A non-asymptotic bandwidth selection method for kernel density estimation of discrete data

Zdravko Botev
Department of Mathematics, University of Queensland, Australia

25 October 2005

Abstract

In this paper we explore a method for modeling of categorical data derived from the principles of the Generalized Cross Entropy method. The method builds on standard kernel density estimation techniques by providing a novel non-asymptotic data-driven bandwidth selection rule. In addition to this, the Entropic approach provides model sparsity not present in the standard kernel approach. Numerical experiments with 10 dimensional binary medical data are conducted. The experiments suggest that the Generalized Cross Entropy approach is a viable method for density estimation, discriminant analysis and classification.

Keywords

Bandwidth selection, kernel density estimator, Generalized Cross Entropy, multivariate binary discrimination, discrete data smoothing.

1 Introduction

Aitchison & Aitken [1] proposed an extension of the kernel density estimation technique to multivariate discrete spaces. Similar to the kernel density estimator, the performance of their proposed estimator depends crucially on a smoothing parameter — usually called the *bandwidth*. Aitchison & Aitken suggest maximizing the cross-validatory (aka as 'leave-one-out') likelihood function as a method of estimating the bandwidth. [4] argues that Aitchison & Aitken's likelihood cross-validation method can behave erratically even for large samples. Since then various different methods for more reliable and consistent estimation of the bandwidth have been suggested. Most of the methods rely on asymptotic approximations and assume sufficient differentiability of the underlying true density. For a survey of the various bandwidth selection methods see [14], [3], [13] and the references therein.

The paper is organized as follows. First, Aitchison & Aitken's likelihood cross-validation method is briefly review and possible remedies for the problems mentioned above are considered. Next, the Generalized Cross Entropy approach is presented and finally in the last section an example of multivariate binary discrimination with real medical data is provided. The example is based on medical data described in [6] and is the same one used by Aitchison & Aitken [1] and some of the follow-up papers (see [4] and [14]) to present the estimation results. The example clearly demonstrates the practical benefits of the Generalized Cross Entropy method (GCE) as a tool for optimal bandwidth selection and more generally for discrete multivariate data modeling.

2 The Kernel Approach

Suppose we are given the observations $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ on a discrete $d-$dimensional space $\mathscr{X}$. To model the data probabilistically we assume that the data is the outcome of a random experiment with probability mass function $q^* : \mathscr{X} \to [0,1]$, i.e.:

$$\mathbf{X}_1, \ldots, \mathbf{X}_n \sim q^*,$$

where the data are not necessarily independent. The kernel method assumes that the true, but unknown, underlying probability mass function q^* can be approximated well by a probability mass function of the form:

$$p(\mathbf{x} \,|\, \sigma, \mathcal{X}) = \frac{1}{n} \sum_{i=1}^{n} K(\mathbf{x} \,|\, \sigma, \mathbf{X}_i) \quad , \mathbf{x} \in \mathscr{X} \tag{1}$$

where:

1. $\mathbf{x} \in \mathscr{X}$ and $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim q^*$ are d-dimensional column vectors.
2. $K : \mathscr{X} \to [0,1]$, $\sum_{\mathbf{x} \in \mathscr{X}} K(\mathbf{x} \,|\, \cdot) = 1$ is a unimodal probability mass function, usually referred to as the *kernel* function.
3. $\sigma \in [0,1]$ is a parameter which controls the "smoothness" of p in a way similar to the bandwidth in kernel density estimation.

Example 1 (Binary Kernel) Suppose that the data $\mathcal{X}$ is binary. Then a simple choice for the kernel function is:

$$\begin{aligned} K(\mathbf{x} \,|\, \sigma, \mathbf{X}_i) &= \prod_{l=1}^{d} \sigma^{I\{\mathbf{x}(l) = \mathbf{X}_i(l)\}} (1-\sigma)^{1 - I\{\mathbf{x}(l) = \mathbf{X}_i(l)\}} && (2) \\ &= \sigma^{\mathrm{d}(\mathbf{x}, \mathbf{X}_i)} (1-\sigma)^{d - \mathrm{d}(\mathbf{x}, \mathbf{X}_i)}, && (3) \end{aligned}$$

where $\mathrm{d}(\mathbf{x}, \mathbf{y}) = \sum_{l=1}^{d} I\{\mathbf{x}(l) = \mathbf{y}(l)\}$ measures the "distance" between the vectors $\mathbf{x}$ and $\mathbf{y}$ and $\sigma \in (.5, 1)$ and $I\{\cdot\}$ is one if the statement inside the brackets is true and zero otherwise. Note that:

$$\begin{aligned} \lim_{\sigma \to 1^-} K(\mathbf{x} \,|\, \sigma, \mathbf{X}_i) &= \begin{cases} 1, & \text{if} \quad \mathbf{x} = \mathbf{X}_i \\ 0, & \text{if} \quad \mathbf{x} \neq \mathbf{X}_i \end{cases}, && (4) \\ K(\mathbf{x} \,|\, 1/2, \mathbf{X}_i) &= \frac{1}{2^d}. && (5) \end{aligned}$$

