Making every bit count: Fast nonlinear axis scaling

Leejay Wu
Carnegie Mellon University
lw2j@cs.cmu.edu

Christos Faloutsos
Carnegie Mellon University
christos@cs.cmu.edu

ABSTRACT

Existing axis scaling and dimensionality methods focus on preserving structure, usually determined via the Euclidean distance. In other words, they inherently assume that the Euclidean distance is *already* correct. We instead propose a novel nonlinear approach driven by an information-theoretic viewpoint, which we show is also strongly linked to intrinsic dimensionality, or degrees of freedom; and uniformity. Nonlinear transformations based on common probability distributions, combined with information-driven selection, simultaneously reduce the number of dimensions required and increase the value of those we retain. Experiments on real data confirm that this approach reveals correlations, finds novel attributes, and scales well.

1. INTRODUCTION

The assumption that the Euclidean distance is already acceptable is inherent in most other scaling problems. We instead focus on determining what is, in fact, a good distance function for data; isomorphically, what space is appropriate for data. Motivating this problem are situations such as the following.

Consider Figure 1, which shows one data set presented in two different sets of scales. The exact scaling methods used need not be specified; suffice it to say that in both cases, the original data could be determined precisely given the methods, and that any mathematical rules learned in either space could be transformed to rules in the original space.

Version (b) shows a strong linear relationship between the axes, while finding any such rule for (a) would be difficult. (b) is also far less dominated by outliers. Not coincidentally, the axes of (b) appear much more uniformly distributed. For these reasons, (b) is a better data space. Traditional scaling methods would have preserved the structure of (a), thus retaining a bad distance function. Distance functions, in turn, are critical to many problems. Trustworthy, well-grounded distance functions are absolutely *vital* – and most algorithms simply assume that this problem has already been solved.

1.1 Intuition

Intuition leads us to three major aspects to bad scaling – skew, or lack of uniformity, along an axis; badly matched ranges between attributes; and redundant expression of hidden factors in the attribute set.

One issue is skew, the opposite of uniformity. In a uniform distribution, distances are simple to interpret; whereas skewed distributions often indicate nonlinearities or severe outliers. Second, it may well be desirable for attributes to have similar ranges. Vastly dissimilar ranges may result in one attribute looking like noise compared to another. Third, redundancy is undesirable. The impact of redundancy on performance and storage is obvious. Its impact on correctness may not be. Consider that a standard Euclidean distance which sees the same basic, underlying distance in ten attributes will take that same distance into account ten times, thus giving it excessive weight.

Mismatched ranges are simply handled with affine normalization schemes. Redundancy can be dealt with by dropping attributes, but only if redundancy can be identified. Skew can only be dealt with via nonlinear transformations – and again, must first be measured. Finding this measure lies at the heart of the problem.

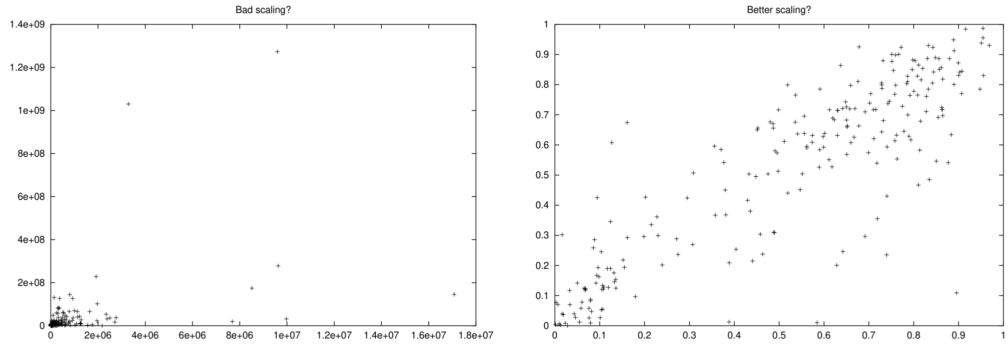
1.2 Informal specification

Given a set of n vectors in $\mathcal{R}^m$, produce through invertible transformations and dimensionality reduction a second set of n vectors in $\mathcal{R}^{m'}$, with $m' \leq m$, with increased uniformity, well-matched ranges, and less redundancy. The underlying problem, of course, is figuring out how to measure uniformity and redundancy.

