(\mathbf{x})$ is non-singular. (d and d_L would then be called equivalent metrics. In practice $\mathbf{J}$ may be singular; then the change of the metric is projective.)

Let us then return to our original goal: measurement of relevant differences. It was assumed that differences in the distribution $p(c|\mathbf{x})$ signify relevant changes. Such differences can be measured by the Kullback-Leibler divergences, and a proof by Kullback [5] implies that for nearby points $\mathbf{x}$ and $\mathbf{x} + d\mathbf{x}$ the divergence can be computed in the form of (1), assuming the densities $p(c|\mathbf{x})$ are differentiable with respect to $\mathbf{x}$.

This makes it possible to *locally measure* the interestingness (signified by changes in $p(c|\mathbf{x})$) while preserving the proximities in the $\mathbb{X}$ space: we plug in a $\mathbf{J}$ such that the d_L locally agree with the Kullback-Leibler distances computed from the conditional probabilities $p(c|\mathbf{x})$. The right form for $\mathbf{J}(\mathbf{x})$ is

$$\mathbf{J}(\mathbf{x}) = E_{p(c|\mathbf{x})} \left\{ \left(\frac{\partial}{\partial \mathbf{x}} \log p(c|\mathbf{x}) \right) \left(\frac{\partial}{\partial \mathbf{x}} \log p(c|\mathbf{x}) \right)^T \right\} \ , \tag{2}$$

where the operator $E_{p(c|\mathbf{x})}$ denotes expectation over the conditional distribution $p(c|\mathbf{x})$. The $\mathbf{J}$ is essentially a *Fisher information matrix*, and distances obtained from the matrix are called (Fisher) information distances or (Fisher) information metrics in the information geometry literature (see, e.g., [4]). Traditionally the arguments of the Fisher information matrix are parameters of generative probabilistic models, and the metric measures distances in the model space. Our new contribution is to use the $\mathbf{x}$ in the role of the parameters to obtain a metric in the data space. In the generated metric the conditional density $p(c|\mathbf{x})$ changes evenly in all directions and at all points $\mathbf{x}$ of the data space. The metric can also be shown to be invariant to a large class of smooth transformations of the space $\mathbb{X}$, called diffeomorphisms.

Above, the probabilities $p(c|\mathbf{x})$ have been assumed known. The derivation of the metric from the $p(c|\mathbf{x})$ is not affected by the original Euclidean structure over $\mathbb{X}$, but in many applications the $p(c|\mathbf{x})$ are unknown and an estimate $\hat{p}(c|\mathbf{x})$ has to be used instead. The estimators necessarily depend somewhat on the Euclidean structure and hence in practice the metric d_L is only asymptotically invariant to the original metric structure of $\mathbb{X}$.

Note that distances could in principle be measured directly as the Kullback-Leibler divergence $D(p(c|\mathbf{x})||p(c|\mathbf{y}))$ between any two points $\mathbf{x}$ and $\mathbf{y}$. The definition would not yield a metric, however, because the divergences are asymmetrical and the triangle equation does not hold. It would also completely ignore the structure of the $\mathbb{X}$-space, and that is not desirable for two reasons: (1) In data analysis applications we often wish to interpret the findings in terms of the original data variables which is harder if the topology has been changed; (2) When estimating the densities $p(c|\mathbf{x})$ from a finite data set the generalization over the data space $\mathbf{x}$ needs be based on some topology (or metric). Usually it is based on the topology of the $\mathbb{X}$-space, which would be inconsistent with the proximity relationships induced by the direct Kullback-Leibler divergence $D(p(c|\mathbf{x})||p(c|\mathbf{y}))$.

Note that the common feature extraction or data transformation methods change the metric as well. If they are diffeomorphisms then our metric is in principle invariant to them. They may, however, change the topology as well. Our metric can, of course, be applied after the transformation, which is beneficial if the change of the topology is.

Below, we assume that only data pairs $\{(c^k, \mathbf{x}^k)\}_k$ of auxiliary and primary data are available.

3 Self-Organizing Maps in Learning Metrics

Below we describe how learning metrics can be used with the Self-Organizing Map (SOM) [6], a method widely used in data analysis and visualization. A SOM consists of a grid of N_{SOM} units, and a model vector $\mathbf{m}_i$ is associated to each unit i. After the SOM has been computed the model vectors follow the input data in an ordered fashion: model vectors of close-by units on the lattice remain close-by in the input space.

