

# Pareto Density Estimation: A Density Estimation for Knowledge Discovery

Alfred Ultsch

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

**Abstract.** Pareto Density Estimation (PDE) as defined in this work is a method for the estimation of probability density functions using hyperspheres. The radius of the hyperspheres is derived from optimizing information while minimizing set size. It is shown, that PDE is a very good estimate for data containing clusters of Gaussian structure. The behavior of the method is demonstrated with respect to cluster overlap, number of clusters, different variances in different clusters and application to high dimensional data. For high dimensional data PDE is found to be appropriate for the purpose of cluster analysis. The method is tested successfully on a difficult high dimensional real world problem: stock picking in falling markets.

## 1 Introduction

Density based clustering algorithms have drawn much attention in the last years within the context of knowledge discovery in databases (Ester et al (1996), Xu et al (1998), Hinneburg and Keim (1998)). All these algorithms rely on methods to estimate the probability density function from the observed data. Methods for density estimation have been studied intensively in mathematics and statistics. Density estimation using the number of points within a hypersphere of a fixed radius around each given data point is used in many of the density clustering algorithms. In this paper we propose a radius for hypersphere density estimation that is optimal in an information theoretic sense. Information optimization calls for a radius such that the hyperspheres contain a maximum of information using minimal volume. Consequence of this approach is that a radius is optimal, if a hypersphere contains in the average about 20% of the data. This gives more than 80% of the possible information any subset of data can have. Since these results coincide with the widely known rule of thumb called "Pareto's 80/20 law" we decided to call this radius the Pareto radius. The hypersphere density estimation method using this radius is called Pareto Density Estimation (PDE). We show in this paper that PDE is an optimal hypersphere density estimation method for data with a mixture of Gaussians as probability density function. It turns out that the method is valid even when the clusters overlap to certain degrees and when the inner cluster variances differ. Furthermore the method scales appropriately with dimensionality of the data set. PDE is tested on a difficult real world problem: the selection of winning stocks in a falling market.

## 2 Methods for Density Estimation

Density estimation means the construction of an estimate of the true probability

density function from the observed data. Methods for density estimation have been studied intensively. See Scott (1992) for an overview. Most density-estimators are based upon one or more of the following techniques: finite mixture models, variable kernel estimates, uniform kernel estimates. Finite mixture models attempt to find a superposition of parameterized functions, typically Gaussians which best account for the sample data. The method can in principle model any shape of cluster, and works best when the data's probability density can be described as a mixture of Gaussians. With kernel based approaches the true probability density function is estimated using local approximations. The local approximations are parameterized such that only data points within a certain distance of the point under consideration have an influence on the shape of the kernel function. This is called (band-) width or radius of the kernel (Scott (1992)). Variable kernel methods adjust the radius of the kernel. Uniform kernel algorithms use a fixed global radius. A special case of uniform kernel estimates is hypersphere density estimation. The number of points within a hypersphere around each data point is used for the density estimation at the center of the hypersphere. Uniform kernel estimates can approximate the true probability up to any desired degree of accuracy, if the true probability is known (Devroye and Lugosi (1996, 1997)). Tuning the bandwidth for optimal variable kernel estimation is computationally expensive and proven to be a computational hard task (Devroye and Lugosi (2000)). This is a clear disadvantage of such methods for large data sets. This is one of the reasons why uniform kernel methods have become popular within the context of knowledge discovery in (large) databases (KDD). Clustering methods as used for KDD usually require the definition of a (dis-)similarity measure between two data points. Density estimation within the context of KDD should therefore use the benefits of a given distance measure in particular for high dimensional data. Regions in data space with high density values are good candidates for clusters if these regions are surrounded by substantially less dense or even empty regions. A suitable estimate for KDD needs therefore to be precise in dense regions and less precise in almost empty regions. Data points in very low density regions are most likely outliers, i.e., are of no concern for clustering. All these requirements make hypersphere density estimation with a global radius a good candidate for density estimation for clustering in databases. For clustering the selection of an appropriate radius should be based on the distribution of the data distances.

