NON-ASYMPTOTIC BANDWIDTH SELECTION FOR DENSITY ESTIMATION OF DISCRETE DATA

B Z I. B D P. K [*]

The University of Queensland

We propose a new method for density estimation of categorical data. The method implements a non-asymptotic data-driven bandwidth selection rule and provides model sparsity not present in the standard kernel density estimation method. Numerical experiments with a well-known 10-dimensional binary medical dataset illustrate the effectiveness of the proposed approach for density estimation, discriminant analysis and classification.

1. Introduction.

Aitchison and Aitken [2] proposed an extension of the standard kernel density estimation techniques for continuous data to multivariate discrete spaces. Similar to the continuous kernel density estimator, the performance of their proposed estimator depends critically on a smoothing parameter — usually called the *bandwidth*. Aitchison and Aitken suggest maximizing the cross-validatory (also known as 'leave-one-out') likelihood function as a method of estimating the bandwidth. Hall [12] argues that Aitchison and 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 sug-

---

[*] Supported by the Australian Research Council, under grant number DP0558957

AMS 2000 subject classifications: Primary 94A17, 60K35; secondary 68Q32, 93E14

Keywords and phrases: Bandwidth selection, kernel density estimator, Generalized Cross Entropy, statistical modeling, discrete data smoothing, multivariate binary discrimination

gested. Most of these methods rely on asymptotic approximations. For a survey of the various bandwidth selection methods see Titterington [34], Faddy [10], Simonoff [31] and the references therein. In this paper we propose a new bandwidth selection rule which does not rely on asymptotic approximations and the concomitant availability of large samples. The bandwidth selection rule is derived from the Generalized Cross Entropy (GCE) method [5]. The GCE method is based on the pioneering work of Kapur and Kesavan, who in a series of papers and books — the most notable of which are [20], [17], [18] and [16] — describe a far-reaching generalization of the maximum entropy method of Jaynes [14], and the information-theoretic concepts of Shannon [29], Kullback and Leibler [21] and Havrda and Charvát [13]. The entropic principles described by Kapur and Kesavan [18] have been recently used in various statistical data modelling contexts [1], in neural networks [9, 24] and with great success in Stochastic Optimization and Rare-Event Simulation [26], [25]. The generalized entropic measures pioneered by Havrda and Charvát [13] and later advocated by Kapur and Kesavan [18] have for a long time been used successfully in statistical mechanics (see Tsallis [35]).

The paper is organized as follows. In Section 2 we briefly review Aitchison and Aitken's likelihood cross-validation method; and consider possible remedies for the problems mentioned above. In Section 3 we present a new framework for density estimation and bandwidth selection for categorical data, based on the GCE approach. In Section 4 we apply the method to a classical example in multivariate binary discrimination. The example is based on medical data described in Anderson [3] and is the same one used by Aitchison and Aitken [2] and some of the follow-up papers (see Hall

[12] and Titterington [34]) to compare the proposed statistical methods. The example clearly demonstrates the practical benefits of the GCE approach as a tool for optimal bandwidth selection and more generally for discrete multivariate data modeling.

2. The Kernel Approach.

In this section we briefly review the kernel density estimation approach to modelling discrete data. We note that the approach is, in essence, the same as for the continuous case (for a comprehensive discussion of the continuous case see, for example, Scott [28], Simonoff [32], Wand and Jones [36] and Silverman [30]).

Suppose we are given the observations $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ on a discrete $d-$dimensional space $\mathscr{X}$, where each $\mathbf{X}_i$ is a row vector of length d. The data is modeled via a probability mass function (pmf) $f : \mathscr{X} \to [0, 1]$, such that $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{i.i.d}{\sim} f$. The kernel method assumes that the true, but unknown, underlying pmf f can be approximated well by a pmf of the form:

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

where:

1. $K : \mathscr{X} \to [0, 1]$, $\sum_{\mathbf{x}\in\mathscr{X}} K(\mathbf{x}\,|\,\cdot) = 1$ is a unimodal pmf, usually referred to as the *kernel* function.
2. $\sigma \in \mathbb{R}^+$ is a bandwidth parameter which controls the “smoothness” of g in a way similar to the bandwidth in continuous kernel density estimation.