3.1 SOM algorithm

The SOM algorithm iterates two steps: winner selection and adaptation. At each iteration t, the index of the winning unit w closest to the current input sample $\mathbf{x}(t)$ is first sought by

$$w(\mathbf{x}(t)) = \arg\min_i d^2(\mathbf{x}(t), \mathbf{m}_i(t)) \ , \tag{3}$$

where d is a distance function, commonly Euclidean. Then the model vectors are adapted according to

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) - \frac{1}{2} h_{wi}(t) \frac{\partial}{\partial \mathbf{m}_i} d^2(\mathbf{x}(t), \mathbf{m}_i(t)) \ . \tag{4}$$

If d is the Euclidean distance, the adaptation becomes

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{wi}(t)(\mathbf{x}(t) - \mathbf{m}_i(t)) \ . \tag{5}$$

Here $h_{wi}(t)$ is the neighbourhood function, a decreasing function of the distance between units w and i on the map lattice. Both the height and width of $h_{wi}(t)$ are decreased gradually as the iteration progresses [6].

In learning metrics, both steps must be modified. The winner is selected by the new distances i.e. d in (3) becomes replaced by d_L which is in principle defined as the minimal path integral. In practice we simplify the computation by relying on the local approximation; the winner selection then becomes

$$w(\mathbf{x}(t)) = \arg\min_i (\mathbf{m}_i(t) - \mathbf{x}(t))^T \mathbf{J}(\mathbf{x}(t))(\mathbf{m}_i(t) - \mathbf{x}(t)) \ , \tag{6}$$

where (1) has been used to approximate the metric locally around $\mathbf{x}(t)$. The approximation is likely to be accurate for model vectors close to $\mathbf{x}$ (by their true distances d_L). We assume it to be accurate enough to keep the selection of the winning unit correct most of the time, but experimental results are of course needed for final conclusion. The case study in Section 4 is favourable.

In the adaptation step, each model vector $\mathbf{m}_i(t)$ is updated in the direction where the distance to $\mathbf{x}(t)$ decreases most rapidly. In the Euclidean metric the direction is that of the negative gradient, $2(\mathbf{m}_i(t) - \mathbf{x}(t))$, but in the more general Riemannian metric the direction is oposite to the natural gradient [7]

$$\mathbf{J}(\mathbf{x})^{-1} \frac{\partial}{\partial \mathbf{m}_i(t)} d_L^2(\mathbf{x}, \mathbf{m}_i(t)) \ . \tag{7}$$

Using again the local approximation this becomes

$$\mathbf{J}(\mathbf{x})^{-1} 2\mathbf{J}(\mathbf{x})(\mathbf{m}_i(t) - \mathbf{x}) = 2(\mathbf{m}_i(t) - \mathbf{x}) \ , \tag{8}$$

which is just the gradient of the Euclidean metric. Thus the model vectors will be adapted with the familiar rule (5).

3.2 Probability Estimation

The winner selection (6) depends on $\mathbf{J}(\mathbf{x})$, i.e. on the auxiliary probabilities $p(c|\mathbf{x})$ and their gradients. In practice these must be estimated from the data. In this paper we use two classical estimates: Parzen kernels and a version of Gaussian mixtures called Mixture Discriminant Analysis 2 (MDA2) [8, 9]. Both estimate the joint probabilities $p(c, \mathbf{x})$ with a similar parametric form, from which $p(c|\mathbf{x})$ are obtained by the Bayes rule.

Consider a generative mixture where the pair $(\mathbf{x}, c)$ is generated by a component chosen by the probabilities $\{\pi_j\}_j$, $j = 1, \ldots, N_U$. The jth component generates $\mathbf{x}$ from a density $b_j(\mathbf{x}; \boldsymbol{\theta}_j)$, parametrized by $\boldsymbol{\theta}_j$, and c from a multinomial distribution $\{\xi_{ji}\}_i$, $i = 1, \ldots, N_C$. The conditional estimates $\hat{p}(c|\mathbf{x})$ are then obtained from the Bayes rule:

$$\hat{p}(c_i|\mathbf{x}) = \frac{\sum_j \pi_j \xi_{ji} b_j(\mathbf{x}; \boldsymbol{\theta}_j)}{\sum_j \pi_j b_j(\mathbf{x}; \boldsymbol{\theta}_j)} \ . \tag{9}$$

