Edgeworth Approximations of the Kullback-Leibler Distance Towards Problems in Image Analysis

Jen-Jen Lin

Associate Professor, Department of Applied Statistics, Ming Chuan University, Taipei 11120, Taiwan
email: jjlin@mcu.edu.tw

Naoki Saito

Associate Professor, Department of Mathematics, University of California, Davis, CA 95616,USA
email: saito@math.ucdavis.edu

Richard A. Levine

Assistant Professor, Department of Statistics, University of California, Davis, CA 95616, USA
email: levine@wald.ucdavis.edu

Summary.
Evaluation of syntheses or simulated data is often done subjectively through visual comparisons with the original samples. This subjective evaluation is particularly dominant in the area of texture modeling and simulation. In order to *objectively* evaluate the similarity (or difference) between original samples and syntheses, we propose an approximation for the Kullback-Leibler distance based on Edgeworth expansions (EKLD). We use this approximation to study the sampling distribution of the original and synthesized images. As part of our development, we present numerical examples to study the behavior of EKLD for sample mean distributions and illustrate the advantages of our approach for evaluating the differential entropy and choosing the least statistically dependent basis from wavelet packet dictionaries. Finally, we introduce how to use EKLD in statistical image processing to validate synthetic representations of images.

Keywords: differential entropy, cumulants, least statistically dependent basis, wavelet packet dictionary, image processing.

1. Introduction

Given a sample of images which obey an unknown distribution, several simulation methods exist to generate synthetic images (see Cross and Jain, 1983; Geman and Geman, 1984; Popat and Picard, 1997; Portilla and Simoncelli, 2000; Simoncelli, 1997; Zhu, Wu and Mumford). Typically in practice, visual comparisons between (and evaluations of) the synthetic and original images are performed. While informative, such subjective, qualitative comparisons can be quite misleading. In this paper, we develop quantitative measures to objectively compare images to improve upon, as well as complement, visual analysis of synthetic images. In information theory, the Kullback-Leibler distance (KLD) has proven to be a useful validation measure for evaluating the similarity (or difference) between original image samples and simulated images. Our proposed quantitative measure for image comparison uses an approximation for the KLD based on Edgeworth expansions (EKLD).

As part of our development, we show that the Edgeworth expansion of the neg-entropy is a useful tool in describing and characterizing image features. Recall that the most difficult problem in image modeling is the "curse of dimensionality." In particular, reliable estimates of probability density functions of images, from a finite number of samples, are hard to obtain in general. It is thus of paramount importance to extract relevant features from the images. The image features can be defined as the expansion coefficients of an image relative to some basis. Therefore, choosing the appropriate basis to the features of the image becomes crucial. Saito (Saito, 1998, 2001) developed an algorithm to find the *least statistically-dependent basis* (LSDB) by quickly selecting a basis from the local basis dictionary. We, in this paper, will use our Edgeworth expansion of neg-entropy to evaluate the differential entropy and choose the LSDB.

Let $\mathbf{X}$ be an m dimensional random vector with density f. The m-dimensional KLD is defined by

$$J(f,g) = \int f(\mathbf{u}) \log \frac{f(\mathbf{u})}{g(\mathbf{u})}\, d\mathbf{u} \tag{1}$$

where g is another m-dimensional density function. We may view the KLD (1) as the expected amount of information in $\mathbf{X}$ with density f for discriminating against g. KLD is thus an appropriate measure of distance in problems of discrimination.

If we let $g = \phi_f$, an m-dimensional multivariate Gaussian distribution of same mean vector and covariance matrix as that of f, then KLD is termed neg-entropy and defined by

$$J(f,\phi_f) = \int f(\mathbf{u}) \log \frac{f(\mathbf{u})}{\phi_f(\mathbf{u})}\, d\mathbf{u}\,. \tag{2}$$

Under regularity conditions, a Gaussian approximation $\hat{f}$ of f may be derived via an Edgeworth expansion (Barndorff-Nielsen and Cox, 1989; Kendall and Stuart, 1977). The neg-entropy in (2) associated with this estimate is the expected neg-entropy $E_{\hat{f}}[\log(\hat{f}/\phi_f)] = J_E(\hat{f},\phi_f)$.