### 3 Pareto Density Estimation

Let  $S$  be a subset of a set of  $n$  points with  $|S| = s$  the number of elements in  $S$ . Then  $p = s/n$  is the relative size of the set. If there is an equal probability that an arbitrary point  $x$  is observed,  $p$  is the probability  $p = p(x \text{ in } S)$ . Information theory calculates the entropy or (partial) information using  $p$ . Scaled to the range  $[0,1]$ , the information of a set is calculated as  $I(S) = -e p \ln(p)$ . To find an optimal set size, define the unrealized potential  $URP(S)$  of a set as the Euclidian distance from the ideal point, i.e. an empty set producing 100% of information. This definition of  $URP(S)$  leads to:  $URP(S) = \sqrt{p^2 + (1 + e p \ln(p))^2}$ . Minimizing the unrealized potential results in an optimal set size of  $p_u = 20.13\%$ . This set size produces 88% of the maximum information. For details see (Ultsch (2001)). The optimality of this set at about

(20%, 80%) might be the reason behind the so called Pareto 80/20 law, which is empirically found in many domains [12]. Subsets or volumes which contain in the average  $p_u$  data points are optimal in the sense that they give as much information as possible with a minimal set size. Define the neighborhood number  $NN(x, r)$  as the number of input data points within a hypersphere (neighborhood) with radius  $r$  around a point  $x$  in data space. Even if the input is drawn from a Normal distribution, the neighborhood numbers are not normally distributed. The Pareto Radius  $r_p$  of a data set is a radius such that for all data points the median of  $NN(x, r)$  equals  $p_u d$ , with  $p_u = 0.2013$  and  $d$  the number of data points in the data set.

Searching among the distance percentiles of the data is a useful way to limit the effort to approximate the Pareto Radius in practical applications. 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.  $p_{par} = \operatorname{argmin}(|pc(p) - r_p|), \forall p \in 1, \dots, 99$ .

#### 4 Adjusting for intra/inter cluster distances

The detection of the cluster structure in an input data set using distance structures is only possible, if most of the data distances within a cluster are smaller than the distances measured between data from different clusters. Let  $v$  denote the ratio of intra cluster distances to inter distances. If this ratio is known a priori the neighborhood radius for the estimation of a suitable data density can be adapted. The Pareto percentile within a cluster can be calculated as  $p = p_{par} / v$ .

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 1 shows  $\bar{v}(k)$  and the 95% confidence interval versus the cluster number  $k$ . For details see Ultsch (2003).

Different experimental setting for variance  $s$  in the range of  $s \in [1, 30]$  and set sizes  $d \in [50, 5000]$  produced results that were within pen point size equal to the results of Figure 1. 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. If the number of clusters is not known, an initial value of  $v=0.33$  is a meaningful starting point for data mining (see the arrows in Figure 1). This value of  $v$  can be typically found in data sets containing from 3 to about 13 clusters. If the number of clusters is known to be  $k$ ,  $v$  can be taken as  $\bar{v}(k)$ . If there are only one or two clusters in the data set  $v_{est} = 0.7$  can be 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  $v$ . If  $k$  is large ( $k > 40$ ), the empirical Pareto Radius converges to the 1-percentile  $pc(1)$  of the data distances.