Here b_j are chosen to be Gaussian with a common covariance $\sigma^2 \mathbf{I}$. For MDA2, the parameters π_j, $\boldsymbol{\theta}_j$ (locations of the Gaussians), and ξ_{ji} are estimated with the EM algorithm. In the Parzen kernel estimate, there is a kernel b_j for each data sample $\mathbf{x}_j$. Thus $\boldsymbol{\theta}_j = \mathbf{x}_j$, $\pi_j = 1/N_U$, and $\xi_{ji} = 1$ for the value of i corresponding to the data sample c_j, and zero otherwise. It can be shown that for these models, the local distances (1) become

$$\sigma^4 d^2(\mathbf{x}, \mathbf{x} + d\mathbf{x}) = E_{\hat{p}(c|\mathbf{x})} \left\{ \left[d\mathbf{x}^T \left(E_{p(u_j|\mathbf{x}, c_i; \boldsymbol{\theta}_j)} \{\boldsymbol{\theta}_j\} - E_{p(u_j|\mathbf{x}; \boldsymbol{\theta}_j)} \{\boldsymbol{\theta}_j\} \right) \right]^2 \right\} \tag{10}$$

where $p(u_j|\mathbf{x}; \boldsymbol{\theta}_j)$ is the probability that the jth mixture component generated the input sample given $\mathbf{x}$ and the parameters, and $p(u_j|\mathbf{x}, c_i; \boldsymbol{\theta}_j)$ is the probability with the auxiliary value c_i also given.

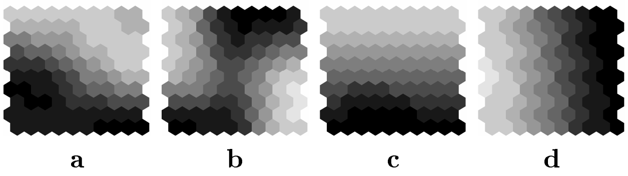

Figure 1: Posterior probabilities of classes c_0 and c_2 on two SOMs representing the same data, in the Euclidean (**a-b**) and the learning metric (**c-d**).

3.3 Summary

To summarize, the computation of a Self-Organizing Map in the learning metric consists of the following steps:

1. Build estimates of the probabilities $p(c|\mathbf{x})$. Use for example the Parzen or MDA2 estimates discussed above.

2. Train the SOM by iterating the following steps: (a) Select a sample $\mathbf{x}(t)$ and winner unit $w(\mathbf{x}(t))$ by (6). For Parzen and MDA2 models, use (10) to compute distances. (b) Adapt model vectors toward $\mathbf{x}(t)$ by (5).

If (10) is used for distance calculation, the computational complexity of the winner selection step becomes $\mathcal{O}(N_{DIM}N_C(N_U + N_{SOM}))$, where N_{DIM} is the dimensionality of $\mathbb{X}$, whereas for Euclidean distances it is $\mathcal{O}(N_{DIM}N_{SOM})$. The complexity of the adaptation step is of course unchanged.

3.4 A Demonstration

Let us demonstrate the change of metric by computing SOMs for a toy data set both in the Euclidean metric (SOM-E) and in the learning metric (SOM-L).

The primary data is evenly distributed within the unit cube. The auxiliary data takes four values, from c_0 to c_3. Their conditional distributions depend linearly on the horizontal dimensions, but are indepedent of the vertical dimension. That is, only the horizontal plane is considered important—representing the vertical dimension just wastes resources.

We trained the SOM-E and SOM-L to this data and visualized the conditional distribution of the auxiliary discrete variable on the SOMs (Fig. 1). The distributions are smoother on the SOM-L, and the true unimodality of the distributions is particularly well visible. By contrast, the SOM-E displays give a false impression of bimodality for c_2. The reason for the difference becomes apparent once we take a look at the distribution of the model vectors inside the cube (Fig. 2).

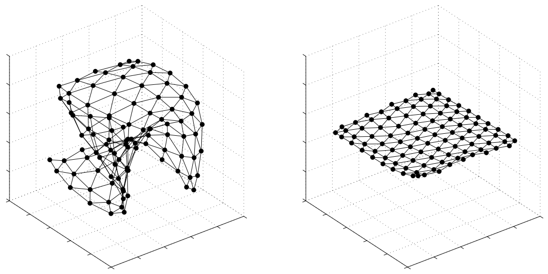

Figure 2: Projections of the model vectors of the two SOMs in Fig. 1. Left: SOM-E, right: SOM-L.

The SOM-E folds to represent the entire data distribution, while the SOM-L represents only the relevant dimensions, as desired.