So the end-points of the interval $[1/2, 1]$ represent two extremes of smoothing. For $\sigma = 1$ there is no smoothing whatsoever and p is simply estimated from corresponding relative frequencies. For $\sigma = 1/2$ the smoothing is maximal, K is not unimodal and p is the uniform probability mass function (pmf) on $\mathscr{X}$. Thus the restriction $\sigma \in (1/2, 1)$ guarantees that $K(\mathbf{x} \,|\, \sigma, \mathbf{X}_i)$ has a single mode at $\mathbf{x} = \mathbf{X}_i$ and that p is "smooth" in the sense of $p(\mathbf{x} \,|\, \sigma, \mathcal{X}) > 0, \ \forall x \in \mathscr{X}$.

Everything in (1) is fixed except the smoothing parameter σ. This is the only parameter over which we have control. We need to adjust σ so that our approximation of q^* is as good as possible— not too small so that the resulting pmf is too peaked and sample dependent, not too large so that the pmf p is too uniform.

2.1 Measuring the performance

The performance of the estimator (1) depends crucially on the parameter σ. Let us assume for the moment that $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ are independent. Then one possible measure of performance is the observed likelihood function:

$$L(\sigma \,|\, \mathcal{X}) = \prod_{i=1}^{n} p(\mathbf{X}_i \,|\, \sigma, \mathcal{X}). \tag{6}$$

Unfortunately maximization of the likelihood function leads to the undesirable value $\sigma = 1$, in much the same way it leads to the empirical Dirac delta measurable density function in the continuous case. Similar to the kernel density estimation method this problem is resolved by considering the *leave-one-out* or cross-validated likelihood:

$$L_{\mathrm{x}}(\sigma \,|\, \mathcal{X}) = \prod_{i=1}^{n} p(\mathbf{X}_i \,|\, \sigma, \{\mathcal{X} \backslash \mathbf{X}_i\}) = \prod_{i=1}^{n} \frac{1}{n-1} \sum_{j \neq i} K(\mathbf{X}_i \,|\, \sigma, \mathbf{X}_j) \tag{7}$$

Thus the Aitchison & Aitken's [1] choice for σ is:

$$\widehat{\sigma} = \underset{\sigma \in (0.5,1)}{\operatorname{argmax}} L_{\mathrm{x}}(\sigma \,|\, \mathcal{X}) \tag{8}$$

The consistency properties of (8) are demonstrated in [1]. Hall [4], however, argues that (8) can behave erratically and often undersmooths by giving a solution close to 1. Occasionally the only maximizer of $L_{\mathrm{x}}(\sigma \,|\, \mathcal{X})$ is 1, in which case the procedure fails to improve on the naive maximum likelihood estimator based on the relative frequencies of the observations[1]. To overcome these problems, Hall [4] proposes to minimize an asymptotic approximation to a global function of the Mean Squared Error. His proposed performance criterion is the analogue of the Integrated Mean Squared Error in the continuous density estimation case. Hall's estimation method is largely asymptotic in nature. For a tretment of the above estimation problem via Markov Chain theory see [3] and for a comparative study of the various kernel-based categorical data smoothing techniques see [14]. We now consider a different approach to the density estimation problem.

[1] This type of behavior is also typical for the Least Squares Cross Validation (LSCV) method, as used in the continuous density estimation case (see [10]).

3 The Generalized Cross Entropy Method

The Generalized Cross Entropy method (GCE) is another possible approach to the problem of density estimation and statistical learning in general. For a detailed derivation of the method and full references see [2].

Let our probabilistic model about the data $\mathcal{X}$ be the probability mass function $p(\mathbf{x})$. Then the entropic formalism can be stated as follows:

1. Minimize a measure of model complexity (or alternatively maximize a measure of model sparsity),

2. subject to agreement with the empirical observations.

In other words we want the function/model p to be as simple as possible while at the same time being truthful and consistent with the observed data. This is in accordance with "Occam's razor" principle which states that one should look for the simplest possible model which explains the observed reality (in our case reality is revealed in the form of statistical data).

