Maximum Leave-one-out Likelihood for Kernel Density Estimation

Etienne Barnard
Multilingual Speech Technologies Group, North-West University
Vanderbijlpark 1900 SOUTH AFRICA
Email: etienne.barnard@gmail.com

***Abstract*—We investigate the application of kernel density estimators to pattern-recognition problems. These estimators have a number of attractive properties for data analysis in pattern recognition, but the particular characteristics of pattern-recognition problems also place some non-trivial requirements on kernel density estimation – especially on the algorithm used to compute bandwidths. We introduce a new algorithm for variable bandwidth estimation, investigate some of its properties, and show that it performs competitively on a wide range of tasks, particularly in spaces of high dimensionality.**

I. Motivation

Kernel density estimators (KDEs) approximate the probability density function of a class characterized by a set of samples with the sum

$$p(x) = \sum_{i=1}^{N} K(x - x_i; H_i)/N, \tag{1}$$

where the x_i are the samples provided, K is a "smoothing" or kernel function and H_i is a scale factor (or "bandwidth") controlling the extent of the kernel. (Note that H is defined as a variable with units of length – thus, the "standard deviation" rather than the "variance" for the scalar case. For multivariate tasks, the variance parameter of the kernel is therefore equal to the "square" of H, usually defined as HH^T.) Such estimators are considered to be amongst the most powerful tools for data analysis [1], because of their ability to model arbitrary functional forms in any number of dimensions, with attractive asymptotic properties. These benefits come at the cost of two significant issues:

- Since the sum in (1) ranges over all samples, the amount of computation required to compute a single density estimate is proportional to the number of training samples.
- The behaviour of the estimator depends strongly on the way that the scale factors are chosen. (The exact shape of the kernel function is much less important in practice [2].)

The first of these issues is less important if data analysis, rather than real-time classification, is the intended application. The second issue is the main focus of the current paper.

In pattern recognition, the vast majority of feature spaces are multi-dimensional, with dozens to hundreds of features being quite common. This implies a number of specific requirements on H_i, which we review below in Section II. This allows us to evaluate the strengths and weaknesses of various existing approaches to bandwidth selection in pattern recognition. In Section III we then propose an alternative approach, the "Maximum leave-one-out Likelihood (MLL)" method, which has a number of benefits compared to current approaches in our domain of interest. Section IV contains experimental investigations into the properties of MLL, including comparisons with several standard methods when applied to both standard statistical problems as well as selected tasks from pattern recognition. Finally, conclusions are drawn and extensions are discussed in Section V.

II. KDEs for Pattern Recognition

The standard approach to statistical pattern recognition is based on a set of features, which are selected to be descriptive of the objects of interest. In most domains – including speech processing, natural language processing, image recognition and bioinformatics – the dimensionalities of practically useful feature spaces are quite high. For example, 39-dimensional feature spaces, derived from mel-frequency cepstral coefficients, are common in speech processing, and protein-folding tasks are often addressed with a feature space consisting of the identities of the amino acids within a window, resulting in feature spaces with 140 or more dimensions. It is widely understood (see, for example [3]) that any usable class of objects in such a space must be clustered around one or more sub-manifolds of a much lower dimensionality. (Otherwise, it would be impossible to delineate the objects with a reasonable number of samples.) In the non-degenerate case, these dimensions will not align with the coordinate axes (since the superfluous dimensions could simply be discarded otherwise). Finally, it is reasonable to expect that the scale of geometric structures may vary through feature space: since the physical phenomena that correspond to these structures may be quite different in different regions of feature space – e.g. specular reflection off a smooth surface opposed to diffuse reflection from a textured material in vision, or slowly and smoothly changing formants opposed to the abrupt changes of a plosive burst in speech recognition – the characteristic lengths and directions of variability within feature space are likely to change significantly across various locations in feature space.

These considerations lead us to the following three requirements of a useful bandwidth estimator in pattern recognition:

1) It must be able to operate successfully in spaces with many dimensions.

2) It must be able to deal with spatial anisotropy.
3) It must be able to model spatial variability in smoothness of the density function (and thus of the length scales employed for smoothing).

These requirements imply that H_i in (1) must be both stable and computationally feasible in many dimensions, that it should not be constrained to a multiple of the unit matrix (as is common practice in conventional KDE), and that it must be allowed to vary substantially with i.

