Optimal density estimation in data containing clusters of unknown structure.

Alfred Ultsch
Databionics Research Group,
University of Marburg,
35032 Marburg, Germany

A method for measuring the density of data sets that contain an unknown number of clusters of unknown sizes is proposed. This method, called Pareto Density Estimation (PDE), uses hyper spheres to estimate data density. The radius of the hyper spheres is derived from information optimal sets. PDE leads to a tool for the visualization of probability density distributions of variables (PDEplot). For Gaussian mixture data this is an optimal empirical density estimation. A new kind of visualization of the density structure of high dimensional data set, the P-Matrix is defined. The P-Matrix for a 79- dimensional data set from DNA array analysis is shown. The P-Matrix reveals local concentrations of data points representing similar gene expressions. The P-Matrix is also a very effective tool in the detection of clusters and outliers in unknown data sets.

1. Introduction

Data mining considers data sets produced by some natural or artificial process. Examples for this are measurements of the number of m-RNA corresponding to expressed genes in cells of living organisms. This is done nowadays using DNA arrays [1]. Such data sets consist of data points in a high dimensional data space. One of the goals of data mining is to discover and describe whether the data producing process operates in different modes. This could be, for the example, different mutants of the organism or different environmental conditions, e.g. aerobic vs. anaerobic.
A mode should identify itself by data points that are members of a cluster in data space. The detection of clusters requires the definition of a meaningful distance measure on the data points. Automatic clustering algorithms, like k-means, or visualization tools, like the U-Matrix [2], use these distances. Distances alone, however are not sufficient to describe clusters properly. Consider, for example, the TwoDiamonds data set depicted in Figure 1. The data consists of two clusters of two dimensional points. Inside each “diamond” the values for each data point were drawn independently from uniform distributions.

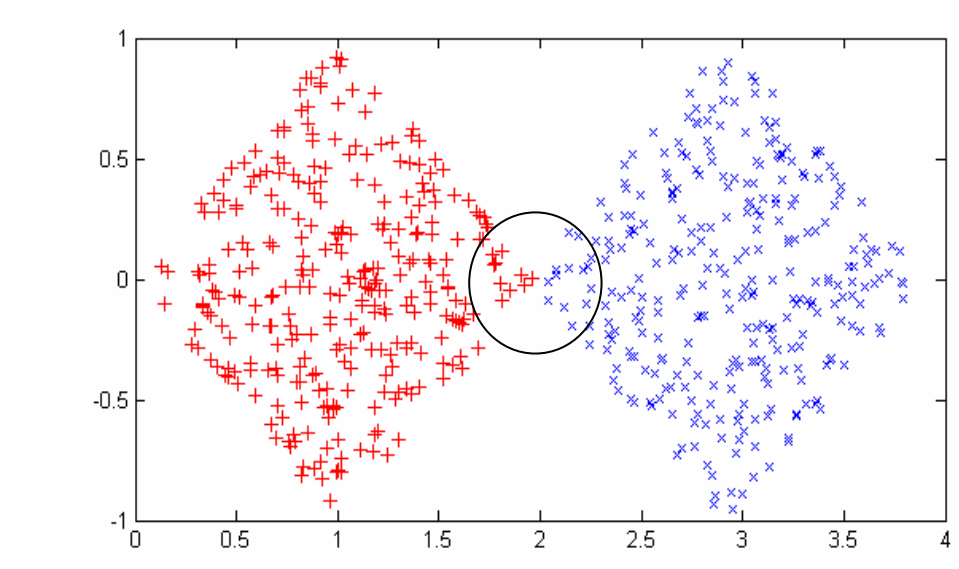

Figure 1: the TwoDiamonds data set

At the central region, marked with a circle in Figure 1, the distances between the data points are very small. For distance based cluster algorithms it is hard to detect correct boundaries for the clusters. Distance oriented clustering methods such as singe linkage, complete linage, Ward etc. produce classification errors. The picture changes, however, when the data's density is regarded. The density at the touching point of the two diamonds is only half as big as the densities in the center regions of the clusters. This information may be used for the clustering of the data. Density based clustering algorithms have drawn much attention in the last years within the context data mining, see for example [3],[4], [5]. All these algorithms call for methods based on the observed data to estimate the density function. In this paper we propose a method for density estimation that is optimal in an information theoretic sense.

The paper is organized as follows: Chapter 2 defines information optimal sets. This is the theoretical foundation of the proposed density estimation. In Chapter 3 the Pareto Number and Pareto Radius are defined. The latter is a kernel bandwidth for density estimation using information optimal sets. In Chapter 4, the intra/inter cluster distance ratios for a wide range of cluster numbers and sizes are determined. Chapter 5 describes the Pareto Density Estimation (PDE) algorithm and gives some hints for a practical implementation. In chapter 6 the data used for the applications of the method are described. The data set consists of DNA arrays for gene expressions in living yeast cells. Chapter 7 describes a PDE based tool for visualization of the probability density distribution of a single variable. In Chapter 8, the usage of PDE for the inspection of density structures in high dimensional data spaces, in the form of a P-Matrix, is described. A P-matrix for the DNA array data is given and it's features are discussed. Chapters 9 and 10 are discussion and summary.

2. Information optimal sets

Let S be a subset of a set of points and p denote the probability that a point belongs to S. The information of the set S can be calculated using Shannon's formula for information. Scaled to the range [0,1], the information of a set I(S) is calculated as I(S) = - e p ln(p) [6]. See Figure 2 for a graph of I(S).

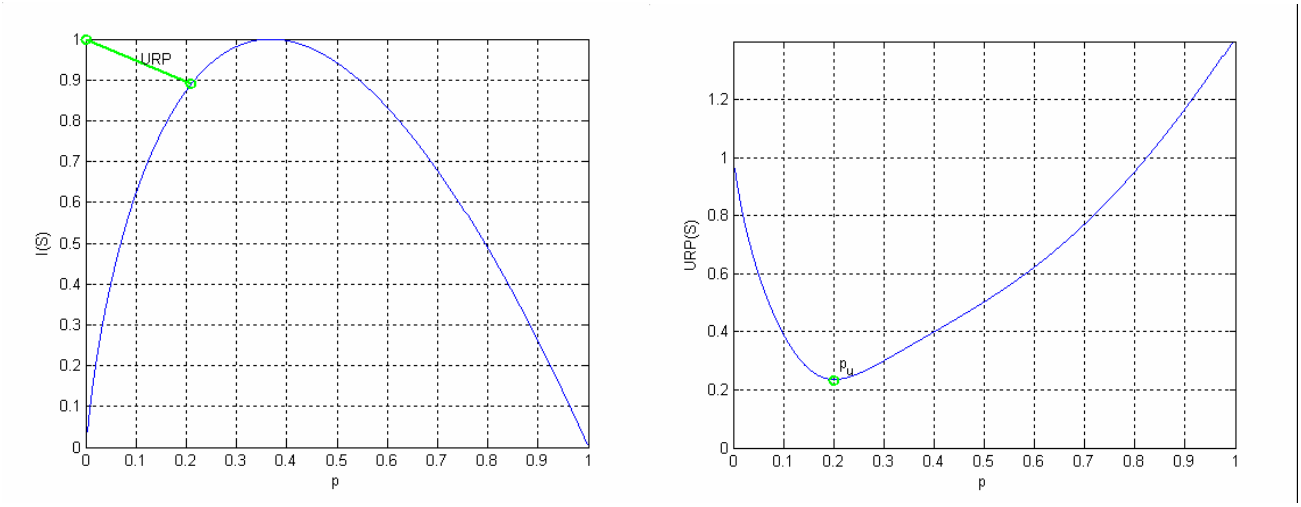

Figure 2: left: Information I(S) and Unrealized Potential URP(S), right: URP(S)

