Multi-scale Kernel Discriminant Analysis

Anil Kumar Ghosh
Theo. Stat. and Math. Unit
Indian Statistical Institute
Calcutta
res9812@isical.ac.in

Probal Chaudhuri
Theo. Stat. and Math. Unit
Indian Statistical Institute
Calcutta
probal@isical.ac.in

Debasis Sengupta
Applied Statistics Unit
Indian Statistical Institute
Calcutta
sdebasis@isical.ac.in

Abstract

The bandwidth that minimizes the mean integrated square error of a kernel density estimator may not always be good when the density estimate is used for classification purpose. On the other hand cross-validation based techniques for choosing bandwidths may not be computationally feasible when there are many competing classes. Instead of concentrating on a single optimum bandwidth for each population density estimate, it would be more useful in practice to look at the results for different scales of smoothing. This paper presents such a multi-scale approach for classification using kernel density estimates along with a graphical device that leads to a more informative discriminant analysis. Usefulness of this proposed methodology has been illustrated using some benchmark data sets.

1. Introduction

Discriminant analysis aims to develop a rule for classifying an observation into one of several competing classes as accurately as possible. When density functions f_j and prior probabilities π_j of all the J competing classes are known, the optimal Bayes rule assigns an observation $\mathbf{x}$ to class j^* where $j^* = \arg\max_j \pi_j f_j(\mathbf{x})$. However, these density functions are usually unknown in practice, and they have to be estimated using the training sample observations. In Kernel discriminant analysis (see e.g., [3], [4], [10]), these class densities are estimated by $\hat{f}_{jh_j}(\mathbf{x}) = n_j^{-1} h_j^{-d} \sum_{k=1}^{n_j} K\left\{h_j^{-1}(\mathbf{x}_{jk} - \mathbf{x})\right\}$, where $\mathbf{x}_{j1}, \mathbf{x}_{j2}, \ldots, \mathbf{x}_{jn_j}$ are training sample observations from the j^{th} population ($j = 1, 2, \ldots, J$), $K(\cdot)$ is a d-dimensional density function and $h_j > 0$ is the associated smoothing parameter commonly known as the bandwidth (see e.g., [15]). Throughout this article, we will use Gaussian kernel $K(t) = (2\pi)^{-1/2} e^{-t^2/2}$ for our analysis.

The value of the bandwidth parameter h_j plays a crucial role in the performance of the kernel density estimate and that of the corresponding classifier. Data based bandwidth selection techniques (see e.g., [11], [15]) that target to minimize the mean integrated square error ($MISE = E[\int \{\hat{f}_{jh}(\mathbf{x}) - f_j(\mathbf{x})\}^2 \, d\mathbf{x}]$) of the density estimate may lead to rather poor misclassification rates for the resulting classifiers (see e.g., [7], [8]). On the other hand, popular cross-validation methods are not quite effective for bandwidth selection in classification problems due to piecewise constant nature of estimated misclassification probabilities with infinitely many minima (see e.g., [7]). One should also keep in mind that the choice of bandwidth should depend on the specific observation to be classified in addition to depending on the population densities, and apart from determining the class of an observation, it is also desirable to assess the strength of the evidence in favor of that population for varying choices of bandwidth parameters.

Instead of going for the classical need of data based bandwidth selection, this article considers a family of density estimates $\{\hat{f}_{jh_j} : h_j \in H_j\}$ for each population to carry out a multi-scale version (see e.g., [2]) of kernel discriminant analysis. Here, we study the effect of various levels of smoothing simultaneously to get more useful information for classification than that obtained in an approach based on a single optimum bandwidth for each class density estimate. The results of this multi-scale analysis are presented using two-dimensional plots, which are specific to an observation to be classified, and there one can visually compare the strength of the evidence in favor of different competing classes over wide ranges of bandwidth parameters. Statistical uncertainties at various locations in the plots are also quantified on the basis of appropriately estimated misclassification probabilities. To arrive at the final decision for classifying an observation, we take some judicious combination of all the information obtained at different levels of smoothing.

In classification problems with many classes, often it is not computationally feasible to use the usual cross-validation based methods for bandwidth selection. Moreover, such methods usually allow a single bandwidth for a population density estimate, which does not vary depending on its competitors. In this article, we have treated such multi-class problem as a combination of several two-class

problems. This pairwise approach not only reduces the computational burden but also provides the flexibility of using different bandwidths for a class density estimate when it is compared with the density estimates for different competing classes.