It is informative to evaluate state-of-the-art approaches to KDE (as summarized, for example, in [1]) against these criteria. We identify four main classes of approaches.

Approaches based on cross validation utilize the fact that the effect of any sample i on the estimator in (1) can easily be removed, by simply removing that sample from the summation. It is therefore possible to compute the likelihood $p(x_i)$ at location x_i as a function of the smoothing matrix efficiently; the expected value of this likelihood over all i is a reasonable criterion to optimize when searching for an optimal bandwidth. Note, however, that this assumes that the same bandwidth is employed everywhere – otherwise, the optimization problem becomes too large and ill-conditioned for practical solution. Cross validation with fixed bandwidths has indeed been found to be quite effective in low-dimensional settings [4], but (as we show below) becomes less attractive as the dimensionality is increased.

Bootstrap-based approaches employ a smoothed bootstrap estimator of the mean integrated square error (MISE) between the estimated and underlying probability density functions, and minimizes this estimator with respect to the bandwidth matrix. These methods turn out to be quite similar to those based on cross validation: again, efficient computational approaches exist, and practical constraints generally limit the use of this method to constant bandwidths. Because of these similarities, we do not consider bootstrap-based approaches below.

Analytic expressions for the MISE are also at the root of *plug-in estimators*. In this case, however, the dependency of the MISE on the estimated density function are retained explicitly in the form of 'higher order' terms, which themselves depend on the density function. At a selected point this chain is broken by using values from a reference density function (typically, a Gaussian), and the resulting values are 'plugged into' the cascade of lower-order terms, finally providing estimates of the optimal bandwidths. As with the previous two approaches, it is not practically feasible to derive spatially varying bandwidths in this way, though Duong and Hazelton have shown [5] that fully anisotropic bandwidth matrices (i.e. more general than a multiple of the unit matrix or a diagonal matrix) can successfully be derived in this fashion.

Bayesian approaches model the entries in the bandwidth matrix as random variables with a presumed prior distribution, which is combined with evidence from the samples in order to estimate the posterior densities of these entries. The approach pioneered by Gangopadhyay and Cheung [6], which was generalized to multivariate problems by De Lima and Atuncar [7], is particularly relevant to our purposes, since it is by nature spatially variable ('adaptive'), and lends itself to full anisotropy, diagonal bandwidth matrices, or spherical bandwidths, depending on the requirements of the application. As is typical of Bayesian approaches, however, the selection of appropriate prior probabilities is a major challenge.

This cursory examination suggests that none of the current approaches to KDE bandwidth estimation is ideally suited to the analysis of data in pattern recognition. The Bayesian methods come closest to our requirements, but the need to specify a prior distribution is a significant obstacle. We therefore examine these methods more closely below, and derive a related approach which does not require such a prior.

III. The Maximum leave-one-out Likelihood (MLL) method

Two key insights motivated the development of the Bayesian method of Gangopadhyay and Cheung [6]. Firstly, the optimal adaptive bandwidth should vary relatively slowly throughout the input (feature) space, which implies that the density function in the neighbourhood of a given sample can be approximated quite well by using its bandwidth matrix at each of the sample points of its neighbours. Since more distant sample points are, by definition, those that do not contribute much to the density at the selected point, that same bandwidth may be associated with *all* sample points for the purpose of selecting the bandwidth at x_i. Secondly, the density function that results from such a convolution of all samples with a selected kernel $K(x;H)$ is closely related to the density $p(x)$ we wish to estimate: it is a sample estimate of $p(x)$ smoothed by $K(x;H)$. If sufficiently many samples are available, it can be assumed that $K(x;H)$ will be quite localized, thus smoothing $p(x)$ to a relatively small degree. Therefore, Gangopadhyay and Cheung suggest that this sample estimate of $p_H(x) = p(x) * K(x;H)$ be used as representative of $p(x)$ for analytic purposes.

For the Bayesian approach, we assume a prior $\pi(H)$ on H, and compute the posterior given a location $X = x$ as

$$\pi(H|x) = \frac{p_H(x)\pi(H)}{\int p_{H_d}(x)\pi(H_d)dH_d}. \tag{2}$$

For each location x, the "optimal" bandwidth can thus be estimated as that H which maximizes this expression. This elegant formulation leads to computationally feasible schemes if convenient forms are assumed for the prior probabilities. Unfortunately, those forms are not convincing choices on physical grounds, and do not lead to particularly good performance in the problems investigated by De Lima and Atuncar [7] (possibly because the practical inappropriateness of the "convenient" priors selected).