The Kullback-Leibler distance, invariant under any invertible linear transformation, can be built through density estimates of g (Joe, 1989; Hall, 1987; Hall and Morton, 1993). Density estimation, however, relies on the choice of kernel function and window size or bandwidth for each estimator. The computational and conceptual complexity in specifying these parameters limits the applicability of density estimation methods for estimating (1).

We propose an alternative method based on Edgeworth expansions to evaluate the Kullback-Leibler distance. Comon (Comon, 1994) and Jones and Sibson (Jones and Sibson, 1987) approximated the neg-entropy in one dimension by

$$J_E(\hat{f},\phi) = \frac{1}{12}\rho_3^2 + \frac{1}{48}\rho_4^2 + \frac{7}{48}\rho_3^4 - \frac{1}{8}\rho_3^2\rho_4 + o(n^{-2})\,, \tag{3}$$

using an Edgeworth expansion. Here, ρ_r is the rth *standardized* cumulant of the random variable Z, the standardized sum of the random variables $X_1,\ldots,X_n$ with independent and identical distribution (i.e. $Z = (\Sigma X_i - n\mu)/\sqrt{n}$, μ is the mean of X_i), and n is the number of available samples. The relationship between ρ_r and the cumulant κ_r of the random variable Z is

$$\rho_r = \kappa_r/\kappa_2^{r/2}\,.$$

We generalize this method towards an approximation of the neg-entropy for an m-dimensional random vector $\mathbf{X}$. In particular, the analogous Edgeworth expansion for neg-entropy of the standardized random vector $\mathbf{Z}$ is

$$J_E(\hat{f},\phi) = \frac{1}{2}\left\{\left(\frac{1}{3!}\right)^2 J_1 + \left(\frac{1}{4!}\right)^2 J_2 + \left(\frac{1}{72}\right)^2 J_3\right\} + o(n^{-2})\,,$$

where J_1, J_2, and J_3 are functions of the moments of the components of $\mathbf{X}$.

Moreover, we may apply the Edgeworth expansion $\hat{f}$ of f and $\hat{g}$ of g to the KLD (1) and obtain the expected KLD, $E_{\hat{f}}[\log(\hat{f}/\hat{g})]{=}J_E(\hat{f},\hat{g})$. Since the expansion is very complicated, we derive the approximation only up to $o(n^{-1})$:

$$J_E(\hat{f},\hat{g}) = a_1 + a_2 - a_3 - a_4 + O(n^{-\frac{3}{2}})$$

where

$$\begin{aligned}
a_1 &= \frac{1}{12}\frac{\kappa_3^2}{\kappa_2^3}\\
a_2 &= \frac{1}{2}\left[\beta^2 - 2\log\beta - 1 + \alpha^2\right]\\
a_3 &= b_1 + b_2 + b_3\\
a_4 &= \frac{1}{36}\frac{\kappa_3\tilde{\kappa}_3}{\tilde{\kappa}_2^2}\left(\frac{1}{\tilde{\kappa}_2} - \alpha^2 + 9\tilde{\kappa}_2\right)\,,\\
b_1 &= \frac{\tilde{\kappa}_3^2}{6\tilde{\kappa}_2^3}(\beta^3(\alpha^3 + 3\alpha) - 3\beta)\,,
\end{aligned}$$

$$
\begin{aligned}
b_2 &= \frac{\tilde{\kappa}_4}{24\tilde{\kappa}_2^2}(\beta^4\delta - 6\beta^2\gamma + 3)\,, \\
b_3 &= \frac{\tilde{\kappa}_3^2\tilde{\kappa}_2^3}{72\tilde{\kappa}_2^3}(\beta^6\eta - 15\beta^4\delta + 45\beta^2\gamma - 15)\,.
\end{aligned}
$$

In practice, we may use the dominant terms and push all the other terms into the error term with $O(n^{-1})$:

$$
J_E(\hat{f}, \hat{g}) \;=\; J(\phi_f, \phi_g) + O(n^{-1})
$$

with

$$
J(\phi_f, \phi_g) = \frac{1}{2}\left[\beta^2 - 2\log\beta - 1 + \alpha^2\right].
$$

Moreover, we may use the distribution of the sample mean instead of the sum of the random variables $X_1, \ldots, X_n$.