2. Description of the methodology

For classifying an observation **x**, we need to compute the density estimates $\hat{f}_{jh_j}(\mathbf{x})$ for all the competing classes. In practice, before this computation, one can standardize the data points in a class using an estimate of the class dispersion matrix to make the data more spherical in nature and thereby making the use of a common bandwidth h_j for all co-ordinate variables more justified. Then, the density estimate for the original data vector can be obtained from that of the standardized data vector by using the simple transformation formula. For a given pair of competing classes, say, class-1 and class-2, and a fixed pair of bandwidths h_1 and h_2 for the two class density estimates, there is an ordering between $\pi_1\hat{f}_{1h_1}(\mathbf{x})$ and $\pi_2\hat{f}_{2h_2}(\mathbf{x})$ that determines the favored class. We now introduce two different measures for the strength of this evidence in favor of one of the two classes.

Posterior probability : Given an observation **x** and a pair of bandwidths (h_1, h_2) for the density estimates of the two classes, the posterior probability in favor of the first population is given by

$$\mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x}) = \frac{\pi_1\hat{f}_{1h_1}(\mathbf{x})}{\pi_1\hat{f}_{1h_1}(\mathbf{x}) + \pi_2\hat{f}_{2h_2}(\mathbf{x})} \ .$$

P-value : For a given pair of bandwidths (h_1, h_2), we classify an observation **x** to population-1 if $\pi_1\hat{f}_{1h_1}(\mathbf{x}) > \pi_2\hat{f}_{2h_2}(\mathbf{x})$. Now, consider the probability $P_{h_1,h_2}(\mathbf{x}) = P\{\pi_1\hat{f}_{1h}(\mathbf{x}) > \pi_2\hat{f}_{2h}(\mathbf{x}) \mid \mathbf{x}\}$. Clearly, high and low values of this probability function indicate the strength in favor of the first and the second population respectively. For fixed (h_1, h_2), the density estimates for different populations are independent, and since they are averages of i.i.d. random variables, we can reasonably assume normality of their distributions to evaluate the above probability even for moderately large training sample sizes. Means and variances of these normal distributions are estimated by $\hat{f}_{jh_j}(\mathbf{x}) = n_j^{-1}h_j^{-d}\sum_{k=1}^{n_j} K\{h_j^{-1}(\mathbf{x}_{jk} - \mathbf{x})\}$ and $s^2_{jh_j}(\mathbf{x}) = \{n_j(n_j-1)\}^{-1}\sum_{k=1}^{n_j}\left\{h_j^{-d}K\{h_j^{-1}(\mathbf{x}_{jk}-\mathbf{x})\} - \hat{f}_{jh_j}(\mathbf{x})\right\}^2$ respectively. Thus, $P_{h_1,h_2}(\mathbf{x})$ is approximated by

$$P_{h_1,h_2}(\mathbf{x}) \simeq \Phi\left(\frac{\pi_1\hat{f}_{1h_1}(\mathbf{x}) - \pi_2\hat{f}_{2h_2}(\mathbf{x})}{\sqrt{\pi_1^2 s^2_{1h_1}(\mathbf{x}) + \pi_2^2 s^2_{2h_2}(\mathbf{x})}}\right),$$

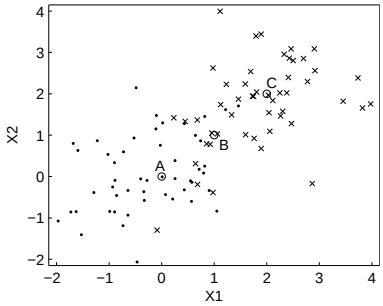

Figure 1: Scatter plot for simulated data.

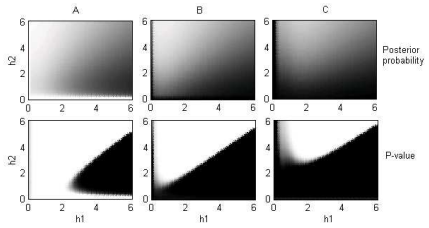

Figure 2: Multi-scale analysis of simulated data.

where Φ is the standard normal distribution function. The above expression can be viewed as the one-sided P-value when the training sample is used to test the hypotheses $H_0 : \pi_1 E\{\hat{f}_{1h}(\mathbf{x})\} \geq \pi_2 E\{\hat{f}_{2h}(\mathbf{x})\}$ against $H_A : \pi_1 E\{\hat{f}_{1h}(\mathbf{x})\} < \pi_2 E\{\hat{f}_{2h}(\mathbf{x})\}$. We can use a wide range of values for h_1 and h_2 to compute these posterior probabilities and P-values, and they are plotted using grey scale in a two-dimensional diagram, where 0 corresponds to black and 1 corresponds to white color.