2. THE IMPORTANCE OF SCALING

This section illustrates the possible benefits of nonlinear scaling, even when applied in a very simple fashion, as compared to the more complex methods elaborated upon later. Here, we'll see that using nonlinear scaling can help in rule-finding.

Figure 2 shows corresponding raw and logarithmic plots for two player statistics for the NBA 1991-1992 season, the number of games played versus the number of field goals made. Linear regression is dubious at first with a correlation coefficient of 0.7569, but after discarding eight vectors corresponding to players who scored zero in either statistic and applying the natural logarithm to the rest, the following rule holds with a greatly improved correlation coefficient of 0.9109:



(a) Bad scaling? (b) Better?

Figure 1: Two versions of the same data. Which is better for data mining? Why?

$$\ln f = 1.5080 * \ln g - 1.0032 \tag{1}$$

or, approximately, $f = g^{1.5}$, where f is the number of field goals and g the number of games. Nonlinear scaling has thus revealed a superlinear rule, for which there is a plausible explanation – more skilled basketball players are more likely to get more playing time, and thus more opportunities to score, and vice versa. If we has relied only on linear scaling, we might have been misled by the 0.7569 correlation on the raw data.

3. RELATED WORK

Other methods of interest include linear projection methods such as principal components [15], FastMap [11] and independent components analysis [12, 13]; robust distance estimators [16]; local methods such as mapping and manifolds [17], [18], [22] and others [8]. Principal components has also been extended via principal curves [9] and nonlinear kernels [19].

The SPARTAN [1] algorithm lossily compresses data via feature selection plus CART trees [5] with the retained attributes as possible inputs [1]. Neural networks have also been used for dimensionality reduction and reconstruction [10].

The above listed methods suffer variously from complexity; the need for human intervention such as choosing kernels; the goal of preserving versus improving structure; and a host of other issues. The system presented in this paper instead relies on a different approach firmly grounded in information theory.

3.1 Shannon information

Shannon information [21], also known as entropy, provides a way to quantify the intuitive concept of information. The formulas

$$H(x) = -\sum_i p_i \log_2 p_i \tag{2}$$

where p_i corresponds to a probability of a discrete outcome i, and

$$H(x) = \int -f(x) \log_2 f(x) dx \tag{3}$$

where $f(x)$ is the probability density function (PDF) evaluated at x, measure the entropy of a random variable x. The second form is also known as *differential entropy.* In each case, entropy corresponds to the theoretical minimum expected number of bits required to transmit individual instances of the random variable.

4. PROPOSED METHOD

While entropy may seem a logical choice for measuring redundancy or uniformity, there are multiple problems with using either of Equations 2 or 3 as a heuristic for determining which scaling method is preferable. The most serious is that continuous data requires either discretization, which is parameter-sensitive; or differential entropy, which requires a probability density function.

Consider a *series* of discretizations with steadily reduced granularity. For self-similar sets, the amount of entropy increases linearly with the number of bits of precision, within limits. Obviously, for a finite point set, there are minimum and maximum amounts of precision beyond which all available information has either been revealed or discarded. This rate of increase is the **marginal information content**.

This metric is *critical* because it objectively compares different spaces in a theoretically well-ground manner, while complying with an intuition that we want to make every bit count by maximizing the information efficiency of our data.