E 1 (Binary Kernel). For the rest of the paper we assume that the data $\mathcal{X}$ is binary and choose the the binary kernel:

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

where $d-\mathrm{d}(\mathbf{x},\mathbf{y}) = d-\sum_{l=1}^{d} I\{\mathbf{x}(l)=\mathbf{y}(l)\}$ measures the number of mismatches between the vectors $\mathbf{x}$ and $\mathbf{y}$ (Hamming distance). Here $I\{\cdot\}$ denotes the indicator function, i.e., it is one if the statement inside the brackets is true and zero otherwise. The kernel is nothing more than a multivariate Bernoulli pmf. Note that:

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

The end-points of the interval $[\frac{1}{2},1]$ represent two extremes of smoothing: for $\sigma=1$ there is no smoothing and g is simply estimated from corresponding relative frequencies; for $\sigma=\frac{1}{2}$, K is not unimodal and g is the uniform pmf on $\mathscr{X}$. For $\sigma=\frac{1}{2}$ the model pmf g is the most uninformative, in the sense that g is closest to the uniform pmf on $\mathscr{X}$. We thus impose the restriction $\sigma\in(\frac{1}{2},1)$ to guarantee that each kernel $K(\mathbf{x}\,|\,\sigma,\mathbf{X}_i)$ has a single mode at $\mathbf{x}=\mathbf{X}_i$ and that g is "smooth" in the sense that $g(\mathbf{x}\,|\,\sigma,\mathcal{X})>0,\ \forall x\in\mathscr{X}$.

Note that everything in (1) is fixed except the bandwidth parameter σ. This is the only parameter over which we have control. The performance of the estimator (1) depends crucially on the parameter σ. We need to adjust σ so that our approximation of f is as good as possible: not too small so that the resulting model pmf g is too peaked and sample dependent, not too large so that the pmf g is too uniform and uninformative about the patterns in the data. To select an optimal value for the bandwidth σ we first need to find a criterion for the performance of (1).

Measuring the estimator's performance. Under the assumption that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent, one possible measure of performance is the observed likelihood function:

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

Unfortunately, maximization of the likelihood function leads to the undesirable value $\sigma = 1$ (i.e. (1) is an estimator based on relative frequencies), in much the same way as it leads to the empirical Dirac delta function in the continuous case. Similar to the kernel density estimation method in the continuous case (see e.g. Loader [23], Stone [33] and Bowman *et al.* [6]) this problem is resolved by considering the *leave-one-out* or cross-validated likelihood:

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

This leads to the Aitchison and Aitken's [2] choice for σ:

$$\hat{\sigma} = \operatorname*{argmax}_{\sigma \in (\frac{1}{2}, 1)} L_{\mathrm{x}}(\sigma \,|\, \mathcal{X}) \, . \tag{8}$$

The consistency properties of (8) are proved in Aitchison [2]. Hall [12], however, argues that (8) can behave erratically and often undersmooths by giving a solution close to 1. Occasionally the global 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.

This type of behavior is also typical for the continuous analogue of (8) known as the Least Squares Cross Validation (LSCV) method (Rudemo [27])). To overcome these problems, Hall [12] 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 (Jones *et al.* [15]). Hall's estimation method is largely asymptotic in nature. For a treatment of the above estimation problem via Markov Chain theory see Faddy and Jones [10] and for a comparative study of the various kernel-based categorical data smoothing techniques see Titterington [34]. We now consider a different approach to the optimal bandwidth choice for the pmf in (1) .

3. The Generalized Cross Entropy Method.

The GCE method (Botev and Kroese [5]) provides an information theoretic approach to the problem of density estimation, statistical learning and Monte Carlo simulation. We briefly summarize the purport of the method.

Let the probabilistic model of the data $\mathcal{X}$ be presented by the pmf $g(\mathbf{x})$. Then the GCE 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 model g to be as simple as possible while at the same time being truthful and consistent with the observed data. This aspiration 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).

The Measure of Complexity. A commonly used measure of complexity in Statistics and Information theory is the negative of Shannon's Entropy [29] $\sum_{\mathbf{x}\in\mathscr{X}} g(\mathbf{x}) \ln g(\mathbf{x})$. Kapur and Kesavan [18] argue that in many cases, instead of considering the Shannon Entropy, it is more convenient to consider the well-known Kullback-Leibler distance from the uniform pmf:

$$\sum_{\mathbf{x}\in\mathscr{X}} g(\mathbf{x}) \ln \frac{g(\mathbf{x})}{p(\mathbf{x})}, \quad p(\mathbf{x}) = \frac{1}{|\mathscr{X}|}.$$