To demonstrate our methodology, we consider a simulated example with a two-class problem where both the classes $[N_2(0, 0, 1, 1, 0.5)$ and $N_2(2, 2, 1, 1, 0.5)]$ are normally distributed. Taking 50 observations from each class, we generated a training sample (see Figure 1) and used it to carry out a multi-scale analysis for observations A=(0,0), B=(1,1) and C=(2,2) (marked by '$\circ$' in Figure 1). We take equal priors for the two populations and use a wide range of bandwidths for each population to evaluate the posterior probabilities and P-values for different levels of smoothing.

Figure 2 gives the grey scale representation of these discrimination measures for the three cases, where the bandwidths of the first and the second populations are plotted along the horizontal and the vertical axes respectively. Though we have allowed h_1 and h_2 to vary in the range $(0, 6)$ one may use longer intervals as well. Since most of the standardized observations are expected to lie within an interval of 6 units (mean $\pm$ 3), it is a fairly good choice for the upper limit of bandwidths. Our empirical experience suggests that further extension of this interval only increases the computational burden but does not reveal any

new pattern. Here, white color indicates the regions in favor of the first population whereas black color points towards the other. Intensity of the color varies with the magnitude of these discrimination measures, and this helps us to find out the regions for strong evidence in favor of one of the two populations. As it is expected, we observe a dominance of light colored region in the case of observation 'A' (which lies at the center of population-1) and that of the dark region in the case of observation 'C' (which lies at the center of population-2). However, for observation 'B', which lies near the class boundary, the evidence is not so clear in favor of any of the two populations. One note-worthy feature of the plots in the two rows in Figure 2 is that the plots corresponding to the P-values at the bottom are much sharper than those corresponding to posterior probabilities on the top. The plots in the second row enable an easier visualization of the strength of classification, and thereby justify the use of P-values as measures of discrimination (see [8] for further theoretical justification).

3. Aggregation of results

The posterior probabilities computed for different choices of (h_1, h_2) may be combined through an appropriate weighted average to arrive at the final decision. Some well known methods like boosting (see e.g, [6]) adopt such a procedure for combining the result of various classifiers, where different weights are assigned to different classifiers based on their corresponding misclassification probabilities. Clearly, the weight function should be higher for those pairs (h_1, h_2) which led to lower misclassification probability $\Delta(h_1, h_2)$. For varying choices of bandwidths, following the idea of [7], we have used normal approximation to the distribution of a kernel density estimate to estimate $\Delta(h_1, h_2)$ by a smooth function $\widehat{\Delta}(h_1, h_2)$, and then consider the weight function given by

$$w(h_1, h_2) = \begin{cases} e^{-\frac{1}{2}\mathcal{D}^2_{h_1,h_2}} & \text{if } \mathcal{D}_{h_1,h_2} \le \tau \text{ and} \\ & \widehat{\Delta}(h_1, h_2) < min\{\pi_1, \pi_2\} \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{D}_{h_1,h_2} = \frac{\widehat{\Delta}(h_1,h_2) - \widehat{\Delta}_0}{[\widehat{\Delta}_0(1-\widehat{\Delta}_0)/N]^{1/2}}$ for $N = n_1 + n_2$ and $\widehat{\Delta}_0 = \min_{h_1,h_2} \widehat{\Delta}(h_1, h_2)$. $\widehat{\Delta}_0$ and $\widehat{\Delta}_0(1 - \widehat{\Delta}_0)/N$ can be viewed as estimates for the mean and the variance of the empirical misclassification rate of the best kernel classifier when it is used to classify N independent observations. The constant τ determines the maximum amount of deviation from $\widehat{\Delta}_0$ in a standardized scale beyond which the weighting scheme ignores the bandwidth pair (h_1, h_2) by putting zero weight on them. It also puts zero weight on those pairs (h_1, h_2), which lead to a poorer performance than that of a trivial classifier (i.e. when $\widehat{\Delta}(h_1, h_2)$ exceeds any of the two prior probabilities). Clearly, $\tau = 0$ corresponds to the situation of putting all the weights only on those bandwidth pairs (h_1, h_2) for which $\widehat{\Delta}(h_1, h_2) = \widehat{\Delta}_0$. Because of the choice of above Gaussian-type weight function, one does not have to consider a value of τ larger than 3 in practice.

