MULTIVARIATE DENSITY ESTIMATION WITH OPTIMAL MARGINAL PARZEN DENSITY ESTIMATION AND GAUSSIANIZATION

Deniz Erdogmus[1], Robert Jenssen[2], Yadunandana N. Rao[1], Jose C. Principe[1]
[1] CNEL, ECE Department, University of Florida, Gainesville, Florida, USA
[2] Department of Physics, University of Tromso, Tromso, Norway
E-mail: derdogmus@ieee.org

Abstract. Multivariate density estimation is an important problem that is frequently encountered in statistical learning and signal processing. One of the most popular techniques is Parzen windowing, also referred to as kernel density estimation. Gaussianization is a procedure that allows one to estimate multivariate densities efficiently from the marginal densities of the individual random variables. In this paper, we present an *optimal* density estimation scheme that combines the desirable properties of Parzen windowing and Gaussianization, using minimum Kullback-Leibler divergence as the optimality criterion for selecting the kernel size in the Parzen windowing step. The performance of the estimate is illustrated in a classifier design example.

INTRODUCTION

In statistical machine learning and statistical signal processing we frequently encounter the problem of estimating the probability distribution of the observed data, which is typically multidimensional. The literature has extensively dealt with this fundamental problem using one of the three main approaches: parametric, semiparametric, and nonparametric. Traditionally, parametric approaches have been adopted widely, which combined with Bayesian techniques such as maximum likelihood (ML) and maximum *a posteriori* (MAP) parameter estimation yield tractable and in many cases useful solutions under the assumptions made [1]. Advances in machine learning and signal processing techniques require less restrictive assumptions, thus parametric techniques became less desirable. Consequently, semiparametric and nonparametric density estimation approaches have become the focus of statistical learning research.

Semiparametric density estimation techniques offer solutions under less restrictive assumptions regarding the data structures. The most commonly used semiparametric method is the so called mixture model, which allows the designer to approximate the data as a two-step mixture of parametric distributions, where each parametric model is also associated with a prior probability of being selected for data generation [2]. Especially the Gaussian mixture model is widely utilized due to its asymptotic universal approximation capability that arises from the theory

0-7803-8608-6/04/$20.00 ©2004 IEEE

of radial basis function networks. However, selecting the number of models becomes a nontrivial problem.

On the other hand, nonparametric approaches often allow the designer to make the least restrictive assumptions regarding the data. Density estimation techniques in this class include histograms, nearest neighbor estimates, and kernel density estimates (referred to as Parzen windowing) [1]. Parzen windowing is a generalization of the histogram technique, where smoother membership functions are used instead of the rectangular volumes typically used in histogram binning. Although asymptotically Parzen windowing can yield unbiased and consistent estimators, in the finite sample case, selecting the kernel function and the kernel size become a challenging problem. Especially in multidimensional density estimation, in general, the full covariance matrix of the kernel must be optimized (assuming elliptically symmetric kernels). Some methods go even further by incorporating the nearest neighbor density estimation approach and try to optimize the kernel size/covariance for each sample based on its nearest neighbor distances [3]. Unfortunately, all these techniques become intractable and ineffective when it comes to on-line adaptive learning and signal processing, due to increasing computational complexity, as well as discontinuities in gradients introduced by switching neighbors.

In this paper, we will describe a multivariate density estimation method that follows Parzen windowing in spirit. We will use Parzen windowing to estimate marginal distributions (nonparametrically) from the samples. The kernel size in Parzen windowing will be optimized to minimize the Kullback-Leibler divergence (KLD) of the true marginal distribution from the estimated marginal density. The estimated marginal densities will be used to transform the random variables to Gaussian-distributed, where joint statistics can be simply determined by sample covariance estimation.