The justification is based on the interpetation of Shannon's Entropy as the Kullback-Leibler distance of the model pmf g from the uniform pmf on the set $\mathscr{X}$:

$$\underbrace{\sum_{\mathbf{x}\in\mathscr{X}} g(\mathbf{x}) \ln g(\mathbf{x})}_{\text{Shannon's Entropy}} + \underbrace{\ln |\mathscr{X}|}_{\text{constant}} = \underbrace{\sum_{\mathbf{x}\in\mathscr{X}} g(\mathbf{x}) \ln \frac{g(\mathbf{x})}{p(\mathbf{x})}}_{\text{Kullback-Leibler distance}} .$$

Moreover the concept of a measure of distance between two pmfs is more intuitive than the concept of the Entropy of a pmf.

It is well known that Kullback-Leibler's distance is a special limiting case of a one-parameter family of distance measures first studied by Havrda and Charvát (see [13]):

$$\mathcal{D}_\alpha(g \to p) = \frac{1}{\alpha(\alpha-1)} \left[-1 + \sum_{\mathbf{x}\in\mathscr{X}} g^\alpha(\mathbf{x})\, p^{1-\alpha}(\mathbf{x}) \right], \quad \alpha \neq 0, 1.$$

It is easy to verify that $\lim_{\alpha\to 1} \mathcal{D}_\alpha\{g \to p\}$ equals Kullback-Leibler's distance. Kesavan and Kapur [19] argue that one can use the Havrda-Charvát one-parameter family as a measure of complexity, taking p as the uniform pmf on

$\mathscr{X}$. This interpretation of $\mathcal{D}$ suits our purpose and we will use, in particular, the so-called χ^2-distance

$$\mathcal{D}_2 = \mathcal{D}_2(g \to p) = \frac{1}{2} \sum_{\mathbf{x} \in \mathscr{X}} \frac{[g(\mathbf{x}) - p(\mathbf{x})]^2}{p(\mathbf{x})}, \quad p(\mathbf{x}) = \frac{1}{|\mathscr{X}|} \tag{9}$$

to measure the complexity of our proposed model for the data $\mathcal{X}$. This choice is motivated by its numerous advantages over the other information measures from the family $\mathcal{D}_\alpha$, including Kullback-Leibler's measure. These advantages include:

1. $\mathcal{D}_2$ enjoys computational tractability shared by none of the other distance measures and the mathematical expressions arising from using $\mathcal{D}_2$ are conceptually simpler and more elegant (see Botev and Kroese [5]).
2. It is easy to show that (see e.g. [4]):
   $$2\,\mathcal{D}_2 \geqslant \sum_{\mathbf{x} \in \mathscr{X}} g(\mathbf{x}) \ln \frac{g(\mathbf{x})}{p(\mathbf{x})} \geqslant \ln\left[1 + 2\,\mathcal{D}_2\right], \qquad 2\,\mathcal{D}_2 \geqslant \left[\sum_{\mathbf{x} \in \mathscr{X}} |g(\mathbf{x}) - p(\mathbf{x})|\right]^2,$$
   i.e., $\mathcal{D}_2$ dominates other useful measures, and minimizing $\mathcal{D}_2$ with respect to the model g is also minimizing an upper bound on the Kullback-Leibler distance and the L_1 metric. The theoretical significance of the L_1 metric is explained in Devroye [8].
3. $\mathcal{D}_2$ is a first order approximation to the Kullback-Leibler distance (see [4]) in the sense that $\sum_{\mathbf{x} \in \mathscr{X}} g(\mathbf{x}) \ln \frac{g(\mathbf{x})}{p(\mathbf{x})} \approx \mathcal{D}_2$ for g and p close to each other. More precisely:
   $$\sum_{\mathbf{x} \in \mathscr{X}} g(\mathbf{x}) \ln \frac{g(\mathbf{x})}{p(\mathbf{x})} = \mathcal{D}_2 - \frac{1}{6} \sum_{\mathbf{x} \in \mathscr{X}} p(\mathbf{x}) \varepsilon^3(\mathbf{x}) + \frac{1}{12} \sum_{\mathbf{x} \in \mathscr{X}} p(\mathbf{x}) \varepsilon^4(\mathbf{x}) - HOT,$$
   where $\varepsilon(\mathbf{x})$ is a perturbation such that $\varepsilon(\mathbf{x}) = g(\mathbf{x})/p(\mathbf{x}) - 1,\ |\varepsilon(\mathbf{x})| < 1,\ \forall \mathbf{x} \in \mathscr{X}$.

