Bankruptcy Analysis with Self-Organizing Maps in Learning Metrics

Samuel Kaski, *Member, IEEE*, Janne Sinkkonen, and Jaakko Peltonen

The authors are with the Neural Networks Research Centre, Helsinki University of Technology, P.O. Box 5400, FIN-02015 HUT, Finland. E-mail: samuel.kaski@hut.fi, janne.sinkkonen@hut.fi, jaakko.peltonen@hut.fi.

Abstract

We introduce a method for deriving a metric, locally based on the Fisher information matrix, into the data space. A Self-Organizing Map is computed in the new metric to explore financial statements of enterprises. The metric measures local distances in terms of changes in the distribution of an auxiliary random variable that reflects what is important in the data. In this paper the variable indicates bankruptcy within the next few years. The conditional density of the auxiliary variable is first estimated, and the change in the estimate resulting from local displacements in the primary data space is measured using the Fisher information matrix. When a Self-Organizing Map is computed in the new metric it still visualizes the data space in a topology-preserving fashion, but represents the (local) directions in which the probability of bankruptcy changes the most.

Keywords

Bankruptcy Analysis, Fisher Information Matrix, Information Metric, Learning Metric, Self-Organizing Map

I. Introduction

Bankruptcies have such a great importance on the financing models and business life in general, that their analysis has become almost its own field of science. It has been widely studied in economics, and most data analysis methods have been suggested to the problem. A traditional benchmark for these methods has been the bankruptcy prediction problem, but we argue that at least as important from the practical point of view is to develop methods for analyzing and understanding the different corporate behavior types and their relation to bankruptcy. In this task, the Self-Organizing Map (SOM) [1], [2] has been found a valuable tool, mainly because of its good visualization capabilities. The present paper introduces a further development of SOM-based data analysis. Our results show that it yields maps with enhanced visualization of bankruptcy risk, and also have a statistically better separation of bankruptcies from healthy companies. The methodology can also be directly utilized in other application areas.

The success of unsupervised algorithms, such as the Self-Organizing Map and clustering methods, depends crucially on the metric, the measure of the distance between the objects of interest. The metric, on the other hand, depends on which kinds of variables have been chosen to represent the objects, i.e., on variable selection and feature extraction. These processing steps affect even supervised methods although many supervised methods are in principle, given unlimited resources, universal approximators. The old problem of feature extraction or variable selection, that is, choosing how to represent the input data, persists as a crucial unsolved research topic in pattern recognition, neural computation, and data analysis.

At its simplest, feature extraction reduces to choosing and scaling the input variables, but more generally it is a nonlinear mapping of the input space to a space that is more suitable for further processing. Successful feature extraction stages are usually tailored for the task at hand using expert knowledge or heuristic rules of thumb. There is often, however, some implicit auxiliary information available about the relevance of the features of the input. For instance, in a classification task the relevant features are those that separate between the classes.

Implicit information about the relevance of the features may also be available for unsupervised descriptive data analysis tasks. A relevant classification of the samples may be known and the goal may be to find a natural grouping for them; a grouping that reflects the classification but may, for example, discover subclasses. Another example is process monitoring in which some indicator of the performance of the process may be associated with each data vector. The quality of the end product could be a suitable indicator. The goal would then be to find out factors affecting the performance of the process.

Our ultimate aim is to develop algorithms that take such auxiliary information into account in order to explicitly transform the original metric of the input space. The space is locally scaled so that the new (local) distances will measure the change of the auxiliary information (for a preliminary account see [3]). Proximity relations or, loosely speaking[1], topology of the input space is still retained. Note that by contrast, a change of the metric that does not preserve the proximity relations would map some close-by points of the input space to very different feature values, and the generalization power originating from the smoothness of the model would be lost.

For computational reasons the new metric is best suited for algorithms that rely mostly on local distances of the input space. The Self-Organizing Map is one example. When an unsupervised algorithm learns using the new metric, the learning process is a useful combination of supervised and unsupervised learning. The proximity relationships of the input space are preserved as is typical of unsupervised methods, while the metric (local scaling of the space) is induced in a supervised manner.

We will apply the new metric to analyze the bankruptcy risk of enterprises on the basis of financial statements. The setting is similar to that of Kiviluoto and Bergius [4], [5], [6]. They

[1]Even though the mapping is continuous it is not topology preserving since it may be projective.

have used Self-Organizing Maps to extend bankruptcy analysis from traditional straightforward prediction of bankruptcy to visual exploratory analyses of the relationship between the financial statements and bankruptcy risk of different kinds of enterprises. We complement their studies by using the new metric in the SOM-based exploratory analyses. The enterprises are organized on a SOM in such a manner that the analysis will concentrate on the (local) factors that affect the probability of bankruptcy most. We will then explore the results to find out the important dimensions for various kinds of enterprises.

II. The metric

We wish to transform the distance measure of the data space so that it will concentrate on the important differences between data samples and disregard irrelevant dimensions. It may be clear that such a metric is impossible to construct without some auxiliary, prior knowledge about the importance of the differences. In this work we assume that there is *auxiliary data* available (more details below), and that the auxiliary data implicitly defines what is important or relevant.

The new metric is to be *learned* based on a data set, and used as a distance measure for subsequent analysis and visualization of the set. The metric is constructed such that it reflects the *local* importance of different directions in the data space. For example, it could measure how much changes in the financial state of a company affect the bankruptcy risk of that kind of enterprise. Due to its locality, the distance measure is capable of revealing different factors for different kinds of enterprises. Moreover, since the distance measure is defined in the original data space, it is straightforward to interpret the results in terms of the original variables, here the indicators of financial states. For example, if the distances for a company type are large along the axis corresponding to the profitability, then profitability contributes to the bankruptcy risk of such companies.

In exploratory data analysis applications the similarity relationships between the data samples, the enterprises, can be visualized with methods such as the Self-Organizing Map, precisely in the same way as previously. The only difference is that the relative distances of the enterprises will change. If they are different along an important dimension (actually the non-linear route of minimal length) their distance is large, whereas if they are different only along an irrelevant dimension they will become very close to each other[2].

[2] In practical computations we will use local approximations to the nonlinear routes, which is sensible for algorithms such as the SOM that depend mostly on local distances.

A. Learning metric: The principle

We seek to describe the similarity relationships of items $\mathbf{x}$ of the data space by utilizing the information within the joint distribution of the data and *auxiliary data* c. Denote the joint probability density function (PDF) by $p(\mathbf{x}, c)$. We will call $\mathbf{x} \in \mathbb{X} \subset \mathbb{R}^n$ the *primary data*, and denote the associated random variable by X.

Denote the random variable that produces the auxiliary data by C. It is assumed that the c or, more specifically, the conditional distributions $p(c|\mathbf{x})$ implicitly convey information about which kinds of similarity relationships are important in the data. In our present application to bankruptcy analysis, the c are binary and indicate whether an enterprise goes bankrupt within the next three years, and the $\mathbf{x}$ are feature vectors derived from the financial statements. The important changes in the financial state $\mathbf{x}$ are then those that change the probability of bankruptcy, the distribution $p(c|\mathbf{x})$.

A change in distributions can be measured by the Kullback-Leibler divergence D. An old result [7] gives a formula for the *local* Kullback-Leibler divergence as

$$D(p(c|\mathbf{x})\|p(c|\mathbf{x}+d\mathbf{x})) = d\mathbf{x}^T \mathbf{J}(\mathbf{x}) d\mathbf{x} \ , \tag{1}$$

where

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

is the Fisher information matrix and $E_{p(c|\mathbf{x})}$ denotes expectation over the possible values of C, conditioned on $\mathbf{x}$. Here the Fisher information matrix $\mathbf{J}(\mathbf{x})$ is the representation of the tensor of the new metric in the original Euclidean coordinates in which $\mathbf{x}$ is also presented. The matrix is positive semidefinite, and it defines the local scaling of the directions of the input space at the point $\mathbf{x}$. We then define the new local metric of the data space as

$$d_F^2(\mathbf{x}, \mathbf{x}+d\mathbf{x}) \equiv D(p(c|\mathbf{x})\|p(c|\mathbf{x}+d\mathbf{x})) = d\mathbf{x}^T \mathbf{J}(\mathbf{x}) d\mathbf{x} \ . \tag{3}$$

In the new metric the conditional density $p(c|\mathbf{x})$ changes evenly in all directions, at all points of the input space.