These complications naturally lead one to consider what happens if we replace the Bayesian formulation with a maximum-likelihood approach – that is, if one optimizes $p_H(x)$ rather than $\pi_H(x)$. This is particularly simple if K is assumed to be a normal density: by setting the derivative of

$p_H(x)$ with respect to $H(x)$ equal to zero, we then obtain

$$H(x_i) = \sqrt{\frac{\sum_{j \neq i}(x_i - x_j)(x_i - x_j)^T K(x_i - x_j; H(x_i))}{\sum_{j \neq i} K(x_i - x_j; H(x_i))}}. \quad (3)$$

Thus, the optimal bandwidth matrix at x, under these assumptions, equals the weighted variance of the samples surrounding x. The weight of each sample is the multivariate normal probability density of the Mahalanobis distance between that sample and x. In this distance measure, $H(x)$ is used as metric. For obvious reasons, we call this the "Maximum Leave-one-out Likelihood" method of bandwidth estimation.

To gain intuition on the implications of this formulation, we consider the one-dimensional case. In that case, it is easy to see that (3) always has at least one solution: as $H(x)$ approaches 0, the right-hand side approaches the squared distance to the nearest neighbour of x, and as $H(x)$ goes to infinity, the right-hand side approximates the mean squared distance from x of all data points. Thus, the expectation value is greater than $H(x)$ for small values of $H(x)$, and less than $H(x)$ for large values. Since the expectation value is a continuous function of $H(x)$, this implies that the two curves(for $H(x)$ and the right-hand side of (3)) must intersect for at least one value of $H(x)$. In fact, the existence of multiple solutions is a common occurrence in practice. We demonstrate this with several examples in Figs. 1 to 3. These figures show the leave-one-out likelihood $p_H(x)$ when evaluated at a number of points in a sample set drawn from a mixture of two Gaussians. (The parameters of these Gaussians were the same as those chosen by [6]: each Gaussian had unit variance, and the means were 0 and 4, respectively. The two components have prior probabilities of 0.4 and 0.6, respectively.) Fig. 3 also demonstrates that the global maximum of $p_H(x)$ is not always the best choice for $H(x)$ – when data points happen to be particularly close, a spurious (and numerically large) maximum exists for small $H(x)$. Thus, a form of regularization is required to eliminate such maxima.

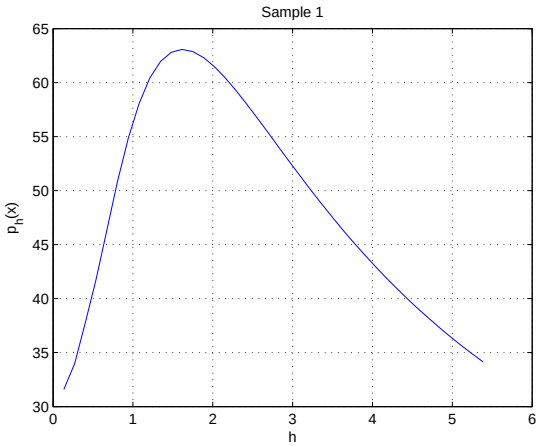

Fig. 1. Example of relationship between p_h and h: a single maximum exists, at a reasonable value of h.

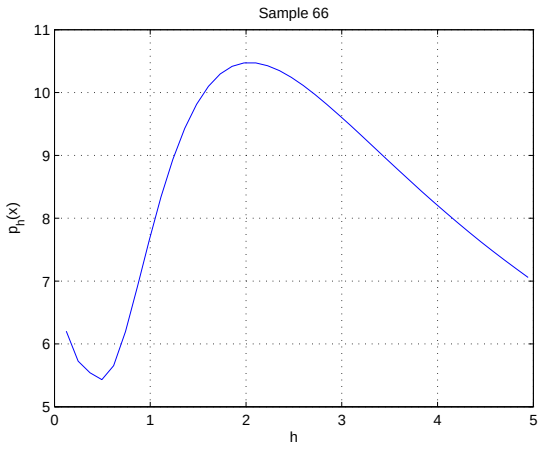

Fig. 2. Example of relationship between p_h and h, with several maxima, the largest of which occurs at a reasonable value of h.