The Edgeworth approximation of KLD in m dimensions is even more complicated than one dimension. Nonetheless, we show that the dominant term is the KLD of two Gaussian distributions ϕ_f and ϕ_g. The Edgeworth approximation of the KLD up to $O(n^{-1})$ is thus analogous to that in one dimension, in that $J_E(\hat{f}, \hat{g}) \;=\; J(\phi_f, \phi_g) + O(n^{-1})$.

This paper elucidates two facts. First, the convergence rate of the corresponding Kullback-Leibler distance based on the Edgeworth expansion is $o(n^{-1})$. On the other hand, the alternative density estimation approach to computing the Kullback-Leibler distance can provide only root-n consistent estimators (Hall and Morton, 1993). Furthermore, the error rate of the histogram estimator not only depends on sample size n, but also on the choice of 'binwidth' value h (Hall, 1987). The total error is, roughly, $O(h^2) + o(n^{-1/2})$. In the case of kernel estimation, the error is $o(n^{-1/2})$ when the dimension is less than (or equal to) 3; the estimator is much less sensitive to choices of the bandwidth h compared to the associated histogram estimator.

Second, the Kullback-Leibler information based on the Edgeworth expansion can be evaluated for any dimensional distribution as compared to density estimation (both histogram and kernel estimator) which can be performed only on low-dimensional distributions (1, 2, and 3 dimensions) in practice.

The paper is organized as follows. Section 2 derives the Edgeworth expansion of the Kullback-Leibler distance in both one and m dimensions. Section 3 discusses estimation of the Kullback-Leibler distance via sample cumulants. In Section 4, we verify numerically that the neg-entropy of the sample mean distribution decreases as the sample size increases and show numerically that discrimination based on the EKLD of arbitrary distributions f and g can be replaced by the EKLD of the sample mean distributions $\bar{f}$ and $\bar{g}$. We also evaluate the differential entropy of high dimensional densities using the Edgeworth expansion. In Section 5, we apply our methods to two image analysis problems:

1) choosing the LSDB from a local basis dictionary and 2) evaluating three methods, INGA (Lin; et. al., 2001), PCA (Watanabe, 1965), and ICA (Jutten and Herault, 1991), for synthesizing/simulating large dimensional images.

2. Kullback-Leibler Distance

The Kullback-Leibler distance (KLD) measure $J(f,g)$, also called relative entropy or cross-entropy, is a measure of the 'distance' between two distributions f and g and is defined by

$$J(f,g) = \int f(\mathbf{u}) \log \frac{f(\mathbf{u})}{g(\mathbf{u})}\, d\mathbf{u}\,. \tag{4}$$

The KLD in (4) can be viewed as the expected amount of information in $\mathbf{X}$ (single m dimensional observation from the distribution with density f) for discriminating against g. If f and g do not share the same first and second moments, under regularity conditions, we may apply the Edgeworth expansion $\hat{f}$ of f and $\hat{g}$ of g to the KLD in (4) and obtain the expected KLD, $E_{\hat{f}}[\log \hat{f}/\hat{g}]{=}J_E(\hat{f},\hat{g})$.

In this section we will derive Edgeworth expansions of the Kullback-Leibler distance (EKLD) useful in image analysis problems presented in later sections. Our development will detail the expansion in the one dimensional case and then present the general m-dimensional expansion as a generalization or extension of the derivations in the one dimensional situation. We conclude the section with a practical implementation of the Edgeworth expansion for the neg-entropy, that is, the KLD with g being an m-dimensional multivariate Gaussian distribution. Throughout the section, we will use the covariant and contravariant system (indexing random variables by lower and upper indices) to denote operations in high dimensional spaces (McCullagh, 1987). See Appendix A for the definition of the covariant-contravariant system and the corresponding properties of cumulants and Appendix B for properties of covariant-contravariant Hermite polynomials.

