Journal of VLSI Signal Processing 45, 67–83, 2006

DOI: 10.1007/s11265-006-9772-7

Gaussianization: An Efficient Multivariate Density Estimation Technique for Statistical Signal Processing

DENIZ ERDOGMUS
CSEE Department, Oregon Health and Science University, Portland, OR, USA

ROBERT JENSSEN
Department of Physics, University of Tromso, Tromso, Norway

YADUNANDANA N. RAO
Motorola Corporation, Plantation, FL, USA

JOSE C. PRINCIPE
CNEL, ECE Department, University of Florida, Gainesville, FL, USA

Received: 28 February 2005; Revised: 16 November 2005; Accepted: 1 December 2005

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 utility of the estimate is illustrated in classifier design, independent components analysis, and Prices' theorem.

Keywords: Gaussianization, multivariate density estimation, statistical signal processing

1. Introduction

In statistical signal processing and machine learning, the problem of estimating the probability distribution of the observed data is frequently encountered. Many situations require this estimation to be carried out for multidimensional data and given a finite set of samples; the solutions are affected negatively by increasing data dimensionality due to the curse of dimensionality. As a course rule-of-thumb, the number of samples required to attain the same level of accuracy in density and other forms of statistical estimation as dimensionality n increases, the number of samples should increase exponentially $\sim N^n$, if N is the number of sample required for the single-dimensional case to achieve the desired accuracy. The literature, therefore, extensively deals with the fundamental problem of density estimation using three main approaches: parametric, semiparametric, and nonparametric. Traditionally, parametric approaches have been adopted widely, since combined with Bayesian techniques (such as maximum likelihood and maximum a posteriori) yield tractable and sometimes useful solutions under the assumptions made [1]. Advances in signal processing and machine learning require less restrictive assumptions,

thus parametric techniques become less desirable for a broad application base. Consequently, semiparametric and nonparametric density estimation approaches have become the focus of statistical learning.

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 (although one could also argue that this is still a parametric model). The mixture model approach 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]. The Gaussian Mixture Model (GMM) has especially attracted much attention and has been widely utilized due to its asymptotic universal approximation capability that arises from the theory of radial basis function networks. In mixture models, selecting the appropriate number of components is still not a trivial problem. Alternative semiparametric models exploit series-expansion approaches such as Edgeworth, or Gram–Charlier, where the unknown data distribution is assumed to be sufficiently close to a reference distribution (typically a Gaussian) and a truncated series expansion is utilized to model the data. For practical reasons, the series are usually truncated at low orders and might not always provide the desired flexibility to model a wide class of distributions that one might encounter.

Nonparametric approaches, on the other hand, often allow the designer to make the least restrictive assumptions regarding the data distribution. Density estimation techniques in this class include histogram (the most crude one), nearest neighbor estimates (better), and kernel density estimates (also known as Parzen windowing) [1]. The variable-size kernel estimates and weighted kernel estimates [1, 3] provide immense flexibility in modeling power with desirable small-sample accuracy levels. Parzen windowing is a generalization of the histogram technique, where smoother membership functions are used instead of the rectangular volumes. Parzen windowing asymptotically yields consistent estimates, but the kernel size selection (similar to bin-size selection) can become a challenging problem. While maximum-likelihood like approaches can be employed for tackling this difficulty, the sample sparsity in high-dimensional situations might force the kernels to be extremely large, creating a high bias in the estimates. Furthermore, assuming variable and full-covariance freedom for multidimensional kernel density estimation might lead to an computationally intractable ML optimization problem. Introducing variable kernel-size further complicates computations and makes the estimator even less desirable. In general, density estimation in high-dimensional spaces is an undesirable and challenging problem and any simplifying procedures are likely to bring both computational and performance improvements.

In this paper, we will exploit the fact that if the joint distribution of the high-dimensional data is Gaussian, then one only needs to estimate the mean and covariance. To exploit this, in general, one needs to nonlinearly transform the original data into a Gaussian distributed data using an appropriate function. Furthermore, we will see that under some circumstances, the nonlinear transformation can be defined elementwise reducing the n-dimensional joint Gaussianization problem to n 1-dimensional Gaussianization problems. In the latter case, the individual Gaussianizing functions for each dimensionality of the original data are determined solely by the marginal distribution of the data along the direction of interest. This marginal distribution will be accurately estimated nonparametrically using Parzen windowing by minimizing the Kullback–Leibler divergence (KLD) [4, 5] with respect to the true marginal distribution of the data. Once the marginal densities are estimated, they will be used to transform the data to Gaussian, where joint statistics can be simply determined by sample covariance estimation.

2. Gaussianization for Density Estimation

Given an n-dimensional random vector $\mathbf{X}$ with joint probability density function (pdf) $f(\mathbf{x})$, our goal is to estimate this pdf nonparametrically such that the KLD between the estimated distribution $\widehat{f}(\mathbf{x})$ and $f(\mathbf{x})$ is minimized; this is equivalent to nonparametric maximum likelihood density estimation:

$$\min_{\widehat{f}} D_{KL}\left(f\middle\|\widehat{f}\right) \equiv \min_{\widehat{f}} \int f(\mathbf{x}) \log \frac{f(\mathbf{x})}{\widehat{f}(\mathbf{x})} d\mathbf{x} \equiv \max_{\widehat{f}} E_f\left[\widehat{f}(\mathbf{X})\right] \tag{1}$$

Since KLD is invariant to monotonic (one-to-one) transformations of the random vector $\mathbf{X}$, the divergence between f and $\widehat{f}$ is identical to the divergence between g and $\widehat{g}$, where the latter are true and estimated Gaussian densities of $\mathbf{h}(\mathbf{X})$. In general, the joint-Gaussianization transform $\mathbf{h}(.)$ is a multi-input multi-output function with a nondiagonal Jacobian. However, in some cases, it is possible to obtain a jointly Gaussian $\boldsymbol{Y}=\mathbf{h}(\mathbf{X})$, where $Y_i=h_i(X_i)$, $i=1,...,n$. We will refer to such distributions as *marginally Gaussianizable* (i.e., employing appropriate marginal transformations achieves joint Gaussianiztion). Specifically, the span of all X_i such that the conditional distribution $p(\mathbf{x}|x_i)$ is unimodal for all x_i will constitute a marginally Gaussianizable subspace. Also note that all distributions that satisfy the linear instantaneous ICA model are marginally Gaussianizable. To illustrate this, we present two examples in Fig. 1; the distribution on the left is marginally Gaussianizable, while the one on the right is not, since the conditional distribution given X_2 is bimodal at some values of X_2. The reason for this is the following: marginal Gaussianizing transformations are invertible function and geometrically they correspond to a local nonlinear stretching/squeezing operation, therefore the nonconvex nature of a conditional distribution as in Fig. 1b cannot be convexified by such transformations and joint Gaussianization is not possible through marginal operations. Nevertheless, the marginal Gaussianization is still useful in many cases, especially if combined with tools that can generate localization of marginally Gaussianizable components in the data such as local principle component analysis (PCA), vector quantization, or clustering. Each local component can be treated under the presented framework to form a global mixture density model.

2.1. *Marginal Gaussianizing Transformations*

Given an n-dimensional random vector $\mathbf{X}$ with joint pdf $f(\mathbf{x})$ that satisfies the convexity condition mentioned above, there exist infinitely many functions $\mathbf{h}:\Re^n\rightarrow\Re^n$ 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 ith marginal of $\mathbf{X}$ is distributed according to $f_i(x_i)$, with a corresponding cumulative distribution function (cdf) $F_i(x_i)$. Let ϕ (.) denote the cdf of a zero-mean unit-variance single dimensional Gaussian variable:

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

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 $\mathbf{h}:\Re^n\rightarrow\Re^n$. Note that after this transformation (whose Jacobian is diagonal everywhere) we obtain a jointly Gaussian vector $\mathbf{Y}$ with zero mean and covariance

$$\boldsymbol{\Sigma}=E\left[\mathbf{Y}\mathbf{Y}^T\right] \tag{3}$$

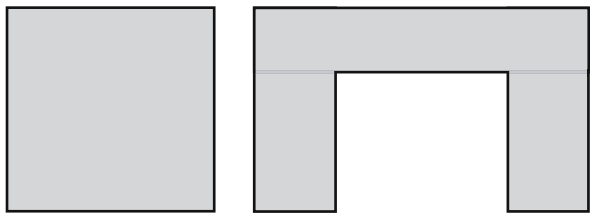

Figure 1. Consider two distributions uniform on the regions shown above. Horizontal and vertical axis correspond to X_1 and X_2, respectively. The distribution in (a) is marginally Gaussianizable, while the one in (b) is not.