3.1 A Suitable Measure of Complexity

In the GCE method a commonly used measure of complexity is the negative of Shannon's Entropy [12]:

$$\sum_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}) \ln p(\mathbf{x}).$$

Shannon's Entropy is a special limiting case of a one-parameter family of Entropy measures first studied by Havrda and Charvat (see [5]):

$$\mathrm{H}_\alpha\{p\} = \frac{1}{1-\alpha}\left[1 - \sum_{\mathbf{x}\in\mathcal{X}} p^\alpha(\mathbf{x})\right], \quad \alpha \neq 1.$$

It is easy to verify that

$$\lim_{\alpha\to 1} \mathrm{H}_\alpha\{p\} = \sum_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}) \ln p(\mathbf{x}).$$

Kesavan and Kapur [8] argue that one can use the Havrda-Charvat one-parameter family as a measure of complexity (or taking the negative of H_α, as a measure of Entropy). This interpretation of H_α suits our purpose and we will use H_α to measure the complexity of our proposed model for the data $\mathcal{X}$, i.e., $p(\mathbf{x})$.

3.2 Fidelity to the Empirical Data

Having chosen a suitable measure of model complexity, we now have to explain what we mean by "agreement with empirical data" in part 2 of the GCE formalism. In the GCE method the agreement of the model with empirical observations is imposed via constraints of the form:

$$\sum_{\mathbf{x}\in\mathscr{X}} p(\mathbf{x})\; K_i(\mathbf{x}) = \mathbb{E}_p[K_i(\mathbf{X})] \geqq \kappa_i^*, \quad i = 1, \ldots, n,$$

where:

- $\{K_i(\mathbf{x}) = K(\mathbf{x}\,|\,\sigma, \mathbf{X}_i)\}_{i=1}^n$ is a set of kernels just like the ones used in the kernel estimation method for categorical data. For concreteness we consider binary data $\mathscr{X}$ and choose $K_i(\mathbf{x}) = K(\mathbf{x}\,|\,\sigma, \mathbf{X}_i)$ to be the binary kernel discussed previously. Note that $K_i(\mathbf{x})$ is a function of the random variable $\mathbf{X}_i$.

- $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim q^*$ is the d-dimensional empirical data coming from the unknown q^*. Note that it is not necessary to assume that $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ are independent.

- $\kappa_i^* = \frac{1}{n-1}\sum_{j\neq i} K_i(\mathbf{X}_j)$ for all $i = 1, \ldots, n$, i.e., κ_i^* is the empirical/sample average which approximates $\mathbb{E}_{q^*}[K_i(\mathbf{X})]$. It is obvious that $\mathbb{E}_{q^*}\left[\frac{1}{n-1}\sum_{j\neq i} K_i(\mathbf{X}_j)\right] = \mathbb{E}_{q^*}[K_i(\mathbf{X})]$ so:

  $$\kappa_i^* = \frac{1}{n-1}\sum_{j\neq i} K_i(\mathbf{X}_j) \approx \mathbb{E}_{q^*}[K_i(\mathbf{X})].$$

  Note that to estimate $\mathbb{E}_{q^*}[K_i(\mathbf{X})]$ we do not use the i-th observation. The reason is that $\frac{1}{n}\sum_{j=1}^n K_i(\mathbf{X}_j)$ is a biased statistical estimator of $\mathbb{E}_{q^*}[K_i(\mathbf{X})]$. We, however, wish to estimate $\mathbb{E}_{q^*}[K_i(\mathbf{X})]$ without bias. To achieve this we drop the i-th observation and obtain the unbiased estimator $\frac{1}{n-1}\sum_{j\neq i} K_i(\mathbf{X}_j)$. This estimator is similar to the cross validatory estimators used in the Least Squares Cross Validation Method for density estimation (see [10]).

- The n constraints could either be strict equalities:

  $$\mathbb{E}_p[K_i(\mathbf{X})] = \kappa_i^*$$

  or inequalities:

  $$\mathbb{E}_p[K_i(\mathbf{X})] \geqslant \kappa_i^*.$$

  Note that the Maximum Entropy Method [7] uses equality constraints only. In the GCE method [2], however, we use inequality constraints instead. Thus the GCE approach differs from the Maximum Entropy method in this respect.

The constraints embody nothing more than the simple concept of *moment matching* first advocated by Karl Pearson [11]. We "match" the moments of our proposed density $\mathbb{E}_p[K_i(\mathbf{X})]$ to the empirical moments κ_i^* (which approximate the true but unknown moments $\mathbb{E}_{q^*}[K_i(\mathbf{X})]$).