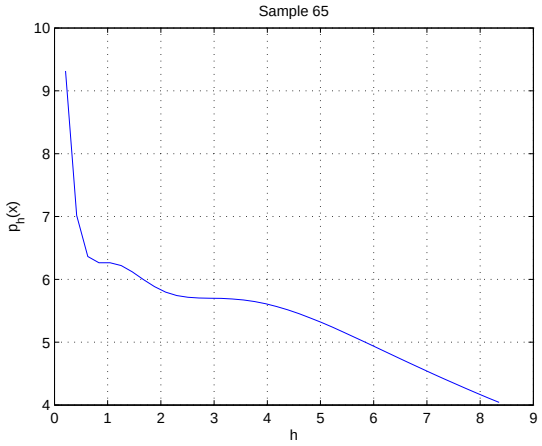

Fig. 3. Example of relationship between p_h and h, with several maxima, the largest of which corresponds to very small h.

Analogously to, for example, the Baum-Welch algorithm, (3) can be used in one of two ways: either in conjunction with a gradient-based optimizer, or as a direct iterative scheme where the value found by calculating the right-hand side for a given value of $H(x)$ is used as the next value for $H(x)$, and so forth. Both approaches will find local solutions based on the initial $H(x)$ selected, and direct iteration is generally found to be somewhat more efficient. Note that the initial $H(x)$ itself serves as a form of regularization: since both local optimizers tend to find solutions in the neighbourhood of the starting point, the process can be guided towards preferable solutions through appropriate selection of this initialization.

It is easy to see that all these findings (existence of a solution, the possibility of several local extrema, the possibility of spurious and large maxima for small $|H(x)|$) also hold for the multivariate case. In fact, in multivariate KDE there are invariably (in addition to other local maxima) singularities corresponding to $|H(x)| \rightarrow 0$ when one of the non-null eigenvectors has a component in a direction parallel to the

vector between x and one of the other training samples. Now, however, another option for regularization presents itself: by restricting the allowed forms of $H(x)$ considered during optimization (e.g. limiting it to multiples of the unit matrix or diagonal matrices), many of the spurious extrema can be avoided.

IV. Experimental results

To evaluate the potential of the MLL approach in practice, we investigate three classes of problems below, namely (a) the estimation of one-dimensional densities using samples drawn from known distributions, (b) known distributions in two dimensions, and (c) selected real-world pattern-recognition tasks. For the known distributions, we are able to estimate the MISE of each estimator, but for the real-world data sets (for which the underlying distribution is not known) we rely on leave-one-out entropy estimates as a basis for comparison.

We compare the MLL algorithm for bandwidth selection to four other methods, as implemented in the *KDE toolkit* [8]:

- *LCV* is the standard leave-one-out cross validation algorithm;
- *local* scales the bandwidth of each sample with the distance to its k-th nearest neighbour (where k is typically the square root of the number of samples, and each dimension is scaled separately) before applying *LCV* bandwidth selection;
- *localp* is identical to *local*, except that the data is pre-sphered (which is only relevant for multivariate tasks); and
- *Hall* is the plug-in estimate suggested by Hall, Sheather, Jones and Marron [9].

A. Known one-dimensional distribution

As discussed in Section III, regularization is expected to play an important role in one-dimensional KDE with MLL. We have not investigated this issue extensively – here, we simply employ the heuristic that the minimal h at any sample is set to 2% of the largest distance between that sample and any other sample in the data set. Fig. 4 shows a typical density function estimated in this way, for the same mixture of Gaussians as in Section III, and also the density function estimated by the locally-adapted LCV method, *local*. (The discontinuities in the MLL bandwidths are quite prominent – these are a consequence of multiple maxima in the likelihood function and the resultant regularization, with significantly different solutions being found for adjacent data points.) Fig. 5 contains the estimated MISE for the four one-dimensional methods on this task, as a function of the number of training samples employed. We see that MLL performs competitively on this task.