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

$$f(\mathbf{x}) = \frac{g_{\mathbf{\Sigma}}(\mathbf{h}(\mathbf{x}))}{\left|\nabla \mathbf{h}^{-1}(\mathbf{h}(\mathbf{x}))\right|} = g_{\mathbf{\Sigma}}(\mathbf{h}(\mathbf{x})) \cdot |\nabla \mathbf{h}(\mathbf{x})|$$

$$= g_{\mathbf{\Sigma}}(\mathbf{h}(\mathbf{x})) \cdot \prod_{i=1}^{n} \frac{f_i(x_i)}{g_1(h_i(x_i))} \tag{4}$$

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.

The proposed joint density estimation is based on Eq. (4). Density estimation is carried out using a set of independent and identically distributed (iid) samples $\{\mathbf{x}_1,...,\mathbf{x}_N\}$ 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 $\widehat{f}_i(.)$. While variable kernel-size and weighted Parzen window estimates provide more flexibility and better asymptotic convergence properties, in this paper, we will employ unweighted and fixed-size kernel density estimates for simplicity. The extension to other density estimation methods is trivial.

Since the marginal Gaussianizing functions $h_i(.)$ require an accurate estimate of the marginal distributions of the data, the kernel sizes in the Parzen window estimates for each dimension must be optimized. A suitable approach is to minimize the KLD as in Eq. (1). This procedure will be described in detail in the next section. From these estimates, approximate Gaussianizing transformations $\widehat{h}_i(.)$ can be easily constructed. Assuming that these estimated transformations convert the joint data distribution to Gaussian, the covariance matrix is simply estimated from the samples using

$$\widehat{\mathbf{\Sigma}} = \frac{1}{N} \sum_{j=1}^{N} \widehat{\mathbf{y}}_j \widehat{\mathbf{y}}_j^T \tag{5}$$

where $\widehat{\mathbf{y}}_j = \widehat{\mathbf{h}}(\mathbf{x}_j)$.[1] In this second phase of the procedure, we basically assume that the samples $\{\widehat{\mathbf{y}}_1, ..., \widehat{\mathbf{y}}_N\}$ 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. Overall, the proposed two-step procedure for estimating the joint distribution of a set of iid samples equivalently minimizes the KLD in an approximate manner as illustrated in Fig. 2. 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 $\widehat{\mathbf{h}}$ is obtained. The corresponding transformed distribution $p_{\widehat{\mathbf{\Sigma}}}$ is projected optimally to the manifold of Gaussian distributions to obtain $g_{\widehat{\mathbf{\Sigma}}}$, which is a better approximation to $g_{\mathbf{\Sigma}}$ due to the Pythagorean Theorem for KLD [5]. The final density estimate is obtained by employing the inverse transformation $\widehat{\mathbf{h}}^{-1}$ to $g_{\widehat{\mathbf{\Sigma}}}$. Clearly, as the number of samples increase, the estimated joint distribution will approach the true underlying data distribution.

Kernel 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 $\{x_1,...,x_N\}$ with true distribution $f(\mathrm{x})$, the Parzen window estimate for this distribution is

$$\widehat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} K_\sigma(x - x_i) \tag{6}$$

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 $E\left[\widehat{f}(x)\right] = f(x) * K_\sigma(x)$, 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

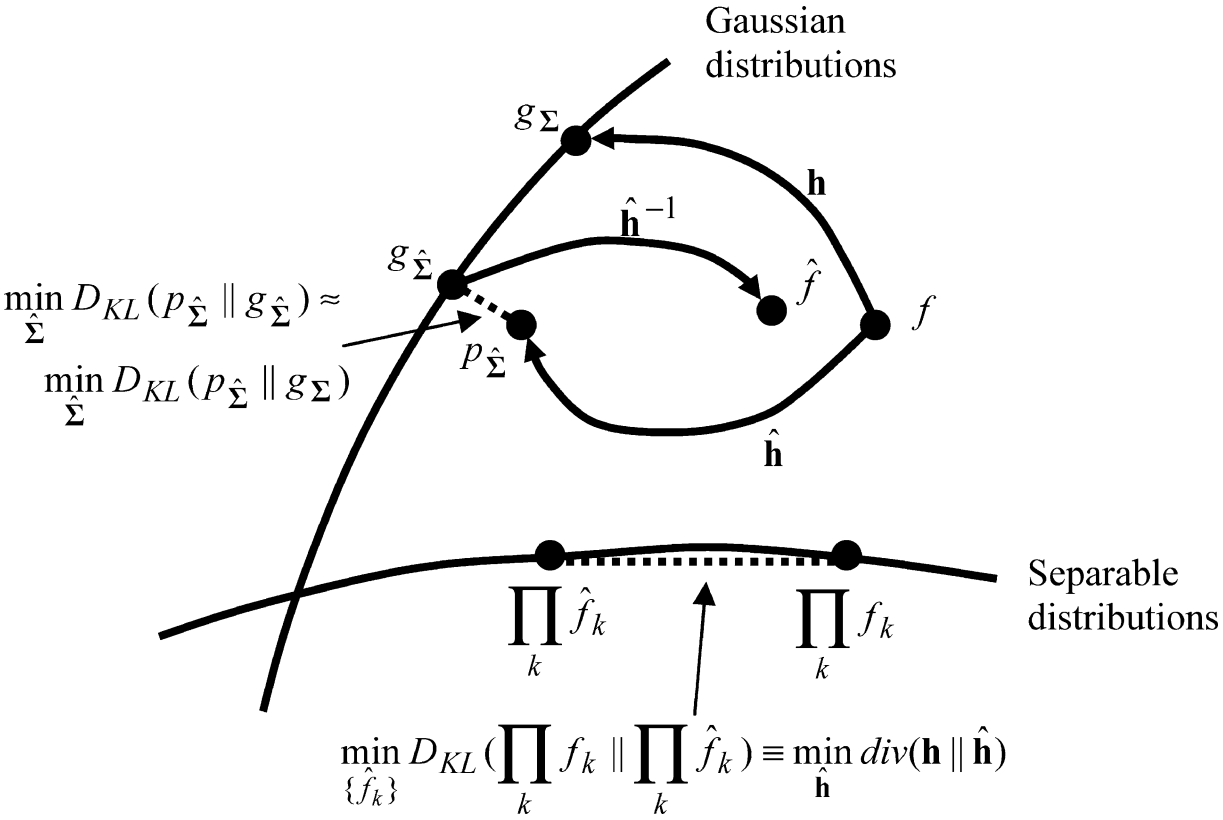

Figure 2. This is an illustration of the proposed joint density estimation procedure. Optimization is carried out in two steps. The marginal density estimates are determined by minimizing the KLD, which is equivalent to minimizing some form of divergence between the estimated and actual Gaussianizing 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.

trade-off between estimation bias and variance: decreasing the kernel size increases the variance, whereas increasing the kernel size increases the bias. In particular, if the following are satisfied, Parzen windowing asymptotically yields an unbiased and consistent estimate: $\lim_{N\to\infty} \sigma(N) = 0$ and $\lim_{N\to\infty} N\sigma(N) = \infty$. 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. 3.[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 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: σ=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].