GAUSSIANIZATION

Given an n-dimensional random vector $\mathbf{X}$ with joint probability density function (pdf) $f(\mathbf{x})$, there exist infinitely many functions $\mathbf{h}:\Re^n \to \Re^n$ such that $\mathbf{Y}=\mathbf{h}(\mathbf{X})$ is jointly Gaussian. We are particularly interested in the elementwise Gaussianization of $\mathbf{X}$. Suppose that the i^{th} marginal of $\mathbf{X}$ is $f_i(x_i)$, with a corresponding cumulative distribution function (cdf) $F_i(x_i)$. Let $\phi(.)$ denote the cdf of a zero-mean unit-variance single dimensional Gaussian variable:

$$\phi(\xi) = \int_{-\infty}^{\xi} \frac{1}{\sqrt{2\pi}} e^{-\alpha^2/2} d\alpha \tag{1}$$

According to the fundamental theorem of probability [4], $Y_i = \phi^{-1}(F_i(X_i))$ is a zero-mean and unit-variance Gaussian random variable. Consequently, we consider the element-wise Gaussianizing functions defined as $h_i(\xi) = \phi^{-1}(F_i(\xi))$. Combining these marginal Gaussianizing functions for each dimension of the data, we obtain the Gaussianizing transformation $\mathbf{h}:\Re^n \to \Re^n$. Note that after this

transformation (whose Jacobian is diagonal everywhere) we obtain a jointly Gaussian vector $\mathbf{Y}$ with zero mean and covariance

$$\mathbf{\Sigma} = E[\mathbf{Y}\mathbf{Y}^T] \tag{2}$$

Hence, if the marginal pdfs of $\mathbf{X}$ and the covariance $\mathbf{\Sigma}$ are known (or estimated from samples), the joint pdf of $\mathbf{X}$ can be obtained using the fundamental theorem of probability as

$$f(\mathbf{x}) = \frac{g_{\mathbf{\Sigma}}(\mathbf{h}(\mathbf{x}))}{\left|\nabla \mathbf{h}^{-1}(\mathbf{h}(\mathbf{x}))\right|} = g_{\mathbf{\Sigma}}(\mathbf{h}(\mathbf{x})) \cdot \left|\nabla \mathbf{h}(\mathbf{x})\right| = g_{\mathbf{\Sigma}}(\mathbf{h}(\mathbf{x})) \cdot \prod_{i=1}^{n} \frac{f_i(x_i)}{g_1(h_i(x_i))} \tag{3}$$

where $g_{\mathbf{\Sigma}}$ denotes a zero-mean multivariate Gaussian distribution with covariance $\mathbf{\Sigma}$ and g_1 denotes a zero-mean univariate Gaussian distribution with unit variance.

OVERVIEW OF THE PROPOSED METHOD

The proposed joint density estimation is based on (3). Density estimation is carried out using a set of independent and identically distributed (iid) samples $\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$ drawn from the joint density $f(\mathbf{x})$. In summary, the marginal distributions $f_i(.)$ are to be approximated using single dimensional Parzen window estimates. The estimated marginal pdfs are denoted by $\hat{f}_i(.)$. The kernel sizes in the Parzen window estimates for each dimension are optimized using the minimum KLD (equivalently maximum likelihood) criterion. This procedure will be described in detail in the next section. From these estimates, approximate Gaussianizing transformations $\hat{h}_i(.)$ can be easily constructed as described in the previous section. Assuming that these estimated transformations convert the joint data distribution to Gaussian, the covariance matrix is simply estimated from the samples using

$$\hat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{j=1}^{N} \hat{\mathbf{y}}_j \hat{\mathbf{y}}_j^T \tag{4}$$

where $\hat{\mathbf{y}}_j = \hat{\mathbf{h}}(\mathbf{x}_j)$.[1] In this second phase of the procedure, we basically assume that the samples $\{\hat{\mathbf{y}}_1,\ldots,\hat{\mathbf{y}}_N\}$ are jointly Gaussian with zero-mean and assign the sample covariance as the parameters of the underlying Gaussian distribution. This is equivalent to selecting the maximum likelihood parameter estimates for the underlying *Gaussian* density, which is also equivalently a minimum KLD estimate, as we show next.

It is well known that asymptotically maximum likelihood density estimation becomes identical to minimum KLD estimation. In particular, if the underlying distribution for $\mathbf{X}$ is $q(\mathbf{x})$ and the parametric family is $p_\theta(\mathbf{x})$, then

[1] Note that the true distribution of the (approximately) Gaussianized samples has a mean of zero. Therefore, the unbiased sample covariance estimate should be as given in (4), without a correction term due to data dimensionality in the denominator.