Fidelity to the Empirical Data. Having chosen the measure of model complexity (9), 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:

$$\sum_{\mathbf{x}\in\mathscr{X}} g(\mathbf{x})\; K_i(\mathbf{x}) = \mathbb{E}_g[K_i(\mathbf{X})] \geqslant \hat{\kappa}_i, \quad i = 1,\ldots,n, \tag{10}$$

where:

- $\{K_i(\mathbf{x}) = K(\mathbf{x}\,|\,\sigma,\mathbf{X}_i)\}_{i=1}^n$ is a set of kernels. For concreteness consider empirical binary data $\mathcal{X} = \{\mathbf{X}_1,\ldots,\mathbf{X}_n\} \overset{i.i.d}{\sim} f$ and choose $K_i(\mathbf{x}) = K(\mathbf{x}\,|\,\sigma,\mathbf{X}_i)$ to be the binary kernel discussed in Example 1. Note that $K_i(\mathbf{x})$ is a function of the random variable $\mathbf{X}_i$ and depends on the smoothing parameter σ.
- $\hat{\kappa}_i$ is a statistical estimate of $\mathbb{E}_f[K_i(\mathbf{X})]$ based on the empirical data $\mathcal{X}$.

The constraints (10) embody nothing more than the simple concept of *moment matching* first advocated by Karl Pearson (see Scott [28]). We "match" the moments $\{\mathbb{E}_g[K_i(\mathbf{X})]\}_{i=1}^n$ of our proposed pmf g to the empirical moments $\{\hat{\kappa}_i\}_{i=1}^n$ (which approximate the moments $\{\mathbb{E}_f[K_i(\mathbf{X})]\}_{i=1}^n$ under the unknown pmf f). Here we use the notion of a moment in its generalized sense, i.e., we call $\mathbb{E}_f[Q(x)]$, for any function Q, a moment of f. We now consider how to estimate the moments $\{\mathbb{E}_f[K_i(\mathbf{X})]\}_{i=1}^n$ from the empirical data $\mathcal{X}$.

Estimation of $\{\mathbb{E}_f[K_i(\mathbf{X})]\}_{i=1}^n$. We use the *same* set $\mathcal{X} = \{\mathbf{X}_1,\ldots,\mathbf{X}_n\}$ as both kernel location parameters and as a sample for the estimation of $\{\mathbb{E}_f K_i(\mathbf{X})\}_{i=1}^n$. Note that $\mathbb{E}_f K_i(\mathbf{X})$ is a function of $\mathbf{X}_i$ and hence is a random variable. Under the assumption $\mathbf{X}_1,\ldots,\mathbf{X}_n \overset{i.i.d}{\sim} f$, each $\mathbf{X}_i$ is independent of all the other $\{\mathbf{X}_j\}_{j\neq i}$

and a simple unbiased estimator of $\mathbb{E}_f K_i(\mathbf{X}) = \mathbb{E}_f K(\mathbf{X} \,|\, \sigma, \mathbf{X}_i)$ is:

$$\hat{\kappa}_i = \hat{\kappa}(\sigma, \mathbf{X}_i) = \frac{1}{n-1} \sum_{j \neq i}^{n} K(\mathbf{X}_j \,|\, \sigma, \mathbf{X}_i). \tag{11}$$

This is the *cross-validatory*, also known as *leave-one-out*, estimator and its consistency properties are established in Bowman [7] and Stone [33]. The estimator (11) is similar to the cross validatory estimators used in the Least Squares Cross Validation Method (LSCV) for continuous density estimation (see Rudemo [27]).

R 1 (Choice of Constraints). In the GCE approach the constraints (10) are inequality constraints. The Maximum Entropy Method (see Kapur [16] and Jaynes [14]), however, uses equality constraints only:

$$\mathbb{E}_g[K_i(\mathbf{X})] = \hat{\kappa}_i, \quad i = 1, \ldots, n.$$

Thus the GCE approach differs from the Maximum Entropy approach in this respect. The advantage of using (10) as opposed to equality constraints is one of computational tractability.