An information optimal set a set is minimal in size but contains as much information as possible. To find such an optimal set size, define the unrealized potential URP(S) of a set S as the Euclidian distance from the ideal point to (p, I(S)) of a given set. The ideal point (0,1) corresponds to a minimal set size producing 100% of information. This definition of URP(S) results in the following formula for URP[7]:

$$\mathrm{URP(S)} = \sqrt{p^2 + (1 + e \cdot p \cdot \ln(p))^2}\ .$$

URP(S) can be seen in the left side of Figure 2 as the length of the line starting at point (0,1) and ending at some point on I(S). Minimizing the unrealized potential URP results in an optimal set with p_u = 20.13%. This set size produces 88% information. Subsets of the relative size p_u are called *information optimal*. The optimality of this set at about (20%, 80%) (see Figure 2 right side) can serve as an explanation for the so called Pareto 80/20 law, which is empirically found in many domains [8]. For a detailed discussion see [9].

3. Pareto Radius and Pareto Numbers

This chapter defines the basic notation for a data density estimation algorithm using information optimal sets. We assume that the process, which generates the data, may be described with n dimensional real valued vectors.

data space: $D \subset R^n$: the subspace of R^n where data points can be observed in principle.

data set: $E = \{x_1, \ldots x_d\}$ with $x_i \in D$ the collected data.

Clustering requires the definition of a meaningful distance measure:

data distance: distance measure defined in the data space $D \times D \rightarrow R^+$: x, y $\in E$, d(x,y) ≥ 0

Define the *neighbourhood number* NN(x,r) as the number of input data points within a hypersphere (neighbourhood) with radius r around a point x in data space :

$$\mathrm{D \times R^+ \rightarrow R^+}: NN(x,r) = \left|\{n \in E \mid d(x,n) \leq r\}\right|.$$

The radius r used to calculate NN(x,r) is called the *neighbourhood radius*.

The density structure of a data set can be arbitrarily complex. Even if the input is drawn from a Normal distribution, the neighbourhood numbers are not normally distributed. For data drawn from a standard normal distribution N(m,s) the neighbour numbers can be precisely calculated as follows:

$$\mathrm{NNnorm(x,r)} = \mathrm{errf(x+r) - errf(x-r)},\ \text{where}\ \mathrm{errf(x)} = \int_{-\infty}^{x} \mathrm{N(m,s)}.$$

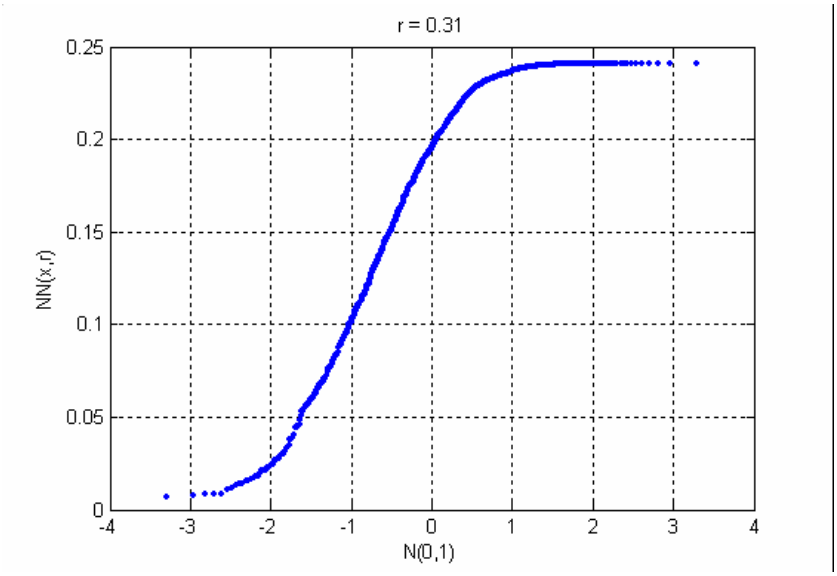

Figure 3: QQ-plot of the neighbour numbers in N(0,1) distributed data.

Figure 3 shows a quantile/quantile plot (QQ-plot) of NNnorm vs. N(0,1) for a neighbourhood radius r=0.31. As expected, the distribution of NNnorm(x,r) is not normal. For data points close to zero the neighbourhood contains on average 23% of the data points. For larger positive and negative points, the neighbourhood numbers decrease following approximately a normal distribution. At the "thin" data space regions the distribution reaches a constant level. In statistics

this S- shaped line in a QQ-plot is associated with distributions having “fat tails”(Durbin 00). The median of NN(x,0.31) is 0.20 whereas the mean is 0.18.

If no further information about the density structure of a data set is given, we propose to use a neighbourhood radius based on information optimality. The neighbourhood radius is chosen such that the average neighbourhood number equals the information optimal set size. As seen in the example above, the density distribution is not normal. We therefore use the median instead of the mean for the calculation of an average in neighbourhood numbers. This leads to the following definition of the Pareto Radius.

The *Pareto Radius* $r_p \in R^+$ of a data set is a neighbourhood radius such that

$$\forall\ x \in E, |E| = d : median(NN(x, r_p)) = p_u \cdot d \text{, with } p_u = 0.2013$$

The set of neighborhood sizes $\{n_i \mid n_i = NN(x_i, r_p), x_i \in E\}$ using the Pareto Radius are called *Pareto Numbers*. The definition of the Pareto Radius assures that it leads to average neighbourhood numbers close to the information optimal set size.

In practical data mining applications it is, however, impossible to calculate neighbourhood numbers for all possible neighbourhood radii in order to find the Pareto Radius. Searching among the distance percentiles of the data is a useful way to limit the effort to approximate the Pareto Radius. Furthermore, percentiles are also useful for considering the data’s cluster structure (see the next chapter). Let pc(p) denote the p-th percentile of the distances between two different points in the data set. The *Pareto Percentile* p_{par} is that percentile of all distances which is closest to the Pareto Radius i.e.

$$p_{par} = \arg\min(|\ pc(p) - r_p|),\ \forall\ p \in 0,1,\ldots 100.$$

It has been shown that Pareto percentiles lead to a density estimation that is closest to the true data density for data consisting of Gaussian mixtures [9]. To limit the effort for very large data sets sampling techniques for the estimation of the distance percentiles can be used.

4. Intra/inter Cluster Distances

A primary goal of data mining is to identify coherent clusters as subsets of the data It is usually not known how many clusters are in the data and what their prior probability is. It is, however, reasonable to assume that the data is collected in such a way that there should be enough data sets for each of the processes’ possible operation modes. For DNA array data, for example, enough data points for the different environmental conditions of cell experiments have to be collected.

Ideally, if the data producing process operates in k modes, the input data should contain k sets of equal size, each containing data from one mode of the process. In particular in the biological domain, some of the modes might be rare events. E.g. a mutant or cancerous type of a cell might be a rare case. In such situations very uneven cluster sizes are to be expected.

Following our line of reasoning the optimal radius for density estimation is the Pareto Radius within each cluster. Since it is, however, not known a priori, how many clusters there are and what their sizes are, this approach is not feasible for data mining. Furthermore to use a different neighborhood radius for each cluster would require volume calculations in high dimensional spaces in order to obtain comparable neighborhood numbers. We follow here the approach to calculate the Pareto percentile for the whole data set and modify this percentile with an estimation of the intra vs. inter cluster distance ratio.