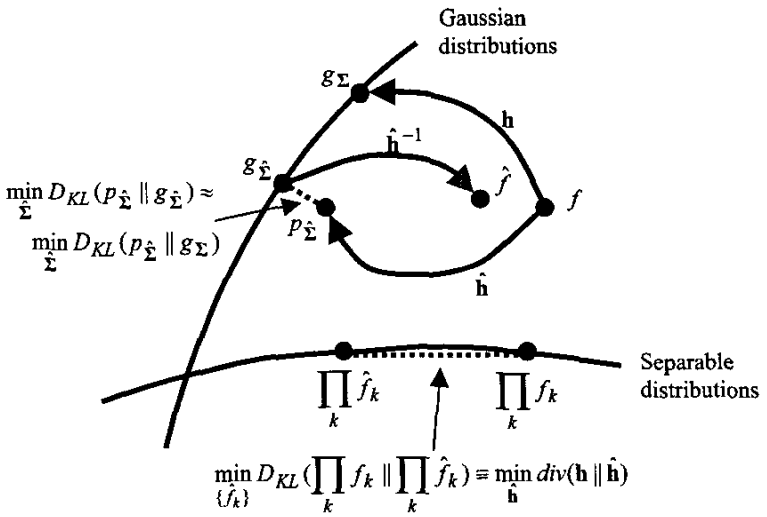

Figure 1. This is an illustration of the proposed joint density estimation procedure. Optimization is carried out in two steps. The marginal density estimates are determined by minimizing the KLD, which is equivalent to minimizing some form of divergence between the estimated and actual Gaussianizing transformations, denoted by $\mathbf{h}$ and $\hat{\mathbf{h}}$. The divergence between the approximately Gaussianized distribution $p_{\hat{\Sigma}}$ and the true Gaussianized distribution g_{Σ} is approximately minimized by projecting $p_{\hat{\Sigma}}$ to the manifold of Gaussian distributions using KLD to obtain $g_{\hat{\Sigma}}$. This is possible due to the Pythagorean theorem for KLD.

$$\min_{\theta} D_{KL}(q(\mathbf{x}) \| p_{\theta}(\mathbf{x})) \equiv \max_{\theta} E_q(p_{\theta}(\mathbf{X})) \tag{5}$$

where we used the following definition for KLD:

$$D_{KL}(q(\mathbf{x}) \| p(\mathbf{x})) = \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x} \tag{6}$$

Consequently, the proposed procedure for estimating the joint distribution of a set of iid samples equivalently minimizes the KLD in an approximate manner as illustrated in Fig. 1. The KLD between the estimated and actual marginal distributions is minimized to obtain an accurate estimate of the true Gaussianizing transformation $\mathbf{h}$. This optimization is performed in a constrained manner in the manifold of separable distributions in the pdf space. However, due to estimation errors, an imperfect transformation $\hat{\mathbf{h}}$ is obtained. The corresponding transformed distribution $p_{\hat{\Sigma}}$ is projected optimally to the manifold of Gaussian distributions to obtain $g_{\hat{\Sigma}}$, which is a better approximation to g_{Σ}, due to the Pythagorean Theorem for KLD [5]. The final density estimate is obtained by employing the

inverse transformation $\hat{\mathbf{h}}^{-1}$ to $g_{\hat{\Sigma}}$. Clearly, as the number of samples increase, the estimated joint distribution will approach the true underlying data distribution.

In this paper, we propose using Parzen windowing to estimate the marginal distributions, which are necessary to determine the Gaussianizing transformation for the data. However, any other density estimation procedure could easily replace Parzen windowing within the general framework presented here.