We first detail the construction of Edgeworth approximations of the KLD (4). Let $\mathbf{X}_1,\ldots,\mathbf{X}_n$ and $\tilde{\mathbf{X}}_1,\ldots,\tilde{\mathbf{X}}_n$ be independent and identically distributed (iid) m-dimensional random vectors. Denote the components of each random vector by $\mathbf{X}_i = (X_i^1,\ldots,X_i^m)$ and $\tilde{\mathbf{X}}_i = (\tilde{X}_i^1,\ldots,\tilde{X}_i^m)$, with means $\boldsymbol{\mu} = (\mu^1,\ldots,\mu^m)$ and $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}^1,\ldots,\tilde{\mu}^m)$ and moments

$$\kappa^{i_1\ldots i_v} = E(X^{i_1}\ldots X^{i_v})\,,$$

$$\tilde{\kappa}^{i_1\ldots i_v} = E(\tilde{X}^{i_1}\ldots \tilde{X}^{i_v})\,,$$

respectively, where $1 \le i_k \le m$. Let $\mathbf{S}_n = \sum_{i=1}^n \mathbf{X}_i$, $\tilde{\mathbf{S}}_n = \sum_{i=1}^n \tilde{\mathbf{X}}_i$, $\mathbf{Z} = (\mathbf{S}_n - n\boldsymbol{\mu})/\sqrt{n}$, and $\tilde{\mathbf{Z}} = (\tilde{\mathbf{S}}_n - n\tilde{\boldsymbol{\mu}})/\sqrt{n}$ such that the cumulants $\kappa^{i_1,\ldots,i_v}$ and $\tilde{\kappa}^{i_1,\ldots,i_v}$ of $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ are of the order $n^{1-\frac{v}{2}}$. Then the Edgeworth expansion of $f_{\mathbf{Z}}$ and $g_{\tilde{\mathbf{Z}}}$, the distributions of $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ respectively, up to order five about its best normal approximate (Barndorff-Nielsen and Cox, 1989; Kendall and Stuart, 1977) are given by

$$\begin{aligned}\hat{f}_{\mathbf{Z}}(\mathbf{z};\kappa) &= \phi_f(\mathbf{z};\kappa)\ [1+v(\mathbf{z};\kappa)]\ +o(n^{-1})\\ \hat{g}_{\tilde{\mathbf{Z}}}(\mathbf{z};\tilde{\kappa}) &= \phi_g(\tilde{\mathbf{z}};\tilde{\kappa})\ [1+u(\tilde{\mathbf{z}};\tilde{\kappa})]\ +o(n^{-1})\end{aligned}$$

where

$$\begin{aligned}\phi_f(\mathbf{z};\kappa) &= (2\pi)^{-m/2}\{\det(\kappa)\}^{-1/2}\exp(-0.5\kappa_{i,j}z^iz^j)\\ \phi_g(\tilde{\mathbf{z}};\tilde{\kappa}) &= (2\pi)^{-m/2}\{\det(\tilde{\kappa})\}^{-1/2}\exp(-0.5\tilde{\kappa}_{i,j}\tilde{z}^i\tilde{z}^j)\end{aligned}$$

denote m-dimensional multivariate normal distributions with zero mean and covariance matrices $\kappa = [\kappa^{i,j}]$ and $\tilde{\kappa} = [\tilde{\kappa}^{i,j}]$ respectively, with $\kappa^{i,j} = E(Z^iZ^j)$, $\tilde{\kappa}^{i,j} = E(\tilde{Z}^i\tilde{Z}^j)$, $[\kappa_{i,j}]$ representing κ^{-1}, and $[\tilde{\kappa}_{i,j}]$ representing $\tilde{\kappa}^{-1}$. Here $v(\mathbf{z};\kappa) = v_1(\mathbf{z};\kappa) + v_2(\mathbf{z};\kappa) + v_3(\mathbf{z};\kappa)$, $u(\tilde{\mathbf{z}};\tilde{\kappa}) = u_1(\tilde{\mathbf{z}};\tilde{\kappa}) + u_2(\tilde{\mathbf{z}};\tilde{\kappa}) + u_3(\tilde{\mathbf{z}};\tilde{\kappa})$, and $v_i(\mathbf{z};\kappa)$ and $u_i(\tilde{\mathbf{z}};\tilde{\kappa})$, $i = 1,2,3$, are the corresponding terms in the sum $\kappa^{i,j,k}h_{ijk}(\mathbf{z})$, $\kappa^{i,j,k,l}h_{ijkl}(\mathbf{z})$, and $\kappa^{i,j,k}\kappa^{l,p,q}h_{ijklpq}(\mathbf{z})$, and $\tilde{\kappa}^{i,j,k}h_{ijk}(\tilde{\mathbf{z}})$, $\tilde{\kappa}^{i,j,k,l}h_{ijkl}(\tilde{\mathbf{z}})$, $\tilde{\kappa}^{i,j,k}\tilde{\kappa}^{l,p,q}h_{ijklpq}(\tilde{\mathbf{z}})$ respectively. Note that there are m^2 terms in $v_1(\mathbf{z};\kappa)$ and $u_1(\tilde{\mathbf{z}};\tilde{\kappa})$, m^3 terms in $v_2(\mathbf{z};\kappa)$ and $u_2(\tilde{\mathbf{z}};\tilde{\kappa})$, and m^6 terms in $v_3(\mathbf{z};\kappa)$ and $u_3(\tilde{\mathbf{z}};\tilde{\kappa})$.