Note 1: The Fisher information matrix was originally derived for measuring the effect that a change in the model *parameters* produces on the probability distributions that the models generate [8]. The resulting distance is called (Fisher) information distance or (Fisher) information metric in the information geometry literature (see, e.g., [9], [10], [11]). Here we measure the effect of a change in the *location* in the primary data space to obtain a metric there. We will call the resulting metric the *Fisher metric* and call the approach *semisupervised* since the auxiliary distribution in a way supervises the construction of the metric.

Note 2: The new metric (3) is defined locally, for close-by points $\mathbf{x}$ and $\mathbf{x}+d\mathbf{x}$, and global distances are defined by path integrals. In principle there exists another more straightforward alternative: to simply measure the distance between *any non-local pair* $\mathbf{x}$ and $\mathbf{x}'$ by $D(p(c|\mathbf{x})\|p(c|\mathbf{x}'))$. Such a measure might be useful for some applications but its disadvantage is that it would completely override the original structure of the data space. Two points with identical density estimates, $p(c|\mathbf{x})$ and $p(c|\mathbf{x}')$, would have a zero distance even if the points were originally far away. If some kind of generalizability exists over $\mathbb{X}$, it would be destroyed by the change of the topology. In fact, the original representations $\mathbf{x}$ would not be needed at all, and the data points could be simply represented by distributions in the C-space. All information contained in the primary data would then be lost.

An additional disadvantage would be that the new representations cannot be interpreted in terms of the original data variables, at least not without further analyses. In the bankruptcy application it is of prime importance to know which aspects of the financial state of a company are related to changes in its bankruptcy risk and our method focuses on this problem.

B. Learning metric: Computation

The conditional probability distribution $p(c|\mathbf{x})$ is usually estimated from a data set $\{\mathbf{x}^k\, c^k\}_k$, $k=1,\ldots,N$. Any method that produces differentiable estimates is potentially useful. The choice of the estimator is discussed in Section II-C; for the moment assume tentatively that we have an estimate $\hat{p}(c|\mathbf{x})$ of the conditional density available.

The estimate $\hat{p}(c|\mathbf{x})$ could in principle be used in place of $p(c|\mathbf{x})$ in (2-3) to approximate the new metric. However, for numerical computations it is not necessary to form the Fisher information matrix explicitly, for one can get the squared local distances directly from

$$d_F^2(\mathbf{x},\mathbf{x}+d\mathbf{x}) = E_{\hat{p}(c|\mathbf{x})}\left\{\left(d\mathbf{x}^T\frac{\partial}{\partial\mathbf{x}}\log\hat{p}(c|\mathbf{x})\right)^2\right\} . \tag{4}$$

The new metric can be used in any supervised or unsupervised method; in Section III we will describe how to use it to compute Self-Organizing Maps for visualization and exploratory data analysis purposes.

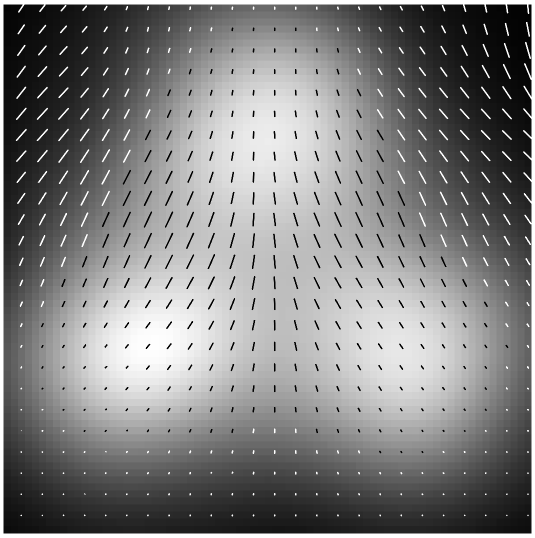


Fig. 1. The metric generated by a PDF estimate for a two-dimensional two-class data set (N=1000). The first class is sampled from a symmetrical Gaussian with $p(c) = \frac{1}{3}$ (the topmost cluster in the figure), the second from a sum of two Gaussians (the bottom clusters). For all the Gaussians, σ=0.6, and the mutual distances of the centers are equal to unity. The gray-scale background illustrates the marginal density $p(\mathbf{x})$, and the small line segments (or dots) depict the dominant direction and relative distances d^2 in the local metric. Distances are nonzero only in the directions where the conditional density changes. The PDF was estimated with a Gaussian Parzen estimator ($\sigma = 0.4$).

If only partially labelled data are available, it is best to use a PDF estimator with the ability to utilize such data. The Gaussian mixture model MDA2, described in Section II-C, can be easily extended for partially labelled data as it is usually optimized by the EM algorithm. After the PDF estimator is fixed, using partially labelled data with the new metric is straightforward, for the metric is defined into the primary data space and therefore can be computed with the knowledge about the primary sample $\mathbf{x}$ only.

A demonstration of the new metric for an artificial, easily visualizable two-dimensional, two-class data set is presented in Figure 1.

Note 1: Non-local distances can be defined as the minimal path integrals of the local distances, minimum taken over all possible paths. This generates a Riemannian metric (for general treatments of the related information geometry, see [9], [10], [11]). In practice the computation of the integrals would be extremely tedious and we will below resort to local approximations which are sensible for methods that rely mostly on local distances (see Section III).

Note 2: If the estimate $\hat{p}(c|\mathbf{x})$ is very uneven or the Fisher metric spans an unpreferably low-dimensional space, the metric can be "regularized" by mixing it with the original Euclidean metric of $\mathbb{X}$, resulting in the metric tensor represented by

$$\mathbf{J}'(\mathbf{x}) = (1-\lambda)\mathbf{J}(\mathbf{x}) + \lambda\mathbf{I} \ , \tag{5}$$

where λ is a small positive constant $(0 < \lambda < 1)$ and $\mathbf{I}$ is the identity matrix.

C. Metrics from two kinds of PDF estimates

Our goal here is to estimate the probability density $p(c|\mathbf{x})$ of the auxiliary random variable C, conditioned on X. Plenty of alternative methods are available. Many of them have been developed for classification purposes (for reviews see e.g. [12], [13]). Most such methods would typically be suboptimal for our purpose, however, because a good classifier optimizes the (sometimes implicit) probability density function (PDF) estimate near the class borders or, more generally, near the area where the decision criterion reaches critical values.

In principle, any estimator which produces differentiable estimates of the conditional densities could be used. In this paper we skip the discussion about the merits of different estimators and rely on two classical methods. The first is a computationally intensive but well-performing non-parametric estimate, the (Parzen) kernel estimator, and the second is a Gaussian mixture model. Both estimators can be expressed within the same general mixture density form.

Let us consider an additive mixture model in which the generating component densities are identified with the discrete random variable U. The value of U is u_j if the jth component generator has generated the current data sample. We assume that c and $\mathbf{x}$ are conditionally independent given the value of U. Then the joint density generated by the jth component is

$$p(c_i, u_j, \mathbf{x}) = p(c_i|u_j)p(\mathbf{x}|u_j)p(u_j) \ .$$

We will model $p(c_i|u_j)$ by a coefficient ξ_{ji}, $p(u_j)$ by π_j, and $p(\mathbf{x}|u_j; \boldsymbol{\theta}_j)$ by a function $b_j(\mathbf{x}; \boldsymbol{\theta}_j)$ parameterized by $\boldsymbol{\theta}_j$. In this notation the model for the joint density of the data is

$$\hat{p}(c_i, \mathbf{x}; \boldsymbol{\Theta}) = \sum_j \pi_j \xi_{ji} b_j(\mathbf{x}; \boldsymbol{\theta}_j) \ , \tag{6}$$

where $\boldsymbol{\Theta}$ has been used to denote the whole set of parameters of the model.

By applying the Bayes rule, an estimate of the conditional density is obtained:

$$\hat{p}(c_i|\mathbf{x}; \boldsymbol{\Theta}) = \frac{\sum_j \pi_j \xi_{ji} b_j(\mathbf{x}; \boldsymbol{\theta}_j)}{\sum_j \pi_j b_j(\mathbf{x}; \boldsymbol{\theta}_j)} \ . \tag{7}$$

The kernel estimator and the mixture density model differ in their parameterizations. Estimation of the parameters in these special cases will be discussed in more detail below. For the moment, assume that the values of all the parameters of the conditional density estimate (7) are known.