The detection of the cluster structure in an input data set using the distance structure is only possible, if most of the data distances within a cluster (intra distances, d_{intra}) are smaller than the distances measured between data from different clusters (inter distances, d_{inter}).

Let v denote the ratio of intra cluster distances to inter distances:

$$v = \frac{d_{Inner}}{d_{Intra}}$$

If this ratio is known a priori the neighborhood radius for the estimation of a suitable data density can be adapted. If p_{par} is the Pareto percentile in the whole data set and v the ratio of inter/intra

distances. Then the Pareto percentile within a cluster can be calculated as follows:

$$p = \frac{p_{par}}{v}.$$

To estimate v for unknown number of clusters and sizes, experiments for a wide range of cluster numbers and sizes were performed. Data set size was set to d =1000 points. For the number of clusters k within the range [1, 40] the relative size p_i of cluster i was randomly drawn from a normal distribution N(m,s) with mean m =k^{-1}, and s = 10. The variance was chosen so large to generate in particular very uneven cluster sizes. Each cluster was required to consist of at least one single data point. For each of the cluster numbers k, 10.000 cases of cluster sizes were generated and the ratio v was calculated. The mean values of $\bar{v}(k)$ for each number of clusters k are given in appendix A. The 95% confidence interval for v was calculated. This is the interval in which the values for v can be found with at most an error probability of 5%. Figure 4 shows $\bar{v}(k)$ and the 95% confidence interval versus the cluster number k.

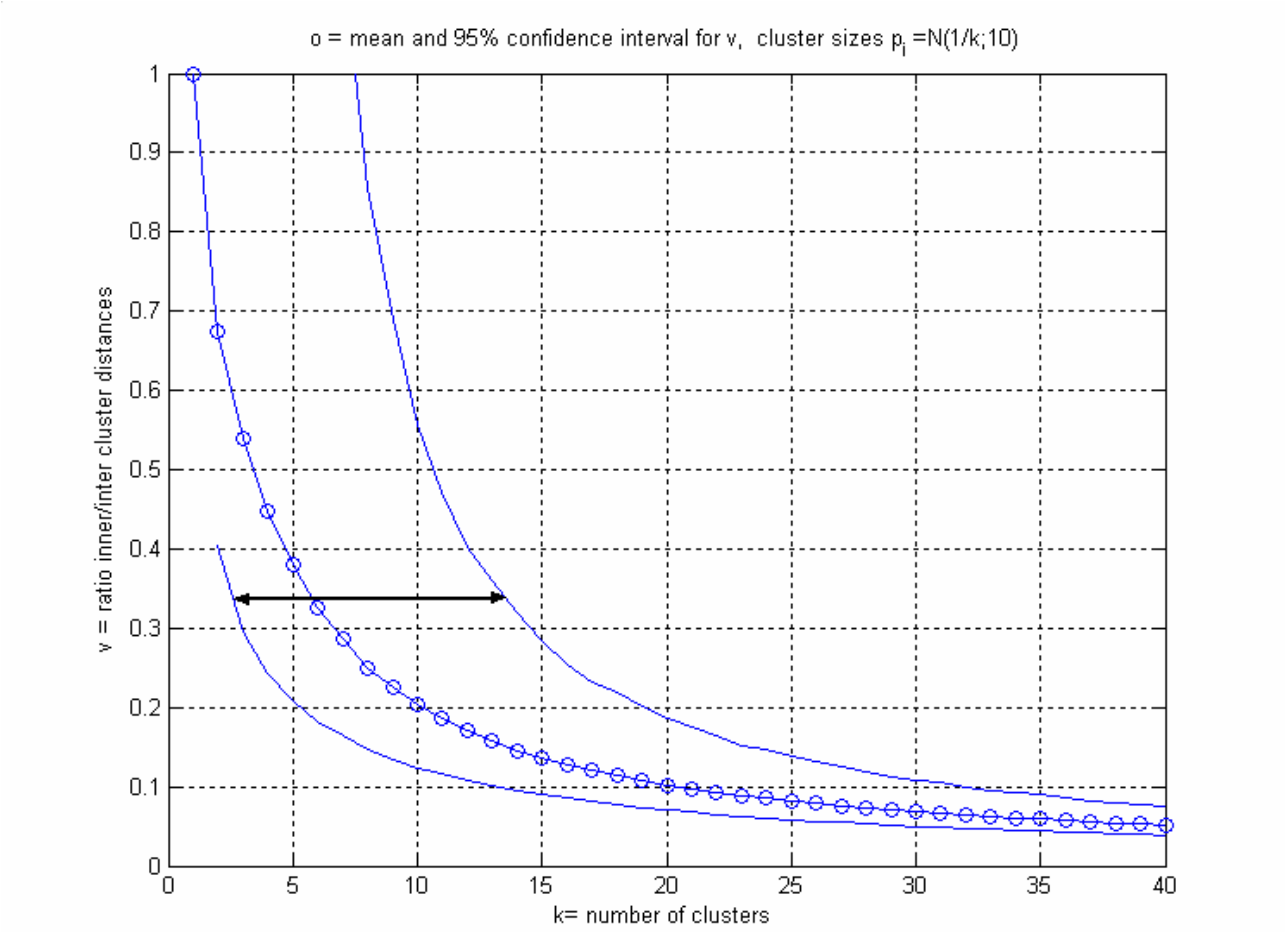

Figure 4: ratio of intra/inter cluster distances

For a given value of v it's confidence interval covers a broad number of clusters. For example a value of v = 0.33 lies within the confidence interval of 3 to 13 clusters. See the arrows in Figure 4.

Different experimental setting for variance s in the range of s ∈ [1,30] and set sizes d ∈ [50, 5000] produced results that were within pen point size equal to the results of Figure 4. So we conclude that $\bar{v}(k)$ is a robust estimation for an initial guess of the intra/inter distance ratio for data mining on typical input data sets containing about k clusters.

5. Pareto Density Estimation

In order to find a suitable radius for density estimation for data mining in data sets containing clusters the following procedure is proposed:

Pareto Density Estimation Algorithm

For a given input data set E containing d data points $\{x_1,\ldots x_d\}$:

1) Search for the Pareto percentile p_{par} among the data set's distance percentiles. This is the distance percentile where the median of all corresponding neighbourhood numbers is closest to the optimal information set size $d_u = 0.2013 * d$.
2) Estimate, the intra/inter cluster distance ratio v. (See below)
   Use the distance corresponding to the percentile $p_u = \text{round}(v * p_{par})$ as the empirical Pareto Radius r_p .
3) The neighbourhood numbers $N(x, r_p)$ for $x \in D$ are then called a Pareto Density Estimation PDE(x).

If the number of clusters is not known, an initial value of v= 0.33 is a meaningful starting point for data mining. This value of v can be typically found in data sets containing from 3 to about 13 clusters. See the arrows in Figure 4.

If the number of clusters is known to be k, v in step 2 of the PDE algorithm can be taken as $\bar{v}(k)$ (see Appendix A).

If there are only one or two clusters in the data set $v_{est} = 0.7$ is used.