The GCE Optimization Problem. We now restate the GCE formalism with the relevant mathematical detail. Given the data $\mathcal{X}$,

1. minimize the measure of model complexity

$$\mathcal{D}_2\{g \to p\} = -\frac{1}{2} + \frac{1}{2} \sum_{\mathbf{x} \in \mathscr{X}} \frac{g^2(\mathbf{x})}{p(\mathbf{x})}, \quad p(\mathbf{x}) = \frac{1}{|\mathscr{X}|},$$

   with respect to the model g; that is, determine

$$\min_g \sum_{\mathbf{x} \in \mathscr{X}} g^2(\mathbf{x}), \tag{12}$$

2. subject to the constraints (agreement with empirical data):

$$\sum_{\mathbf{x}\in\mathscr{X}} g(\mathbf{x})\; K(\mathbf{x}\,|\,\sigma,\mathbf{X}_i) = \mathbb{E}_g[K_i(\mathbf{X})] \geqslant \hat{\kappa}(\sigma,\mathbf{X}_i) = \hat{\kappa}_i, \quad i = 1,\ldots,n. \tag{13}$$

The solution to this *variational optimization problem* is given in the following theorem, which is proved in the appendix.

Theorem 1 (Optimal GCE Solution). The optimal solution of the convex optimization problem (12)+(13) is given by the kernel mixture:

$$g(\mathbf{x}) = \sum_{j=1}^{n} \lambda_j\; K(\mathbf{x}\,|\,\sigma,\mathbf{X}_j), \tag{14}$$

where $\boldsymbol{\lambda} = [\lambda_1,\ldots,\lambda_n]^T$ are *Lagrange multipliers* determined by the solution of the convex *Quadratic Programming Problem* (QPP):

$$\begin{aligned} \min_{\boldsymbol{\lambda}} \quad & \frac{1}{2}\boldsymbol{\lambda}^T C\boldsymbol{\lambda} - \boldsymbol{\lambda}^T\hat{\boldsymbol{\kappa}} && (15)\\ \text{subject to:} \quad & \boldsymbol{\lambda} \geqslant \mathbf{0}. && (16) \end{aligned}$$

Here C is the $n\times n$ matrix with entries

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

and $\hat{\boldsymbol{\kappa}}(\sigma,\mathcal{X}) = [\hat{\kappa}(\sigma,\mathbf{X}_1),\ldots,\hat{\kappa}(\sigma,\mathbf{X}_n)]^T$.

Note that both the moments $\hat{\boldsymbol{\kappa}}$ and the matrix C are functions of the smoothing parameter σ and the empirical data $\mathcal{X}$. The theorem provides the optimal solution of the convex problem (12)+(13) for a given value of the bandwidth parameter σ and thus does not provide an appropriate value for the bandwidth. We now consider the important problem of automatic (data-driven) bandwidth selection.

Selecting the bandwidth σ. A data-driven bandwidth selection rule is provided as follows: Let $\mathscr{G} = \{g : \sum_{\mathbf{x}\in\mathscr{X}} g(\mathbf{x}) = 1,\ g(\mathbf{x}) \geqslant 0\ \forall \mathbf{x}\}$ be the set of all *bona fide* probability density functions on $\mathscr{X}$. Observe that the integral (sum) of (14) is

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

The solution (14) is always a non-negative function, therefore $g \in \mathscr{G}$ is equivalent to $\sum_{j=1}^{n} \lambda_j = 1$. We now select the as yet unspecified bandwidth σ to meet the constraint $\sum_{j=1}^{n} \lambda_j = 1$, i.e., we choose the bandwidth parameter σ such that:

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

Now all of the parameters in the model are specified and the GCE approach is summarized in the following algorithm:

A 1.

1. Solve the root-finding program:

$$(\sigma^*, \boldsymbol{\lambda}^*) = \left\{(\sigma, \boldsymbol{\lambda}) : \ \mathbf{1}^T \boldsymbol{\lambda}(\sigma) = 1,\ \boldsymbol{\lambda}(\sigma) = \operatorname*{argmin}_{\boldsymbol{\lambda}\geqslant\mathbf{0}} \left(\frac{1}{2}\boldsymbol{\lambda}^T C(\sigma)\boldsymbol{\lambda} - \boldsymbol{\lambda}^T \hat{\boldsymbol{\kappa}}(\sigma)\right)\right\},$$

where matrix C and vector $\hat{\boldsymbol{\kappa}}$ are the elements of the QPP (15)+(16) in theorem 1.