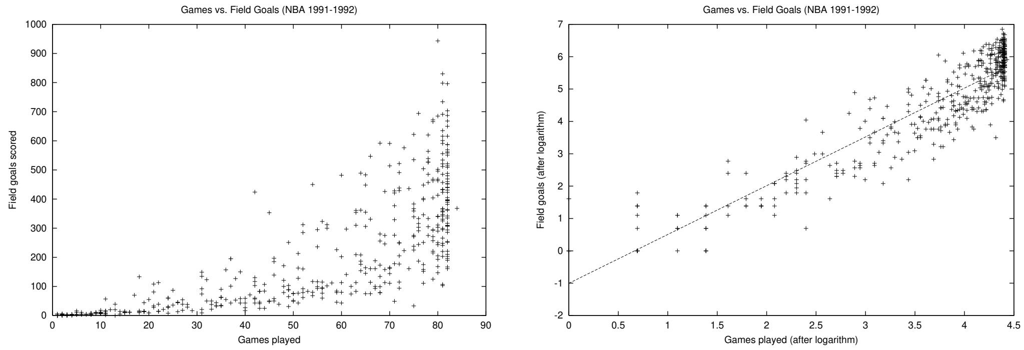
4.1 Fractal dimension

This metric is also related to the fractal dimension – in particular, the D_1 fractal dimension [3, 23]

$$D_1 = \frac{\partial \sum_i p_{i,r} \log_2 p_{i,r}}{\partial \log_2 r} \tag{4}$$

where $p_{i,r}$ is the fraction of pairs of vectors within distance r of each other. Suppose we normalize continuous data so it fits within a unit hypercube. Then, for any given radius r discretize it by dividing that unit hypercube into smaller hypercubes of side length r along each dimension. Let H_r be the entropy of the discretized version with respect to side length r. Then,

$$H_r(x) = -\sum_i p_{i,r} \log_2 p_{i,r} \tag{5}$$



(a) Raw data. (b) Logarithmic axes.

Figure 2: Games played versus field goals scored by players in the NBA 1991-1992 season, (a) raw and (b) scaled, with least-squares linear regression.

As per [20], we can replace 4 with:

$$D_1 = \frac{-\partial H_r(x)}{\partial \log_2 r} \tag{6}$$

$\log_2 r$ is the number of bits of precision per attribute. Hence, D_1 is the derivative of information with respect to precision, or the marginal information content.

The D_1 fractal dimension is but one of a series of generalized fractal dimensions [3], all of which can be estimated in time that scales linearly with cardinality and dimensionality. They have been used as estimators for intrinsic dimensionality [2], another concept which intuitively corresponds to the number of degrees of freedom present within data. That a single number corresponds to both marginal information and to degrees of freedom reinforces our belief that this is a suitable measure for the goodness of a given scale.

4.2 Complete problem definition

Given the MIC metric for uniformity and redundancy, we may now formally define the problem.

Version 1: Axis Scaling. Given a set of n vectors in $\mathcal{R}^m$, produce through invertible transformations a second set of n vectors in $\mathcal{R}^m$, with maximal MIC.

Version 2: Scaling and Reduction. Identical to Version 1, except that we reduce to $\mathcal{R}^{m'}$ where $m' \leq m$. This means that need to retain only independent, high-information attributes.

m' may be chosen in two different ways.

1. m' is supplied by the user.
2. m' is inferred from the data.

We focus on axis scaling and dimensionality reduction with implicitly specified m'.

4.3 Transformations

Now that we have established MIC as the logical metric, we describe how we create attribute spaces to compare. For computational and implementation reasons, we constrain our search to the space of transformation methods in which transformations are applied to individual attributes and without rotation. The MIC will determine the value of different combinations.

Recall that we wish to address skew, mismatched ranges, and redundancy. The first of these problems dictate that we use nonlinear transformations. The second is solvable via simple affine transformations if not already dealt with via nonlinear ones. The third is a bit orthogonal, and is instead handled through dropping attributes.