In case the minimum number of clusters in the data set is known, the lower of the 95% confidence interval boundaries is a good choice for v. For example, if it is known that there should be at least 13 clusters in the data, $v_{est} = 0.1$ can be taken from Figure 4. If upper bounds on k are known the upper bounds of the confidence interval are used

If k is large (k>40), the empirical Pareto Radius converges to the 1-percentile pc(1) of the data distances.

In large input data sets, it might be costly to calculate the neighbourhood numbers for all distance percentiles. To estimate where the Pareto percentile is usually found, we have calculated the Pareto distance percentiles for set sizes between 50 and 1000 points and dimensions between 1 and 1000. Table 1 gives the Pareto distance percentiles for set sizes between 50 and 1000 points and dimensions between 1 and 1000.

	Dimension of the data points					
Points in input data	1	3	10	50	100	1000
50	19	22	23	22	23	25
100	18	20	21	23	24	25
500	19	21	23	24	23	24
1000	18	21	22	23	23	24

Table 1: Pareto distance percentiles for Gaussian distributed data

For this type of data the Pareto percentile is within the range of [18…25]. If the cluster number is unknown, a percentile in the range 0.33 [18…25], this is the 6-th to18-th percentile of distances, could be taken as an approximation of a suitable Pareto Radius. Furthermore, the 20 percentile is a good starting point for the search for an empirical Pareto percentile.

Determining the Pareto Radius from the distance distribution of a data set takes the particularities of the distances structures of the data into account. Our experience with quite different data sets led us to the conviction that any priori assumptions about distributions and cluster distributions are invalidated by the praxis. The procedure outlined above is a good starting point for gaining an unbiased impression of the structure of given data sets. If more insight is gained on the data set, the Pareto Radius can be calculated more precisely.

6. DNA Array Data

A fundamental problem in data mining is to estimate the distribution of the variables in the data set. Histograms are often used in order to get an impression on the distribution of singular variables of the data. Figure 5 shows a histogram for DNA array data. The data set stems from a collection of DNA microarray hybridization experiments [10]. Each data point represents the logarithm of base 2 of the ratio of expression level of a gene under experimental conditions (LogRatio data). The expression levels are measured as intensities of two fluorescent colours (Cy3 and Cy5). The data consists of a set of 2465 gene expressions of yeast. The genes were selected by Eisen et al [10]. The data is available from the web site "http://www-genome.stanford.edu".

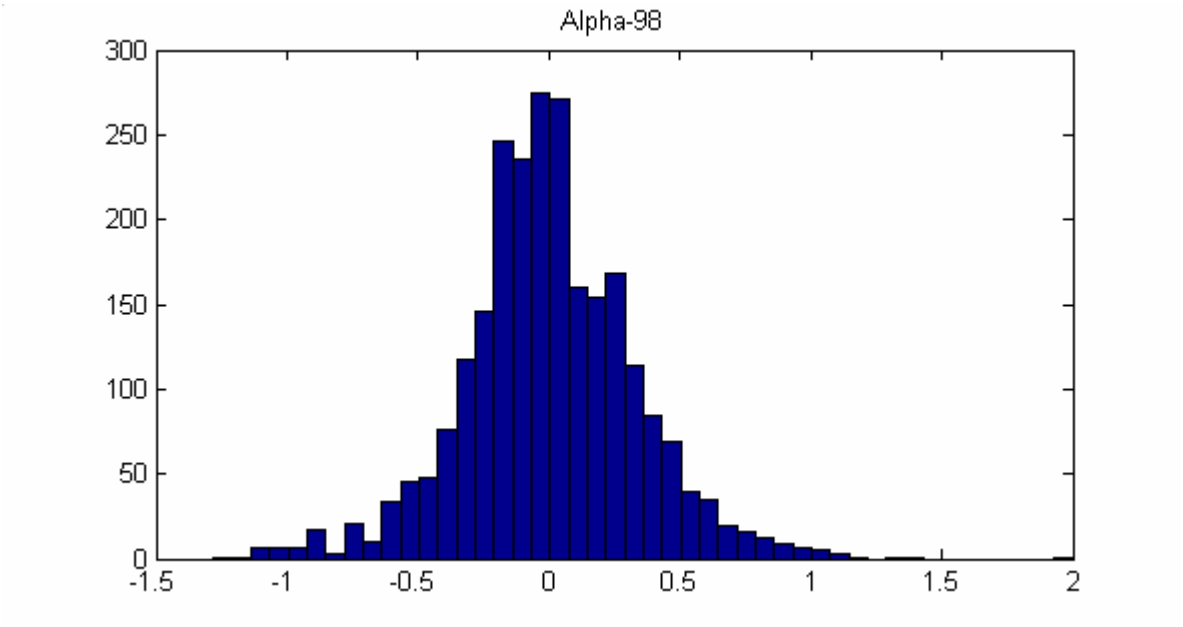

Figure 5: Histogram using an optimal number of bins

As it is expected in this type data, most of the data is concentrated around zero, corresponding to an average expression level of a gene. For pre-processing such as filtering, denoising, normalization and scaling of the data it has to be assumed that most of the data is normal distributed around zero. A QQplot of the data versus a normal distribution seems to confirm this hypothesis. See the following Figure 6. We want to demonstrate in the following that N(m,s) where m is the measured mean and a the empirical variance is not a good approximation for this distribution.

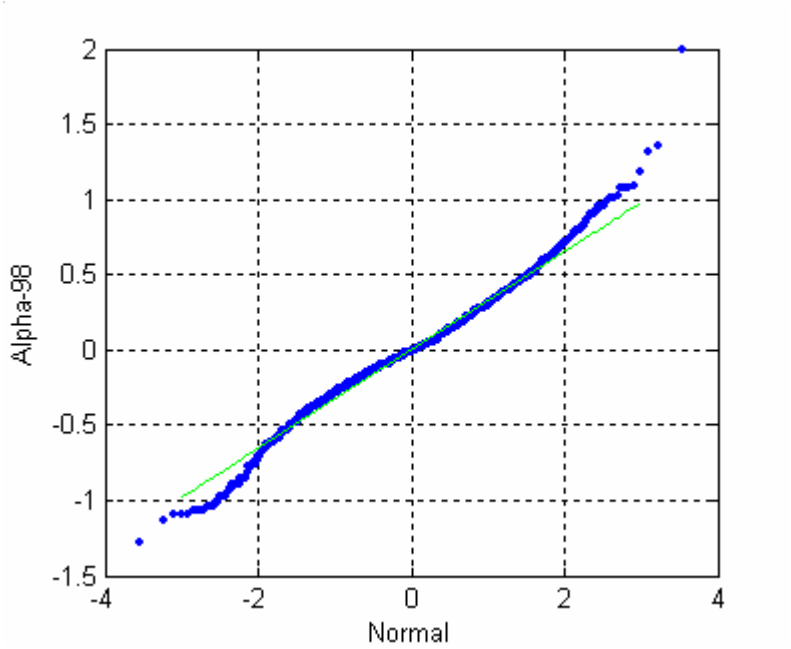

Figure 6: Q/Q Plot of a DNA array variable against the Normal distribution.

7. PDE plots

If only one variable of the data set is considered PDE(x) can be normalized to be an estimation of the probability density function of the variable. This chapter introduces a tool that visualizes the probability density distribution of a variable more accurately than histograms.

For Gaussian distributed one dimensional data it is has been demonstrated that a density estimation using the PDE leads to the best density estimation for data containing clusters [9]. This density estimation is appropriate, even if the clusters overlap to a large extent. Differences in variance in the clusters are also tolerated (see [9] for details).