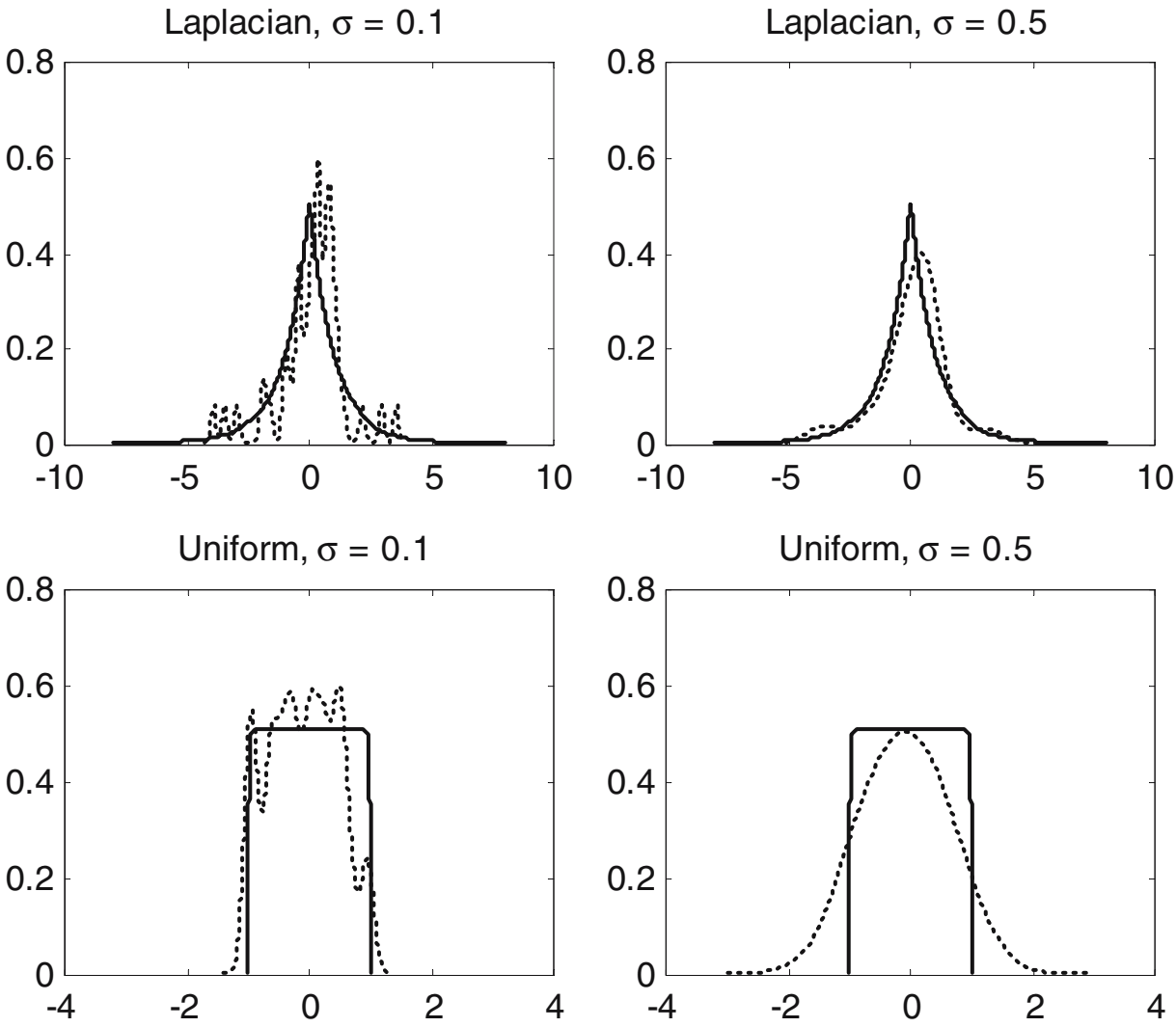

Figure 3. Laplacian and uniform distributions estimated using Parzen windowing with Gaussian kernels (kernel size *indicated in titles*) with 50 samples from each distribution.

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

$$\widehat{f}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\prod_{k=1}^{n} K_{\sigma_k}\left(x^k - x_i^k\right) \tag{7}$$

where x^k denotes the kth entry of the vector $\mathbf{x}$ and the multidimensional kernel is the product of unidimensional kernels, all using appropriately selected widths—referred to as product kernel-based Parzen windowing.

In this paper, motivated by the graphical description of the method in Fig. 2, 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 Eq. (1), the ML approach turns out to be optimal in an information theoretic sense after all.

Maximum Likelihood Kernel Size Optimization
Here, 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 Eq. (6). Our goal is to minimize the KLD between the true and the estimated densities $f(x)$ and $\widehat{f}(x)$. Equivalently we will maximize the log-likelihood of the observed data, i.e.,$E_X\left[\log\widehat{f}(X)\right]$. The expectation is approximated by the sample mean, resulting in

$$J(\sigma) = \frac{1}{N}\sum_{\mathrm{j}=1}^{N}\log\widehat{f}\left(x_j\right) \tag{8}$$

For Parzen windowing this becomes

$$J(\sigma) = \frac{1}{N}\sum_{j=1}^{N}\log\left(\frac{1}{N}\sum_{i=1}^{N}K_\sigma\left(x_j - x_i\right)\right) \tag{9}$$

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

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

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: $\{x_1,...,x_M\}$ and $\{x_{M+1},...,x_N\}$. While one set was used in the density estimation, the other was used in the sample mean. If desired, a generalized version of Eq. (10) 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.

The kernel size can be optimized by maximizing Eq. (10) using standard iterative procedures such as a gradient ascent or an EM-like fixed-point algorithm. Alternatively, (semi-) Newton methods could be utilized for faster convergence. Silverman's rule-of-thumb is a suitable initial estimate for the optimal kernel size.

We illustrate the utility of the kernel size optimization procedure described above by demonstrating how the solution 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 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 Eq. (10) 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).

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
β=1	0.56	0.48	0.45	0.41
β=2	0.50	0.38	0.38	0.38
β=3	0.43	0.37	0.34	0.30
β=5	0.34	0.27	0.25	0.24

2.2. *Joint Gaussianizing Transformations*

The marginal Gaussianizing transformations have the drawback of being unsuitable for some situations such as the example shown in Fig. 1. In general, a

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

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

joint Gaussianization procedure is necessary and a neural network could be employed for this purpose. Consider a multiplayer perceptron (MLP) for this purpose. Given a random vector $\mathbf{X}$, there exists an MLP $\mathbf{g}(.)$ such that $\boldsymbol{Y}=\mathbf{g}(\mathbf{X})$ is jointly Gaussian with zero-mean and identity-covariance. This MLP could be determined by optimizing its coefficients with respect to a suitable criterion under a fixed-output-covariance constraint. As it is well known, under the fixed-covariance constraint, the Gaussian distribution maximizes entropy [5]. Consequently, the weights of the MLP are optimized according to the following:

$$\max_{\mathbf{w}} H_S(\mathbf{Y}) \text{ subject to } E[\mathbf{Y}]=\mathbf{0}, E\left[\mathbf{Y}\mathbf{Y}^T\right]=\mathbf{I} \quad (11)$$

This is similar to the Infomax principle [19] where the entropy at the output of a sigmoid nonlinearity is maximized to estimate the joint entropy of a distribution. Infomax, however, relies on the accurate estimation of appropriate sigmoid nonlinearities for the proper estimation of the joint distribution. In many situations, these nonlinear functions may be difficult to *guess*.

In Eq. (11), the entropy of the network output can be estimated using Parzen windowing with multidimensional kernels. If these kernels are selected to be separable as in Eq. (7) (e.g., in the case of a Gaussian kernel, with a diagonal kernel covariance matrix) the maximum likelihood procedure described in the previous section can be employed to optimize the kernel size individually for each dimension. If the topology is a 2-layer MLP, the constraints can be incorporated by selecting the linear second (output) layer weight matrix to satisfy the constraints (i.e., as the whitening matrix of the hidden layer outputs) after every update of the first layer weight matrix. We will not study this possibility in detail here, since the focus of this paper is the marginal Gaussianization case.

3. Applications

The Gaussianization procedure described above is applicable to all problems where the solution can be formalized based on the joint density estimate of the data. In this section, we will present the following applications: nonparametric classifier design, independent component analysis, and extending Price's theorem.

3.1. 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 Eq. (4) or the product-kernel-based Parzen windowing method presented in Eq. (7). 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 $f_i(\mathbf{x})=c_1 e^{-c_2\|\mathbf{x}-\boldsymbol{\mu}_i\|_\infty}$ whose means were selected as $\boldsymbol{\mu}_1=[-1\,-1]^T$ and $\boldsymbol{\mu}_2=[1\;\;1]^T$. 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.

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. 4a. 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 two 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. 4b 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.

3.2. *Independent Components Analysis*

Independent components analysis (ICA) is now a mature field with numerous approaches and algorithms to solve the basic instantaneous linear mixture case as well as a variety of extensions of these basic principles to solve the more complicated problems involving convolutive or nonlinear mixtures [20–22]. Due to the existence of a wide literature and excellent survey papers [23, 24], in addition to the books listed above, we shall not go into a detailed literature survey. In this section, we will demonstrate the utility of Gaussianization in ICA and establish its relationship with nonlinear principal components analysis (NPCA) [25]. We would like to stress that the goal of this section is *not* to present yet another ICA algorithm, but to demonstrate an interesting selection of the nonlinearity in NPCA as this method is applied to solving the linear ICA problem [26], as well as to illustrate the applicability of Gaussianization to nonlinear ICA (which will be called Homomorphic ICA) [27]. For the latter problem, certain existence and uniqueness criteria have recently been demonstrated by Hyvarinen and Pajunen [28]. Several different techniques include minimum mutual information [29], variational Bayesian learning [30], symplectic transformations and nonparametric entropy estimates [31], higher order statistics [32], temporal decorrelation [33], and kernel-based methods [34]. A review of the current state-of-the-art in nonlinear ICA is provided recently by Jutten and Karhunen [24].