It is shown in Appendix A that if the component densities $b_j(\mathbf{x};\boldsymbol{\theta}_j)$ are Gaussians with equal diagonal covariance matrices $\sigma^2\mathbf{I}$ and means $\boldsymbol{\theta}_j$ then the distance in (4) becomes

$$\sigma^4 d^2(\mathbf{x},\mathbf{x}+d\mathbf{x}) = E_{\hat{p}(c|\mathbf{x})}\left\{\left[d\mathbf{x}^T\left(E_{p(u_j|\mathbf{x},c_i;\boldsymbol{\theta}_j)}\{\boldsymbol{\theta}_j\} - E_{p(u_j|\mathbf{x};\boldsymbol{\theta}_j)}\{\boldsymbol{\theta}_j\}\right)\right]^2\right\} . \quad (8)$$

The parameter σ governs the width of the Gaussians and therefore the smoothness of the resulting PDF estimates. A method for choosing the value of σ when the new metric is used for learning Self-Organizing Maps will be described later in Section IV-B.1: A suitable likelihood measure is proposed, and the value of the sigma can be selected to maximize the measure in the learning or validation set.

C.1 Kernel estimation

In kernel density estimation the component densities $b_j(\mathbf{x};\boldsymbol{\theta}_j)$ are called kernels; the number of kernels is equal to the number of data points n, and the parameters $\boldsymbol{\theta}_j$ are set to the data samples, $\boldsymbol{\theta}_j = \mathbf{x}_j$. The prior probabilities are set to $\pi_j = 1/N$. The parameter $\xi_{ji} = 1$ if in the jth data pair $(\mathbf{x}_j, c'_j)$ the value of C is $c'_j = c_i$. Otherwise $\xi_{ji} = 0$. The only free parameter left to be estimated is the variance σ^2 of the kernels.

C.2 Gaussian mixture

When the component densities in the model (7) are chosen to be Gaussians parameterized by their means, the model is equivalent to the Mixture Discriminant Analysis 2 in [14] (cf. also [15]; the relation between mixture discriminant analysis and our work will be discussed in more detail in Section II-D). Now π_j, ξ_{ji}, and $\boldsymbol{\theta}_j$ will all be estimated from the data. Formulas for estimating the model with the EM algorithm [16] are presented in Appendix B.

D. Related works

According to our knowledge the introduced principle is new. Works in which some aspects resemble our approach exist, however. Amari and Wu [17] have augmented support vector machines by making an isotropic change to the metric near the class border. In contrast to this, our metric is non-isotropic and changes the metric everywhere. Jaakkola and Haussler [18] induced a distance measure into a discrete input space using a generative probability model. The crucial differences are that they do not use external information, and that they do not constrain the metric to preserve the topology.

In some earlier works auxiliary information has been incorporated directly into the representations of the data (see, e.g., [2], [19]; note, however, that the goal in these works is different from ours). The auxiliary information can be encoded for example in the 1-out-of-C manner and concatenated to the data vectors $\mathbf{x}$. The main problem of this approach, for our purposes, is the arbitrary relative scale of the primary and auxiliary data. If the relative scale of the auxiliary data is too small, the primary data will dominate in the distance measure, whereas our goal is to measure changes in the auxiliary data and represent these changes as the distance measure of the primary data space. If the relative scale of the (discrete) auxiliary data is too large, on the other hand, then the data vectors will effectively be divided into separate clusters, each corresponding to one possible value of the auxiliary variable. The proximity relations (topology) of the original data space will then be destroyed.

Mappings from the original space to a new lower- or equal-dimensional space, which is the general definition of feature extraction, have a relation to our method. Automatic methods for optimizing such mappings, for example by maximizing mutual information have been proposed [20], [21]. Unlike in a standard separate feature extraction stage, however, the change of the metric in our method defines a manifold which cannot in general be projected to a Euclidean space of the same or lower dimensionality. Therefore, no dimensionality-preserving or dimensionality-reducing mapping with the same local properties exists which means that the change of the metric is a more general operation than feature selection by a dimensionality-preserving (or dimensionality-reducing) nonlinear mapping.

The change of the metric can additionally be interpreted as a kind of nonlinear version of linear discriminant analysis (LDA; for applications of LDA in finance see, e.g., [22]). The LDA finds a linear transformation, defined globally for the whole data space, that aims at maximizing class separability. In a more recently proposed variant called mixture discriminant analysis [14], [15], a set of Gaussian kernels are fitted to data by optionally constraining the dimensionality of the subspace within which the kernels are allowed to reside. In contrast to LDA and the newer variants, we transform the input space locally to make the class distribution change isotropically, or with the same rate in every direction. This allows inspection of the class distributions even more closely.

Note that the discriminant analysis is commonly used for two tasks: acquiring classifications and understanding the relationships between classes by visualizations and other means. Our model is more closely related to the latter task, whereas LDA usually emphasizes the former.

The classical canonical correlation analysis has recently been generalized by replacing the linear combinations with nonlinear functions [23], [24]. Our framework could as well be adapted to the task of finding statistical dependencies between two data sets by replacing the discrete auxiliary random variable with a parametrized set of features computed from an auxiliary continuous random variable, which will be explored in future work.

III. Self-Organizing Maps in the Fisher metric

In principle any model that utilizes local distances could be adapted to use the Fisher metric (4). In this work we derive the on-line Self-Organizing Map algorithm for the new metric and use it in data analysis.

A. The Self-Organizing Map

The Self-Organizing Map (SOM) [1], [2] is a regular grid of units, with a model vector $\mathbf{m}_i$ associated with each unit i. During the learning process the model vectors are gradually modified to follow the distribution of the input data in an ordered fashion: model vectors close-by on the map lattice attain close-by locations in the input space. Two-dimensional map grids can be used to visualize various properties of the input data in data analysis applications.

The SOM algorithm iterates two steps. The index of the winning unit w closest to the current input sample $\mathbf{x}(t)$ at time t is first sought by

$$w(\mathbf{x}(t)) = \arg\min_i d^2(\mathbf{x}(t), \mathbf{m}_i(t)) \ , \tag{9}$$

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

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) - \frac{1}{2} h_{wi}(t) \frac{\partial}{\partial \mathbf{m}_i} d^2(\mathbf{x}, \mathbf{m}_i) \ . \tag{10}$$

If d is the Euclidean distance, the adaptation rule becomes the familiar

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{wi}(t)(\mathbf{x} - \mathbf{m}_i) \ . \tag{11}$$

Here $h_{wi}(t)$ is the so-called neighborhood function, a decreasing function of the distance between the units w and i on the map lattice. The height and width of $h_{wi}(t)$ decrease gradually in time. For more details see [2].

B. SOM in the new metric

To organize SOMs in the Fisher metric that is determined by the differences between the estimated posterior distributions $\hat{p}(c|\mathbf{x})$, we first construct a PDF estimator for the distributions. To find the winning unit for the data sample $\mathbf{x}(t)$, we then calculate distances to the set of model vectors, and find the one closest to $\mathbf{x}$. In general, the distances will be non-local.

To compute non-local distances, a search for the minimal path integral would be required, where differential distances along the paths would be defined by (4). Here we will approximate the non-local distances by the local distance measure (4), computed around the data sample $\mathbf{x}(t)$. The winning unit will then be found with (9). If $d\mathbf{x}$ is small, this approximation will be fair, whereas for far-off points the approximation will be rougher.

The assumption is that the approximation is locally accurate enough to preserve the order of the distances, so that the model vector $\mathbf{m}_i$ actually closest in the Riemannian metric (defined as the shortest path from $\mathbf{x}$ to $\mathbf{m}_i$) is equal to the $\mathbf{m}_w$ computed from the local approximation. This is sensible since model vectors that are close to $\mathbf{x}$ (as measured by the true non-local distances) are also likely to have small $d\mathbf{x}$, so the local approximation will not usually affect which unit becomes the winner. Occasionally this may still happen, so the true test of the goodness of the approximation will be the experimental results. The results of the case study in Section IV are favourable.