UNIVARIATE PARZEN WINDOW DENSITY ESTIMATION

Parzen windowing is a kernel-based density estimation method, where the resulting estimate is continuous and differentiable provided that the selected kernel is continuous and differentiable [3,6]. Given a set of iid scalar samples $\{x_1,\ldots,x_N\}$ with true distribution $f(x)$, the Parzen window estimate for this distribution is

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} K_\sigma(x - x_i) \tag{7}$$

In this expression, the kernel function $K_\sigma(.)$ is a continuous and smooth, zero-mean pdf itself, typically a Gaussian. The parameter σ controls the *width* of the kernel and it is referred to as the kernel size. This pdf estimate is, in general, biased, since its expected value is $E[\hat{f}(x)] = f(x) * K_\sigma(x)$, where * denotes convolution. The bias can be asymptotically reduced to zero by selecting a unimodal symmetric kernel function (such as the Gaussian) and reducing the kernel size monotonically with increasing number of samples, so that the kernel asymptotically approaches a Dirac-delta function. In the finite sample case, the kernel size must be selected according to a trade-off between estimation bias and variance: decreasing the kernel size increases the variance, whereas increasing the kernel size increases the bias. In particular, the kernel size should be reduced slower than $1/N$, in order to guarantee asymptotic consistency. To illustrate the effect of kernel size on the estimated density, Parzen pdf estimates of 50-sample sets of Laplacian and uniformly distributed samples with small and large kernel sizes are shown in Fig. 2.[2]

For accurate density estimation, variable kernel size methods are proposed in the statistics literature [3]. However, for our purposes (i.e., adaptive signal processing) such approaches to density estimation are not feasible due to increased computational complexity. The complexity of information theoretic methods based on Parzen density estimates are already $O(N^2)$ in batch operation mode [7-12]. Assigning and optimizing a different kernel size for each sample would make the algorithmic complexity even higher.

Therefore, we will only consider the fixed kernel size approach where the same kernel size is used for each sample. This parameter can be optimized based

---

[2] The generalized Gaussian density family is described by $G_\beta(x) = C_1 \exp(-C_2 |x|^\beta)$, where C_1 and C_2 are positive constants and β is the order of the distribution. Laplacian and uniform distributions are special cases corresponding to $\beta = 1$ and $\beta = \infty$.

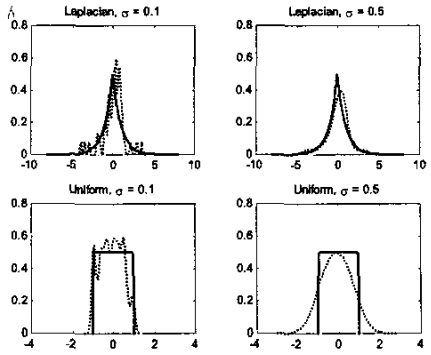

Figure 2. Laplacian and uniform distributions estimated using Parzen windowing with Gaussian kernels (kernel size indicated in titles) with 50 samples from each distribution.

on various metrics, such as the integrated square error (ISE) between the estimated and the actual pdf, as discussed by Fukunaga [13]. In actuality, the ISE approach is not practical, since the actual pdf is unknown. However, certain approximations exist. For a Gaussian kernel, Silverman provides the following rule-of-thumb, which is based on ISE and the assumption of a Gaussian underlying density: $\sigma = 1.06\sigma_X N^{-1/5}$, where σ_X denotes the sample variance of the data [14]. More advanced approximations to the ISE solution are reviewed in [15].

Maximum likelihood (ML) methods for kernel size selection have also been investigated by researchers. For example, Duin used the ML principle to select the kernel size of a circularly symmetric Gaussian kernel for joint density estimation with Parzen windowing [16]. More recently, Schraudolph suggested optimizing the full covariance matrix of the Gaussian kernel using the ML approach [12]. In joint density estimation, another option is to assume a separable multidimensional kernel (whose covariance is diagonal in the case of Gaussian kernels). Then, one only needs to optimize the size of each marginal kernel using single dimensional samples corresponding to the marginals of the joint density being estimated. The latter approach has the desirable property that the kernel functions used for marginal density estimation uniquely determine the kernel function that is used for joint density estimation, in addition to the fact that the marginal of the estimated joint density is identical to the estimated marginal density using this type of separable kernels [10]. In this latter approach, the joint density estimate becomes

$$\hat{f}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\prod_{k=1}^{n} K_{\sigma_k}(x^k - x_i^k) \tag{8}$$

where x^k denotes the k^{th} entry of the vector $\mathbf{x}$ and the multidimensional kernel is the product of unidimensional kernels, all using appropriately selected widths.