Nonlinear ICA The nonlinear ICA problem is described by a generative signal model that assumes the observed signals, denoted by $\mathbf{x}$, are a nonlinear

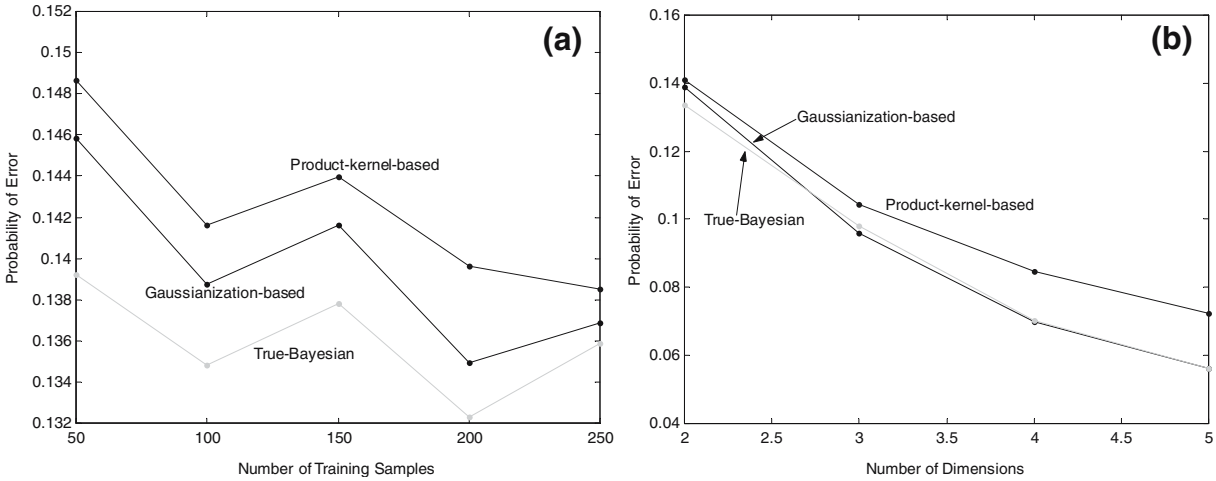

Figure 4. Probability of error for the three classifiers on a test set of 100 samples averaged over 100 Monte Carlo runs for (a) different sizes of training set with fixed dimensionality (b) different dimensionalities of training set using fixed number of training samples.

instantaneous function of some unknown independent source signals, denoted by $\boldsymbol{s}$. In particular, $\mathbf{x}_k=\mathbf{h}(\mathbf{s}_k)$, where k is the sample index. Let the observation vector be n-dimensional, $\mathbf{x}_k \in \Re^n$. Then, according to the existence results on nonlinear ICA, it is always possible to construct a function $\mathbf{g}: \Re^n \to \Re^n$, such that the outputs $\mathbf{y}=\mathbf{g}(\mathbf{x})$ are mutually independent [28]. Furthermore, this separation function is not unique. Clearly, there exist a number of operations that one might employ to change the distributions of these outputs individually without introducing mutual dependence; thus an uncertainty regarding the independent component densities exists. Furthermore, as will be shown later, in accordance with the rotation uncertainty reported in [28], the Homomorphic ICA solution will separate the observation into independent components, which are possibly a related to the original sources by an unknown rotation matrix. Also, by partitioning the variables in $\mathbf{y}$ to disjoint sets and taking various nonlinear combinations of the variables in these partitions, it is possible to generate a random vector $\mathbf{z} \in \Re^m$, where $m<n$ is the number of partitions. Thus, $\mathbf{z}=\mathbf{f}(\mathbf{y})=\mathbf{f}(\mathbf{g}(\mathbf{x}))$ also has independent components. Hence it is, in fact, possible to come up with infinitely many separating solutions that result in a smaller number of outputs than the inputs. A number of possible regularization conditions have been proposed before [28, 30] to ensure uniqueness and the actual separation of the unknown sources.

Due to these uncertainties, we will consider the problem of determining n independent components from $\mathbf{x} \in \Re^n$, which is a necessary condition for independent source separation, but not sufficient. In particular, the essence of the proposed solution is to generate n independent Gaussian distributed outputs and this can be achieved quite easily. Consider the ideal case where an observation vector $\mathbf{x} \in \Re^n$ is available and the marginal cumulative distribution functions (cdf) of each observed signal is known. Let $\mathbf{x}=[x^1,...,x^n]^T$ and let $F_d(.)$ denote the cdf of x^d. Also let $\phi_\sigma(.)$ denote the cdf of a zero-mean Gaussian random variable with variance σ^2. According to Section 2.1, z^d has a zero-mean, unit-variance Gaussian pdf: $z^d = \phi_1^{-1}\left(F_d\left(x^d\right)\right) = g_d\left(x^d\right)$. Combining these random variables into a random vector $\mathbf{z}=[z^1,...,z^n]^T$, we observe that the joint distribution of $\mathbf{z}$ is also zero-mean Gaussian with covariance $\mathbf{\Sigma_z}$. Now consider the principal components of $\mathbf{z}$. Let $\mathbf{y}=\mathbf{Q}^T\mathbf{z}$, where $\mathbf{Q}$ is the orthonormal eigenvector of $\mathbf{\Sigma_z}$, such that $\mathbf{\Sigma_z}=\mathbf{Q\Delta Q}^T$ and $\mathbf{\Delta}$ is the diagonal eigenvalue matrix. Then the covariance of $\boldsymbol{y}$ is $\mathbf{\Sigma_y}=\mathbf{\Delta}$. Hence, since $\mathbf{z}$ is zero-mean jointly Gaussian, $\mathbf{y}$ is zero-mean and jointly Gaussian with covariance $\mathbf{\Delta}$. It is well known that uncorrelated Gaussian random variables are also independent. Therefore, the components of $\mathbf{y}$ are mutually independent.[4] The overall scheme of the proposed nonlinear ICA topology is illustrated in Fig. 5.

Certain conditions must be met by the nonlinear mixing function for the separated outputs and the original sources to have maximal mutual information. In the most restrictive case, for the reconstruction of independent components that are related to the original sources by an invertible function, the mixing function must be invertible, i.e., its Jacobian must be nonsingular when evaluated at any point in its input space.[5] The following theorem summarizes this fact.

Theorem 3.2.1. If the source distribution obeys the convexity condition of Section 2.1, the mixing nonlinearity is invertible, and the marginal probability distributions of the observed vector are always positive except possibly at a set of points whose measure is zero, then, with probability one, there is a one-to-one function between the source signals and the independent components when the outputs are constructed according to Homomorphic ICA rules.

Proof By the first two assumptions, the joint mixture distribution obeys the convexity condition. By construction the PCA matrix $\mathbf{Q}^T$ is invertible and the Gaussianizing function g is monotonically increasing in all principal directions with probability one since the measure of the set on which its Jacobian has zero eigenvalues is zero. Similarly, due to the same reason, the probability of having source signals in this zero-measure set is zero. Therefore, with probability one, the Jacobian of the overall nonlinear function from $\boldsymbol{s}$ to $\boldsymbol{y}$ is invertible. Hence, there is a one-to-one relationship between these two vectors. ■

Another possible scenario is that the mixing nonlinearity is only locally invertible (i.e., its Jacobian is invertible in a set $S \subset \Re^n$). In this case, if S is the support of the source distribution, one can achieve maximum mutual information between the

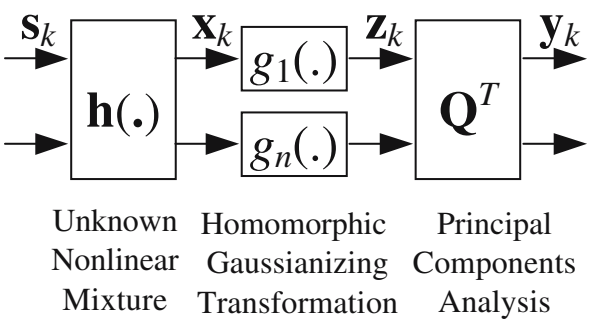

Figure 5. A schematic diagram of the proposed homomorphic independent components analysis topology.