3.3 The GCE Optimization Problem

For clarity we now restate the GCE approach. Given the data $\mathcal{X}$,

1. minimize the measure of complexity $\mathrm{H}_2\{p\} = \sum_{\mathbf{x}\in\mathscr{X}} p^2(\mathbf{x}) - 1$, i.e.,

$$\min_{p\in\mathscr{P}} \sum_{\mathbf{x}\in\mathscr{X}} p^2(\mathbf{x}), \tag{9}$$

2. subject to the constraints:

$$\sum_{\mathbf{x}\in\mathscr{X}} p(\mathbf{x})\; K_i(\mathbf{x}) = \mathbb{E}_p[K_i(\mathbf{X})] \geqslant \kappa_i^*, \quad i = 1,\ldots,n. \tag{10}$$

Here $\mathscr{P} = \{p : \sum_{\mathbf{x}\in\mathscr{X}} p(\mathbf{x}) = 1,\ p(\mathbf{x}) \geqslant 0\ \forall \mathbf{x}\}$ denotes the set of all probability density functions on $\mathscr{X}$. We have chosen $\alpha = 2$ because $\mathrm{H}_2\{p\}$ is easier to manipulate than other choices for $\mathrm{H}_\alpha\{p\}$. Using the Lagrange multiplier technique [15] it can be shown that the optimal solution of the GCE problem (9)+(10) has the form:

$$p(\mathbf{x}) = \sum_{j=1}^{n} \lambda_j\; K_j(\mathbf{x}), \tag{11}$$

where $\boldsymbol{\lambda} = [\lambda_1,\ldots,\lambda_n]^T$ are positive Lagrange multipliers each of which takes care of the n constraints (10). Thus the optimal solution to (9)+(10) is a linear combination of the n kernels. All that remains is to determine the actual Lagrange multipliers $\boldsymbol{\lambda}$. Fortunately finding $\boldsymbol{\lambda}$ is not a difficult task as it only requires to solve a common optimization problem.

3.4 The Quadratic Programming Problem

To find the Lagrange multipliers substitute (11) into (9) and (10). This gives the following quadratic programming problem (QPP):

$$\text{(9) transforms to} \rightarrow \qquad \min_{\boldsymbol{\lambda}} \quad \frac{1}{2}\boldsymbol{\lambda}^T C\boldsymbol{\lambda} \tag{12}$$

$$\text{(10) transforms to} \rightarrow \qquad \text{subject to:} \quad C\boldsymbol{\lambda} \geqslant \boldsymbol{\kappa}^*, \tag{13}$$

where C is the $n \times n$ matrix with

$$C_{ij} = \sum_{\mathbf{x}\in\mathscr{X}} K_i(\mathbf{x})\; K_j(\mathbf{x}) = K(\mathbf{X}_i \,|\, \varsigma, \mathbf{X}_j), \quad \varsigma = \sigma^2 + (1-\sigma)^2$$

for the simple binary kernel and $\boldsymbol{\kappa}^* = [\kappa_1^*, \ldots, \kappa_n^*]^T$. The only quantity that is still unspecified is the bandwidth (aka as scale, spread or concentration) parameter σ. This is the tricky part of the method. First note that the integral of (11) is

$$\sum_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}) = \sum_{j=1}^{n} \lambda_j \sum_{\mathbf{x}\in\mathcal{X}} K_j(\mathbf{x}) = \sum_{j=1}^{n} \lambda_j \, .$$

It follows that $p \in \mathscr{P}$ is equivalent to $\sum_{j=1}^{n} \lambda_j = 1$. Second, note that the condition $\sum_{j=1}^{n} \lambda_j = 1$ is not present in the quadratic programming problem (12)+(13). We now select the as yet unspecified bandwidth σ to meet this constraint. We thus choose the bandwidth parameter σ such that:

- $p \in \mathcal{P}$, which is equivalent to
- $\sum_{j=1}^{n} \lambda_j = 1$, which in turn is equivalent to
- $p(\mathbf{x})$ integrates to one.

Now all of the unknown variables in the model are specified and we can summarize the GCE approach:

1. Find a σ such that $\sum_{j=1}^{n} \lambda_j^* = 1$, where $\{\lambda_j^*\}_{j=1}^{n}$ is the solution of the QPP, i.e., find:

$$\sigma^* = \left\{ \sigma : \; \mathbf{1}^T \boldsymbol{\lambda}^*(\sigma) = 1, \; \boldsymbol{\lambda}^*(\sigma) = \underset{\boldsymbol{\lambda}:\, C(\sigma)\boldsymbol{\lambda} \geqslant \boldsymbol{\kappa}^*(\sigma)}{\operatorname{argmin}} \; \boldsymbol{\lambda}^T C(\sigma) \boldsymbol{\lambda} \right\}$$

2. Present the mixture pmf:

$$p(\mathbf{x}) = \sum_{j=1}^{n} \lambda_j^* \, K(\mathbf{x} \,|\, \sigma^*, \mathbf{X}_j)$$

   as the GCE pmf which models the data $\mathcal{X}$.