As in the original SOM, the model vector of the winning unit and units in its neighborhood are updated into the direction where the distance $d^2(\mathbf{x}, \mathbf{m}_i)$ decreases most rapidly, and proportionally to the magnitude of the change. In a Euclidean metric, the update is given by the gradient $\frac{\partial}{\partial \mathbf{m}_i}\|\mathbf{x} - \mathbf{m}_i\|^2$. The Fisher metric, however, is a Riemannian metric, and steepest descent in a Riemannian metric is given by the so-called natural gradient [25]. Generally, the natural gradient is equal to the conventional gradient multiplied by the representation of the metric tensor (a matrix). Because in the original coordinate system the metric tensor is represented by the Fisher information matrix, the natural gradient in these same coordinates is given by

$$\mathbf{J}^{-1}(\mathbf{x})\frac{\partial}{\partial \mathbf{m}_i} d^2(\mathbf{x}, \mathbf{m}_i) = \mathbf{J}^{-1}(\mathbf{x})\frac{\partial}{\partial \mathbf{m}_i}[(\mathbf{m}_i - \mathbf{x})^T \mathbf{J}(\mathbf{x})(\mathbf{m}_i - \mathbf{x})] = 2(\mathbf{m}_i - \mathbf{x}) \ . \qquad (12)$$

Assuming that $\mathbf{x}$ and $\mathbf{m}_i$ are close to each other, (12) coincides with the direction of the shortest path from $\mathbf{x}$ to $\mathbf{m}_i$.

In conclusion, the update rule in the Fisher metric is the same as in the Euclidean SOM, (11). The difference lies in the definition of the winner, (9), where the distance measure is in general defined by (4), and in the case of Gaussian kernel-based PDF estimators by (8).

C. A demonstration

The effect of the change of the metric on SOM is demonstrated in Figure 2 for a six-dimensional, three-class toy data set. We computed a Self-organizing Map in both the original Euclidean metric and in the Fisher metric, and visualized the posterior class distribution on the SOMs. As can be seen in the figure, the classes are more distinctly and orderly separated on the SOM computed in the Fisher metric. Most notably the unimodality of the distribution of each class is clearly visible.

In data analysis applications the same SOM grid can be used for visualizing other aspects of the data as well. Such displays will be used in Section IV.

D. Computational complexity

Each iteration of the SOM algorithm consists of the selection of the winning SOM unit for the current input, and an update of the model vectors. Since the update rule in the Fisher metric is unchanged from the Euclidean case, the computational complexity of the update is the same, i.e., $\mathcal{O}(N_{DIM}N_{SOM})$ for a neighborhood function that covers the whole map grid. Here N_{DIM} is the dimensionality of the input and N_{SOM} is the number of SOM units.

To select the winner, distances must be calculated from the input to each SOM unit. Using the local distance approximation, this can be done by computing the Fisher information matrix first, or by directly calculating the distances. The first alternative requires $\mathcal{O}(N_{DIM}N_C N_U + N_{DIM}^2(N_C + N_{SOM}))$ operations, where N_C is the number of classes and N_U is the number of mixture components in the PDF estimate. The second alternative requires $\mathcal{O}(N_{DIM}N_C(N_U + N_{SOM}))$ operations. If the dimensionality is small compared to the number of classes and the size of the SOM, then computing the Fisher information matrix explicitly may be faster; otherwise it is preferable to calculate the distances directly.

In the better method (MDA2) of the present case study there are 10 kernels, the number of classes is 2, the dimensionality is 23, and the winner search therefore requires about twice the

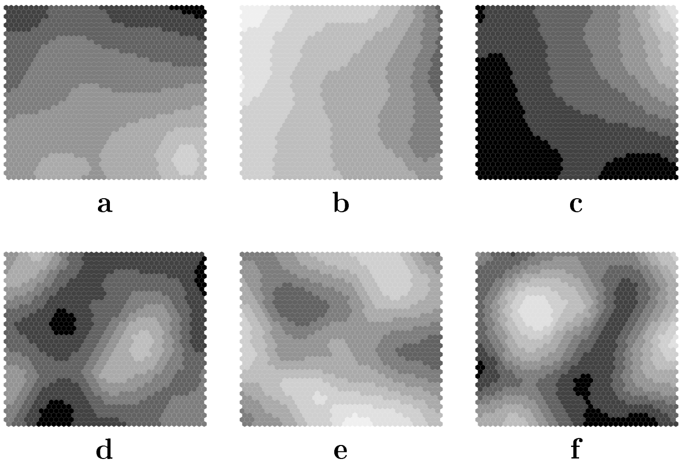


Fig. 2. A demonstration of the difference between SOMs computed in the Fisher metric and in the Euclidean metric. The primary data was six-dimensional and multinormally distributed, i.e. $\mathbf{x} \sim \mathcal{N}(0, 1)$. The auxiliary data was divided into three smoothly changing Gaussian classes, i.e. $p(c|\mathbf{x}) = G(\mathbf{x}-\mathbf{m}_c, \rho^2)/\sum_l G(\mathbf{x}-\mathbf{m}_l, \rho^2)$, where $G(\mathbf{x}, \rho^2)$ is the probability density at $\mathbf{x}$ given by the distribution $\mathcal{N}(0, \rho^2)$. The class centers $\mathbf{m}_c$ were placed evenly around the origin so that $\|\mathbf{m}_c\| = 1$, and the variance ρ^2 was 0.81. The size of the data set was 3382 points. A PDF estimate was generated using the Parzen model, with $\sigma = 1.0$. SOMs were then trained to the data with the stochastic algorithm (9), (11). Posterior probabilities of the classes (according to (7) for the Parzen estimate) evaluated at the model vectors of the SOM are shown for the two SOMs (size: 40 by 40 units) organized to represent the same data set in the Fisher (**a**: class 0, **b**: class 1, **c**: class 2) and in the standard Euclidean (**d**–**f**) metric. The probability 0.767 is shown with the lightest shade and the probability 0.040 as pure black.

amount of computation required for the simple Euclidean metric.

Note that there exist several speedup methods for the SOM (see, e.g., [26]). We have not investigated in detail their use with the Fisher metric but many of them are applicable.

IV. Application to bankruptcy analysis

The method presented in the previous chapters is applied below to a bankruptcy analysis task. Traditionally, most of the quantitative studies on bankruptcy have been directed towards prediction. The two dominating approaches in the bankruptcy prediction problem have been classification and probability estimation. In a classification task, based on the present and possibly also past data, the companies are divided into two groups: those that are likely to go bankrupt within a certain time interval, and those that are not. In probability estimation, the aim is to get estimates of the probability of bankruptcy within certain time interval—a simplified version of this is to rate the companies according to their bankruptcy risk, without requiring the ratings to be true probabilities. Naturally, probability estimation (and risk rating) models also offer a basis for classification.

A seminal work on bankruptcy prediction was performed by Altman and co-workers (summarized in [22]), who applied Linear Discriminant Analysis to this problem. Later, almost every statistical method, including neural network approaches, has been proposed (see e.g. [27], [28], [29], [30], [31], [32], [33], [34], [35], [36]. Generally, it has been observed that some of these methods, especially the more "advanced" ones such as neural network models, have slightly over-performed LDA. However, in all cases the improvement has been quite small, excluding the studies where only training set performance has been reported, or where the data set has been very small.

Another view, complementary to the bankruptcy prediction problem, is here referred to as bankruptcy analysis: trying to understand the different corporate behaviors and their relation to the risk of bankruptcy. A very influential qualitative work in this area has been carried out by Argenti [37]. One of his observations was that there are several different bankruptcy types ("failure trajectories") that differ in their causes, symptoms, and length. Along these lines of thought, a research project in Helsinki University of Technology has attempted to quantify and visualize these different behavior patterns [5], [6], [38], [39]. Because the present study is closely related to this project, some of its findings and also challenges are briefly summarized below.

First, the SOM does not increase the accuracy in bankruptcy prediction, but is very useful in visualizing the present state of a company and possible directions of its future development—the analyst gets a much more accurate idea of the state of the company from the visualization on a SOM than from a single scalar estimating the bankruptcy risk.

Second, different types of corporate behavior (trajectories) can be identified with the SOM.

Third, one problem with the visualization using an SOM is that when the data has an intrinsic dimensionality higher than that of the SOM grid, discontinuities in the mapping sometimes result. For instance, a single cluster of high bankruptcy risk may appear multimodal on the SOM. The Fisher metric approach, described in previous chapters, is likely to help with this problem, for the manifold spanned by the Fisher metric is of lower dimensionality than the original data space.

Thus, the primary goal in this section is to use the new methods to better understand the (nonlinear) dependencies between bankruptcies and financial indicators. The dependencies are converted into a metric of the input space, and we use the SOM to visualize the dependencies in a concise form. Because the metric is chosen to describe changes in the bankruptcy sensitivity, the SOM should emphasize features of the input space that are (locally) contributing to bankruptcies.