B. Known multivariate distributions

In the multivariate case, we do not employ explicit regularization: instead, we limit our bandwidth matrices to be diagonal, and initialize $H(x)$ in the iterative computation of

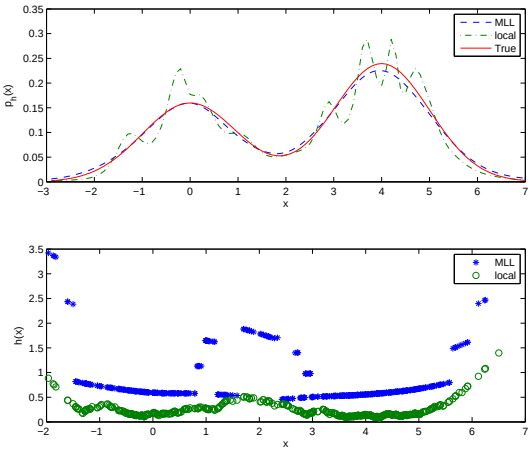

Fig. 4. A typical density estimate with the MLL method compared with the local LCV estimate, along with the bandwidths selected by the two methods. 100 training samples are employed.

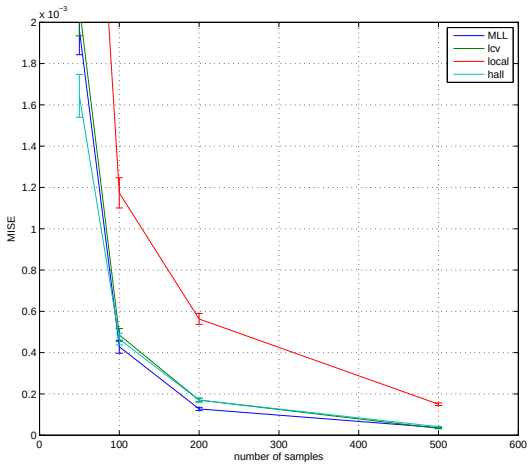

Fig. 5. Mean integrated square error of four methods on one-dimensional task. Error bars correspond to one standard error.

equation 3 with the *Hall* plug-in estimate. We have experimented with two density functions proposed in [7], which we call "DeLimaA" and "DeLimaC", respectively. Each is a mixture of two components, and, though anisotropic and spatially variable, quite smooth. Figs. 6 and 7 show the MISE curves for the five multivariate methods on these two tasks. As several authors have found (see, for example, [4]), LCV-based methods are highly competitive on this class of problem. The flexibility of the MLL method explains its poor performance for small sample sets; as the number of samples increases, its performance is second to only the *LCV* method on both tasks.

C. Pattern-recognition data sets

As is discussed more comprehensively elsewhere [10], the detailed analysis of real-world data sets typically requires a sequence of processing steps including scaling, principal

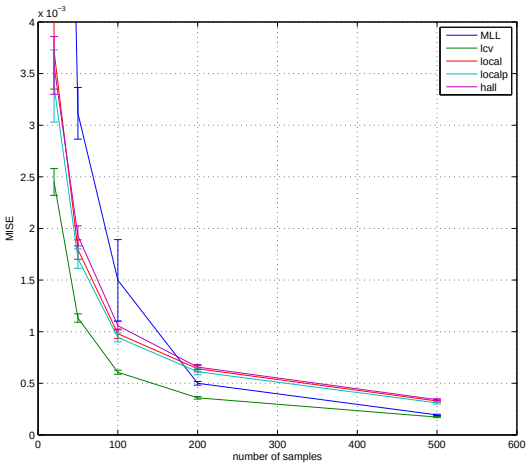

Fig. 6. MISE of five methods on two-dimensional problem "DeLimaA"

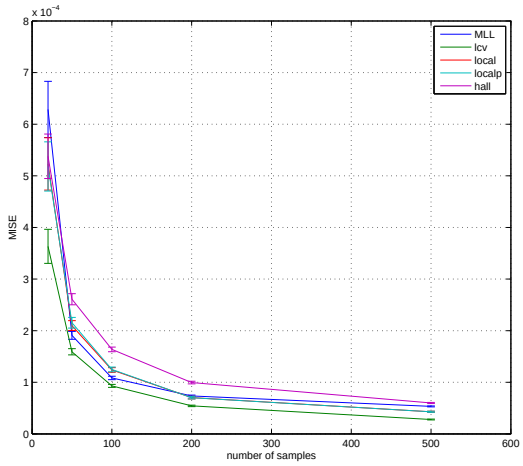

Fig. 7. MISE of five methods on two-dimensional problem "DeLimaC"