Below we describe the transformations we chose. This framework by no means requires the same choices we made, nor excludes other invertible transformations; where domain knowledge makes possible better guesses, we would suggest using them. The goal of uniformizing data clearly points towards using cumulative distribution functions (CDFs) of real-world distributions. We thus chose the CDFs of uniform, normal, gamma, lognormal and Pareto distributions, all of which have have common, actual real-world applications [14]. In addition, we allow a natural logarithm transformation, composed with an affine normalization method to alter the range to $[0, 1]$ – the same as that of CDFs.

Additional transformations could be added if one desires. We note that we specifically reject the quantile transformation, as the interpolation demands a potentially ridiculous number of parameters and makes it impossible to meaningfully transform rules from transformed space to the original space. Multi-attribute transformations are theoretically possible, but may be too computationally intensive to be practical.

4.4 Selection

Suppose the data set is $A = \{\hat{a}_i\}$ where each $\hat{a}_i$ is a continuous attribute. Define the transformation set as $T = \{t_j\}$, where $t_j : \mathcal{R} \mapsto \mathcal{R}$. Then, each attribute $\hat{a}$ may be retained in exactly one transformed version $t_j(\hat{a})$, or it may be dropped entirely. This search space scales linearly with both $|T|$ and $|A|$.

We chose forwards-selection search; specifically, a more flexible variation of what was used before [23]. Our algorithm starts with no attributes retained. At each iteration the search may add a transformed attribute; greedily drop a retained transformed attribute and replace it with another; or stop. Additions and exchanges are constrained so that no two versions of the same attribute are ever simultaneously retained. The higher the MIC gain, the better an addition or exchange. When the best addition matches the best exchange, we prefer the exchange as it leaves dimensionality untouched.

If neither additions nor exchanges produces a reasonable gain, then the algorithm terminates. In all our tests, we required a minimum MIC gain of 0.1.

Overall, the time cost is $O(tmnk^2)$, with t transformations, m attributes, n vectors and k retained attributes. Exchanges do not impact performance that much, as currently the drop is greedy and independent of the replacement – it uses two linear searches, rather than a quadratic search through the space of ¡drop,add¿ pairs.

5. EXPERIMENTS

Our experiments were conducted with a focus on answering three questions.

1. What is the impact of our MIC-based selection method on MIC and overall dimensionality?
2. What about the quality of our chosen attributes?
3. Does our method scale well?

The first question will be easy to answer; we need simply present before-and-after comparisons of the attribute counts and MIC values. Likewise, the third presents no particular difficulties where cardinality is concerned; elapsed time will suffice. Answering the second requires some explanation.

5.1 Measuring selection quality

Intuitively, the information that dropped attributes provide should be largely covered by the retained ones. Attributes retained early should be more novel than attributes retained later which should be more novel than dropped attributes.

Consider a retained attribute x. The greedy search imposes an ordering on retained attributes. Let y and z be the two previously selected attributes greedily chosen to maximize relative redundancy $R_r(x|y,z)$,

$$R_r(x|y,z) = \frac{H_r(x) + H_r(y,z) - H_r(x,y,z)}{H_r(x)} \tag{7}$$

Then, the estimated value of x given y and z is the information that is not redundant. In absolute terms, this is $R_r(x|y,z)H_r(x)$. We can define the normalized form novelty $\mathcal{N}_r(x)$ as

$$\mathcal{N}_r(x) = \frac{R_r(x|y,z)H_r(x)}{-\log_2 r} \tag{8}$$

A novelty of 0 means that the attribute is completely explainable in terms of previous attributes – yielding no additional information – and a novelty of 1 means that it provided the maximum possible information for that level of precision.

Non-retained attributes, by which we mean attributes that are not retained in any transformed form at all, are grouped by underlying original attribute. We choose the (x,y,z) triple that maximizes $R_r(x|y,z)$ where x may be any transformed version of the same original attribute, and y and z are two accepted versions of attributes.

For both retained and non-retained attributes, we set r to be $\frac{1}{16}$, resulting in a 4096 total possible discretized outcomes per triple. The larger the cardinality of the data, the more reasonable a smaller, more precise r would be.