In the case of one dimension, we use the Edgeworth expansion of f_Z and $g_{\tilde{Z}}$ up to order five about its best normal approximates given by (Barndorff-Nielsen and Cox, 1989):

$$\begin{aligned}\hat{f}_Z(z) &= \phi_f(z)(1+v(z))\ +o(n^{-1})\\ \hat{g}_{\tilde{Z}}(\tilde{z}) &= \phi_g(\tilde{z})(1+u(\tilde{z}))\ +o(n^{-1})\,,\end{aligned} \tag{5}$$

where $\phi_f(z)$ and $\phi_g(\tilde{z})$ denote normal distributions with zero means and variances $\kappa^{i,i} = E(Z^iZ^i)$ and $\tilde{\kappa}^{i,i} = E(\tilde{Z}^i\tilde{Z}^i)$ respectively. Here

$$v(z) = \frac{1}{3!}\rho_3H_3(z) + \frac{1}{4!}\rho_4H_4(z) + \frac{10}{6!}{\rho_3}^2H_6(z)\,,$$

$$u(\tilde{z}) = \frac{1}{3!}\tilde{\rho}_3H_3(\tilde{z}) + \frac{1}{4!}\tilde{\rho}_4H_4(\tilde{z}) + \frac{10}{6!}{\tilde{\rho}_3}^2H_6(\tilde{z})\,.$$

Substituting the Edgeworth expansions (5) into the KLD (4) and using the equality

$$\frac{\hat{f}}{\hat{g}} = \frac{\hat{f}}{\phi_f}\frac{\phi_f}{\phi_g}\frac{\phi_g}{\hat{g}}$$

we obtain the following expansion

$$\begin{aligned}J_E(\hat{f},\hat{g}) &= \int \hat{f}(y)\log\frac{\hat{f}(y)}{\hat{g}(y)}\,dy\\ &= \int \hat{f}(y)\log\frac{\hat{f}(y)}{\phi_f(y)}\,dy + \int \hat{f}(y)\log\frac{\phi_f(y)}{\phi_g(y)}\,dy + \int \hat{f}(y)\log\frac{\phi_g(y)}{\hat{g}(y)}\,dy.\end{aligned} \tag{6}$$

The first term of (6) is the expected neg-entropy $J_E(\hat{f}, \phi_f)$. The second term, after substituting $\hat{f}$ and $\hat{g}$ in (5), the expression becomes

$$\int \hat{f}(y) \log \frac{\phi_f(y)}{\phi_g(y)}\, dy = \int \phi_f(y) \log \frac{\phi_f(y)}{\phi_g(y)}\, dy + \int \phi_f(y) v(y) \log \frac{\phi_f(y)}{\phi_g(y)}\, dy. \tag{7}$$

The first term of (7), denoted as $J(\phi_f, \phi_g)$, is the KLD of two Gaussian distributions, and the second term, using the properties of the Hermite polynomials (Appendix B), is zero.