component analysis and dimension reduction. The details of these processes are somewhat problem dependent, and here we present results on a variety of data sets after appropriate preprocessing has been performed. Since each of the problems contains data from several classes, we present results for a representative sampling of these classes; as mentioned above, we do not have "ground truth" for any of these classes, and therefore present our results as leave-one-out estimates of the entropy for each of these classes. (Despite appearances, this measure is not unduly biased towards the MLL method, which also uses a leave-one-out strategy: for the MLL method, the left-out sample x_i occurs within a probability estimate using the same bandwidth $H_i(x)$ everywhere, which is distinct from the actual estimate used for evaluation. We have confirmed that all methods experience comparable increases in entropy when the leave-one-out method is used to counter training-set bias.)

Our experiments involved two tasks from speech processing:

- Phone recognition on the TIMIT corpus [11], using a 39-dimensional feature set consisting of mel-scale cepstral coefficients along with the usual delta and double-delta coefficients. We employ 40 phonetic classes in our experiments, and report here on 5 classes drawn from different broad phonetic categories.
- Age classification on the Deutsche Telekom (DT) corpus [12], using a 20-dimensional feature space of "long-term" features to distinguish between 7 age-and-gender classes, as described in [13].

We also experimented with one image-recognition task, namely the "Vehicle Silhouettes (VS)" [14] task from the UCI repository. In this data set, each of 7 classes is described using 18 features designed to capture various geometric aspects of three-dimensional objects in a two-dimensional image.

For each task, we experimented with various numbers of training samples and numbers of principal components retained. We compared the leave-one-out entropies obtained with the five multivariate methods mentioned above, as well as a full-covariance Gaussian density. Results are summarized in Table I; the "N.A" entries in that Table refer to cases where an estimator did not provide a meaningful bandwidth, and the column labelled "Dimensions" lists the number of principal components retained.

When interpreting these values, it is important to keep in mind that entropies correspond to log-likelihoods: thus, an entropy difference of as little as 0.2 implies a difference in expected likelihoods of more than 20%, and entropy differences of 0.5 or more are highly significant. The results in Table I (and similar results for experiments not reported here) suggest a number of conclusions:

- With few exceptions, all KDEs outperform the Gaussian density on these tasks, showing the importance of non-parametric estimation.
- As expected, the scale differences between the various regions in feature space severely impact on the methods that employ a fixed bandwidth matrix throughout, namely *LCV* and *Hall*.
- Of the remaining methods, *MLL* generally performs best, with *localp* achieving the lowest entropy on one task and *local* on two tasks.
- The relative outperformance of *MLL* is increased as the number of dimensions rises.
- The relative performance of the various methods does not seem to depend much on the number of training samples employed, for the ranges investigated here.

V. Conclusion and Outlook

We have motivated and derived the MLL algorithm for KDE bandwidth estimation. This algorithm has a number of attractive properties: it is spatially adaptive, scales naturally to high dimensions and performs well on a wide set of tasks as summarized in Section IV, especially for high-dimensional problems. The least satisfactory aspect of the current version of the algorithm is the ad-hoc fashion in which regularization is performed. Currently, we rely on the restrictions imposed by a diagonal bandwidth matrix, and this is certainly not optimal.