A final note is that $\boldsymbol{\lambda}^T C \boldsymbol{\lambda} = \sum_{\mathbf{x}\in\mathcal{X}} p^2(\mathbf{x}) > 0, \; \forall \boldsymbol{\lambda} \neq \mathbf{0}$ implies that C is a positive definite matrix and hence the QPP (12)+(13) has a unique global solution for each value of $\sigma \in (1/2, 1)$.

4 Application to Multivariate Binary Discrimination

In this section we apply the GCE method to the diagnosis of *Keratoconjunctivitis sicca* (KCS) based on the medical data reported in [6]. The same data set was used by Aitchison & Aitken [1] and we can compare the results of the two studies. The description of the data set is:

1. 40 patients suffering from KCS given in the first two columns of Table 1. Each patient may or may not have any of the 10 possible symptoms of the disease. The presence of the symptoms is represented as binary row vectors of length 10. A 1 means that the symptom is present and a 0 stands for no clinically obvious pathology.

2. 37 non-KCS patients given in the first two columns of Table 2.

3. Table 1 and Table 2 form the first group of 77 patients, referred to as group-1.

4. The same 10 symptoms are recorded for another group of 41 patients, henceforth referred to as the group-2 patients. Of this group the 24 KCS patients are given in the first two columns of Table 3 and the 17 non-KCS patients in the first two columns of Table 4.

We now use the GCE method to estimate the pmf of the "group-1 KCS" observations (denoted by p_{KCS}) and the pmf of the "group-1 non-KCS" observations (denoted $p_{\text{non-KCS}}$). The estimated mixture pmf $p_{\text{non-KCS}}$ is summarized in Table 5. Kernels associated with zero Lagrange multiplier are not listed as they do not contribute toward the value of $p_{\text{non-KCS}}(\mathbf{x})$ for any $\mathbf{x} \in \mathscr{X}$. Similarly the mixture pmf p_{KCS} is summarized in Table 6. This time the number of observations (and hence kernels) associated with a zero multiplier is much largeer. The GCE smoothing parameters $\sigma^*_{\text{KCS}} = 0.79275$ and $\sigma^*_{\text{non-KCS}} = 0.947666$ are close to the Aitchison and Aitken's leave-one-out maximum likelihood parameter values of 0.843 and 0.96 respectively. Next, based on the training data in group-1 we classify the patients in group-2. Each patient is classified in the following way. If the *odds ratio*

$$\frac{p_{\text{non-KCS}}(\mathbf{X}_i^{(2)})}{p_{\text{KCS}}(\mathbf{X}_i^{(2)})} > 1,$$

where $\mathbf{X}_i^{(2)}$ denotes the i-th observation from group-2, then the i-th patient from group-2 is classified as a non-KCS patient. Alternatively if

$$\frac{p_{\text{KCS}}(\mathbf{X}_i^{(2)})}{p_{\text{non-KCS}}(\mathbf{X}_i^{(2)})} > 1,$$

then the i-th patient from group-2 is classified as KCS patient. The same classification procedure is used by Aitchison & Aitken [1]. The odds ratios computed using the GCE and Aitchison & Aitken's method are given in the third and fourth column, respectively, of Table 3 and 4. Both methods classify all of the patients in group-2 correctly. There is some element of doubt, however, for patients 1 and 3 in Table 4 as the odds ratios are smaller than 10.

In addition to classifying the patients in group-2 the effectiveness of the GCE method is tested using the approach of Lachenbruch and Mickey [9]. Their approach is another example of cross validation. In it each patient is omitted from the training set in turn. Suppose, for example, that we omit a KCE patient from group-1. A new estimate for $\sigma^{\text{new}}_{\text{KCE}}$ is calculated from the reduced set and then the omitted patient is classified using $\sigma^{\text{new}}_{\text{KCE}}$ and $\sigma^{*}_{\text{non-KCE}}$. The results of this procedure for each patient are presented in the third columns of Table 1 and 2. Of those suffering from KCS, patients $10, 21, 26, 38, 39$ are misclassified. The misclassification is lower for the non-KCS patients with patients 3 and 25 being the only misclassified ones. Misclassification in this case does not necessarily signify problems with the statistical model but rather it may point to some anomalous symptoms in the diagnosis of Keratoconjunctivitis Sicca. For example, the misclassified KCS patient 39 does not exhibit any of the symptoms of the disease, yet full medical tests (see [6]) confirm the KCS diagnosis.