4 Case Study: Bankruptcy Analysis

In this section we apply the SOM and learning metrics to bankruptcy analysis. Bankruptcies are widely studied, for they have a tangible impact on business life. Most of the quantitative studies have aimed at prediction, the main approaches being classification and probability estimation on the basis of the financial statements given by the companies. A complementary approach is the analysis of the effects of corporate behaviour on the bankruptcy risk. A qualitative work was done by Argenti [10], and recently the SOM has been applied to this problem by Kiviluoto and Bergius [11]. Our research complements their work by using the company status (bankrupt or not) to build learning metrics for SOM training. Since the metric describes changes in bankruptcy risk, the SOM should emphasize the most interesting features of the financial statements, i.e. those that contribute locally to the bankruptcies.

4.1 Data

The data consisted of financial statements from about 1500 Finnish companies. Multiple statements from different years were treated as independent samples; of all the 6195 statements 158 concerned companies which later collapsed. 23 financial indicators were extracted from the statements, including measures of growth, profitability and liquidity. The auxiliary variable indicated whether the company went bankrupt within 3 years of the statement.

The data were randomly divided into an estimation set and a test set of roughly equal sizes. For the estimation set, the Parzen estimate and the Gaussian mixture model ($N_U = 10$) of the previous section were used to compute hexagonal SOMs of 20×10 units in the learning metrics. A Euclidean SOM of similar size was computed for reference.

4.2 Goodness measures

At least the following factors contribute to the goodness of a learning metric SOM as a description of the bankruptcy data:

(1) The quality of the probability estimate. In this paper we will not measure this, and just resort to standard estimators.

(2) The accuracy of the SOMs in representing the bankruptcy risk. The SOM units can be regarded as local probability estimators of their Voronoi regions by assigning to them the estimated probabilities of the auxiliary data at the point of the model vectors. The SOM as an estimator can then be evaluated by the conditional log-likelihood of the test data, estimated at the winner unit locations:

$$\sum_k \log \hat{p}(c^k | \mathbf{m}_{w(\mathbf{x}^k)}) \ . \tag{11}$$

(3) Visualization quality, i.e., smoothness and organization. Here we will resort to visual comparisons.

4.3 Results

The test-set likelihoods for the probability estimators and the SOMs were computed over a wide range of the parameter σ that governs the smoothness of the estimates. The likelihoods of the probability estimators approximately indicate the best possible SOM performance, and a "model" always predicting prior bankruptcy probabilities served as a lower limit of useful results.

The accuracies of SOMs in describing the test set are shown in Fig. 3. As expected, the SOM-L performs better than the SOM-E; the results are roughly equal only for the Parzen estimator with very small σ; then the estimates of the conditional probabilities are presumably very uneven, resulting in an uneven metric. The accuracy difference between SOM-E and SOM-L was statistically significant ($p < 0.002$; sign test for the peaks of the accuracy curves with 10-fold cross-validation).

Visually, the SOM-L displays were comparable or better than the SOM-E displays. Sample visualizations of the data made by the SOM-L are shown in Fig. 4. There is a novel kind of a display included, depicting the relevance of a data variable at different locations

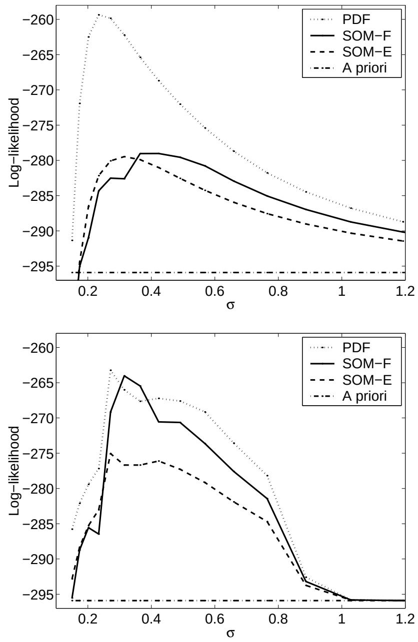

Figure 3: The accuracy (11) of SOM-Es and SOM-Ls in representing bankruptcy risk. The $p(c|\mathbf{x})$ are estimated by Gaussian kernels (top) and Gaussian mixtures with $N_U = 10$ (bottom).

of the map. The relevance $r_l(\mathbf{x})$ of the variable l at $\mathbf{x}$ is computed as the contribution of the variable to the distance,

$$r_l(\mathbf{x}) = \sqrt{\frac{\mathbf{e}_l^T \mathbf{J}(\mathbf{x}) \mathbf{e}_l}{\sum_m \mathbf{e}_m^T \mathbf{J}(\mathbf{x}) \mathbf{e}_m}} \ , \tag{12}$$

where $\mathbf{e}_l$ is the unit vector parallel to the axis corresponding to the lth input variable.