For the visualization of the density function of one variable therefore the 18 distance percentile is used as Pareto Radius. In order to calculate an estimation of the variable's probability density function the Pareto Numbers have to be normalized. Define the Pareto Probability Density Estimation function PPDE(x) as follows

$$\text{PPDE}(x) = \frac{\text{NN}(x, r_p)}{\text{area}} \quad \text{where area is} \quad \int_{-\infty}^{\infty} \text{NN}(x, r_p)\, dx \quad \text{(i)}$$

The denominator 'area' is approximated using the trapezoidal method on $(x_i, \text{NN}(x_i, r_p))$. The formula (i) assures that the integral on PPDEplot(x) is equal to 1 to get a valid probability density function. Plotting PPDE(x) against x is the PDEplot. Figure 7 shows a PDE plot of the DNA array data together with N(m,s) (dashed line).

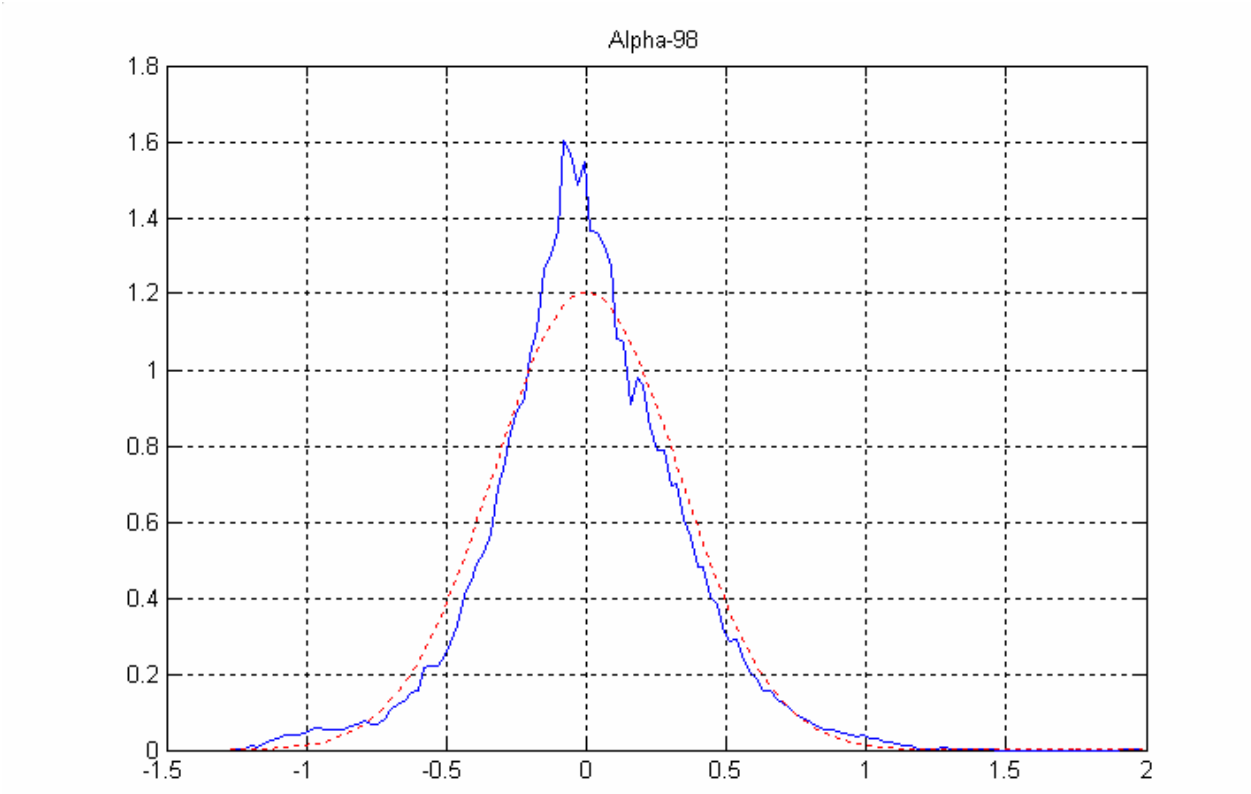

Figure 7: PDE plot of DNA array data and empirical Gauss

Compared to the histogram in Figure 5 and the QQplot of Figure 6 this allows a better analysis of the variable's distribution. The PDEplot reveals that N(m,s), the empirical Gauss, is not a good model of the data. The actual data concentrates more around a central point than the empirical Gauss. At both sides of the distribution there is substantial more data than in a normal distribution These data points have in fact a biological meaning. They belong to genes that are over- respectively under- expressed relative to the normal gene expression level. This means that there is

an over- or under population of the corresponding proteins in the cell. This is the important information sought in DNA array experiments.

PDEplots reveal that for DNA array data the empirical variance is not a good estimation for the true variance of the central distribution. A better solution is to use a robust measurement for the empirical variance [13]. This estimation derives the variance from the inter quartile range as follows:

$\hat{s} = \min(s, \frac{IQR}{1.349})$ with the empirical standard deviation s and the inter quartile range IQR.

See [13] for a derivation of this formula. The PDEplot of the data vs. this distribution $N(m, \hat{s})$ is shown in Figure 8. It can be seen, that this Gaussian leads to a better description of the central distribution of the data.