Finally, after substituting $\hat{f}$ and $\hat{g}$ in (5), the third term of (6) becomes

$$\int \hat{f}(y) \log \frac{\phi_g(y)}{\hat{g}(y)}\, dy = -\int \phi_f u(y)\, dy - \int \phi_f v(y) u(y)\, dy$$

where

$$\begin{aligned} \int \phi_f u(y)\, dy &= b_1 + b_2 + b_3\,, \\ \int \phi_f v(y) u(y)\, dy &= \frac{1}{36} \frac{\kappa_3 \tilde{\kappa}_3}{\tilde{\kappa}_2^2} \left(\frac{1}{\tilde{\kappa}_2} - \alpha^2 + 9\tilde{\kappa}_2 \right), \end{aligned}$$

with

$$\begin{aligned} b_1 &= \frac{\tilde{\kappa}_3}{6\tilde{\kappa}_2^{\frac{3}{2}}} (\beta^3(\alpha^3 + 3\alpha) - 3\beta)\,, \\ b_2 &= \frac{\tilde{\kappa}_4}{24\tilde{\kappa}_2^2} (\beta^4\delta - 6\beta^2\gamma + 3)\,, \\ b_3 &= \frac{\tilde{\kappa}_3^2 \tilde{\kappa}_2^3}{72\tilde{\kappa}_2^3} (\beta^6\eta - 15\beta^4\delta + 45\beta^2\gamma - 15). \end{aligned}$$

Here, $\alpha = \tilde{\kappa}_2^{-\frac{1}{2}}(\kappa_1 - \tilde{\kappa}_1)$, $\beta = (\kappa_2 \tilde{\kappa}_2^{-1})^{\frac{1}{2}}$, $\gamma = \alpha^2 + 1$, $\delta = \alpha^4 + 6\alpha^2 + 3$, and $\eta = \alpha^6 + 15\alpha^4 + 45\alpha^2 + 15$.

Let $a_1 = J(\hat{f}, \phi_f)$, $a_2 = J(\phi_f, \phi_g)$, $a_3 = \int \phi_f(y) u(y)\, dy$, and $a_4 = \int \phi_f(y) v(y) u(y)\, dy$. Then the one-dimensional Edgeworth KLD approximation is

$$J_E(\hat{f}, \hat{g}) = a_1 + a_2 - a_3 - a_4 + o(n^{-1})$$

where

$$\begin{aligned} a_1 &= \frac{1}{12} \frac{\kappa_3^2}{\kappa_2^3} \\ a_2 &= \frac{1}{2} \left[\beta^2 - 2\log\beta - 1 + \alpha^2 \right] \\ a_3 &= b_1 + b_2 + b_3 \\ a_4 &= \frac{1}{36} \frac{\kappa_3 \tilde{\kappa}_3}{\tilde{\kappa}_2^2} \left(\frac{1}{\tilde{\kappa}_2} - \alpha^2 + 9\tilde{\kappa}_2 \right), \\ b_1 &= \frac{\tilde{\kappa}_3^2}{6\tilde{\kappa}_2^3} (\beta^3(\alpha^3 + 3\alpha) - 3\beta)\,, \\ b_2 &= \frac{\tilde{\kappa}_4}{24\tilde{\kappa}_2^2} (\beta^4\delta - 6\beta^2\gamma + 3)\,, \\ b_3 &= \frac{\tilde{\kappa}_3^2 \tilde{\kappa}_2^3}{72\tilde{\kappa}_2^3} (\beta^6\eta - 15\beta^4\delta + 45\beta^2\gamma - 15)\,. \end{aligned}$$

In practice, we may use the dominant terms and push all the other terms into the error term with $O(n^{-1})$:

$$J_E(\hat{f},\hat{g}) \;=\; J(\phi_f,\phi_g) + O(n^{-1})$$

with

$$J(\phi_f,\phi_g) = \frac{1}{2}\left[\beta^2 - 2\log\beta - 1 + \alpha^2\right].$$