In this section, we will for brevity call a SOM computed in the Euclidean metric SOM-E, and a SOM computed in the Fisher metric SOM-F.

A. Data

The financial statements were from Finnish small and medium-sized enterprises. The line of business, age, size, and completeness of the available data were used as the selection criteria, but no data was otherwise rejected on the basis of 'atypicality'. In the data set there were 6 195 financial statements given by about 1 500 companies. Of these statements, 158 concerned companies which have gone bankrupt.

In this paper we do not take into account the development of companies in time. Multiple statements from the same enterprise but from different years are treated as independent samples.

We used a set of 23 common financial indicators including measures of growth, profitability, solidity, liquidity, and operational efficiency; the samples of the primary data space $\mathbf{x}$ were 23-dimensional real vectors. The indicators were preprocessed (each separately) using histogram equalization. The auxiliary random variable C was binary, indicating whether the statement was followed by a bankruptcy within 3 years.

B. Methods

The data was randomly divided into an estimation set and a test set of roughly equal sizes. Two PDF estimates, the first based on Parzen estimation with Gaussian kernels and the second on a Gaussian mixture with 10 mixture components, were fitted to the estimation set. Hexagonal SOMs of the size of 20×10 units were then computed both in the Euclidean and in the Fisher metric, the latter derived from the PDF estimates.

B.1 Verification measures

In this section we present a measure of goodness for verifying that the SOM-F reflects aspects of the input data that are relevant to the risk of bankruptcy. There are three components affecting the goodness: (1) the quality of the PDF estimator, (2) the accuracy by which the SOMs represent the probability of bankruptcy[3], and (3) the quality of the visualizations, i.e., the smoothness and quality of organization of the SOMs.

We will not measure the first component; it is assumed that the standard PDF estimators are adequate. The accuracy of representation will be measured by the log-likelihood of the test data given the estimates at the locations of the winner units,

$$\sum_k \log \hat{p}(c^k|\mathbf{m}_{w(\mathbf{x}^k)}) \ . \tag{13}$$

Regarding the quality of visualizations we will resort to visual comparisons between visualizations obtained by SOM-E and SOM-F.

The likelihoods obtained by the two PDF estimates and SOM-E and SOM-F were computed for a wide range of values of the parameter σ, from the order of the average distance between two closest data points to the order of the maximal distance. The likelihood obtained directly from the PDF estimator was computed to find out an approximation of the best possible performance, and a model always predicting prior probabilities of the classes served as a lower limit of useful results.

Note: The likelihood used for measuring the accuracy of representation has a connection to the quantization error that is commonly used for measuring the quality of SOMs. The quantization error is defined to be the average distance from the original data to the winning SOM units, $E\{\|\mathbf{x}-\mathbf{m}_{w(\mathbf{x})}\|\}$. In SOM-F, the corresponding measure would be the Kullback-Leibler divergence between the posterior distributions $\hat{p}(c|\mathbf{x})$ and $\hat{p}(c|\mathbf{m}_{w(\mathbf{x})})$. Assuming that $\mathbf{x}$ and $\mathbf{m}_{w(\mathbf{x})}$ are close to each other, the squared difference can be computed using the Fisher information matrix as in (4). However, since the measure is based solely on estimates of PDFs and not on the data itself, minimizing the quantization error in the estimated Fisher metric does not guarantee that the map represents the real data. There exists a simple remedy: If the estimate $\hat{p}(c|\mathbf{x})$ is replaced by the real distribution $p(c|\mathbf{x})$, then the Kullback-Leibler divergence measures deviance of the representation from the true PDF. It can be easily shown that the average divergence between $p(c|\mathbf{x})$ and $\hat{p}(c|\mathbf{m}_{w(\mathbf{x})})$ is approximated by a linear function of the likelihood (13).

[3]Note that although this accuracy can be measured by the prediction accuracy, our goal is not to simply maximize prediction accuracy but to quantify accuracy of visualizations.

B.2 Visualization of the results

In addition to the usual visualization methods available for all SOMs, with SOM-F one can visualize correlations between bankruptcy sensitivity and directions of the data space. The amount of scaling of a direction $d\mathbf{x}$, revealed by the quadratic form $d\mathbf{x}^T\mathbf{J}d\mathbf{x}$, measures the effect of the direction on the bankruptcy sensitivity. We will visualize the magnitudes of these scalings at the most easily interpretable directions of the data space, the original variables of the data.

The relative amount of scaling in the direction of coordinate axis l is given by

$$r_l(\mathbf{x}) = \sqrt{\frac{\mathbf{e}_l^T\mathbf{J}(\mathbf{x})\mathbf{e}_l}{\sum_m \mathbf{e}_m^T\mathbf{J}(\mathbf{x})\mathbf{e}_m}} \tag{14}$$

where $\mathbf{e}_l$ is the unit vector parallel to the axis. A large value of $r_l(\mathbf{x})$ indicates a strong effect by the variable l, locally around $\mathbf{x}$.

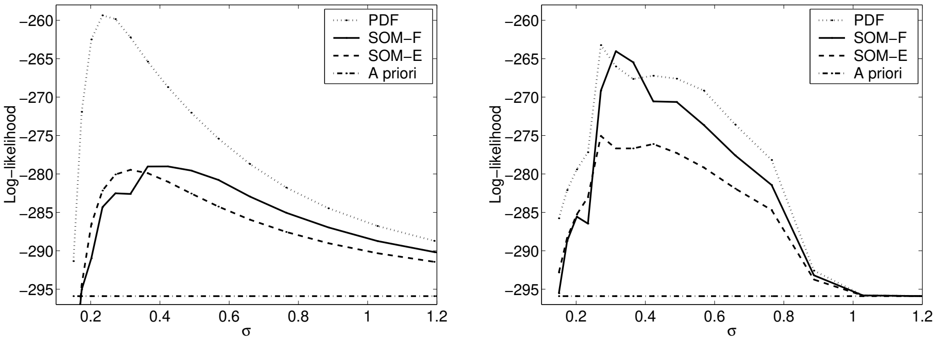
C. Results

The likelihoods of SOM-E and SOM-F in the test set are shown in Figure 3 as a function of the parameter σ (which governs the smoothness of the PDF estimates). The SOM-F, as expected, performs clearly better than the Euclidean SOM. The SOM-E is roughly equal only for the kernel-based PDF estimate when σ is very small—then the PDF estimate and the resulting Fisher metric are probably very uneven. The location of a financial statement on the SOM-F is thus a more accurate predictor of bankruptcy than the location of the statement on the SOM-E.

To test the statistical significance of the performance difference, the data was divided into 10 separate sets. At each test round, one set was used as the test data and the other sets as the training data. The likelihood curves of the SOM-E and SOM-F were calculated for the test sets (using the MDA2 estimate), and the peaks of the curves were compared with the sign test. The SOM-F outperformed the SOM-E ($p < 0.002$).

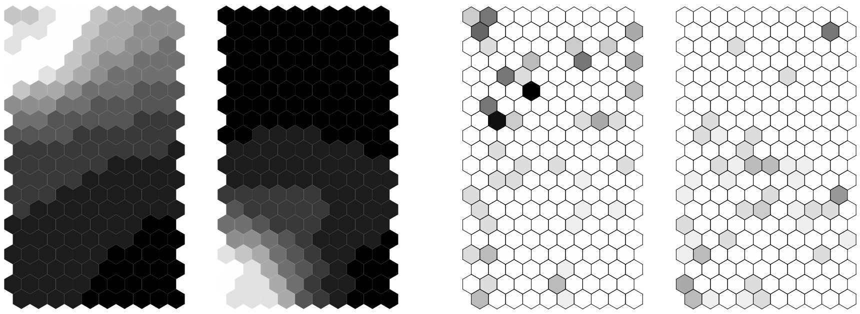
Still, the variation of the financial indicators on SOM-F displays (Fig. 5) is remarkably smooth, comparable to the smoothness of the SOM-E displays. Moreover, the bankrupt companies are visually clearly separated on both SOMs, in the sense that their distribution is unimodal and that the posterior class densities change smoothly on the map (Fig. 4). In summary, the good organization and visualization capabilities of the SOM have been maintained or even improved while the Fisher metric has increased the prediction accuracy.