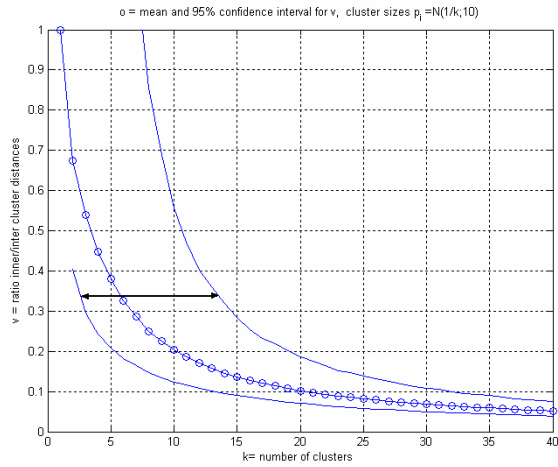


Figure 1: ratio of intra/inter cluster distances

## 5 Pareto Probability Density Estimation

For one dimensional data the PDE can be scaled such that its integral is one. The trapezoidal method on  $(x_i, NN(x_i, r_p))$  can be used as the scaling factor. This leads to a probability density estimation PPDE. To measure the quality of the probability density estimation the mean of the sum of squared errors (MSSE) is used. Error is the difference between PPDE and the true probability density. Two sets containing 500 data points with  $N(0,1)$  and  $N(20,1)$  were generated. The union of these sets represents data points with two clusters of distance 20. As Pareto radius the 18<sup>th</sup> percentile of all distances was found in all experiments. For 200 such sets the density estimation using hyperspheres with the 1.99 percentiles of distances were measured. Figure 2 shows the MSSE  $\pm$  the standard deviation of these experiments. Minimum error was encountered at the Pareto radius, i.e. the 18<sup>th</sup> percentile.

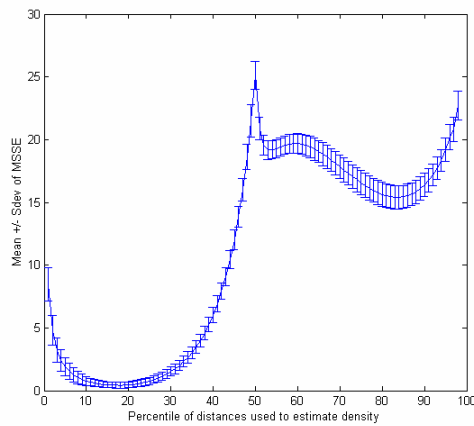


Figure 2: Quality of density estimation using all distance percentiles

## 6 Robustness of the Method

In this chapter we investigate the properties of PDE with regard to number of clusters, higher dimensionality and cluster overlap. The first question is how well PDE generalizes to more than two clusters. For one to 50 clusters with a distance of the cluster centers of 20, 100 Experiments with 1000 data points were performed. A t-test was used to test the null hypothesis that Pareto Probability Density Estimation is different from the best radius to measure density. It turned out that for a 2% error level the null hypothesis - PPDE being different from true density - could be rejected for all these cluster numbers.

The next question concerns the overlapping of clusters. In the two-cluster data set as described in the last chapter, the distances between the clusters were decreased stepwise from 20 to 1. For non overlapping clusters the Pareto radius is best. A t-test with alpha level of 5% rejects the hypothesis that the Pareto Radius is different from the best radius starting with a distance of 2.7 between the cluster centers. This corresponds to about 20% of common points. The distribution of distances is strongly influenced by inner cluster variances. In order to investigate the dependency of PDE on inner cluster variance 1000 data sets from a distribution with  $N(0,1)$  and  $N(20,s^2)$  were generated. For  $s^2$  in the range of  $[0.2, 20]$ , the best distance estimation was compared with PPDE. The experiments were repeated 100 times. It turned out the MSSE for PPDE differs less than 5% compared to the mean plus standard deviation of the best of the hypersphere density estimation. For small values of  $s^2$ , i.e.  $s^2 < 0.1$ , PDE overestimates the true density. The next question concerns the dimensionality of the data. 100 experiments were performed with a two dimensional data set with 200 MMI distributed data points.

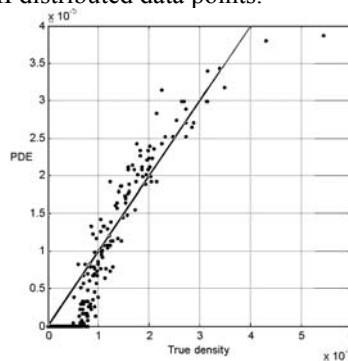


Fig. 3 PDE vs. true density for dimensional data

The points were within two clusters with cluster center distance 6. As in the one dimensional case the MSSE decreases rapidly from the 1 percentile to the 10 percentile. The MSSE increases significant with percentiles greater than 25. The Pareto radius is again the best radius for density estimation. Figure 3 shows the correlation between true density and PDE for the same experimental setting as in the two dimensional case for a ten dimensional data set.