Name	Source	Attrs.	Vecs.	Notes
`baseball`	MLB '96	17	365	
`basketball`	NBA '91-92	45	459	1
`CIA-1992`	CIA	2	215	2,3
`CIA-2001`	CIA	2	235	2,4
`machine`	UCI/ML	8	209	5,6
`page-blocks`	UCI/ML	11	5473	5
`synthia`	synthetic	28	5000	1,8
`wine`	UCI/ML	13	178	5,7

Note	Meaning
1	Has related attributes.
2	Area versus population, in km^2.
3	CIA World Factbook 1992[6].
4	CIA World Factbook 2001[7].
5	From the UCI Machine Learning Repository [4]
6	`cpu-performance` database, minus nominal attributes.
7	Minus the class attribute.
8	Generated specifically for this work as three IID Gaussian variables, and 25 linear combinations of cubic polynomials of the Gaussians.

Table 1: Summary of the data used.

	Before		After	
Name	Attrs.	MIC	Attrs.	MIC
`baseball`	17	2.0540	6	3.6987
`basketball`	45	1.6310	8	4.8211
`CIA-1992`	2	0.3160	2	1.5840
`CIA-2001`	2	0.2591	2	1.5724
`machine`	8	0.5858	4	2.1718
`page-blocks`	11	1.6726	5	3.2933
`synthia`	28	2.3918	4	3.0966
`wine`	13	0.9909	4	3.0425

Table 2: Summary of the dimensionality and MIC changes.

5.2 Data

The data sets used are listed in Table 1. Of these, only `synthia` was specifically generated by us subsequent to the design of the algorithm tested.

5.3 Impact on dimensionality and MIC

Table 2 lists the effects of greedily selecting attributes from the transformed versions as previously described. In *all* cases except for the two-attribute sets – the 1992 and 2001 versions of national area and population from the CIA World Factbooks – the number of attributes retained was significantly lower than the number of original attributes. In addition, in *each case* the MIC increased, in most cases significantly. This is possible because MIC-based aims to improve rather than preserve structure.

5.4 Quality of chosen attributes

Here we present graphs illustrating the novelty estimates of each additional attribute. Sets in which all attributes are retained in some form – the two CIA World Factbook

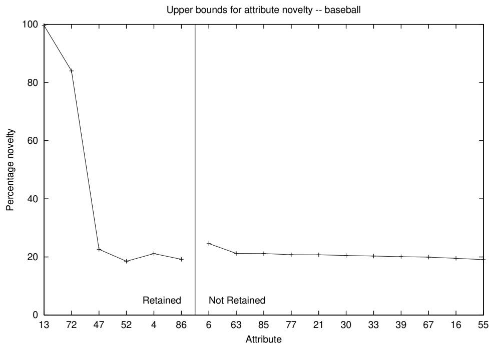


Figure 3: New information in the `baseball` data.

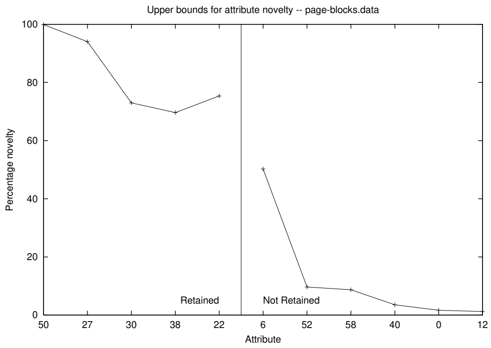


Figure 4: New information in the `page-blocks` data.

sets – are not represented here. For other sets, each graph is as follows. The x-axis is split into two halves; left of the vertical line are retained attributes; right, non-retained. The retained attributes are listed in order of selection, while the non-retained are sorted in order of decreasing value.

Value on the y-axis is $\mathcal{N}$, "percentage novelty", which is the estimated amount of new information they provide versus previously selected attributes, relative to the theoretical maximum based on the number of intervals.