In order to classify an observation $\mathbf{x}$, it would be meaningful to incorporate the corresponding P-value $P_{h_1,h_2}(\mathbf{x})$ in weights. It makes sense to rely more on those bandwidth pairs, which lead to stronger evidence for one of the two classes and adjust the weight function accordingly. Then these adjusted weights will not only depend on the estimated overall misclassification probabilities but also on the particular observation to be classified. In all our numerical work, we have used the adjusted weight function $w_{\mathbf{x}}(h_1, h_2) = w(h_1, h_2)\,|P_{h1,h2}(\mathbf{x}) - 0.5|$. This choice of the adjusted weight function is somewhat subjective, and one may use many other suitable functions as well. However, our empirical experience suggests that the final result is not much sensitive to the weighting procedure as long as any reasonable weight function is used.

Now, consider once again the simulated data for the purpose of illustration. For $\tau = 0$, weighted average of posterior probabilities for population-1 were found to be 0.521, 0.497 and 0.464 respectively for A, B and C, from which it is difficult to visualize the difference in the strength of classification. However, $\tau = 3$ gives a much better result. In the case of observations A and C, it led to 0.665 and 0.299 as the weighted averages, which give a clear indication about the classes to which they belong. In the case of observation B, however, this value was found to be 0.484, which is very close to 0.5 as one would expect since the observation lies near the class boundary where both the classes have almost equal strength.

We conclude this section with another simulated example on a six-dimensional data set, which shows the utility of these weight functions in visualization of strength in classification in addition to its use for aggregating the posteriors. As before, we consider both the populations to be multivariate normal with the same dispersion matrix $\mathbf{\Sigma} = \mathbf{I}_6$, but different location parameters $\boldsymbol{\mu}_1 = (2, 0, \ldots, 0)$ and $\boldsymbol{\mu}_2 = (0, 0, \ldots, 0)$. We generate a training sample taking 50 observations from each class and choose the priors to be equal for our analysis.

Next, consider an observation $\mathbf{x} = (x_1, 0, 0, 0, 0, 0)$. Clearly, $x_1 = 0$ and $x_1 = 2$ give the center for population-2 and population-1 respectively while $x_1 = 1$ represents a point near the class boundary. So, one should expect to have three different behavior of the classification methodology at these three points. As before, the weighted method with $\tau = 0$ fails to reflect the difference in strength of classification, where $\mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x})$ is observed to be 0.471, 0.499 and 0.528 respectively for $x_1 = 0, 1$ and 2. Once again, $\tau = 3$ gives a better result (weighted posteriors = 0.383,

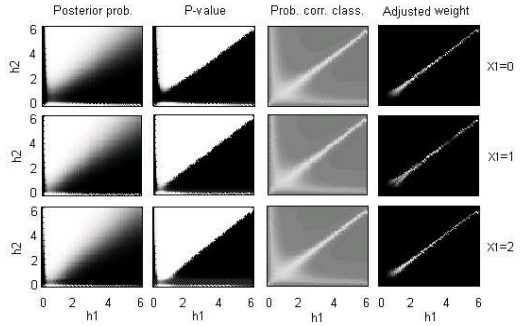

Figure 3: Multi-scale analysis on six-dimensional simulated data.

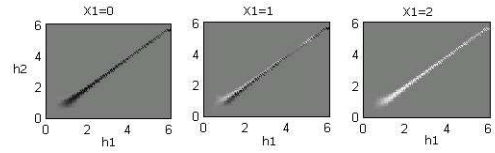

Figure 4: Linearly transformed signed adjusted weight function.

0.493 and 0.609 respectively). In Figure 3, the posterior probabilities and P-values for these observations are plotted for various choices of bandwidth parameters along with the corresponding probabilities of correct classification, where white color points out the regions having low misclassification rates. Adjusted weight functions (re-scaled to have maximum value 1 and minimum value 0) are also presented in the last column of the figure to identify the regions of high reliability (indicated by white color). In the plots both of posterior probabilities and P-values, the white region extends as we move on from $x_1 = 0$ to $x_1 = 2$. However, the strength of the decisions are not quite clear from these figures since in all the cases we have almost equal split in favor of the classes indicated by white and black regions.