separated outputs and the original sources. It is well known that the nonlinear ICA problem is ill-posed and the original sources can be at most resolved up to a rotation uncertainty with the independence assumptions alone. That is, even if the mixing function is invertible, one can arrive at independent components that are not necessarily the separated versions of the original sources. This can easily be observed by examining the Homomorphic ICA output. Suppose a set of independent components are obtained from an observed vector $\mathbf{x}$ by $\mathbf{y}=\mathbf{Q}^T\mathbf{g}(\mathbf{x})$, where $\mathbf{g}(.)$ consists of individual Gaussianizing functions for each components of $\mathbf{x}$ and $\mathbf{Q}$ is the orthonormal eigenvector that is the solution to the PCA problem after Gaussianization. If the covariance of $\mathbf{y}$ is $\mathbf{\Lambda}$, by selecting an arbitrary orthonormal matrix $\mathbf{R}$, one can generate the output $\mathbf{z}=\mathbf{R}\mathbf{\Lambda}^{-1/2}$, which still has independent components (since it is jointly Gaussian with identity covariance matrix), however, different choices of $\mathbf{R}$ result in different independent components. In order to resolve this ambiguity, one needs additional information about the sources or the mixing process.

The principle can be applied to complex-valued nonlinear mixtures as well. Consider the following complex-signal-complex-mixture model: $\mathbf{x}_r + i\mathbf{x}_i = \mathbf{h}_r(s) + i\mathbf{h}_r(\mathbf{s})$, where $\mathbf{s} = \mathbf{s}_r + i\mathbf{s}_i$. The Gaussianizing homomorphic transformations are denoted by $g_{dr}(.)$ and $g_{di}(.)$ for the real and imaginary parts of the dth observed signal in $\mathbf{x}$. The result of Gaussianization is the complex Gaussian vector $\mathbf{z} = \mathbf{z}_r + i\mathbf{z}_i$, whose covariance is $\mathbf{\Sigma}_z=\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathrm{H}}$. The separated outputs are given by $\boldsymbol{y}=\mathbf{Q}^H\mathbf{z}$. A theorem similar to the one above can be proven for the complex-valued case as well.

Theorem 3.2.2 If the source distribution obeys the convexity condition of Section 2.1, the marginal probability distributions of the observed vector are always positive except possibly at a set of points whose measure is zero, and the function $\overline{\mathbf{h}}(\mathbf{s}_r, \mathbf{s}_i) = \left[\mathbf{h}_r^T(\mathbf{s}_r, \mathbf{s}_i)\, \mathbf{h}_i^T(\mathbf{s}_r, \mathbf{s}_i)\right]^T$ is invertible then, with probability one, the mutual information between the original source vector $\mathbf{s}$ and the separated output vector $\mathbf{y}$ is maximized.

Proof Note that, the output is explicitly given by $\mathbf{y}_r + i\mathbf{y}_i = \left(\mathbf{Q}_r^T - i\mathbf{Q}_i^T\right)(\mathbf{g}_r(\mathbf{h}_r(\mathbf{s})) + i\mathbf{g}_i(\mathbf{h}_i(\mathbf{s})))$. We construct the vectors $\overline{\mathbf{y}} = \left[\mathbf{y}_r^T\, \mathbf{y}_i^T\right]^T$ and $\overline{\mathbf{s}} = \left[\mathbf{s}_r^T\, \mathbf{s}_i^T\right]^T$. The Jacobian of $\overline{\mathbf{y}}$ with respect to $\overline{\mathbf{s}}$ is

$$\frac{\partial \overline{\mathbf{y}}}{\partial \overline{\mathbf{s}}} = \begin{bmatrix} \mathbf{Q}_r^T & \mathbf{Q}_i^T \\ -\mathbf{Q}_i^T & \mathbf{Q}_r^T \end{bmatrix} \cdot \begin{bmatrix} \nabla \mathbf{g}_r(\mathbf{h}_r(\mathbf{s})) & \mathbf{0} \\ \mathbf{0} & \nabla \mathbf{g}_i(\mathbf{h}_i(\mathbf{s})) \end{bmatrix} \cdot \begin{bmatrix} \partial \mathbf{h}_r(\mathbf{s})/\partial \mathbf{s}_r & \partial \mathbf{h}_r(\mathbf{s})/\partial \mathbf{s}_i \\ \partial \mathbf{h}_i(\mathbf{s})/\partial \mathbf{s}_r & \partial \mathbf{h}_i(\mathbf{s})/\partial \mathbf{s}_i \end{bmatrix} \tag{12}$$

This Jacobian is nonsingular at every possible value of $\boldsymbol{s}$ if and only if the third term on the right hand side of Eq. (12) is nonsingular for every value, since the other two terms are nonsingular (the second term is nonsingular with probability one as discussed in Theorem 3.2.1). Thus with Homomorphic ICA, the function from the original sources to the outputs is invertible with probability one, which equivalently means maximum mutual information between these vector signals. ■

Linear ICA The linear ICA problem is described by a generative signal model that assumes the observed signals, denoted by $\mathbf{x}$, and the sources, denoted by $\mathbf{s}$, are obtained by a *square* linear system of equations. The sources are assumed to be statistically independent. In summary, assuming an unknown mixing matrix $\mathbf{H}$, we have $\mathbf{x}_k=\mathbf{H}\mathbf{s}_k$ where the subscript k is the sample/time index. The linear ICA problem exhibits permutation and scaling ambiguities, which cannot be resolved by the independence assumption. For the sake of simplicity in the following arguments, we will assume that the marginal pdfs of the sources and the mixtures are known and all are strictly positive valued (to guarantee the invertibility of Gaussinizing transformations). It is assumed without loss of generality

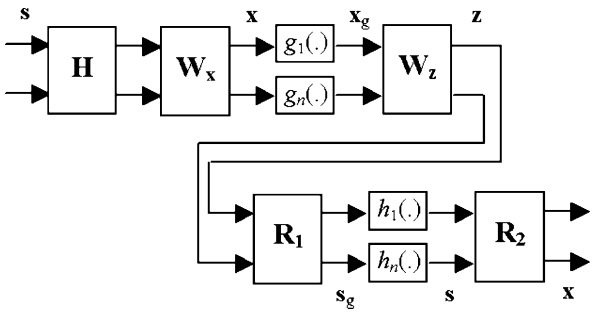

Figure 6. Schematic diagram of the proposed homomorphic linear ICA topology.

that the sources are already zero-mean. Consider the topology shown in Fig. 6 as a solution to linear ICA. The observed mixtures are first spatially whitened by $\mathbf{W_x}$ to generate the whitened mixture vector $\mathbf{x}$. Since whitening reduces the mixing matrix to only a coordinate rotation, without loss of generality, we can always focus on mixing matrices that are orthonormal. In this case, we assume that the mixing matrix is $\mathbf{R_2}=\mathbf{W_x}\mathbf{H}$. Since the marginal pdfs of the mixtures are known, one can construct the Gaussianizing functions $g_i(.)$ according to the previous section to obtain the Gaussianized mixtures $\mathbf{x}_g$. Whitening the Gaussianized mixtures will yield zero-mean unit-variance and uncorrelated signals $\mathbf{z}$. Since $\mathbf{z}$ is jointly Gaussian, uncorrelatedness corresponds to mutual independence. However, considering the function from the sources ($\mathbf{s}$) to the Gaussianized mixtures ($\mathbf{x_g}$) as a post-nonlinear mixture, we notice that although by obtaining $\mathbf{z}$ we have obtained independent components, due to the inherent rotation ambiguity of nonlinear mixtures in the ICA framework [28], we have not yet achieved source separation. Consequently, there is still an unknown orthonormal matrix $\mathbf{R_1}$ that will transform $\mathbf{z}$ into Gaussianized versions of the original sources. If the marginal source pdfs are known, the inverse of the Gaussianizing transformations for the sources could be obtained in accordance with the previous section (denoted by $h_i(.)$ in the figure), which would transform $\mathbf{s_g}$ to the original source distribution, thus yield the separated source signals (at least their estimates).