4.1 Matlab Implementation

Some issues concerning the implementation of the GCE are :

1. The Matlab routine "mosekopt" is used to solve the QPP. "mosekopt" was downloaded from this webpage:

   http://www.mosek.com/trials.html# students

2. To find a σ such that $\sum_{j=1}^{n} \lambda_j^* = 1$, where $\left\{\lambda_j^*\right\}_{j=1}^{n}$ is the solution of the QPP, the Matlab build-in root-finding function "fzero.m" was used. Each iteration of "fzero.m" requires the solution of a QPP and hence calls "mosekopt".

	Symptoms										
KCS Patient	1	2	3	4	5	6	7	8	9	10	Misclassification odds ratio
1	1	1	1	0	1	0	1	0	0	1	8551.4
2	1	1	1	1	1	0	0	1	0	0	4.1789×10^5
3	1	1	0	1	1	1	0	0	1	0	64775
4	1	1	0	1	1	0	0	1	1	0	83398
5	1	1	1	1	0	0	1	0	0	1	26506
6	1	1	0	0	1	0	0	0	0	1	121.05
7	1	1	1	0	0	1	0	1	0	0	5394.3
8	1	1	0	1	0	1	0	0	1	0	1132.5
9	1	1	1	1	1	0	1	1	0	0	1.8521×10^5
10	1	0	0	0	0	0	0	0	0	0	0.046952
11	1	1	1	1	0	1	0	1	0	0	3.2027×10^5
12	1	1	0	0	0	0	1	1	1	0	218.29
13	1	1	1	1	1	1	1	0	0	1	5.5853×10^6
14	1	1	1	1	1	0	1	1	0	1	3.4814×10^6
15	0	0	1	1	0	0	1	1	0	0	26.729
16	1	1	0	1	0	0	0	0	1	0	310.05
17	0	1	1	0	0	1	0	1	0	1	663.26
18	1	0	1	1	1	0	0	1	0	0	11933
19	1	0	1	1	1	0	1	0	0	0	991.85
20	1	1	1	1	0	0	1	0	0	1	26506
21	0	0	0	0	1	0	0	0	0	0	0.033421
22	1	1	1	1	1	1	1	0	0	0	2.4328×10^5
23	1	1	1	0	1	0	0	0	0	1	4041.5
24	1	1	0	1	0	0	1	1	1	0	12271
25	1	1	1	1	0	0	0	1	0	0	21215
26	0	0	0	1	0	0	1	0	0	0	0.083477
27	1	1	0	1	1	0	0	1	1	1	5.371×10^5
28	1	1	1	1	0	0	0	0	0	1	26392
29	1	0	1	0	1	0	0	0	1	0	63.845
30	1	1	0	1	0	1	0	0	0	1	4774.9
31	0	1	1	1	0	0	0	0	0	1	416.46
32	1	1	1	1	1	1	1	0	0	1	5.5853×10^6
33	0	0	1	1	1	0	1	0	1	0	83.169
34	1	1	1	1	0	1	1	0	0	1	4.8692×10^5
35	1	0	1	0	1	0	0	1	0	0	249.01
36	1	1	1	1	0	0	0	1	0	0	21215
37	1	1	1	1	1	0	0	0	0	0	13902
38	1	1	0	0	0	0	0	0	0	0	0.80488
39	0	0	0	0	0	0	0	0	0	0	0.0018276
40	0	1	1	1	0	0	1	0	0	1	860.13

Table 1: Group 1-KCS

	Symptoms										
Non-KCS Patient	1	2	3	4	5	6	7	8	9	10	Misclassification odds ratio
1	1	0	0	0	0	0	1	0	0	0	3.848
2	0	0	0	0	0	0	0	0	0	0	71.673
3	0	1	1	0	0	0	1	0	0	0	0.46621
4	0	0	0	0	0	0	0	0	0	0	71.673
5	0	0	0	0	0	0	0	1	0	0	14.068
6	0	0	0	0	0	0	0	0	0	0	71.673
7	0	0	1	0	0	0	1	0	0	0	10.048
8	0	0	0	0	0	0	0	0	0	0	71.673
9	0	0	0	0	0	0	0	0	0	0	71.673
10	0	0	0	0	0	0	0	0	1	0	14.611
11	0	1	0	0	0	0	1	0	0	0	9.5114
12	0	0	0	0	0	0	0	0	0	0	71.673
13	0	0	0	0	0	0	1	0	0	0	50.653
14	0	0	0	0	0	0	0	0	0	0	71.673
15	0	0	0	0	0	0	1	0	0	0	50.653
16	0	0	0	0	0	0	0	0	0	0	71.673
17	0	0	0	0	0	0	1	0	0	0	50.653
18	0	0	0	0	0	0	0	0	0	0	71.673
19	0	0	0	0	1	0	0	0	0	0	9.9515
20	0	0	0	0	0	0	0	0	0	0	71.673
21	0	0	0	0	0	0	0	0	0	0	71.673
22	0	0	0	0	0	1	0	0	0	0	14.769
23	0	0	0	0	0	0	0	0	0	0	71.673
24	0	0	0	0	1	0	0	0	0	0	9.9515
25	1	0	0	0	0	0	0	0	1	0	0.8358
26	0	0	0	0	0	0	0	0	0	0	71.673
27	0	0	0	0	0	0	0	0	0	0	71.673
28	0	0	0	0	0	0	0	0	0	0	71.673
29	0	0	0	0	0	0	1	0	0	0	50.653
30	0	0	0	0	0	0	0	0	0	0	71.673
31	0	0	0	1	0	0	0	0	0	0	13.597
32	0	0	0	0	0	0	0	0	0	0	71.673
33	0	0	0	0	0	0	1	0	0	0	50.653
34	0	0	0	0	0	0	0	0	0	0	71.673
35	0	0	0	0	0	0	0	0	0	1	14.359
36	0	0	0	0	0	0	0	0	0	0	71.673
37	0	0	0	0	0	0	1	0	0	1	10.352