Retained attributes may be surprising, but logical. Figures 3 and 4 show these results. Figure 3 demonstrates a pattern also found in the `basketball` and `wine` sets (graphs omitted for brevity); the first few selected attributes are highly novel, but attribute value rapidly decreases until it levels out – perhaps due to independent noise. In this particular case, `baseball`, the first two statistics to be selected, OBA (opponents batting average) and SO (strikeouts). These choices may be counterintuitive, but when one realizes that these characterize the offensive skills of the teams, they are in fact quite logical choices: the offensive skill of one player has little to do with that of his opponents, but many of the other attributes such as the number of hits will obviously be related to either of these two.

Results for the `machine` set (graph omitted for brevity) are similar, although the novelty trend levels off more gradually. The first two attributes selected from `machine` correspond to the published and estimated relative CPU performance ratings – the goal attribute and a linear regression estimate guess, respectively. These attributes are closely related to the other continuous attributes such as maximum main memory in kilobytes.

Redundant attributes get rejected. Figure 4 shows a pattern in `page_blocks` that is also present for `synthia` (graph omitted for brevity), in which there is a significant gap in value between the selected and non-selected attributes. The non-selected attributes are chosen reasonably. For instance, the most redundant attribute in `page_blocks`, `wb_trans`, is actually completely redundant since it is the ratio between two of the retained attributes, `blackpix` and `mean_tr`.

5.5 Scalability

In addition to testing the quality of results, we also tested the scalability. The synthetic nature of `synthia` allowed us to easily and fairly test for scalability with respect to cardinality. We generated variations of `synthia` with cardinalities of 1000, 2000, 3000 through 10,000. Figure 5 shows time

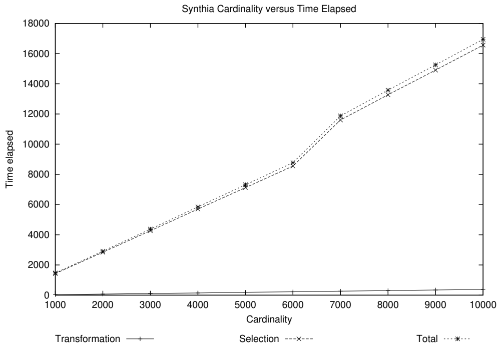


Figure 5: Performance results. Both the transformation and scaling phases scale linearly with cardinality.

required for the transformation phase, the selection phase, and the total of those two. The total clearly scales linearly with cardinality, as confirmed by a correlation coefficient of 0.9958.

6. DISCUSSION

Consider how our algorithm performed. Our experiments were performed to answer questions regarding the algorithm's effect on MIC and dimensionality; whether or not the algorithm chose informative attributes, and dropped redundant or low-content ones; and its overall scalability.

First, as shown in Table 2, reducing the embedding dimensionality is compatible with substantial increases in marginal information content, meaning that with transformations and selection one expects a greater gain per additional bit of precision.

Second, as shown in Figures 3 and 4, using MIC as an attribute selection criterion does appear to prefer attributes that provide more novel, non-mutual information. This is desirable as we wish to capture as much information as we can without succumbing to noise or inefficiently accepting overly redundant data.

Third, the algorithm scales well. Explicit scalability testing on `synthia` variants further showed linear scalability with respect to cardinality. In addition, in real cases it may be possible to use domain knowledge to trim the transformation set, which would increase performance significantly.

7. CONCLUSIONS

Preserving distances and structure is a common theme in principal components analysis, random projection, FASTMAP, and many other methods. However, these and many other tasks make a potentially fatal assumption: that the distances and corresponding structure are already reasonable. When data is distributed in a skewed manner, when axes are badly weighted, when the distance function no longer makes sense – they fail. This is implicit not only in many dimensionality reduction methods, but also other problems – *distance*-based outlier detection, *nearest-neighbor* methods, and so forth.