2. Present the mixture pmf:

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

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

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 [3]. The same data set was used by Aitchison & Aitken [2] and we compare the results of the two studies. The data set is described in four tables:

1. Table 1 gives the symptoms of 40 patients suffering from KCS. 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. Table 2 describes the symptoms of 37 patients who do not suffer from the disease, i.e., non-KCS patients.
3. The patients in Table 1 and Table 2 form the first group of 77 patients, referred to as group-1. This group forms the *training* sample from which the statistical model is build.
4. The same 10 symptoms are recorded for another group of 41 patients, henceforth referred to as the group-2 patients. This group represents the *test* sample which is used to validate the statistical model build from the *training* sample. Table 3 describes the symptoms of the 24 KCS-afflicted patients from group 2 and Table 4 describes the symptoms the 17 non-KCS patients from group 2.

We now use the GCE method to estimate the pmf of the “group-1 KCS” observations as g_{KCS} and the pmf of the “group-1 non-KCS” observations as $g_{\mathrm{non-KCS}}$. The estimated mixture pmf $g_{\mathrm{non-KCS}}$ is summarized in Table 5. Kernels associated with zero Lagrange multiplier are not listed as they do not contribute toward the value of $g_{\text{non-KCS}}(\mathbf{x})$ for any $\mathbf{x} \in \mathscr{X}$. Similarly the mixture pmf g_{KCS} is summarized in Table 6. This time the number of observations (and hence kernels) associated with a zero multiplier is much larger. The GCE smoothing parameters obtained via Algorithm 1 are $\sigma^*_{\text{KCS}} = 0.79275$ and $\sigma^*_{\text{non-KCS}} = 0.947666$. For comparison, Aitchison and Aitken [2] report bandwidth values of 0.843 (KCS) and 0.96 (non KCS) obtained via the leave-one-out maximum likelihood approach. 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*

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

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

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

then the i-th patient from group-2 is classified as KCS patient. The same classification procedure is used by Aitchison & Aitken [2]. The odds ratios computed using the GCE and Aitchison & Aitken's method are given in the last two columns 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 [22]. 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 $\sigma_{\text{KCE}}^{\text{new}}$ is calculated from the reduced set and then the omitted patient is classified using $\sigma_{\text{KCE}}^{\text{new}}$ 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 [3]) confirm the KCS diagnosis.

Implementation Issues. Some issues concerning the implementation of Algorithm 1 are:

1. The `mosekopt` software was used to solve the Quadratic Programming Problem arising at each step of the root-finding problem in step 1 of Algorithm 1. The software was downloaded from

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

   An advantage of the `mosekopt` solver is that it recognizes any linearly dependent rows of the matrix C and removes them before applying a primal-dual interior point algorithm to the convex QPP problem.
2. To find the bandwidth σ such that $\sum_{j=1}^{n} \lambda_j^* = 1$, where $\left\{\lambda_j^*\right\}_{j=1}^{n}$ is the solution of a convex QPP (this is step 1 of Algorithm 1), the Matlab root-finding function `fzero.m` was used. Each iteration of `fzero.m` requires the solution of a QPP (as per step 1 of Algorithm 1) and

hence calls the `mosekopt` QPP solver.

Discussion and Conclusions. We have presented a new approach to the problem of density estimation and discrimination based on simple information theoretic concepts. Some advantages of the proposed GCE approach relative to the traditional non-parametric kernel estimation approach are:

- The GCE mixture pmf which models the data is sparse in the sense that the number of non-zero mixture components is usually much smaller than the number of observations. This is in sharp contrast to the traditional kernel density estimation techniques where the model pmf is an equally weighted mixture with as many components as the number of observations. The sparsity of the model pmf makes it computationally less expensive to evaluate and manipulate the pmf, e.g., in the context of discrimination as described in this paper.
- The Maximum Entropy Method of Jaynes [14] has for a long time been the only method to provide a density estimator derived from the solution of a functional optimization problem. Our proposed model is thus the second desnity estimator to be derived from the solution of a functional optimization problem. In contrast to the popular Maximum Entropy Method, we use Pearson's divergence measure (as opposed to Kullback-Leibler's measure) in combination with generalized moment constraints to construct the model. To the best of our knowledge, Pearson's divergence measure has not been used in a functional optimization context to produce a probability model.
- Whilst the popular maximum entropy programs [16] in probability and statistical mechanics utilize the Euler-Lagrange techniques of op-

timization, we exploit for the first time the more general Kuhn-Tucker duality to provide a bandwidth selection rule for the resulting estimator. The proposed approach gives rise to a bandwidth selection rule which is justified without appealing to asymptotic arguments. The only other currently existing non-asymptotic bandwidth selection rule is derived from the (least squares or likelihood) cross validation method. Non-asymptotic methods make it possible to construct a good statistical model in situations where the sample size is not large enough and models based on asymptotic approximations fail.