This shows the expected property of uniform kernel estimation: large densities tend to be underestimated, low densities are overestimated. For the ten dimensional space PDE is close to the true density for dense regions down to about one tenth of the maximum measured density.

## 7 Stock picking: a difficult real world problem

Selecting stocks for a portfolio that have a high potential for rising stock prices is in general a difficult problem(O'Neill (1995)). Today there are more than 7000 stocks traded at the US stock exchanges. The problem is to pick between six and fifteen stocks that have a high potential for rising prizes (Maranjian (2002)). Each stock is characterized by 15 variables, from the company's fundamental data and the stock's performance (Deboeck and Ultsch (2002)).

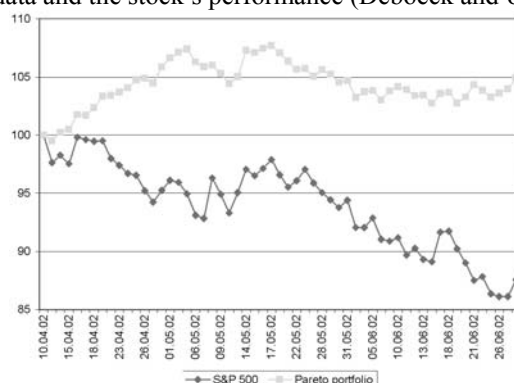


Figure 4: PDE selected Portfolio vs. S&P 500 stock market index

A stock is classified as a winner when the stocks price rises more than 10% within a period of 60 marked days compared to the mean price level of the 10th to 15th market day following the beginning of a quarter. It is classified a loser if the price falls more than 5%. For the first quarter (q1) of 2002 (January 1st to March 31st) the percentages of winners within the Pareto spheres of all data points were calculated. Using a binomial with 0.4% alpha level one single Pareto sphere contained a significantly high winner percentage. This Pareto sphere consisted of 15 stocks with 14 winners. The same Pareto sphere was used to select a portfolio in the second quarter (q2) of 2002 (April 1st to July 1st). Within this Pareto sphere were 11 stocks with 7 (64%) winners. An investment of equal value in each stock of this Pareto portfolio starting April 10th 2002 gave a steady increase in the total value of the portfolio up to 108% until May 3rd. Here after the portfolio's value has a mean of 106% with a range of 104% to 108%. At the first of July the Portfolio closed with 5% gain compared to the -12% loss of the S&P500 index. Figure 4 shows the total portfolio value compared to the development of the S&P 500 index.

## 8 Discussion

Density based cluster algorithms for large high dimensional data as used for

KDD impose special requirements on density estimation. For this kind of application efficiency of the density estimation is very important. It is well known that variable kernel methods estimate the density function with more precision than uniform density methods (Hall (1992)). The tuning of the radius according to the true density has, however, been proven to be intractable for large data sets (Devroye and Lugosi (2000)). The main disadvantage of uniform kernel methods is the overestimation of the true density in low density regions. With the focus on clustering thin regions are typically of no concern. They may even be regarded to contain “outliers”. Fixing a global radius for density estimation has the advantage that the density estimation can be efficiently calculated using the number of points within the hypersphere. In practical applications clusters are often not well separated but overlap to a certain extend. We showed here that PDE is optimal up to 20% of overlapping points of two clusters. This is a very high overlap. It is questionable that data with this percentage of overlap would be considered to be from different clusters. Therefore PDE seems to work even for a reasonable overlap in clusters. Since a distance measure is required for clustering, the density estimation should also make use of this measure. PDE is measured using the 18 percentile of the distances. For large data sets it might be infeasible to calculate the distance percentiles by first calculating the  $O(n^2)$  distances between all data points. The URP function is rather flat around the minimum  $p_u$  (see Ultsch 2001). In Ultsch (2001) it is demonstrated that set sizes in the range 16% to 24% give more than 90% of the information obtained using the optimal  $p_u$ . Furthermore the findings on Gaussian inner cluster structure presented in this work showed also, that the exact value of the Pareto radius is not really critical for these cluster structures. We expect therefore that a computational cheap estimation for the Pareto radius is sufficiently good enough for PDE. Sampling techniques combined with hashing methods may be used for an efficient calculation of the Pareto radius. The practicability of the method was demonstrated on a large database of high dimensionality ( $d=15$ ). The search for a portfolio containing winning stocks is by itself difficult. In the quarter from which the Pareto sphere was constructed the prior winner probability was 40%. In the test quarter this probability dropped to 22%. The performance of indices showed also that the market situation became more difficult in the test quarter. Finding a portfolio with more than 44% winners would have been significant for a binomial model with a 2% alpha level. The Pareto portfolio containing 64% winners surpassed this substantially.