In summary, given the whitened mixtures, their marginal pdfs and the marginal pdfs of the sources (up to permutation and scaling ambiguities in accordance with the theory of linear ICA), it is possible to obtain an estimate of the orthonormal mixing matrix $\mathbf{R_2}$ and the sources $\boldsymbol{s}$ by training a constrained multilayer perceptron (MLP) topology with first layer weights given by $\mathbf{R_1}$ and second layer weights given by $\mathbf{R_2}$. The nonlinear functions of the hidden layer processing elements (PE) are determined by the inverse Gaussianizing transformations of the source signals. This MLP with square first and second layer weight matrices would be trained according to the following constrained optimization problem:

$$\min_{\mathbf{R_1},\mathbf{R_2}} E\left[\|\mathbf{x}-\mathbf{R_2}\mathbf{h}(\mathbf{R_1}\mathbf{z})\|^2\right] \text{subject to } \mathbf{R_1}\mathbf{R_1^T} = \mathbf{I}, \mathbf{R_2}\mathbf{R_2^T} = \mathbf{I}. \tag{13}$$

Constrained neural structures of this type have been considered previously by Fiori [35]. Interested readers are referred to his work and the references therein to gain a detailed understanding of this subject.

This technique is, in fact, a special case based on mutual information of the nonlinear PCA approach for solving linear ICA using properly selected nonlinear projection functions. Various choices of these functions correspond to different ICA criteria ranging from kurtosis to maximum likelihood (ML) [20]. In the most general sense, the NPCA problem is compactly defined by the following optimization problem:

$$\min_{\mathbf{W}} E\left[\left\|\mathbf{x}-\mathbf{W}\mathbf{f}(\mathbf{W}^T\mathbf{x})\right\|^2\right] \tag{14}$$

where $\mathbf{f}(.)$ is an elementwise function (i.e., with a diagonal Jacobian at every point) that is selected *a priori*. For the special case of $\mathbf{f}(\mathbf{z})=\mathbf{z}$, this optimization problem reduces to the linear bottleneck topology, which is utilized by Xu to obtain the LMSER algorithm for linear PCA [36]. Returning to the topology in Fig. 6, under the assumptions of invertibility (which is satisfied if and only if the source pdfs are strictly greater than zero[6]) we observe that $\mathbf{z}=\mathbf{W_z}\mathbf{g}(\mathbf{x})$ and $\mathbf{x}=\mathbf{R_2}\boldsymbol{s}$, therefore, the cost function in Eq. (13) is $E\left[\|\mathbf{R_2}\mathbf{s}-\mathbf{R_2}\mathbf{h}(\mathbf{R_1}\mathbf{W_z}\mathbf{g}(\mathbf{R_{2s}}))\|^2\right]$. Being orthonormal, $\mathbf{R_2}$ does not affect the Euclidean norm, and the cost becomes $E\left[\|\mathbf{s}-\mathbf{h}(\mathbf{R_1}\mathbf{W_z}\mathbf{g}(\mathbf{R_2}\mathbf{s}))\|^2\right]$. In the ICA setting, $\mathbf{s}$ is approximated by its estimate, the separated outputs $\mathbf{y}$, which is the output of the $\mathbf{h}(.)$ stage of Fig. 1. In the same setting, assuming whitened mixtures, NPCA would optimize

$$\min_{\mathbf{W}} E\left[\|\mathbf{y}-\mathbf{f}(\mathbf{y})\|^2\right] \tag{15}$$

where $\mathbf{y}=\mathbf{w}\mathbf{x}$, in accordance with Eq. (14) [20]. A direct comparison of Eq. (15) and the expression

given above that is equivalent to Eq. (13) yields $\mathbf{f}(\mathbf{y}) = \mathbf{h}(\mathbf{R_1}\mathbf{W}_z\mathbf{g}(\mathbf{R_2}\mathbf{y}))$. In summary, the homomorphic linear ICA approach tries to determine a nonlinear subspace projection of the separated outputs such that the projections become independent. While an arbitrary selection of the nonlinear projection functions would not necessarily imply independence of the separated outputs, the proposed approach specifically exploits homomorphic Gaussianizing transformations of the signals such that orthogonality (uncorelatedness of zero-mean signals) is equivalent to mutual independence.

3.3. *Extending Price's Theorem to Non-Gaussian Input Distributions*

In nonlinear information processing, Price's theorem plays an important role [37]. It allows calculating the expected value of a nonlinear function of jointly Gaussian input random variables by facilitating the construction of a set of ordinary or partial differential equations relating the sought quantity to the correlation coefficients between pairs of input variables. While the original theorem deals with the class of separable nonlinear functions, several extensions to Price's theorem have been proposed to generalize the theorem. Specifically, while Price's original theorem dealt with separable functions on multiple input arguments, McMahon provided a generalization to bivariate jointly Gaussian inputs processed by nonlinear functions that are not necessarily separable [38]. Pawula extended McMahon's result to arbitrary number of input arguments [39]. While the original theorem and these extensions relied on the use of Laplace transforms, which introduced restrictive existence conditions for the integrals, the condition on the nonlinear function for the existence of the expectation was relaxed by Papoulis in the bivariate case [40]. Papoulis' idea was also utilized later by Brown in determining the most general form of Price's theorem to date including its converse statement with a weak convergence condition on the nonlinearity involved [41]. Recently, Price's theorem was also generalized for functions of any number of jointly Gaussian complex random variables [42]. Price's theorem and almost all of its extensions deal with the problem of information processing by nonlinear memoryless systems acting on jointly Gaussian inputs (with the exception of McGraw and Wagner's extension of the result to elliptically symmetric distributions [43], which is a special case of marginally Gaussianizable distributions that we will discuss here. Since nonlinear systems with finite-memory can be regarded as memoryless provided that the input vector definition is extended to encompass all past input values within the memory depth, the application of Price's theorem to finite-memory systems such as finite impulse response (FIR) filters and time-delay neural networks (TDNN) is trivial. All one needs to do is to modify the input vector and the associated covariance matrix by considering the temporal correlations in the input signal. In this section, we will extend Price's theorem such that nonlinear finite-memory information processing systems acting on stationary inputs with arbitrary probability distributions can be analyzed. For simplicity, only the case of real-valued signals will be considered here. Extensions of the idea presented here to complex signals can be easily accomplished following the same general principles and utilizing previously derived results on complex valued Gaussian inputs [42] and complex homomorphic ICA mentioned above. For completeness, we first present Brown's extension of Price's theorem.

Theorem 3.3.1 Assume that $\mathbf{X}$ is a random vector with components $X_1,...,X_n$. Without loss of generality, suppose that $E[X_i]=0$ and $E[X_i^2]=1$ for $i=1,...,n$. Let ρ_{ij} be the correlation coefficient between X_i and X_j, i.e., $\rho_{ij}=E[X_iX_j]$, where $i\neq j$. Suppose the joint probability density function (pdf) of $\mathbf{X}$ is $p_{\mathbf{X}}(\mathbf{x},\boldsymbol{\rho})$, where $\boldsymbol{\rho}$ denotes dependency of the joint distribution on the inter-variable correlations. The pdf $p_{\mathbf{X}}(\mathbf{x},\boldsymbol{\rho})$ is Gaussian if and only if the condition

$$\frac{\partial E_{\mathbf{X}}[f(\mathbf{X})]}{\partial \rho_{ij}} = E_{\mathbf{X}}\left[\frac{\partial^2 f(\mathbf{X})}{\partial X_i \partial X_j}\right] \tag{16}$$

holds identically for all real valued functions $f(\mathbf{X})$ defined on the n-dimensional Euclidean space having bounded continuous second partial derivatives with respect to its arguments X_i.