The relative scaling of the coordinate axes in the new metric can be visualized as easily un-



a **b**

Fig. 3. The accuracy of the SOMs computed in the Euclidean metric (SOM-E) and in the Fisher metric (SOM-F) in representing the probability of bankruptcy, measured by the likelihood of data at the locations of the best-matching SOM units (13). **a**: The PDF is estimated with the Gaussian kernel (Parzen) estimate. **b**: The PDF is estimated by a Gaussian mixture having 10 mixture components. The curve marked by "PDF" provides an approximate upper limit: it is the likelihood at the data points instead of at the best matching units. The curve marked by "A priori" provides the lower limit of sensible results, obtained by the best constant estimates. The parameter σ governs the smoothness of the PDF estimates.



a **b** **c** **d**

Fig. 4. The separation of bankruptcy-prone and healthy companies on the SOMs. **a**-**b**: The estimate of the probability of bankruptcy at each map unit in **a** SOM-F and **b** SOM-E. The estimate is the posterior density of the Gaussian mixture model at $\sigma = 0.31$. The darkest shade denotes probability 0.12; the lightest denotes probability 0.002. (Note that the prior probability of bankruptcy is small, just 0.022.) The actual relative frequency of bankruptcies in the test set for each map unit is shown in **c** for SOM-F and in **d** for SOM-E. The frequency graphs are noisy since the number of bankrupt companies was small. White: no bankruptcies, black: two thirds of all companies have gone bankrupt.

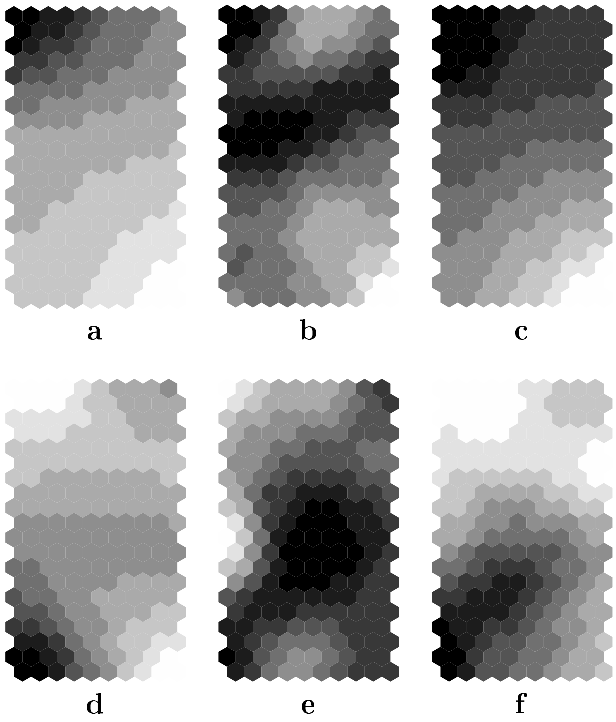


Fig. 5. The distribution of the values of three financial indicators on (**a**-**c**) SOM-F and (**d**-**f**) SOM-E. An index of (**a** and **d**) profitability; (**b** and **e**) liquidity; (**c** and **f**) capital structure. The Fisher metric for SOM-F has been computed from a Gaussian mixture estimate with $\sigma = 0.31$.

derstandable overviews of the relative importance of the input variables. Some examples are presented in Figure 6. The non-constant values of the $r(\mathbf{x})$ suggest that nonlinear effects exist, which would justify the use of nonlinear models for this data set.

V. Discussion

We have introduced a new method for deriving metrics from the estimated posterior distribution of an auxiliary relevance-inducing variable, and used it in computing Self-Organizing Maps. The metric is based on the Fisher information matrix, which results from a local approximation of the Kullback-Leibler divergence between the posterior densities at close-by points in the primary data space. In the new metric the estimated posterior probabilities change evenly in all directions. In other words, the metric represents local contribution of the directions of the data space to changes in the relevance-indicating random variable.

We computed a SOM in the Fisher metric and applied it to the visualization of bankruptcy sensitivity as a function of several quantitative financial indicators. The SOM was more accurate

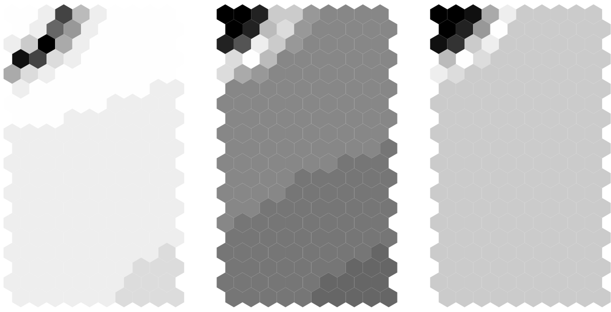


Fig. 6. The relative contributions $r_l^2(\mathbf{x})$ to the change in the bankruptcy sensitivity, plotted as gray levels on the map display, for the indicators of Figure 5. The relative contribution of the profitability indicator (**a**: scale from 0.007 to 0.080) decreases and the contribution of the capital structure indicator (**c**: scale from 0.0002 to 0.215) increases at the bankruptcy zone (the stripe in the top left corner), while the contribution of the liquidity indicator (**b**: scale from 0.001 to 0.013) is very low.

in representing the (estimated) probability of bankruptcy than an Euclidean SOM while the visual quality of the maps was comparable or improved.

The Fisher metric can be used for discovering and visualizing locally relevant dimensions, and as a kind of automatic feature-extraction stage. We have presented one way of visualizing the contributions of the input variables to the Fisher metric. In general the visualization of the metric tensor as a function of the primary data space is a subject for further experimentation and research.

When used as a kind of feature extraction stage the method has the nice property that it changes the metric while still preserving the proximity relations of the original data space. For a PDF estimator which approaches the real PDF when the number of data grows, the results are independent of the original coordinate system, and therefore of the metric, of the data. Preservation of the proximity relations is, of course, a natural requirement for sensible operation of further processing stages like the SOM; if the topology of the original space is not worth preserving then it is best to use a suitable (discontinuous) preprocessing stage.

It may be worth noting that the change of the metric affects the density of $\mathbb{X}$. In the Fisher metric, the density of $p(\mathbf{x})$ changes to $|\mathbf{J}(\mathbf{x})|^{-1/2}p(\mathbf{x})$. This change reduces the density of data at the points of the $\mathbb{X}$-space where the posterior probabilities $p(c|\mathbf{x})$ change rapidly. If this is undesirable, a modified Fisher metric with a constant magnification factor can be used.

In finding the relevant local features of the input space, the extraction of the Fisher metric is similar to the recently introduced Kullback-Leibler clustering algorithm [40], [41]. This connection will be detailed in future papers.

In summary, we have extended the SOM-based exploratory analyses of the factors affecting bankruptcy risk in different kinds of companies by the new learning metric. The Fisher metric derived from PDFs improved the accuracy with which the visual maps represent bankruptcy and even the quality of the visualizations. Bankruptcy analysis from financial statements is a common task, and it is relatively well known which features are meaningful; moreover, the effective dimensionality of meaningful data spaces is small. It is therefore hard to improve on the methods that are already in use in this field. Hence, the Fisher metric is likely to be even more useful when the structure of the data is less known and there is little justification for manual feature selection.

Appendix A
Derivation of the distance d^2 for the mixture model

The gradient of (7) is

$$\frac{\partial}{\partial \mathbf{x}} \hat{p}(c_i|\mathbf{x};\mathbf{\Theta}) = \sum_j \pi_j \frac{\xi_{ji} - \hat{p}(c_i|\mathbf{x})}{\sum_k \pi_k b_k(\mathbf{x};\boldsymbol{\theta}_k)} \frac{\partial}{\partial \mathbf{x}} b_j(\mathbf{x};\boldsymbol{\theta}_j) \ ,$$

and hence

$$\frac{\partial}{\partial \mathbf{x}} \log \hat{p}(c_i|\mathbf{x};\mathbf{\Theta}) = \sum_j \pi_j \frac{\xi_{ji}/\hat{p}(c_i|\mathbf{x}) - 1}{\sum_k \pi_k b_k(\mathbf{x};\boldsymbol{\theta}_k)} \frac{\partial}{\partial \mathbf{x}} b_j(\mathbf{x};\boldsymbol{\theta}_j) \ .$$