In high dimensions, the Edgeworth approximation of the KLD up to $O(n^{-1})$ is analogous to that in one dimension

$$J_E(\hat{f},\hat{g}) \;=\; J(\phi_f,\phi_g) + O(n^{-1}) \tag{8}$$

where

$$\begin{aligned}
J(\phi_f,\phi_g) &= \frac{1}{2}(a+b+c-1)\,,\\
a &= \log\frac{\det(\tilde{\kappa})}{\det(\kappa)}\,,\\
b &= \sum_i \frac{(\kappa^{i,i})^2 + (\kappa^i - \tilde{\kappa}^i)^2}{(\tilde{\kappa}^{i,i})^2}\,,\\
c &= \sum_{i\neq j} \frac{2\tilde{\kappa}^{i,j}}{\tilde{\kappa}^{i,i}\tilde{\kappa}^{j,j}}\{\kappa^{i,i}\kappa^{i,j} + (\kappa^i - \tilde{\kappa}^i)(\kappa^j - \tilde{\kappa}^j)\}\,.
\end{aligned}$$

In particular, if $g = \phi_f$ in (4) is the m-dimensional multivariate Gaussian distribution of same mean vector and covariance matrix as those of f, then KLD is termed neg-entropy and defined by

$$J(f,\phi_f) = \int f(\mathbf{u})\log\frac{f(\mathbf{u})}{\phi_f(\mathbf{u})}\,d\mathbf{u}\,. \tag{9}$$

The corresponding Edgeworth neg-entropy, similar to Edgeworth KLD in (8), can be shown up to $o(n^{-2})$ as

$$J_E(\hat{f},\phi) = \frac{1}{2}\left\{\left(\frac{1}{3!}\right)^2 J_1 + \left(\frac{1}{4!}\right)^2 J_2 + \left(\frac{1}{72}\right)^2 J_3\right\} + o(n^{-2})\,,$$

where

$$\begin{aligned}
J_1 &= (\kappa^{i,j,s})^2\kappa_{i,\pi3}\kappa_{j,\pi3}\kappa_{s,\pi3}[3!],\\
J_2 &= (\kappa^{i,j,s,l})^2\kappa_{i,\pi4}\kappa_{j,\pi4}\kappa_{s,\pi4}\kappa_{l,\pi4}[4!],\\
J_3 &= (\kappa^{i,j,s})^2(\kappa^{l,m,n})^2\kappa_{i,\pi6}\kappa_{j,\pi6}\kappa_{s,\pi6}\kappa_{l,\pi6}\kappa_{m,\pi6}\kappa_{n,\pi6}[6!],
\end{aligned}$$

and $\pi3$, $\pi4$, and $\pi6$ represent the permutations of (i,j,s), (i,j,s,l) and (i,j,s,l,m,n) respectively. Here, J_1, J_2, and J_3 denote the tensor notation over the index (i,j,s), (i,j,s,l), and (i,j,s,l,m,n).

3. Sample Cumulants

The Edgeworth expansion of the Kullback-Leibler distance $J(f,g)$ involves the third order cumulants $\kappa^{i,j,s}$ and $\tilde{\kappa}^{i,j,s}$ of the random vectors $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$, corresponding to f and g respectively, where

$$\begin{aligned}\kappa^{i,j,s} &= E(Z^i-\mu^i)(Z^j-\mu^j)(Z^s-\mu^s)\\ \tilde{\kappa}^{i,j,s} &= E(\tilde{Z}^i-\tilde{\mu}^i)(\tilde{Z}^j-\tilde{\mu}^j)(\tilde{Z}^s-\tilde{\mu}^s)\,.\end{aligned}$$

In the case of one dimension, the third order standardized cumulants ρ_3 and $\tilde{\rho}_3$ are needed. To apply all the approximations in Section 2, we need to estimate these third order cumulants.

The sample cumulants, the so-called k-statistics, are unbiased estimates of the cumulants. For each cumulant of $\mathbf{X}_i$, κ, with appropriate superscripts, there is a unique polynomial symmetric function, denoted by k with matching superscripts, such that k is an unbiased estimate of κ. For example,

$$\begin{aligned}k^r &= n^{-1}\sum_{i=1}^{n} x_i^r\,,\\ k^{r,t} &= \frac{1}{n}\phi^{ij}x_i^r x_j^t,\\ k^{r,t,u} &= \frac{1}{n}\phi^{ijs}x_i^r x_j^t x_s^u,\end{aligned} \tag{10}$$