Finally the method provides another method for bandwidth selection which is useful in complementing the currently existing techniques. A thorough modern analysis of a given data set will usually involve the application of all existing methods and accounting for any discrepancies in the results and conclusions derived from the various techniques.

APPENDIX A: PROOF OF THEOREM 1

To solve the (convex) constraint optimization problem (12)+(13) we use the Lagrange multiplier method. Define the Lagrangian:

$$\begin{aligned} &\mathcal{L}(g;\boldsymbol{\lambda}) = \\ =\; & -\frac{1}{2}+\frac{1}{2}\sum_{\mathbf{x}\in\mathscr{X}}\frac{g^2(\mathbf{x})}{p(\mathbf{x})}-\sum_{i=1}^{n}\lambda_i\left(\sum_{\mathbf{x}\in\mathscr{X}} g(\mathbf{x})\, K_i(\mathbf{x})-\hat{\kappa}_i\right) \\ =\; & -\frac{1}{2}+\sum_{i=1}^{n}\lambda_i\,\hat{\kappa}_i+\sum_{\mathbf{x}\in\mathscr{X}}\left(\frac{1}{2}\frac{g^2(\mathbf{x})}{p(\mathbf{x})}-g(\mathbf{x})\sum_{i=1}^{n}\lambda_i\, K_i(\mathbf{x})\right), \end{aligned}$$

where $\boldsymbol{\lambda}=[\lambda_1,\ldots,\lambda_n]^T$. If we think of (12)+(13) as the **Primal** optimization problem (see [11]) then its corresponding **Dual** is:

$$\begin{array}{rl}\max\limits_{\boldsymbol{\lambda}} & \left\{\inf\limits_{g} \mathcal{L}(g;\boldsymbol{\lambda})\right\} \\ \text{subject to:} & \boldsymbol{\lambda} \geqslant \mathbf{0}\end{array}$$

We now simplify the dual. First $\inf_g \mathcal{L}(g;\boldsymbol{\lambda})$ can be calculated explicitly by setting $\frac{\partial \mathcal{L}}{\partial g} = 0$. This gives:

$$\frac{g(\mathbf{x})}{p(\mathbf{x})} = \sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{x}) \; . \tag{17}$$

We can then substitute this $g(\mathbf{x})$ into the Lagrangian to obtain:

$$\begin{aligned}\mathcal{L}^*(\boldsymbol{\lambda}) &= \inf_g \mathcal{L}(g;\boldsymbol{\lambda}) \\ &= -\frac{1}{2} + \sum_{i=1}^{n} \lambda_i \, \hat{\kappa}_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \, \lambda_j \, \mathbb{E}_p K_i(\mathbf{X}) K_j(\mathbf{X})\end{aligned}$$

Thus, the dual becomes:

$$\begin{array}{rl}\max\limits_{\boldsymbol{\lambda}} & \mathcal{L}^*(\boldsymbol{\lambda}) \\ \text{subject to:} & \boldsymbol{\lambda} \geqslant \mathbf{0} \; .\end{array}$$

Since $p = 1$ is the uniform pmf on $\mathscr{X}$, it can be treated as a scaling constant in the optimization. Written in matrix notation with $p = 1$, the above dual is equivalent to the QPP (15)+(16). Thus the solution of the primal (12)+(13) is obtained via:

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

where $\{\lambda_j\}$ is the unique solution vector of (15)+(16). Note that the primal problem is strictly convex. It is well known (see Fletcher [11]) that if the primal problem is (strictly) convex the dual problem is (strictly) concave and the solution of the primal, which is a (unique) minimizer, coincides exactly with the solution of the dual— a (unique) maximizer. Thus the solution of our dual problem must be unique.

	Symptoms of KCS-afflicted patient										
Patient number	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

T 1

Group 1-KCS: The ten symptoms for each of the 40 KCS-afflicted patients from group 1. The misclassification odds ratio for each patient is given in the last column.

	Symptoms of non-KCS patient										
Patient number	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

T 2

Group 1-non KCS: The ten symptoms for each of the 37 non-KCS patients from group 1. The misclassification odds ratio for each patient is given in the last column.