Proof Please see [41]. ■

Now, we extend Price's theorem to non-Gaussian inputs. Consider a memoryless nonlinear system $g(\mathbf{Z})$, where the input vector $\mathbf{Z}$ has an arbitrary joint distribution $p_{\mathbf{Z}}(.)$. In the case of a causal nonlinear

system with known finite memory depth, all we need to do is to define a new input vector consisting of all past inputs upto the memory depth of the system. Suppose that the Gaussianizing transformations for Z_i are known to be $h_i(.)$. Notice that this procedure does not require the knowledge of the full joint pdf $p_{\mathbf{Z}}(.)$, but only the marginal pdfs $p_i(.)$. In accordance with Price's theorem, we are interested in evaluating the following:

$$E_{\mathbf{Z}}[g(\mathbf{Z})] = \int g(\mathbf{z}) p_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z} \tag{17}$$

Using marginal Gaussianization, with a change of variables we observe that

$$\begin{aligned} E_{\mathbf{Z}}[g(\mathbf{Z})] &= \int g\left(\mathbf{h}^{-1}(\mathbf{x})\right) p_{\mathbf{Z}}\left(\mathbf{h}^{-1}(\mathbf{x})\right) \left|\nabla \mathbf{h}^{-1}(\mathbf{x})\right| d\mathbf{x} \\ &= \int g\left(\mathbf{h}^{-1}(\mathbf{x})\right) \frac{p_{\mathbf{Z}}\left(\mathbf{h}^{-1}(\mathbf{x})\right)}{\left|\nabla \mathbf{h}\left(\mathbf{h}^{-1}(\mathbf{x})\right)\right|} d\mathbf{x} \\ &= \int g\left(\mathbf{h}^{-1}(\mathbf{x})\right) \mathbf{G}(\mathbf{x}, \boldsymbol{\Sigma}) d\mathbf{x} \\ &= E_{\mathbf{X}}\left[g\left(\mathbf{h}^{-1}(\mathbf{X})\right)\right] . \end{aligned} \tag{18}$$

In Eq. (18), we assumed $\mathbf{h}(.)$ is invertible. Defining $f(\mathbf{X})=g(\mathbf{h}^{-1}(\mathbf{X}))$, where $\mathbf{X}$ is jointly Gaussian, we can employ Theorem 3.3.1 immediately. Thus, for arbitrary functions of inputs with arbitrary distributions that obey the convexity condition, we obtain the following theorem.

Theorem 3.3.2 Assume that $\mathbf{Z}$ is a random vector with components $Z_1,...,Z_n$. Suppose that the marginal Gaussianizing function $\mathbf{h}(.)$ for $\mathbf{Z}$ is known. Let $\mathbf{X}=\mathbf{h}(\mathbf{Z})$ be the corresponding jointly Gaussian random vector with distribution $\mathbf{G}(\mathbf{x},\boldsymbol{\Sigma})$. If $\mathbf{h}(.)$ is invertible then

$$\begin{aligned} \frac{\partial E_{\mathbf{Z}}[g(\mathbf{Z})]}{\partial \rho_{ij}} &= \frac{\partial E_{\mathbf{X}}\left[g\left(\mathbf{h}^{-1}(\mathbf{X})\right)\right]}{\partial \rho_{ij}} \\ &= E_{\mathbf{X}}\left[\frac{\partial^2 g\left(\mathbf{h}^{-1}(\mathbf{X})\right)}{\partial X_i \partial X_j}\right] \\ &= E_{\mathbf{X}}\left[\frac{g_{ij}\left(\mathbf{h}^{-1}(\mathbf{X})\right)}{h_i'\left(h_i^{-1}(X_i)\right) h_j'\left(h_j^{-1}(X_j)\right)}\right] \end{aligned} \tag{19}$$

for all real valued functions $\mathbf{g}: \Re^n \to \Re^n$ with bounded continuous second partial derivatives, $g_{ij}(\mathbf{z}) = \partial^2 g(\mathbf{z}) / \left(\partial z_i \partial z_j\right)$, such that

$$\left|\mathbf{g}(\mathbf{Z})\right| < A e^{\sum_k h_k^{\alpha}(Z_k)} \text{ for some } 0 < \alpha < 2, A > 0 \tag{20}$$

Conversely, for given invertible $h_i(.)$, i=1,...,n, if the equality in Eq. (19) is satisfied for all $\mathbf{g}(.)$ as described above, then $\mathbf{X}$ is jointly Gaussian with pdf $\mathbf{G}(\mathbf{x},\boldsymbol{\Sigma})$. Thus, the joint distribution of $\mathbf{Z}$ is given by

$$p_{\mathbf{Z}}(\mathbf{z}) = \frac{\mathbf{G}(\mathbf{h}(\mathbf{z}), \boldsymbol{\Sigma})}{\left|\nabla \mathbf{h}^{-1}(\mathbf{h}(\mathbf{z}))\right|} = \mathbf{G}(\mathbf{h}(\mathbf{z}), \boldsymbol{\Sigma}) \left|\nabla \mathbf{h}(\mathbf{z})\right|. \tag{21}$$

Proof Given the conditions stated in the theorem, the derivation in Eq. (19) is easily obtained using Eq. (18) in the first equation, Eq. (16) in the second equation, and chain rule of differentiation in the third equation. The existence condition in Eq. (20) is also obtained easily by an invertible change of variables from the relaxed existence condition for $E_{\mathbf{X}}[f(\mathbf{X})]$ pointed out by Papoulis [40], which is

$$|f(\mathbf{X})| < A e^{\sum_k X_k^{\alpha}} \text{ for some } 0 < \alpha < 2, A > 0 \tag{22}$$

The proof of the converse statement follows directly from the converse statement of Theorem 3.3.1 using the fundamental theorem of probability and the invertibility of $\mathbf{h}(.)$. ■

4. Conclusion

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. The practical and theoretical utility of the Gaussianization procedure is illustrated in MAP classifier design, linear and nonlinear ICA, and extending Price's theorem to arbitrary distributions.

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.

Acknowledgment

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

Notes

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 Eq. (5), without a correction term due to data dimensionality in the denominator.
2. The generalized Gaussian density family is described by $G_\beta(x) = C_1 \exp\left(-C_2|x|^\beta\right)$, 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 β=1 and β=∞.
3. To minimize Eq. (10), 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.
4. After developing this principle for nonlinear ICA, it came to the authors' attention that the importance of Gaussianization for breaking the curse of dimensionality was independently recognized earlier by Chen et al. [44].
5. Notice that for a broad class of nonlinear mixtures, the condition that at most one source can have a Gaussian distribution is not necessary, as the nonlinear mixture will not preserve the Gaussianity. The commonly considered post-nonlinear mixtures are easily excluded from this group. In fact, to the best knowledge of authors, there is no result available in the literature about the general conditions that the nonlinear mixture should satisfy for the non-Gaussianity condition to be lifted. Clearly, when applying the Homomorphic ICA principle to linear source separation using ICA, the non-Gaussianity conditions must still hold.
6. In the case of zero probability densities, the Gaussianizing functions will not be invertible in general, since locally at these points the Jacobian might become singular. However, since the probability of occurrence of such points is also zero for the same reason, for the given signal-mixture case global invertibility is not necessary. However, it is assumed for simplicity.

References

1. R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification," 2nd ed., Wiley, New York, 2001.
2. S. Theodoridis and K. Koutroumbas, "Pattern Recognition," Academic, New York, 2003.
3. L. Devroye and G. Lugosi, "Combinatorial Methods in Density Estimation," Springer, Berlin Heidelberg New York, 2001.
4. A. Papoulis, "Probability, Random Variables, Stochastic Processes," McGraw-Hill, New York, 1991.
5. T. M. Cover and J. A. Thomas, "Elements of Information Theory," Wiley, New York, 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, J. C. Principe, and T. Eltoft, "Towards a Unification of Information Theoretic Learning and Kernel Methods," in *Proceedings of MLSP'04*, Sao Luis, Brazil, 2004.
8. K. E. Hild II, D. Erdogmus, and J. C. Principe, "Blind Source Separation Using Renyi's Mutual Information," in *IEEE Signal Processing Letters*, no. 8, 2001, pp. 174–176.
9. K. Torkkola, "Visualizing Class Structure in Data Using Mutual Information," in *Proceedings of NNSP'00*, Sydney, Australia, 2000, pp. 376–385.
10. D. Erdogmus, "Information Theoretic Learning: Renyi's Entropy and its Applications to Adaptive System Training," Ph.D. Dissertation, University of Florida, Gainesville, Florida, 2002.
11. M. M. Van Hulle, "Kernel-Based Topographic Map Formation Achieved with an Information–Theoretic Approach," *Neural Netw.*, vol. 15, 2002, pp. 1029–1039.
12. N. N. Schraudolph, "Gradient-Based Manipulation of Nonparametric Entropy Estimates," *IEEE Trans. Neural Netw.*, vol. 15, no. 4, 2004, pp. 828–837.
13. K. Fukunaga, "Statistical Pattern Recognition," Academic, New York, 1990.
14. B. W. Silverman, "Density Estimation for Statistics and Data Analysis," Chapman & Hall, London, 1986.
15. M. C. Jones, J. S. Marron, and S. J. Sheather, "A Brief Survey of Bandwidth Selection for Density Estimation," *J. Am. Stat. Assoc.*, vol. 87, 1996, pp. 227–233.