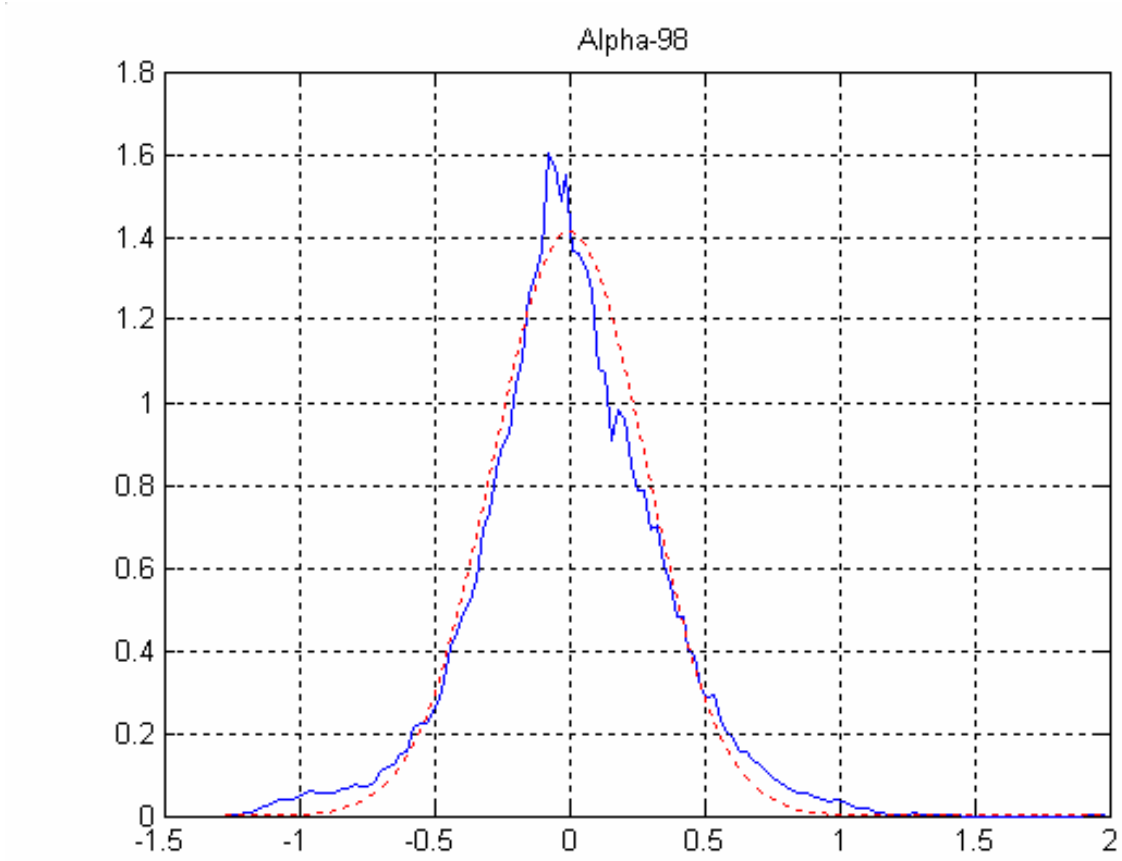

Figure 7: PDEplot for DNA data and $N(m, \hat{s})$

In particular the regions that are of special interest for DNA array data, the "fat tails" of the data's distribution can be seen more clearly. For this and other distribution models of the data a PDEplot allows a better judgment on the quality of the approximation than histograms or QQplots.

An implementation of PDEplots as MATLAB routine may be obtained from the author (http://www.mathematik.uni-marburg.de/~databionics/)

8. The Visualization of Density Structures in high dimensional Data

The U-Matrix is the canonical tool for the display of the distance structures of the input data on self organizing feature maps (SOM) [12]. The U-matrix shows a "landscape" of the distance relationships of the input data in the data space. Properties of the U-Matrix are:

- the projections of the data on the neuronal map (bestmatches) reflect the topology of the input space, this is inherited from the underlying SOM algorithm
- weight vectors of neurons with **large** U-heights are very distant from other vectors in the

data space
- weight vectors of neurons with **small** U-heights are surrounded by other vectors in the data space
- bestmatches are typically found in depressions
- outliers in the input space are found in „funnels“.
- “mountain ranges“ on a U-Matrix point to cluster boundaries
- „valleys“ on a U-Matrix point to cluster centers

The right side of figure 8 shows a U-Matrix for the Hepta data set depicted at the left side.

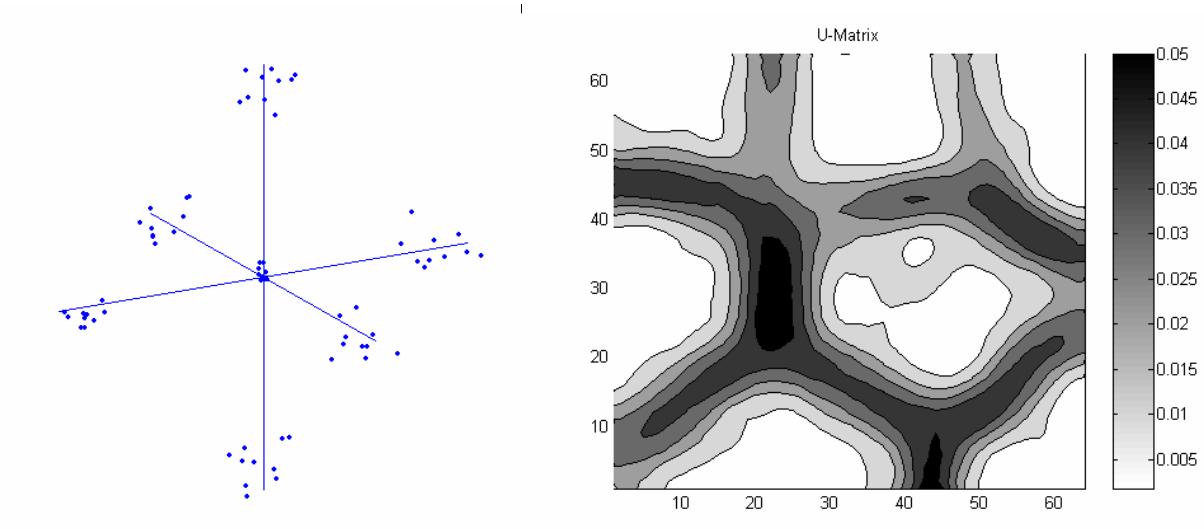

Figure 8: The Hepta data set and it’s U-Matrix

The central cluster of the Hepta data set posesses a much higher density than the other clusters. This is not shown on the U-matrix.

In [7] the P-matrix has been introduced. It consists of a display of the Pareto Numbers associated with each weight vector of the mapping space of a emergent SOM (ESOM). See [7] for details. The Pareto Numbers are drawn as height values. This gives the following P-Map for the hepta data set.

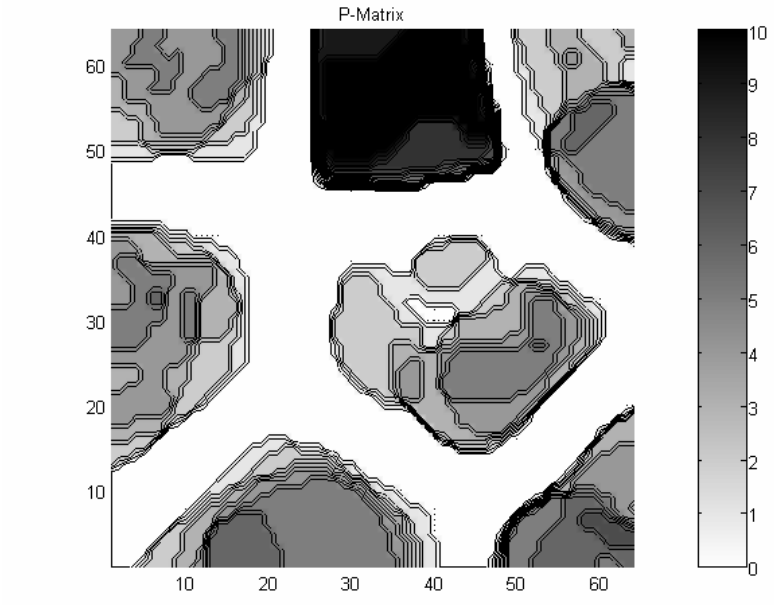

Figure 9 : P-Matrix for the Hepta data.

The P- matrix shows the larger density of the central cluster. P-matrices can be used to gain helpful

insights into the otherwise inaccessible density situations of high dimensional data sets. We demonstrate this below on the DNA array data.

Properties of a P-Matrix are:

- the positions of the data points images on the SOM reflect the topology of the input space
- neurons with **large** P-heights are situated in dense regions of the data space
- neurons with **small** P-height are "lonesome" in the data space
- outliers in the input space are found in „funnels".
- "ditches" on a P-Matrix point to cluster boundaries
- „plateaus" on a P-Matrix point to cluster centers

One can see, that many, but not all, properties of the P-matrix are the inverse of the U-Matrix. In contrast to the U-Matrix , which visualizes the distance structure of the data space, the P-Matrix visualizes the data's density structure. This gives a new and complementary insight into the high dimensional data space. To illustrate this, we show a picture of a P-Matrix for the DNA-array data (figure 10).