Table 2: Group 1-non KCS

	Symptoms											
KCS Patient	1	2	3	4	5	6	7	8	9	10	GCE	Maximum Likelihood
1	0	1	1	1	1	1	1	0	1	1	1.2927×10^{6}	1.8381×10^{5}
2	1	1	1	1	0	0	0	0	1	0	9277.4	1323.9
3	1	0	1	1	1	1	1	0	1	1	7.5692×10^{6}	1.115×10^{6}
4	0	1	1	1	1	1	0	0	0	0	2545.9	1835.3
5	1	1	0	1	0	0	0	0	1	0	652.68	180.17
6	1	1	1	0	1	0	1	0	0	0	206.79	56.087
7	1	1	1	1	0	1	1	1	0	1	2.9445×10^{6}	7.8095×10^{5}
8	1	0	1	0	1	0	0	0	1	0	75.469	136.48
9	1	1	1	1	0	1	0	0	0	0	9569.2	11940
10	1	1	1	0	1	1	1	1	1	0	2.039×10^{5}	70854
11	1	1	0	1	1	0	0	0	0	0	379.53	438.74
12	1	1	1	1	0	1	0	1	0	0	3.7973×10^{5}	6.7367×10^{5}
13	1	1	1	0	0	0	0	1	0	0	1165.3	896.3
14	0	1	1	1	0	0	0	0	0	0	25.447	21.869
15	1	1	1	0	1	0	0	0	0	0	476.64	579.01
16	1	1	0	0	0	0	0	1	0	0	24.154	31.903
17	1	1	1	0	0	1	1	0	0	0	141.13	21.43
18	1	0	1	1	0	0	0	0	0	0	28.576	25.058
19	1	1	1	1	0	1	1	1	0	1	2.9445×10^{6}	7.8095×10^{5}
20	1	1	1	0	0	0	0	0	0	0	28.424	37.81
21	0	1	0	1	1	0	0	1	0	1	1633	1189.1
22	1	1	0	0	1	1	1	0	0	0	447.05	129.59
23	1	1	0	1	1	0	0	0	0	0	379.53	438.74
24	1	0	0	1	1	0	0	1	0	0	381.13	254.75

Table 3: Group 2-KCS

	Symptoms											
Non-KCS Patient	1	2	3	4	5	6	7	8	9	10	GCE	Maximum Likelihood
1	0	0	1	0	0	1	0	0	0	0	2.9395	4.2407
2	0	0	0	0	0	0	0	1	0	0	14.625	24.687
3	0	0	0	0	1	1	0	0	0	0	2.1732	2.6514
4	0	0	0	0	0	0	0	1	0	0	14.625	24.687
5	0	0	1	0	0	0	0	0	0	0	14.246	10.971
6	0	0	0	0	0	1	0	0	0	0	15.354	26.05
7	0	0	0	0	0	0	0	0	0	0	73.767	53.258
8	0	0	0	0	0	0	0	0	0	0	73.767	53.258
9	0	0	0	0	0	1	0	0	0	0	15.354	26.05
10	0	0	0	0	0	0	0	0	0	0	73.767	53.258
11	0	0	0	0	0	0	0	0	0	0	73.767	53.258
12	0	0	1	0	0	0	0	0	0	0	14.246	10.971
13	0	0	0	0	0	0	0	0	0	0	73.767	53.258
14	0	0	0	0	0	0	0	1	0	0	14.625	24.687
15	0	0	0	0	0	0	0	0	0	0	73.767	53.258
16	0	0	0	0	0	0	0	0	0	0	73.767	53.258
17	0	0	0	0	0	1	0	0	0	0	15.354	26.05