However, the difference in the strength of classification become evident if we look at the adjusted weight function with the sign same as that of $P_{h_1,h_2}(\mathbf{x}) - 0.5$. In Figure 4, we have plotted a linearly transformed version of this signed weight function which have a maximum value of 1 and a minimum value of 0. This version can be expressed as $W^*_{\mathbf{X}}(h_1, h_2) = 0.5 + Sign\{P_{h_1,h_2}(\mathbf{x}) - 0.5\} w^*_{\mathbf{X}}(h_1, h_2)$, where $w^*_{\mathbf{X}}$ is the re-scaled version of the adjusted weight function plotted in the last column of Figure 3. $W^*_{\mathbf{X}}(h_1, h_2)$ can also be viewed as a super-imposition of the plots of P-values over that of the weight functions $w(h_1, h_2)$. It is quite clear from the definition that when the pairs (h_1, h_2) have low weights, $W^*_{\mathbf{X}}(h_1, h_2)$ is expected to be very close to 0.5, which is indicated by the grey regions in the plots. However, in more reliable regions (pairs having high weights), we get stronger evidence as $P_{h_1,h_2}(\mathbf{x})$ moves away (in either direction) from 0.5. When $x_1 = 0$ (or $x_1 = 2$), we observe a black (or white) shade in this region, which gives a clear idea about the direction and the strength of the decision. Evidence for classification is very strong in these cases. For $x_1 = 1$, we observe some white as well as some black shades of almost equal intensity. Clearly, the evidence is poor in this case, and the figure gives a clear indication about a border line case.

In the plots of posterior probabilities and P-values, one may notice a white or a black streak near each of the two axes. This is because for the given sample sizes use of very small bandwidth makes one density estimate very close to zero and therefore the competing class density estimate turns out to be the winner. However, these streaks appear in a region of the plot where misclassification rates are high (see third column of Figure 3). Consequently, the weight function becomes almost zero in those regions, and the aggregation procedure does not get affected.

4. Case studies

We now consider some benchmark data sets that illustrate the utility of the proposed method. Results of the kernel discriminant analysis based on $\mathbf{h}_0$ (the bandwidths that minimize $MISE$) and those based on the weighted averaging of posteriors are presented to compare their performance. In multi-class problems, we adopt the pairwise classification method and proceed in the same way as before. The results of all these pairwise classification are combined by the method of majority voting (see e.g., [5]) as well as by the method of pairwise coupling (see e.g., [9]). Voting method in some cases may end up with a tied situation, which is considered as misclassification here. Hence, the reported results on voting are the error rates in the worst possible cases. Misclassification rates for usual linear and quadratic discriminant analysis (LDA and QDA) are reported for each data set. Error rates for some well known nonparametric methods like classification trees (CART, see [1]), nearest neighbors (see e.g., [4]), and neural nets (see e.g., [13]) are also given to facilitate the comparison. Throughout these experiments, sample proportions for different classes are used as their priors.

Chemical and overt diabetes data : This data set contains information on five measurement variables (fasting plasma glucose level, steady state plasma glucose level, glucose area, insulin area and relative weight) and three classes of individuals ("overt diabetic", "chemical diabetic" and "normal") reported in [12]. There are 145 individuals with 33, 36 and 76 in the three classes according to clinical classification. For this data set, leave one out cross-validated error rates for different classifiers are given in Table 1. Due to computational difficulties, these error rates could not be computed for CART and neural network. Clearly, the weighted averaging methods outperformed the other classifiers when majority voting is used to reach the final result.

Image segmentation data : This data set contains 19 different measurements on each image of one of seven different objects. There are 210 observations in the training and 2100 observations in the test set which are equally distributed in those 7 classes. The data set and the description of the variables are available at UCI repository. The value of the variable 'region pixel count' is '9' for all observations. For the two variables, 'short line density-5' and 'short line density-2', almost 95% of the values are zero. We did not consider these variables in our study. There are some variables in the data set which are linear or nonlinear functions of R ('raw red mean'), B ('raw blue mean') and G ('raw green mean'). We have deleted those variables too and carried out our analysis using the remaining 9 variables. Among different classifiers, weighted averaging methods and LDA had better misclassification rates.