For a Gaussian b_j having a diagonal covariance matrix $\sigma^2\mathbf{I}$,

$$\frac{\partial}{\partial \mathbf{x}} b_j(\mathbf{x};\boldsymbol{\theta}_j) = -\frac{1}{\sigma^2} b_j(\mathbf{x};\boldsymbol{\theta}_j)(\mathbf{x} - \boldsymbol{\theta}_j) \ ,$$

and hence

$$\sigma^2 \frac{\partial}{\partial \mathbf{x}} \log \hat{p}(c_i|\mathbf{x};\mathbf{\Theta}) = \sum_j \left[\frac{\xi_{ji}\pi_j b_j(\mathbf{x};\boldsymbol{\theta}_j)}{\hat{p}(c_i|x) \sum_k \pi_k b_k(\mathbf{x};\boldsymbol{\theta}_k)} - \frac{\pi_j b_j(\mathbf{x};\boldsymbol{\theta}_j)}{\sum_k \pi_k b_k(\mathbf{x};\boldsymbol{\theta}_k)} \right] (\boldsymbol{\theta}_j - \mathbf{x}) \ . \quad (15)$$

Using (7), the expression in brackets can be simplified to

$$\frac{\sum_k \pi_k b_k(\mathbf{x};\boldsymbol{\theta}_k)}{\sum_k \pi_k \xi_{ki} b_k(\mathbf{x};\boldsymbol{\theta}_k)} \frac{\pi_j \xi_{ji} b_j(\mathbf{x};\boldsymbol{\theta}_j)}{\sum_k \pi_k b_k(\mathbf{x};\boldsymbol{\theta}_k)} - \frac{\pi_j b_j(\mathbf{x};\boldsymbol{\theta}_j)}{\sum_k \pi_k b_k(\mathbf{x};\boldsymbol{\theta}_k)} =$$
$$\frac{\pi_j \xi_{ji} b_j(\mathbf{x};\boldsymbol{\theta}_j)}{\sum_k \pi_k \xi_{ki} b_k(\mathbf{x};\boldsymbol{\theta}_k)} - \frac{\pi_j b_j(\mathbf{x};\boldsymbol{\theta}_j)}{\sum_k \pi_k b_k(\mathbf{x};\boldsymbol{\theta}_k)} =$$
$$\frac{p(c_i, u_j, \mathbf{x};\boldsymbol{\theta}_j)}{\sum_k p(c_i, u_k, \mathbf{x};\boldsymbol{\theta}_k)} - \frac{p(\mathbf{x}, u_j;\boldsymbol{\theta}_j)}{\sum_k p(\mathbf{x}, u_k;\boldsymbol{\theta}_k)} =$$
$$p(u_j|\mathbf{x}, c_i;\boldsymbol{\theta}_j) - p(u_j|\mathbf{x};\boldsymbol{\theta}_j) \ . \quad (16)$$

Plugging (16) into (15) yields

$$\sigma^2 \frac{\partial}{\partial \mathbf{x}} \log \hat{p}(c_i|\mathbf{x};\boldsymbol{\Theta}) =$$
$$\sum_j [p(u_j|\mathbf{x}, c_i;\boldsymbol{\theta}_j) - p(u_j|\mathbf{x};\boldsymbol{\theta}_j)](\boldsymbol{\theta}_j - \mathbf{x}) =$$
$$E_{p(u_j|\mathbf{x},c_i;\boldsymbol{\theta}_j)}\{\boldsymbol{\theta}_j - \mathbf{x}\} - E_{p(u_j|\mathbf{x};\boldsymbol{\theta}_j)}\{\boldsymbol{\theta}_j - \mathbf{x}\} =$$
$$E_{p(u_j|\mathbf{x},c_i;\boldsymbol{\theta}_j)}\{\boldsymbol{\theta}_j\} - E_{p(u_j|\mathbf{x};\boldsymbol{\theta}_j)}\{\boldsymbol{\theta}_j\} \ , \quad (17)$$

and plugging (17) into (4) yields (8).

Appendix B

EM estimation of the Gaussian mixture model of joint densities

We used the EM algorithm to maximize the likelihood of the model (6). The value of the random variable U that indicates which generator has produced each data item is considered as the missing data. The value of U for the data sample $(\mathbf{x}^k, c^k)$ is denoted by u^k. Both the data and the u^k are assumed to be independent and identically distributed.

As an initialization we set $\pi_j = 1/N_U$ and $\xi_{ji} = 1/(N_U N_C)$, where N_U denotes the number of component generators and N_C denotes the number of possible values of C. The $\boldsymbol{\theta}_j$ are initialized by the K-means algorithm.

The E-step consists of two sub-steps. First the joint distribution of the missing data is inferred, and then the expected log-likelihood of the model with respect to this distribution is computed, conditioned on the old parameters and the data.

Given the old set of parameters $\boldsymbol{\Theta}^{(0)}$, the joint distribution of the missing data is

$$p(\{u_j^k\}_k|\{(\mathbf{x}^k, c^k)\}_k;\boldsymbol{\Theta}^{(0)}) = \prod_k p(u_j|\mathbf{x}^k, c^k;\boldsymbol{\Theta}^{(0)}) \ ; \quad (18)$$

below we will generally use the superscript (0) to refer to the old parameters. The probability $p(u_j|\mathbf{x}^k, c^k; \mathbf{\Theta}^{(0)})$ that the mixture component u_j has generated the data sample $(\mathbf{x}^k, c^k)$ is

$$p(u_j|\mathbf{x}^k, c^k; \mathbf{\Theta}^{(0)}) = \frac{b_j(\mathbf{x}^k; \boldsymbol{\theta}_j^{(0)})\xi^{(0)}_{ji:c_i=c^k}\pi_j^{(0)}}{p(\mathbf{x}^k, c^k|\mathbf{\Theta}^{(0)})} \equiv \tau_{jk} \; .$$

It can be shown that the expected log-likelihood $\mathcal{L}(\mathbf{\Theta})$ of the model (6) with respect to the distribution (18) is

$$E\{\mathcal{L}(\mathbf{\Theta})\} = \sum_{j,k} \tau_{jk} \left[\log b_j(\mathbf{x}^k; \boldsymbol{\theta}_j) + \log \xi_{ji:c_i=c^k} + \log \pi_j\right] \; .$$

In the M-step the expected log-likelihood is maximized. It can be shown that with respect to the ξ_{ji} the maximum is at

$$\xi_{ji} = \frac{\sum_{k:c^k=c_i} \tau_{jk}}{\sum_k \tau_{jk}} \; . \tag{19}$$

For the π_j the maximum is at

$$\pi_j = \sum_k \tau_{jk}/N \tag{20}$$

where N is the number of data samples, and for the $\boldsymbol{\theta}_j$ at

$$\boldsymbol{\theta}_j = \frac{\sum_k \tau_{jk}\mathbf{x}^k}{\sum_k \tau_{jk}} \; . \tag{21}$$

Acknowledgments

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

References

[1] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, pp. 59–69, 1982.

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

[3] S. Kaski and J. Sinkkonen, "Metrics that learn relevance," in *Proc. IJCNN-2000, International Joint Conference on Neural Networks*, vol. V, pp. 547–552. IEEE, 2000.

[4] K. Kiviluoto and P. Bergius, "Analyzing financial statements with the self-organizing map," in *Proceedings of WSOM'97, Workshop on Self-Organizing Maps*, pp. 362–367. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, 1997.

[5] K. Kiviluoto, "Predicting bankruptcies with the self-organizing map," *Neurocomputing*, vol. 21, no. 1–3, pp. 191–201, 1998.

[6] 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.

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

[8] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters," *Bull. Calcutta Math. Soc.*, vol. 37, pp. 81–91, 1945.

[9] M. K. Murray and J. W. Rice, *Differential Geometry and Statistics*, Chapman & Hall, London, 1993.

[10] S.-I. Amari, *Differential-Geometrical Methods in Statistics*, Springer, New York, 1990.

[11] R. E. Kass and P. W. Vos, *Geometrical Foundations of Asymptotic Inference*, Wiley, New York, 1997.

[12] L. Holmström, P. Koistinen, J. Laaksonen, and E. Oja, "Neural and statistical classifiers—taxonomy and two case studies," *IEEE Transactions on Neural Networks*, vol. 8, pp. 5–17, 1997.

[13] B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, Great Britain, 1996.

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

[15] Trevor Hastie and Robert Tibshirani, "Discriminant analysis by Gaussian mixtures," *Journal of the Royal Statistical Society Series B*, vol. 58, pp. 155–176, 1996.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[17] S. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks*, vol. 12, pp. 783–789, 1999.

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

[19] H. Ritter and T. Kohonen, "Self-organizing semantic maps," *Biological Cybernetics*, vol. 61, pp. 241–254, 1989.