Table 4: Group 2-non KCS

i-th non-KCS observation	$K(\mathbf{x} \mid \sigma^*, \mathbf{X}_i)$ with $\mathbf{X}_i$ given below										i-th weight λ_i
36	0	0	0	0	0	0	0	0	0	0	0.84474
33	0	0	0	0	0	0	1	0	0	0	0.15115
24	0	0	0	0	1	0	0	0	0	0	0.0011155
3	0	1	1	0	0	0	1	0	0	0	0.0030032

Table 5: "Group 1-non KCS" mixture pmf with $\sigma^* = 0.947666$

Index of KCS patient	$K(\mathbf{x} \mid \sigma^*, \mathbf{X}_i)$ with $\mathbf{X}_i$ given by										λ_i
39	0	0	0	0	0	0	0	0	0	0	0.055622
21	0	0	0	0	1	0	0	0	0	0	0.0093176
31	0	1	1	1	0	0	0	0	0	1	0.010266
10	1	0	0	0	0	0	0	0	0	0	0.039071
35	1	0	1	0	1	0	0	1	0	0	0.0057993
18	1	0	1	1	1	0	0	1	0	0	0.019046
12	1	1	0	0	0	0	1	1	1	0	0.012942
6	1	1	0	0	1	0	0	0	0	1	0.011055
16	1	1	0	1	0	0	0	0	1	0	0.0072441
24	1	1	0	1	0	0	1	1	1	0	0.0066307
8	1	1	0	1	0	1	0	0	1	0	0.05691
4	1	1	0	1	1	0	0	1	1	0	0.025143
27	1	1	0	1	1	0	0	1	1	1	2.3899×10^{-5}
3	1	1	0	1	1	1	0	0	1	0	0.0053602
23	1	1	1	0	1	0	0	0	0	1	0.037531
36	1	1	1	1	0	0	0	1	0	0	0.18358
20	1	1	1	1	0	0	1	0	0	1	0.22707
2	1	1	1	1	1	0	0	1	0	0	0.095159
9	1	1	1	1	1	0	1	1	0	0	0.0024979
32	1	1	1	1	1	1	1	0	0	1	0.18974

Table 6: "Group 1-KCS" mixture pmf with $\sigma^* = 0.79275$

References

[1] J. Aitchison and C.G.G. Aitken. Multivariate binary discrimination by the kernel method. *Biometrika*, 63:413–420, 1976.

[2] Z. I. Botev. *Stochastic Methods for Optimization and Machine Learning*. ePrintsUQ, http://eprint.uq.edu.au/archive/00003377/, BSc (Hons) Thesis, Department of Mathematics, School of Physical Sciences, The University of Queensland, 2005.

[3] M. J. Faddy and M. C. Jones. Semiparametric smoothing for discrete data. *Biometrika*, 85:131–138, 1998.

[4] P. Hall. On nonparametric multivariate binary discrimination. *Biometrika*, 68:287–294, 1981.

[5] J. H. Havrda and F. Charvat. Quantification methods of classification processes:concepts of structural α entropy. *Kybernatica*, 3:30–35, 1967.

[6] J. Williamson J. A. Anderson, K. Whaley and W. W. Buchanan. A statistical aid to the diagnosis of keratoconjunctivitis sicca. *Quarterly Journal of Medicine*, 41:175–189, April,1972.

[7] J. N. Kapur. *Maximum Entropy Models in Science and Engineering*. Wiley Eastern, New Delhi, India, 1989.

[8] J. N. Kapur and H. K. Kesavan. The generalized maximum entropy principle. *IEEE Transactions on Syst., Man., and Cybernetics*, 19:1042–1052, 1989.

[9] P. A. Lachenbruch and M. R. Mickey. Estimation of error rates in discriminant analysis. *Technometrics*, 10:1–10, 1968.

[10] M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scan. J. Statist.*, 9:65–78, 1982.

[11] D. W. Scott. *Multivariate Density Estimation. Theory, Practice and Visualization*. John Wiley & Sons, 1992.

[12] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423;623–659, 1948.

[13] J. S. Simonoff. Smoothing categorical data. *Journal of Statistical Planning and Inference*, 47:41–69, 1995.

[14] D.M. Titterington. A comparative study of kernel-based density estimates for categorical data. *Technometrics*, 22:259–268, 1980.

[15] Frederick Y.M. Wan. *Introduction To The Calculus of Variations and Its Applications*. Chapman and Hall, 1995.