In this paper, motivated by the graphical description of the method in Fig. 1, and the fact that optimality of density estimates need to consider the information geometry of certain manifolds in the pdf space [17], we assume the minimum

KLD criterion. Recalling the equivalence between minimum KLD and ML principles pointed out in (5), the ML approach turns out to be optimal in an information theoretic sense after all.

OPTIMIZING THE KERNEL SIZE

In this section, we will focus on the optimization of the kernel size in Parzen window density estimates for single-dimensional variables. Consider the density estimator given in (7). Our goal is to minimize the KLD between the true and the estimated densities $f(x)$ and $\hat{f}(x)$. Equivalently we will maximize the log-likelihood of the observed data, i.e., $E_X[\log \hat{f}(X)]$. The expectation is approximated by the sample mean, resulting in

$$J(\sigma) = \frac{1}{N}\sum_{j=1}^{N} \log \hat{f}(x_j) \tag{9}$$

For Parzen windowing this becomes

$$J(\sigma) = \frac{1}{N}\sum_{j=1}^{N} \log\left(\frac{1}{N}\sum_{i=1}^{N} K_\sigma(x_j - x_i)\right) \tag{10}$$

If a unimodal and symmetric kernel function (such as Gaussian) is used, this criterion exhibits an undesirable global maximum at the null kernel size, since as σ approaches zero, the kernel approaches a Dirac-δ function and the criterion attains a value of infinity. To avoid this situation, the criterion needs to be modified in accordance with the leave-one-out technique. This yields

$$J(\sigma) = \frac{1}{N}\sum_{j=1}^{N} \log\left(\frac{1}{N-1}\sum_{i=1, i\neq j}^{N} K_\sigma(x_j - x_i)\right) \tag{11}$$

A similar approach for optimizing the kernel size was previously proposed by Viola *et al.* [18], where the available samples were partitioned to two disjoint sets: $\{x_1,\ldots,x_M\}$ and $\{x_{M+1},\ldots,x_N\}$. While one set was used in the density estimation, the other was used in the sample mean. If desired, a generalized version of (11) could be obtained along these lines using a leave-M out strategy; however, this would increase the computational complexity of evaluating the cost function in a combinatorial way in proportion with M.

EXPERIMENTS

Kernel size optimization: In our first experiment, we will investigate how well the kernel size optimization procedure described above approximates the actual optimal kernel size according to the minimum KLD measure. For this purpose, we have performed a series of Monte Carlo experiments to evaluate the value of the proposed kernel size optimization procedure for marginal density

	$N=$ 50	$N=$ 100	$N=$ 150	$N=$ 200
$\beta=1$	0.56	0.48	0.45	0.41
$\beta=2$	0.50	0.38	0.38	0.38
$\beta=3$	0.43	0.37	0.34	0.30
$\beta=5$	0.34	0.27	0.25	0.24

Table 1. Average optimal Gaussian kernel sizes for unit-variance generalized Gaussian distributions of order β for Parzen estimates *using N samples.*

	$N=$ 50	$N=$ 100	$N=$ 150	$N=$ 200
$\beta=1$	0.51	0.38	0.30	0.31
$\beta=2$	0.49	0.41	0.41	0.36
$\beta=3$	0.43	0.35	0.34	0.31
$\beta=5$	0.34	0.28	0.26	0.23

Table 2. Average optimal Gaussian kernel sizes for unit-variance generalized Gaussian distributions of order β for the true KLD.

estimation. For generalized Gaussian densities of order 1, 2, 3, and 5 (all set to be unit-variance), using 20 independent experiments for each, the optimal kernel size that minimizes (11) for a range of sample sizes were determined.[3] Since the true distributions are known, for each case, the true optimal kernel size values minimizing the actual KLD were also numerically determined. Tables 1 and 2 summarize the results, which demonstrate that the estimated kernel size values match their theoretical values (within reasonable statistical variations).