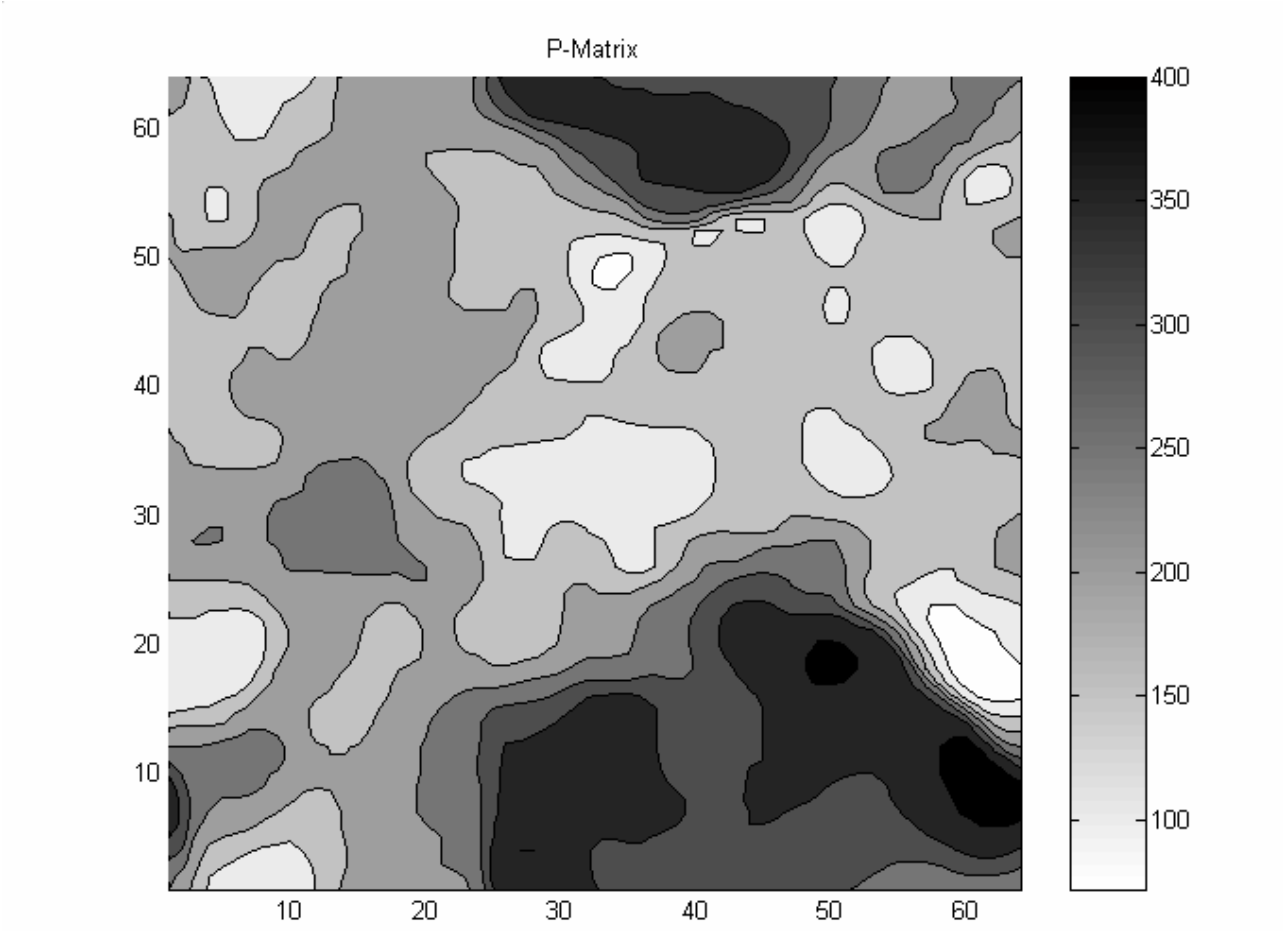

Figure 10: P Matrix of DNA data

The data set contains 2465 data points of dimension 79. The ESOM mapping preserves the data's distance topology. On the P-Matrix it can be seen that a substantial subset of the data points are mapped to locations where there is a big concentration of points. Compare the dark regions in Figure 10. There, the neighbourhood numbers are around 400. Other regions, distant from the first have also a local density maximum of more than 250. This points to possible cluster structures. Some regions on the ESOM are also very under-populated. This is an indication for "outliers", i.e. singular special situations in the data set.

P-Matrices can also be used to enhance the visibility of cluster borders in an U-Matrix and to detect clusters in data sets. This is described in [14]. Pareto Density Estimation, PDEplots for one dimensional data and the construction of P-matrices for high dimensional data have been implemented as MATLAB® routines. These routines may be obtained from the author (see http://www.mathematik.uni-marburg.de/~databionics/).

9. Discussion

Neighbourhood numbers are uniform kernel estimates. Uniform kernel estimates can approximate the true density up to any desired degree of accuracy, if the true density is known [14], [15]. In general, uniform kernel estimates tend to overfit the sample for small bandwidths (=radius) and oversmooth the estimation for large bandwidths. See the figure 11 for this effect.

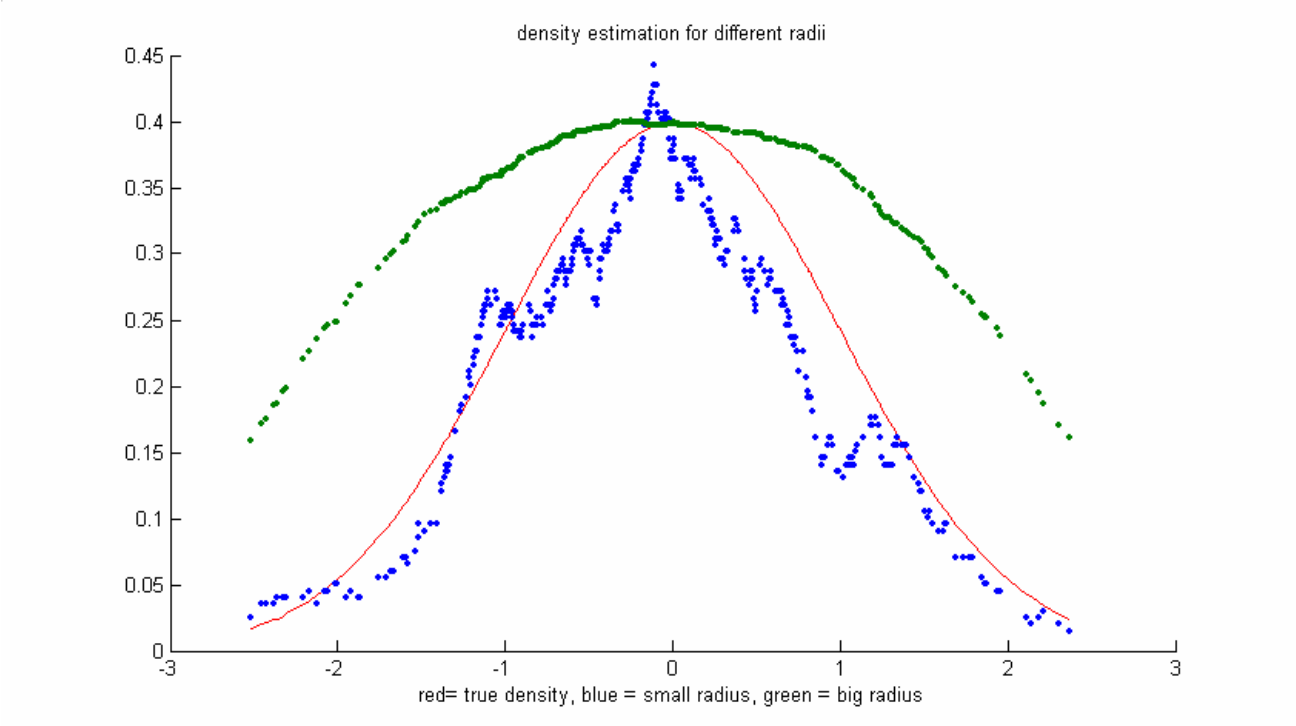

Figure 11: The effect of kernel bandwidth

Tuning the bandwidth for optimal variable kernel estimation is computational expensive and proven to be an exceptional hard task [15]. This is a clear disadvantage of such methods for large data sets. This might be one of the reasons why uniform kernel methods have become in particular popular for data mining ([3], [4],[5]).