[20] J. W. Fisher III and J. Principe, "A methodology for information theoretic feature extraction," in *Proc. IJCNN'98, International Joint Conference on Neural Networks*, vol. 3, pp. 1712–1716. IEEE, 1998.

[21] K. Torkkola and W. M. Campbell, "Mutual information in learning feature transformations," in *Proc. ICML'2000, The Seventeenth International Conference on Machine Learning*, Morgan Kaufmann, pp. 1015–1022, 2000.

[22] E. I. Altman, *A complete guide to predicting, avoiding, and dealing with bankruptcy*, John Wiley & Sons, Inc., 1983.

[23] S. Becker, "Mutual information maximization: models of cortical self-organization," *Network: Computation in Neural Systems*, vol. 7, pp. 7–31, 1996.

[24] P. L. Lai and C. Fyfe, "A neural implementation of canonical correlation analysis," *Neural Networks*, vol. 12, pp. 1391–1397, 1999.

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

[26] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela, "Self organization of a massive document collection," *IEEE Transactions on Neural Networks*, vol. 11, pp. 574–585, 2000.

[27] J. A. Ohlson, "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of Accounting Research*, vol. 18, no. 1, pp. 109–131, 1980.

[28] Gordon V. Karels and Arun J. Prakash, "Multivariate normality and forecasting of business bankruptcy," *Journal of Business Finance & Accounting*, pp. 573–592, Winter 1987.

[29] K.Y. Tam and M.Y. Kiang, "Managerial applications of neural networks: the case of bank failure predictions," *Management Science*, vol. 38, no. 7, pp. 926–947, July 1992.

[30] R.L. Wilson and R. Sharda, "Bankruptcy prediction using neural networks," *Decision Support Systems*, vol. 11, no. 5, pp. 545–557, June 1994.

[31] Youngohc Yoon, G. Swales, Jr., and T.M. Margavio, "A comparison of discriminant analysis versus artificial neural networks," *Journal of the Operational Research Society*, vol. 44, no. 1, pp. 51–60, January 1993.

[32] D. Fletcher and E. Goss, "Forecasting with neural networks: an application using bankruptcy data," *Information and Management*, vol. 24, no. 3, pp. 159–167, March 1993.

[33] J. Tsukuda and S.-I. Baba, "Predicting Japanese corporate bankruptcy in terms of financial data using neural networks," *Computers & Industrial Engineering*, vol. 27, pp. 445–448, September 1994.

[34] Godwin Udo, "Neural network performance on the bankruptcy classification problem," *Computers and Industrial Engineering*, vol. 25, no. 1-4, pp. 377–380, 1993.

[35] Bonifacio Martín del Brío and Carlos Serrano-Cinca, "Self-organizing neural networks for the analysis and representation of data: Some financial cases," *Neural Computing & Applications*, vol. 1, pp. 193–206, 1993.

[36] S. A. Shumsky and A.V. Yarovoy, "Neural network analysis of Russian banks," in *Proceedings of the workshop on self-organizing maps (WSOM'97)*, Espoo, Finland, June 1997, Neural Networks Research Centre, Helsinki University of Technology.

[37] John Argenti, *Corporate collapse – the causes and symptoms*, McGraw-Hill, 1976.

[38] Kimmo Kiviluoto and Pentti Bergius, "Two-level self-organizing maps for analysis of financial statements," in *Proceedings of the 1998 IEEE International Joint Conference on Neural Networks (IJCNN'98)*, Piscataway,

New Jersey, USA, May 1998, IEEE Neural Networks Council, vol. 1, pp. 189–192.

[39] Kimmo Kiviluoto, "Comparing 2D and 3D self-organizing maps in financial data visualization," in *Methodologies for the Conception, Design and Application of Soft Computing – Proceedings of the 5th International Conference on Soft Computing and Information/Intelligent Systems (IIZUKA'98). Iizuka, Fukuoka, Japan.*, Takeshi Yamakawa and Gen Matsumoto, Eds., Singapore, Oct. 1998, vol. 1, pp. 68–71, World Scientific.

[40] 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.

[41] J. Sinkkonen and S. Kaski, "Semisupervised clustering based on conditional distributions in an auxiliary space," Tech. Rep. A60, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 2000.

Figure Legends:

Figure 1:

The metric generated by a PDF estimate for a two-dimensional two-class data set (N=1000). The first class is sampled from a symmetrical Gaussian with $p(c) = \frac{1}{3}$ (the topmost cluster in the figure), the second from a sum of two Gaussians (the bottom clusters). For all the Gaussians, σ=0.6, and the mutual distances of the centers are equal to unity. The gray-scale background illustrates the marginal density $p(\mathbf{x})$, and the small line segments (or dots) depict the dominant direction and relative distances d^2 in the local metric. Distances are nonzero only in the directions where the conditional density changes. The PDF was estimated with a Gaussian Parzen estimator ($\sigma = 0.4$).

Figure 2:

A demonstration of the difference between SOMs computed in the Fisher metric and in the Euclidean metric. The primary data was six-dimensional and multinormally distributed, i.e. $\mathbf{x} \sim \mathcal{N}(0, 1)$. The auxiliary data was divided into three smoothly changing Gaussian classes, i.e. $p(c|\mathbf{x}) = G(\mathbf{x} - \mathbf{m}_c, \rho^2)/\sum_l G(\mathbf{x} - \mathbf{m}_l, \rho^2)$, where $G(\mathbf{x}, \rho^2)$ is the probability density at $\mathbf{x}$ given by the distribution $\mathcal{N}(0, \rho^2)$. The class centers $\mathbf{m}_c$ were placed evenly around the origin so that $\|\mathbf{m}_c\| = 1$, and the variance ρ^2 was 0.81. The size of the data set was 3382 points. A PDF estimate was generated using the Parzen model, with $\sigma = 1.0$. SOMs were then trained to the data with the stochastic algorithm (9), (11). Posterior probabilities of the classes (according to (7) for the Parzen estimate) evaluated at the model vectors of the SOM are shown for the two SOMs (size: 40 by 40 units) organized to represent the same data set in the Fisher (**a**: class 0, **b**: class 1, **c**: class 2) and in the standard Euclidean (**d**–**f**) metric. The probability 0.767 is shown with the lightest shade and the probability 0.040 as pure black.

Figure Legends:

Figure 3:

The accuracy of the SOMs computed in the Euclidean metric (SOM-E) and in the Fisher metric (SOM-F) in representing the probability of bankruptcy, measured by the likelihood of data at the locations of the best-matching SOM units (13). **a**: The PDF is estimated with the Gaussian kernel (Parzen) estimate. **b**: The PDF is estimated by a Gaussian mixture having 10 mixture components. The curve marked by "PDF" provides an approximate upper limit: it is the likelihood at the data points instead of at the best matching units. The curve marked by "A priori" provides the lower limit of sensible results, obtained by the best constant estimates. The parameter σ governs the smoothness of the PDF estimates.

Figure 4:

The separation of bankruptcy-prone and healthy companies on the SOMs. **a**-**b**: The estimate of the probability of bankruptcy at each map unit in **a** SOM-F and **b** SOM-E. The estimate is the posterior density of the Gaussian mixture model at $\sigma = 0.31$. The darkest shade denotes probability 0.12; the lightest denotes probability 0.002. (Note that the prior probability of bankruptcy is small, just 0.022.) The actual relative frequency of bankruptcies in the test set for each map unit is shown in **c** for SOM-F and in **d** for SOM-E. The frequency graphs are noisy since the number of bankrupt companies was small. White: no bankruptcies, black: two thirds of all companies have gone bankrupt.

Figure Legends:

Figure 5:

The distribution of the values of three financial indicators on (**a**-**c**) SOM-F and (**d**-**f**) SOM-E. An index of (**a** and **d**) profitability; (**b** and **e**) liquidity; (**c** and **f**) capital structure. The Fisher metric for SOM-F has been computed from a Gaussian mixture estimate with $\sigma = 0.31$.

Figure 6:

The relative contributions $r_l^2(\mathbf{x})$ to the change in the bankruptcy sensitivity, plotted as gray levels on the map display, for the indicators of Figure 5. The relative contribution of the profitability indicator (**a**: scale from 0.007 to 0.080) decreases and the contribution of the capital structure indicator (**c**: scale from 0.0002 to 0.215) increases at the bankruptcy zone (the stripe in the top left corner), while the contribution of the liquidity indicator (**b**: scale from 0.001 to 0.013) is very low.