16. R. P. W. Duin, "On the Choice of the Smoothing Parameters for Parzen Estimators of Probability Density Functions," *IEEE Trans. Comput.*, vol. 25, no. 11, 1976, pp. 1175–1179.
17. S. Amari, "Differential–Geometrical Methods in Statistics," Springer, Berlin Heidelberg New York, 1985.
18. P. Viola, N. Schraudolph, and T. Sejnowski, "Empirical Entropy Manipulation for Real-World Problems," in *Proceedings of NIPS'95*, 1996, pp. 851–857.
19. T. Bell and T. Sejnowski, "An Information–Maximization Approach to Blind Separation and Blind Deconvolution," *Neural Comput.*, vol. 7, 1995, pp. 1129–1159.
20. A. Hyvarinen, J. Karhunen, and E. Oja, "Independent Component Analysis," Wiley, New York, 2001.
21. A. Cichocki and S. I. Amari, "Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications," Wiley, New York, 2002.
22. T. W. Lee, "Independent Component Analysis: Theory and Applications," Kluwer, New York, 1998.
23. A. Hyvarinen, "Survey on Independent Component Analysis," *Neural Comput. Surv.*, vol. 2, 1999, pp. 94–128.
24. C. Jutten and J. Karhunen, "Advances in Nonlinear Blind Source Separation," in *Proceedings of ICA'03*, Nara, Japan, 2003, pp. 245–256.
25. J. Karhunen and J. Joutsensalo, "Representation and Separation of Signals Using Nonlinear PCA Type Learning," *Neural Netw.*, vol. 7, 1994, pp. 113–127.
26. D. Erdogmus, Y. N. Rao, and J. C. Principe, "Gaussianizing Transformations for ICA," in *Proceedings of ICA'04*, Granada, Spain, 2004, pp. 26–32.
27. D. Erdogmus, Y. N. Rao, and J. C. Principe, "Nonlinear Independent Component Analysis by Homomorphic Transformation of the Mixtures," in *Proceedings of IJCNN'04*, Budapest, Hungary, 2004, pp. 47–52.
28. A. Hyvarinen and P. Pajunen, "Nonlinear Independent Component Analysis: Existence and Uniqueness Results," *Neural Netw.*, vol. 12, no. 3, 1999, pp. 429–439.
29. L. B. Almeida, "MISEP—Linear and Nonlinear ICA Based on Mutual Information," *J. Mach. Learn. Res.*, vol. 4, 2003, pp. 1297–1318.
30. H. Valpola, E. Oja, A. Ilin, A. Honkela, and J. Karhunen, "Nonlinear Blind Source Separation by Variational Bayesian Learning," *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. 86, no. 3, 2003, pp. 532–541.
31. L. Parra, "Symplectic Nonlinear Independent Component Analysis," in *Proceedings of NIPS'96*, 1997, pp. 437–443.
32. Y. Tan and J. Wang, "Nonlinear Blind Source Separation Using Higher Order Statistics and a Genetic Algorithm," *IEEE Trans. Evol. Comput.*, vol. 5, no. 6, 2001.
33. A. Ziehe, M. Kawanabe, S. Harmeling, and K. R. Muller, "Blind Separation of Post-Nonlinear Mixtures Using Linearizing Transformations and Temporal Decorrelation," *J. Mach. Learn. Res.*, vol. 4, 2003, pp. 1319–1338.
34. S. Harmeling, A. Ziehe, M. Kawanabem, and K. R. Müller, "Kernel-Based Nonlinear Blind Source Separation," *Neural Comput.*, vol. 15, 2003, pp. 1089–1124.
35. S. Fiori, "A Theory for Learning by Weight Flow on Stiefel–Grassman Manifold," *Neural Comput.*, vol. 13, 2001, pp. 1625–1647.
36. L. Xu, "Least Mean Square Error Reconstruction Principle for Self-Organizing Neural Nets," *Neural Netw.*, vol. 6, 1993, pp. 627–648.
37. R. Price, "A Useful Theorem for Nonlinear Devices Having Gaussian Inputs," *IRE Trans. Inf. Theory*, vol. 4, 1958, pp. 69–72.
38. E. L. McMahon, "An Extension of Price's Theorem," *IEEE Trans. Inf. Theory*, vol. 10, 1964, p. 168.
39. R. F. Pawula, "A Modified Version of Price's Theorem," *IEEE Trans. Inf. Theory*, vol. 13, no. 2, 1967, pp. 285–288.
40. A. Papoulis, "Comment on 'An Extension of Price's Theorem'," *IEEE Trans. Inf. Theory*, vol. 11, 1965, p. 154.
41. J. L. Brown, "A Generalized Form of Price's Theorem and Its Converse," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, 1967, pp. 27–30.
42. A. van den Bos, "Price's Theorem for Complex Variates," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, 1996, pp. 286–287.
43. D. McGraw and J. Wagner, "Elliptically Symmetric Distributions," *IEEE Trans. Inf. Theory*, vol. 14, no. 1, 1968, pp. 110–120.
44. S. S. Chen and R. A. Gopinath, "Gausianization," in *Proceedings of NIPS*, 2000.

Deniz Erdogmus received the B.S. in Electrical & Electronics Engineering (EEE), and the B.S. in Mathematics both in 1997, and the M.S. in EEE in 1999 from the Middle East Technical University, Turkey. He received his PhD in Electrical & Computer Engineering from the University of Florida (UF) in 2002. He worked as a research engineer at TUBITAK-SAGE, Turkey from 1997 to 1999, focusing on the design of navigation, guidance, and flight control systems. He was also a research assistant and a postdoctoral research associate at UF from 1999 to 2004, concentrating on signal processing, adaptive systems, machine learning, and information theory, specifically with applications in biomedical engineering including brain machine interfaces. Currently, he is holding an Assistant Professor position jointly at the Computer Science and Electrical Engineering Department and the Biomedical Engineering Department of the Oregon Health and Science University. His research focuses on information theoretic adaptive signal processing and its applications to biomedical signal processing problems. Dr. Erdogmus has over 35 articles in international scientific journals and numerous conference papers and book chapters. He has also served as associate editor and guest editor for various journals, participated in various conference organization and scientific committees, and he is a member of Tau Beta Pi, Eta Kappa Nu, IEEE, and IEE. He was the recipient of the IEEE-SPS 2003 Best Young Author Paper Award and 2004 INNS Young Investigator Award.

Robert Jenssen received the MS and PhD in Electrical Engineering (EE), in 2001 and 2005, respectively, from the University of Tromso, Norway. In his research he has focused on an information theoretic approach to machine learning, including kernel methods, spectral clustering and independent component analysis. Jenssen spent the academic year 2002/2003 and March/April 2004 at the University of Florida, as a visitor at the Computational NeuroEngineering Laboratory. In 2005/2006, he is employed as an associate professor in EE at the University of Tromso. Starting August 2006, he assumes a three-year postdoctoral position funded by the Norwegian research council. Jenssen received the 2003 outstanding paper honor from the Pattern Recognition Journal, and the 2005 ICASSP outstanding student paper award.

Yadunandana N. Rao was born in Mysore, India. He received his BE in Electronics and Communication Engineering from the University of Mysore, India in 1997 and MS and PhD in Electrical and Computer Engineering from the University of Florida, Gainesville, FL in 2000 and 2004, respectively. Between August 1997 and July 1998, he worked a Software Engineer in Bangalore, India. From May 2000 to January 2001, he was a Design Engineer at GE Medical Systems, WI. He is currently with Motorola, Inc. His research interests include adaptive signal processing theory, algorithms and analysis, neural networks for signal processing, communications, and biomedical applications.

Jose C. Principe is Distinguished Professor of Electrical and Biomedical Engineering at the University of Florida since 2002. He joined the University of Florida in 1987, after an eight-year appointment as Professor at the University of Aveiro, in Portugal. Dr. Principe holds degrees in electrical engineering from the University of Porto (Bachelors), Portugal, University of Florida (Master and Ph.D.), USA and a Laurea Honoris Causa degree from the Universita Mediterranea in Reggio Calabria, Italy. Dr. Principe interests lie in nonlinear non-Gaussian optimal signal processing and modeling and in biomedical engineering. He created in 1991 the Computational NeuroEngineering Laboratory to synergistically focus the research in biological information processing models.

Dr. Principe is a Fellow of the IEEE, past President of the International Neural Network Society, and Editor in Chief of the Transactions of Biomedical Engineering since 2001, as well as a former member of the Advisory Science Board of the FDA. He holds 5 patents and has submitted seven more. Dr. Principe was supervisory committee chair of 47 Ph.D. and 61 Master students, and he is author of more than 400 refereed publications (3 books, 4 edited books, 14 book chapters, 116 journal papers and 276 conference proceedings).