Since the actual density distribution of a data set in data mining is unknown and may be arbitrarily complex it is futile to aim at the perfect density estimation in terms of approximation of a supposedly “true” density estimation. This might be done at a later stage in data mining, when tested hypotheses about the true data density exist. Pareto Density Estimation relies instead on information optimality. PDE shares with all uniform kernel estimates the property that thin data regions are overestimated. Since it is sometimes cumbersome to collect sufficient data for rare modes of the data generating process, this might be tolerable for data mining.

For the central parts of uniform distributions the Pareto Radius leads to an average of 20% of the data points of a set in the neighbourhood hypersphere. It can be expected that with a smaller kernel bandwidth more of the fine structures of the density distributions are discovered. For a first goal, the identification of very dense and very thin regions in the data space indicating the location of (large) cluster centers and their borderlines, the Pareto Radius is an appropriate choice.

We propose to use an estimation on the ratio of intra to inter cluster distances for the calculation of an empirical Pareto Radius for a data set containing clusters. The estimations are derived in this work from a large number of randomly generated cluster sizes. It is clear that for every practical situation the de facto ratio may be quite different. With this approach, however, partial knowledge, such as the minimal, maximal or average number of clusters can be incorporated in the calculation.

The more information about the true number of clusters is available the better the Pareto percentile can be adapted to the cluster structure of the data using the factor v. The broad limits of the confidence intervals for v suggest, that the choice of this factor is not critical (see Figure 4).

It is known that histograms have to be regarded critical for the visualization of a variable's probability density estimation [11]. QQ-plots allow to judge whether a transformation to the data should be applied in order to reach a known type of distribution. It has been demonstrated elsewhere that probability density estimation with PDE gives the smallest errors compared with the true probability density function for Gaussian data. In particular this has been showen for Gaussian data containing clusters that overlap and have different variances. Using a PDEplot for the visualization of a variable's probability density is therefore better to reveal the true structure of the distribution than histograms.

We are not aware of any approaches for the visualization of density for the high dimensional data sets that consider both the distance and the density relationships in the data. The P-Matrix is therefore a unique tool for the inspection of high dimensional data sets.

10. Summary

One of the goals of data mining is to discover clusters in empirical data. Distances are a prerequisite for the detection of clusters, but sometimes not enough for an automatic clustering. Data density is an alternative viewpoint on the data leading often to better cluster definition. The combination of both methods is hardly attempted. In this work a method for an efficient measurement of data density is presented. Pareto Density Estimation (PDE) is a method for the estimation of density functions using hyper spheres. The radius of the hyper spheres is derived from information optimal sets. The construction of the PDE from an empirical data set takes in particular into account that there might be an unknown number of clusters of also unknown size in the set. Starting at an educated initial guess the information on clusters discovered during the process of data mining can be employed in the method. A tool for the visualization of probability density distributions of variables, the PDEplot is defined. The usefulness of this tool is demonstrated on DNA array data. The visualization guides the search for better models for empirical distributions for this type of data.

The usage of PDE to visualize the density relationships of high dimensional data sets leads to so called P-Matrices which are defined on the mapping space of emergent self-organizing maps (ESOM). A P-Matrix for the 79-dimensional DNA array data set is shown. The ESOM mapping preserves the data's topology. The P-Matrix reveals local concentrations of data points. This is a very useful tool in the detection of clusters and outliers in unknown data sets.

Acknowledgements

Special thanks thel members of the Databionics Research Group for their comments on this paper.

11. References

[1] J.Quackenbush, Computational Analysis of Microarray Data, Nat. Rev. Genet. 2, 418 -427, 2001

[2] S. Kaski, J. Nikkila, and T. Kohonen. Methods for interpreting a self-organized map in data analysis. In Proc. 6th European Symposium on Artificial Neural Networks (ESANN98). D-Facto, Brugfes, Belgium, 1998.

[3] M.Ester, H.-P. Kriegel, J. Sander, and X.Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD), 226-231, 1996.

[4] X. Xu, M. Ester, H.-P. Kriegel, J. Sander. "A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases", Proceedings of the ICDE Conference, 1998.

[5] A.Hinneburg, Keim,D.A,An Efficient Approach to Clustering in Large Multimedia Databases with Noise', Proc. 4rd Int. Conf. on Knowledge Discovery and Data Mining, AAAI Press, 1998.

[6] C.E.Shannon , A Mathematical Theory of Communication, The Bell System Technical Journal, Vol 27, pp 379-423, 1948

[7] A.Ultsch, Maps for the Visualization of high-dimensional Data Spaces,Proc. Workshop on Self Organizing Maps WSOM03, pp 225-230,2003

[8] J.M.Juran, Pareto, Lorenz, Carnot, Bernoulli, Juran and Others, Industrial Quality Control, October 1950, p. 25

[9] A.Ultsch, “The reasons behind Pareto’s 80/20 law and limits for an ABC analysis (in German), Technical Report, Nr 30, Department of Computer Science, Univerisity of Marburg, 2001

[10] M. Eisen, P. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression´patterns. Proceedings of the National Academy of Sciences of the United States of America, 95:14863–14868,1998.

[11] Keating J.P., D.W. Scott: A Primer on Density Estimation for the Great Home Run Race of 98 in Stats #25, 16-22, 1999

[12] J. Vesanto et al.,”Self-organizing map in matlab: the SOM toolbox”, Proceedings of the Matlab DSP Conference, pp 35--40, Espoo, Finland, November, 1999

[13] A. Ultsch, U*Clustering: automatic clustering on Emergent Self Organizing Feature Maps, Technical Report, University of Marburg, 2003

[14] L.Devroye, G.Lugosi, A universally acceptable smoothing factor for kernel density estimation, Annals of Statistics, vol. 24, pp. 2499– 2512, 1996.

[15] L.Devroye, G.Lugosi, Non-asymptotic universal smoothing factors kernel complexity and Yatracos classes, Annals of Statistics, vol. 25, pp. 2626–2637, 1997.

Appendix A

Mean intra vs. inter cluster distance ratio measured in 10.000 randomized experiments for each cluster number.

number of clusters	mean of ratio intra/inter cluster distances
1	1
2	0,673655
3	0,540071
4	0,448394
5	0,380795
6	0,3263
7	0,286768
8	0,25035
9	0,225546
10	0,202865
11	0,185515
12	0,170194
13	0,157708
14	0,145808
15	0,136427
16	0,127417
17	0,119922
18	0,113884
19	0,10755
20	0,102236
21	0,097573
22	0,09319
23	0,088828
24	0,085517
25	0,081595
26	0,078621
27	0,075995
28	0,072974
29	0,070437
30	0.101487
31	0.097475
32	0.09462
33	0.091768
34	0.089551
35	0.08696
36	0.084779
37	0.082454
38	0.080437
39	0.078333
40	0.076277