5 Discussion

In this paper, a novel approach to data analysis is described. A relevance-indicating signal is used to guide distance-based unsupervised learning methods to concentrate on relevant properties of data. Put in another way, we have introduced a way to carry out "semisupervised" exploratory data analysis.

The metric of the data space is modified to measure relevant changes in the data. As a case study, we use the resulting non-Euclidean metric in the Self-Organizing Map algorithm. The modified SOM method was applied to financial statements of enterprises, and the indicator of whether the company went bankrupt or not guided the analysis. The resulting SOM then describes

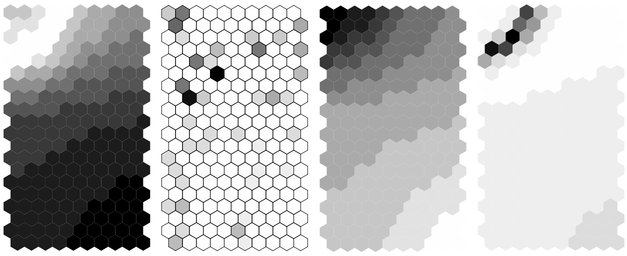

Figure 4: SOM-L displays of the bankruptcy data. **a** Posterior bankruptcy probability, **b** empirical ratio of healthy to bankrupt companies, **c** distribution and **d** relevance of a profitability indicator. The hexagons correspond to SOM units and light shades denote high values.

only such variation of the financial statements that correlates with the bankruptcy sensitivity of the companies. The results were satisfactory both qualitatively and quantitatively in that the companies that later went bankrupt became better separated on the learning metric SOM, and the factors affecting the bankruptcy were well presented.

The goodness of the method depends on the estimator used to approximate the relevance-indicating signal or auxiliary data. We will later investigate the proper choice of the estimator in more detail.

The method described in this paper is a product of a larger research project where the aim is to develop learning metric methods for exploratory data analysis. So far we have developed a related clustering method and applied it to gene expression data [12] and text documents [13]. More details of the bankruptcy prediction application described in this paper are available in [3].

Acknowledgments

The authors would like to thank Finnvera Ltd. and particularly Pentti Bergius for the data set, Kimmo Kiviluoto for his help regarding its interpretation, and the Academy of Finland for financial support.

References

[1] T. S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Advances in Neural Information Processing Systems 11*, Michael S. Kearns, Sara A. Solla, and David A. Cohn, Eds., pp. 487–493. Morgan Kauffmann Publishers, San Mateo, CA, 1999.

[2] M. E. Tipping, "Deriving cluster analytic distance functions from Gaussian mixture models," in *Proceedings of ICANN99, Ninth International Conference on Artificial Neural Networks*, pp. 815–820. IEE, London, 1999.

[3] S. Kaski, J. Sinkkonen, and J. Peltonen, "Bankruptcy analysis with self-organizing maps in learning metrics," *IEEE Transactions on Neural Networks*, 2001, Accepted for publication.

[4] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry*, Translations of Mathematical Monographs 191, American Mathematical Society and Oxford University Press, Providence, Rhode Island, 2000.

[5] S. Kullback, *Information Theory and Statistics*, Wiley, New York, 1959.

[6] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, 1995, (Third, extended edition 2001).

[7] S.-I. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, pp. 251–276, 1998.

[8] T. Hastie, R. Tibshirani, and A. Buja, "Flexible discriminant and mixture models," in *Neural Networks and Statistics*, J. Kay and D. Titterington, Eds. Oxford University Press, 1995.

[9] T. Hastie and R. Tibshirani, "Discriminant analysis by Gaussian mixtures," *Journal of the Royal Statistical Society Series B: Methodological*, vol. 58, pp. 155–176, 1996.

[10] J. Argenti, *Corporate collapse – the causes and symptoms*, McGraw-Hill, London, 1976.

[11] K. Kiviluoto and P. Bergius, "Exploring corporate bankruptcy with two-level self-organizing maps. Decision technologies for computational management science," in *Proceedings of Fifth International Conference on Computational Finance*, pp. 373–380. Kluwer Academic Publishers, Boston, 1998.

[12] J. Sinkkonen and S. Kaski, "Clustering based on conditional distributions in an auxiliary space," *Neural Computation*, 2001, Accepted for publication.

[13] J. Sinkkonen and S. Kaski, "Clustering by similarity in an auxiliary space," in *Proceedings of IDEAL 2000, Second International Conference on Intelligent Data Engineering and Automated Learning*, Kwong Sak Leung, Lai-Wan Chan, and Helen Meng, Eds., pp. 3–8. Springer, Berlin, 2000.