Instead of ignoring this problem, we have presented an effective, new approach. Instead of preserving distances that we may have little *a priori* reason to trust, we aim at maximizing information via reversible transformations, while reducing embedding dimensionality by discarding redundant attributes. Assessing how close any given solution is to optimal, however, may be well be impossible; instead, we rely on marginal information content (MIC), intrinsic dimensionality, and estimates of attribute worth based on entropy. We simply focus on *making every bit count*, and we do this by improving instead of merely preserving.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grants No. IIS-9910606, IIS-9988876, IIS-0083148, IIS-0113089, IIS-0209107 and by the Defense Advanced Research Projects Agency under Contract No. N66001-00-1-8936.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, DARPA, or other funding parties.

8. REFERENCES

[1] S. Babu, M. Garofalakis, R. Rastogi, and A. Silberschatz. Model-based semantic compression for network-data tables. In *Proc. of NRDM 2001*, May 2001.

[2] S. D. Backer, A. Naud, and P. Scheunders. Nonlinear dimensionality reduction techniques for unsupervised feature extraction. *Pattern Recognition Letters*, 19:711–720, 1998.

[3] A. Belussi and C. Faloutsos. Estimating the selectivity of spatial queries using the 'correlation' fractal dimension. In U. Dayal, P. M. D. Gray, and S. Nishio, editors, *Proc. of 21th International Conference on Very Large Data Bases*, pages 299–310. Morgan Kaufmann, September 1995.

[4] C. Blake and C. Merz. UCI repository of machine learning databases, 1998.

[5] L. Breiman, J. H. Freidman, R. A. Olshen, and C. J. Stone. *CART: Classification and Regression Trees.* Chapman & Hall / CRC Press, 1984.

[6] Central Intelligence Agency, editor. *The World Factbook.* U.S. Government Printing Office, 1992. http://www.cia.gov/cia/publications/factbook/.

[7] Central Intelligence Agency, editor. *The World Factbook.* U.S. Government Printing Office, 2001. http://www.cia.gov/cia/publications/factbook/.

[8] K. Chakrabarti and S. Mehrotra. Local dimensionality reduction: A new approach to indexing high dimensional spaces. In A. E. Abbadi, M. L. Brodie, S. Chakravarthy, U. Dayal, N. Kamel, G. Schlageter, and K.-Y. Whang, editors, *Proc. of 26th International Conference on Very Large Data Bases*, pages 89–100. Morgan Kaufmann, September 2000.

[9] K. Chang and J. Ghosh. Principal curves for nonlinear feature extraction and classification. *SPIE Applications of Artificial Neural Networks in Image Processing III*, 3307:120–129, 1998.

[10] D. DeMers and G. Cottrell. Non-linear dimensionality reduction. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 580–587. Morgan Kaufmann, San Mateo, CA, 1993.

[11] C. Faloutsos and K.-I. D. Lin. Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. *ACM SIGMOD*, pages 163–174, May 23-25 1995.

[12] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

[13] A. Hyvärinen, J. Karunen, and E. Oja. *Independent Component Analysis.* John Wiley & Sons, 2001.

[14] N. Johnson and S. Kotz. *Continuous univariate distributions.* Houghton Mifflin, 1970.

[15] I. T. Jolliffe. *Principal Components Analysis.* Springer-Verlag, New York, 1986.

[16] E. M. Knorr, R. T. Ng, and R. Zamar. Robust space transformations for distance-based operations. In *Proc. of 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2001.

[17] T. Kohonen. The self-organizing map. In *Proceedings of the IEEE*, volume 78, 1990.

[18] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290, December 2000.

[19] Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[20] G. Schuster. *Deterministic Chaos an Introduction.* Verlagsgesellschaft, Weinheim, Germany, 3rd edition, 1995.

[21] C. Shannon. A mathematical theory of communcation. *Bell Systems Technical Journal*, 1948.

[22] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. 290:2319–2322, December 2000.

[23] C. Traina Jr, A. Traina, L. Wu, and C. Faloutsos. Fast feature selection using fractal dimension. Simpósio Brasileiro de Banco de Dados, Oct. 2000.