	Symptoms of KCE-afflicted patient											
Patient number	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

T 3

Group 2-KCS: The ten symptoms for each of the 24 KCS-afflicted patients from group 2. The odds ratios for each patient provided by the GCE and Maximum Likelihood methods are given in the last two columns.

	Symptoms of non-KCE patient											
Patient number	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

T 4

Group 2-non KCS: The ten symptoms for each of the 17 non-KCS patients from group 2. The odds ratios for each patient provided by the GCE and Maximum Likelihood methods are given in the last two columns.

i-th non-KCS patient	$K(\mathbf{x} \,|\, \sigma^*, \mathbf{X}_i)$ 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

T 5

"Group 1-non KCS" mixture pmf with $\sigma^* = 0.947666$. *The table presents the mixture weight and location for each of the four kernels constituting the mixture model for the "Group 1-non KCS" patients.*

$i-$th KCS patient	$K(\mathbf{x}\,\vert\,\sigma^{*},\mathbf{X}_i)$ with $\mathbf{X}_i$ given by										$i-$th weight 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

T 6

"Group 1-KCS" mixture pmf with $\sigma^{*} = 0.79275$*. The table presents the mixture weight and location for each of the twenty kernels constituting the mixture model for the "Group 1-KCS" patients.*

REFERENCES

[1] G. Judge A. Golan and D. Miller. *Maximum Entropy Econometrics: robust estimation with limited data*. New York : Wiley Series in Financial Economics and Quantitative Analysis, 1996.

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

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

[4] Z. I. Botev. *Stochastic Methods for Optimization and Machine Learning*. ePrintsUQ, http://eprint.uq.edu.au/archive/00003377/, Technical Report, 2005.

[5] Z. I. Botev and D. P. Kroese. The generalized cross entropy method, with applications to probability density estimation. *Electronic Preprint*, 2006, http://espace.library.uq.edu.au/.

[6] A. W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71:353–360, 1984.

[7] A. W. Bowman. A comparative study of some kernel-based nonparametric density estimators. *J. Statist. Comput. Simul.*, 21:313–327, 1985.

[8] L. Devroye and L. Gyofri. *Nonparametric Density Estimation: The L_1 View*. Wiley Series In Probability And Mathematical Statistics, 1985.

[9] D. Erdogmus and J.C. Principe. An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems. *IEEE Transactions on Signal Processing*, 50, 2002.

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

[11] R. Fletcher. *Practical Methods of Optimization*. John Wiley and Sons, 1987.

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

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

[14] E. T. Jaynes. Information theory and statistical mechanics. *Physical Reviews*, 106:621–630, 1957.

[15] M. C. Jones, J. S. Marron, and S. J. Sheather. Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics*, 11:337–381, 1996.

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

[17] J. N. Kapur. *Measures of Information and Their Applications*. John Wiley & Sons, New Delhi, India, 1994.

[18] J. N. Kapur and H. K. Kesavan. *Generalized Maximum Entropy Prin-*

ciple (With applications). Standford Educational Press, University of Waterloo, Waterloo, Ontario, Canada, 1987.

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

[20] J. N. Kapur and H. K. Kesavan. *Entropy Optimization Principles with Applications*. Academic Press, New York, 1992.

[21] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann.Math.Stat.*, 22:79–86, 1951.

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

[23] C. R. Loader. Bandwidth selection: Classical or plug-in. *The Annals of Statistics*, 27:415–438, 1999.

[24] R. A. Morejon and J. C. Principe. Advanced search algorithms for information-theoretic learning with kernel-based estimators. *IEEE TRANSACTIONS ON NEURAL NETWORKS*, 15, 2004.

[25] R. Y. Rubinstein. The stochastic minimum cross-entropy method for combinatorial optimization and rare-event estimation. *Methodology and Computing in Applied Probability*, 7:5–50, 2005.

[26] R. Y. Rubinstein and D. P. Kroese. *The Cross-Entropy Method*. Springer, 2004.

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

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

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

[30] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

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

[32] J. S. Simonoff. *Smoothing Methods in Statistics*. Springer, 1996.

[33] C. J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, 12, 1984.

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

[35] C. Tsallis. Possible generalization of boltzmann-gibbs statistics. *J. Stat. Phys.*, 52:479, 1988.

[36] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman & Hall, 1995.

D M
T U Q
B 4072
AUSTRALIA
E- : {botev,kroese}@maths.uq.edu.au