MAP Classifier Design: In this experiment, we demonstrate the utility of the proposed Gaussianization-based joint density estimation scheme for classifier design. According to the theory of Bayesian risk minimization for pattern recognition, a classifier that selects the class for which the *a posteriori* probability of the feature vector sample is maximized asymptotically minimizes the probability of classification error (denoted by p_e). That is, in a two-class scenario with class priors $\{p_1,p_2\}$ and conditional class distributions $\{f_1(\mathbf{x}),f_2(\mathbf{x})\}$, the optimal strategy to minimize p_e is to select the class with larger $\{p_i f_i(\mathbf{x})\}$, i=1,2.

In practice, however, the class priors and the data distributions have to be estimated from samples. In the nonparametric framework we pursued in this paper, one could use either the Gaussianization-based estimate provided in (3) or the product-kernel-based Parzen windowing method presented in (8). Both methods could use the same KL-optimized marginal density estimates with the corresponding univariate kernels. The difference is in the way they estimate the joint distribution using the knowledge provided by the marginal density estimates. At this point, we expect the former technique to be more data-efficient than the latter, and the results we will show next confirm this hypothesis.

A set of Monte Carlo simulations is designed as follows. A finite number of training samples are generated from two 2-dimensional class distributions, which are both Laplacian. Specifically, we used equal-prior identical distributions $f_i(\mathbf{x}) \propto c_1 e^{-c_2\|\mathbf{x}-\boldsymbol{\mu}_i\|_\infty}$ whose means were selected as $\mu_1=[-1\ -1]^T$ and $\mu_2=[1\ 1]^T$. Due to symmetry, the optimal Bayesian classifier has a linear boundary passing through the origin and has a slope of –1 in the 2-dimensional feature space.

[3] *To minimize (11), first the samples of the scalar random variable under consideration are normalized* to unit variance. Then gradient descent is employed starting from a reasonable initial condition, which is in the interval [0.5,1] for most unimodal data distributions.

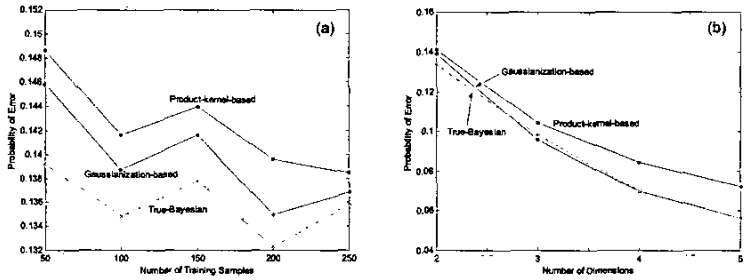

Figure 3. Probability of error for the three classifiers on a test set of 100 samples averaged over 100 Monte Carlo runs for (a) different sizes of training set with fixed dimensionality (b) different dimensionalities of training set using fixed number of training samples.

For each of the training data set sizes of 50 to 250, we conducted 100 Monte Carlo simulations. Three classifiers are designed using each training data set: Gaussianization-based, Product-kernel-based, and True-Bayesian. All classifiers were tested on an independent set of 100 samples (generated randomly in each experiment). Average probability error plots of these classifiers on the testing set are shown in Fig. 3a. As expected, the True-Bayesian classifier yields the lower bound, while the Gaussianization-based classifier outperforms the Product-kernel-based classifier. These results demonstrate that the Gaussianization-based joint density estimation procedure is extracting the higher-order statistical information about the joint distribution more effectively than the product-kernel estimator.

In order to test the hypothesis that this method will avoid the so called *curse of dimensionality* the experiment is generalized to more than 2 dimensions while maintaining the same symmetry conditions. A set of 100 Monte Carlo simulations under similar training and testing conditions are repeated for each data dimensionality (using 100 training samples in every case). The results summarized in Fig. 3b demonstrate that the Gaussianization-based density is able to cope with the increasing dimensionality of the features given the same number of training samples, while the product-kernel approach starts breaking down.