Vowel recognition data : This data set is related to a vowel recognition problem where ten measurements on speech signal are taken on each observation corresponds to one of the 11 vowels (see [14] for detail description of this data set). There are 528 observations in the training set while the test set consisting of 468 cases. This is a difficult data set and a part of this difficulty arises due to the presence of a fairly large number of competing classes. In this data set, many well-known classifiers misclassified more than 50% of the test set observations. The best error rate was observed for nearest neighbor method. Error rates for the multi-scale methods were quite competitive as compared to the other nonparametric classifiers.

Classification Methods.	Diabetes Data	Image Data	Vowel Data
LDA	11.0	11.4	55.6
QDA	9.7	14.6	52.8
CART	—	12.6	56.4
Neural Net	—	12.1	50.9
Nearest Neighbor	9.0	18.2	43.7
Kernel (with $\mathbf{h}_0$)	12.4	15.7	62.1
Kernel (Wt. avg.)			
Voting ($\tau = 0$)	6.2	10.5	50.6
Coupling ($\tau = 0$)	15.2	11.7	47.2
Voting ($\tau = 3$)	6.2	11.0	51.9
Coupling ($\tau = 3$)	8.3	10.6	48.9

Table1 : Misclassification rates (in %) for different classifiers.

5. Conclusions

In usual kernel discriminant analysis, one looks at the estimated posteriors both for classifying an observation and assessing the strength of the decision. But the multi-scale approach proposed here is a more informative classification procedure. Using the information obtained from various levels of smoothing, it gives a clear idea about the strength of classification and the related statistical uncertainties present there. Unlike the usual kernel method based on single bandwidth, this multi-scale method uses a data dependent adjusted weight function to arrive at the final result. This case specific emphasis on various bandwidths provides more flexibility to the classification methodology.

The pairwise classification method seems to be a useful technique for multi-class problems, when it is computationally difficult to find out the optimal bandwidths by minimizing $\widehat{\Delta}(h_1, h_2, \ldots, h_J)$. It not only reduces the computational burden significantly, but also provides the flexibility of using different bandwidths for a class when we compare it to different competing classes. The plots of the discrimination measures and that of the weight functions enable an easier visualization of the strength of classification.

References

[1] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classifi cation and Regression Trees*. Chapman and Hall, New York.

[2] Chaudhuri, P. and Marron, J. S. (2000) Scale space view of curve estimation. *Ann. Statist.*, **28**, 408-428.

[3] Devijer, P. A. and Kittler, J. (1982) *Pattern Recognition: a Statistical Approach.* Prentice Hall, London.

[4] Duda, R., Hart, P. and Stork, D. G. (2000) *Pattern Classifi - cation*. Wiley, New York.

[5] Friedman, J. H. (1996) Another approach to ploychotomous classifi cation. *Tech. Rep., Dept. of Stat., Stanford University*.

[6] Friedman, J. H., Hastie, T. and Tibshirani, R. (2000) Additive logistic regression : a statistical view of boosting (with discussion). *Ann. Statist.*, **28**, 337-407.

[7] Ghosh, A. K. and Chaudhuri, P. (2003) Optimal smoothing in kernel discriminant analysis. To appear in *Statist. Sinica*.

[8] Ghosh, A. K., Chaudhuri, P. and Sengupta, D. (2003) Classifi cation using kernel density estimates : multi-scale analysis and visualization. Submitted for publication.

[9] Hastie, T. and Tibshirani, R. (1998) Classifi cation by pairwise coupling. *Ann. Statist.*, **26**, 451-471.

[10] Hastie, T., Tibshirani, R. and Friedman, J. H. (2001) *The elements of statistical learning : data mining, inference and prediction.* Springer Verlag, New York.

[11] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996) A brief summary of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.*, **91**, 401-407.

[12] Reaven, G. M. and Miller, R. G. (1979) An Attempt to Defi ne the Nature of Chemical Diabetes using a Multidimensional Analysis. *Diabetologia*, **16**, 17-24.

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

[14] Robinson, A. (1989) Dynamic error propagation networks. *Tech. Rep., Ph.D. Thesis, Elec. Eng., Cambridge University*.

[15] Silverman, B. W. (1986) *Density Estimation for Statistics and Data analysis*. Chapman and Hall, London.