Task	Class	Dimensions	Samples	LCV	localp	local	Hall	MLL	Gaussian
TIMIT	eh	5	500	7.5537	7.6381	7.5715	7.8968	**7.1809**	8.4228
	f	5	500	8.0834	8.0427	8.0455	8.7798	**7.8688**	8.0598
	m	5	500	7.5707	7.5401	7.5382	8.0619	**7.2962**	8.3426
	ay	5	500	7.3585	7.3800	7.2856	7.6419	**7.0629**	8.3153
	sil	5	500	8.1805	8.1751	8.1451	9.0109	**7.9337**	8.3599
DT	0	5	500	9.4375	9.2581	9.2068	10.3715	**8.7871**	9.5902
	1	5	500	9.2841	8.9847	8.9544	10.4290	**8.6347**	9.4482
	2	5	500	8.8123	8.6066	8.6213	9.9735	**8.1686**	9.3996
	3	5	500	9.6525	9.4852	9.5097	10.4898	**9.0139**	9.6425
	4	5	500	9.2758	9.1036	9.0993	10.1581	**8.5998**	9.4869
	6	5	500	9.3038	9.1186	9.1455	10.1808	**8.6651**	9.3935
	0	10	500	15.0478	15.3535	14.2261	14.9965	**11.7518**	14.8114
	1	10	500	14.8897	15.3749	14.0206	15.2921	**11.6279**	14.8711
	0	5	1000	9.3753	9.1197	9.1358	10.9304	**8.8658**	9.5986
	1	5	1000	9.2580	8.9730	8.9839	10.3653	**8.7005**	9.5069
VS	0	5	297	5.7392	**4.7743**	4.8511	8.5778	5.0618	8.2155
	1	5	300	8.3679	7.5953	7.3949	13.2202	**7.0653**	8.8257
	2	5	299	6.0426	**4.0075**	4.1183	10.5294	4.0914	8.6148
	3	5	300	7.1626	5.4190	5.8222	N.A	**5.1394**	8.7665
	4	5	298	6.6018	4.3974	**4.1763**	N.A	4.2539	8.7621
	5	5	292	6.8485	6.1951	6.5128	6.9231	**5.7538**	8.6636
	6	5	300	8.4700	7.7218	7.7265	12.3537	**7.5359**	8.7661

TABLE I
LEAVE-ONE-OUT ENTROPY ESTIMATES FOR SEVERAL PATTERN-RECOGNITION TASKS

The derivation in Section III suggests a more principled approach: since we apply the bandwidth of sample x_i at each of its neighbours, it is reasonable to insist that the entries of the bandwidth matrix should change reasonably slowly throughout feature space. A model such as a Markov Random Field, defined on the neighbourhood graph, may provide an appropriate structure for enforcing this smoothness requirement; we are currently investigating the development of a regularizer based on this structure.

REFERENCES

[1] D. Scott and S. Sain, "Multidimensional density estimation," *Handbook of Statistics*, vol. 24, pp. 229–261, 2005.

[2] J. Liao, Y. Wu, and Y. Lin, "Improving Sheather and Jones' bandwidth selector for difficult densities in kernel density estimation," *Journal of Nonparametric Statistics*, vol. 22, no. 1, pp. 105–114, 2010.

[3] P. Niyogi, S. Smale, and S. Weinberger, "Finding the Homology of Submanifolds with High Confidence from Random Samples," *Discrete & Computational Geometry*, vol. 39, no. 1, pp. 419–441, 2008.

[4] T. Duong and M. Hazelton, "Cross-validation bandwidth matrices for multivariate kernel density estimation," *Scandinavian Journal of Statistics*, vol. 32, no. 3, pp. 485–506, 2005.

[5] ——, "Plug-in bandwidth matrices for bivariate kernel density estimation," *Journal of Nonparametric Statistics*, vol. 15, no. 1, pp. 17–30, 2003.

[6] A. Gangopadhyay and K. Cheung, "Bayesian approach to the choice of smoothing parameter in kernel density estimation," *Journal of Nonparametric Statistics*, vol. 14, no. 6, pp. 655–664, 2002.

[7] M. de Lima and G. Atuncar, "A Bayesian method to estimate the optimal bandwidth for multivariate kernel estimator," *Journal of Nonparametric Statistics*, vol. 22, 2010.

[8] A. Ihler and M. Mandel, "Kernel density estimation toolbox for matlab," http://www.ics.uci.edu/ ihler/code/.

[9] P. Hall, S. Sheather, M. Jones, and J. Marron, "On optimal data-based bandwidth selection in kernel density estimation," *Biometrika*, vol. 78, no. 2, p. 263, 1991.

[10] E. Barnard, "Visualizing data in high-dimensional spaces," in *PRASA*, 2010.

[11] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," *NTIS order number PB91-100354*, 1993.

[12] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Müller, R. Huber, B. Andrassy, J. G. Bauer, and B. Littel, "Comparison of four approaches to age and gender recognition for telephone applications," in *ICASSP*, Honolulu, Hawaii, April 2007, pp. 1089–1092.

[13] C. Müller, "Automatic recognition of speakers' age and gender on the basis of empirical studies," in *Interspeech*, Pittsburgh, Pennsylvania, September 2006.

[14] J. Siebert, "Vehicle recognition using rule based methods," Turing Institute Research Memorandum, Tech. Rep. TIRM-87-018, 1987.