CONCLUSIONS

Nonparametric multivariate density estimation is an important and very difficult ill-posed problem that has fundamental consequences in statistical signal processing and machine learning. Here we proposed a joint density estimation methodology that combines the Gaussianization principle with Parzen windowing. The former effectively concentrates all higher-order statistical information in the data to second-order statistics. The latter is a simple, yet useful density estimation technique based on the use of smooth kernel functions, especially in univariate density estimation. Here, the kernel size in Parzen windowing is optimized using the minimum KLD principle.

The proposed density estimation method, which approximately minimizes the KLD by a two-step procedure, is shown to be more data efficient than Parzen windowing with a structured multidimensional kernel. It is also demonstrated that the curse of dimensionality is beaten (at least to the extent investigated here) by the proposed method. Further investigation will be conducted to test these hypotheses on real data. Finally, note that although we have imposed the constraint of a fixed kernel size with Parzen windowing for the estimation of marginal distributions here, the overall estimation philosophy could be utilized with any (and possibly more advanced) univariate density estimation techniques. Our concern in making this selection was simple and tractable applicability to adaptive signal processing and machine learning, rather than obtaining the *best* density estimate.

Acknowledgements. This work was partially supported by the National Science Foundation under Grant ECS-0300340.

REFERENCES

1. R.O. Duda, P.E. Hart, D.G. Stork, **Pattern Classification**, 2nd ed., Wiley, NY, 2001.
2. S. Theodoridis, K. Koutroumbas, **Pattern Recognition**, Academic Press, NY, 2003.
3. L. Devroye, G. Lugosi, **Combinatorial Methods in Density Estimation**, Springer, NY, 2001.
4. A. Papoulis, **Probability, Random Variables, Stochastic Processes**, McGraw-Hill, NY, 1991.
5. T.M. Cover, J.A. Thomas, **Elements of Information Theory**, Wiley, NY, 1991.
6. E. Parzen, "On Estimation of a Probability Density Function and Mode", in **Time Series Analysis Papers**, Holden-Day, CA, 1967.
7. R. Jenssen, D. Erdogmus, K.E. Hild II, J.C. Principe, T. Eltoft, "Efficient Information Theoretic Clustering Using Stochastic Approximation," submitted to MLSP'04, Sao Luis, Brazil, 2004.
8. K.E. Hild II, D. Erdogmus, J.C. Principe, "Blind Source Separation Using Renyi's Mutual Information," **IEEE Signal Processing Letters**, no. 8, pp. 174-176, 2001.
9. K. Torkkola, "Visualizing Class Structure in Data Using Mutual Information," **Proceedings of NNSP'00**, pp. 376-385, Sydney, Australia, 2000.
10. D. Erdogmus, **Information Theoretic Learning: Renyi's Entropy and its Applications to Adaptive System Training**, PhD Dissertation, University of Florida, Gainesville, Florida, 2002.
11. M.M. Van Hulle, "Kernel-Based Topographic Map Formation Achieved with an Information-Theoretic Approach", **Neural Networks**, vol. 15, pp. 1029-1039, 2002.
12. N.N. Schraudolph, "Gradient-Based Manipulation of Nonparametric Entropy Estimates," **IEEE Transactions on Neural Networks**, 2004. (to appear)
13. K. Fukunaga, **Statistical Pattern Recognition**, Academic Press, NY, 1990.
14. B.W. Silverman, **Density Estimation for Statistics and Data Analysis**, Chapman & Hall, London, 1986.
15. M.C. Jones, J.S. Marron, S.J. Sheather, "A Brief Survey of Bandwidth Selection for Density Estimation," **Journal of American Statistical Association**, vol. 87, pp. 227-233, 1996.
16. R.P.W. Duin, "On the Choice of the Smoothing Parameters for Parzen Estimators of Probability Density Functions," **IEEE Transactions on Computers**, vol. 25, no. 11, pp. 1175-1179, 1976.
17. S. Amari, **Differential-Geometrical Methods in Statistics**, Springer, Berlin, 1985.
18. P. Viola, N. Schraudolph, T. Sejnowski, "Empirical Entropy Manipulation for Real-World Problems", **Proceedings of NIPS'95**, pp. 851-857, 1995.