## 9 Conclusion

The aim to discover new and useful knowledge in large sets of data has brought up the requirements for efficient cluster algorithms. Density based clustering methods have been proposed for this. Density estimation using hyperspheres with a global radius are a simple and efficient way to estimate data density. In this paper the radius for such density estimation is optimized according to an information theoretic criterion. The radius is adjusted such that in the average

information optimal subsets are used for density calculations. This density estimation is called Pareto Density Estimation(PDE). PDE is a robust against a wide range of inner cluster variance, cluster numbers and dimensionality. This allows an efficient implementation of density based clustering algorithms for clustering in large databases with high dimensional data. PDE was tested on a difficult real world problem: the selection of stocks in a falling market. This problem consisted of a large number of high dimensional data. Although the market situation deteriorated from the period when the parameters for PDE were constructed, the predicted portfolio substantially outperformed a broad market index.

## 10 References

- Deboeck, G. J., Ultsch, A., (2002) Picking Stocks with Emergent Self-Organizing Value Maps, in: Novak, M.(Ed): Neural Networks World, Vol. 10, Nr. 1-2, pp 203 - 216.
- Devroye, L., Lugosi, G. (1996), A universally acceptable smoothing factor for kernel density estimation, *Annals of Statistics*, vol. 24, pp. 2499– 2512,.
- Devroye, L., Lugosi, G. (1997) Non-asymptotic universal smoothing factors kernel complexity and Yatracos classes, *Annals of Stat.*, 25, pp. 2626–2637.
- Devroye, L., Lugosi, G. (2000) Variable kernel estimates: on the impossibility of tuning the parameters, in: E. Giné and D. Mason (editors), *High- Dimensional Probability*, Springer-Verlag, New York.
- Hall, P.,( 1992) On global properties of variable bandwidth density estimators, *Annals of Statistics*, vol. 20, pp. 762–778.
- Hinneburg, A., Keim , D.A. (1998) An Efficient Approach to Clustering in Large Multimedia Databases with Noise, *Proc. 4th Int.Conf. on Knowledge Discovery and Data Mining*.
- M. Ester, H.-P. Kriegel, J. Sander (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *Proc. 2nd Int. Conf. On Knowledge Discovery and Data Mining*.
- Maranjian, S. (2002) *The Best Number of Stocks*, The Motley Fool, 26.
- O'Neill, W.J. (1995) *How to make money in stocks*, Mc Graw Hill.
- Scott, D.W. (1992) *Multivariate Density Estimation*. Wiley-Interscience.
- Ultsch, A., (2001) Eine Begründung der Pareto 80/20 Regel und Grenzwerte für die ABC- Analyse, Technical Report Nr. 30, Department of Computer Science, University of Marburg.
- Ultsch,A. (2003) Optimal density estimation in data containing clusters of unknown structure, Technical Report Nr. 34, Department of Computer Science, University of Marburg.
- Xu, X., et al.(1998) Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases, *Proc. Conf. on Data Engineering*, pp. 324-331.