where x_i, x_j are samples of the process $\mathbf{X_i}, \mathbf{X_j}$ and

$$\phi^{ij} = \begin{cases} 1\,, & \text{if } i=j\\ -\frac{1}{n-1}\,, & i\neq j\ ,\end{cases}$$

$$\phi^{ijs} = \begin{cases} 1\,, & \text{if } i=j=s\\ -\frac{1}{n-1}\,, & i=j\neq s\\ \frac{2}{(n-1)(n-2)}\,, & i\neq j\neq s\end{cases}$$

ensure the estimators are unbiased (McCullagh, 1987).

Another way to calculate the sample cumulants is to use the sample moments $k^i = \frac{1}{n}\sum_{r=1}^n x_r^i$, $k^{ij} = \frac{1}{n}\sum_{r=1}^n x_r^i x_r^j$, and $k^{ijs} = \frac{1}{n}\sum_{r=1}^n x_r^i x_r^j x_r^s$, and the relationship between cumulants and moments from Appendix A. Then the third order cumulant can be expressed in terms of moments as

$$k^{i,j,s} = \frac{n^2}{(n-1)(n-2)}\left[k^{ijs} - k^i k^{js} - k^j k^{is} - k^s k^{ij} + 2\,k^i k^j k^s\right].$$

In this paper, we use the sample cumulants defined in (10). In the two dimensional case, there are four terms: $k^{1,1,1}$, $k^{2,2,2}$, $k^{1,1,2}$, and $k^{1,2,2}$. In the general case of m dimensions, there are m terms of $k^{i,i,i}$, $m(m-1)$ terms of $k^{i,i,j}$, and $\frac{m}{6}(m-1)(m-2)$ terms of $k^{i,j,s}$. For applications of k-statistics in detecting departures from the usual linear model assumption (see Anscombe, 1961; Bickel, 1978; Hinkley, 1985; McCullagh and Pregibon, 1987; Brillinger, 1994).

4. Numerical Examples

The Edgeworth approximations of Section 2 require that the distributions f and g of interest are "not far from the Gaussian distribution" (Hall, 1987). Though the approximations are relatively robust to the Gaussian constraint, we propose application of these approximations in image analysis to the sample mean distributions with respect to f and g which, by the central limit theorem, will be normally distributed up to order $n^{1/2}$. For example, in discriminating between images drawn from the distributions f and g, we compute the EKLD between the sampling distributions of the sample mean image from each of these image distributions. Given the relatively large sample sizes in our applications, the use of the sampling distributions overcomes any sensitivity of our procedures to extreme violations of the Gaussian assumption by the distributions f and g under study.

The central limit theorem guarantees that our approximations for the sampling distributions are valid, however reliance on the sampling distributions of images, as opposed to the true underlying stochastic process, may bias our image evaluation procedures. In this section, we numerically study the sensitivity of our proposed method for computing neg-entropy and differential entropy and discriminating between distributions using the EKLD. In Section 4.1, we verify numerically that the neg-entropy of the sample mean distribution decreases as the sample size increases. In Section 4.2, we show numerically that discrimination based on the EKLD of arbitrary distributions f and g can be replaced by the EKLD of the sample mean distribution denoted by $\bar{f}$ and $\bar{g}$. We conclude our numerical studies in Section 4.3 with a comparison of our EKLD approximations of differential entropy with the commonly used density estimation approach.

4.1. The neg-entropy of the sample mean distribution

In order to investigate the neg-entropy of the sample mean distribution, we generate data sets of sample size $20, 25, \ldots, 100$, from the exponential distribution with means 1, 0.1, and 0.01 and uniform distribution with interval $(0, 1)$, $(0, 10)$, and $(0, 100)$. Figure 1 shows the six neg-entropies of the sample mean generated from the distributions $\exp(1)$, $\exp(10)$, $\exp(100)$, $U(0, 1)$, $U(0, 10)$, and $U(0, 100)$. It is clear that the larger the sample size, the less the neg-entropy. This shows that the larger the sample size, the closer the distribution is to that of a Gaussian distribution, despite the non-Gaussian distributions from which the data is generated.

This study indicates that conclusions based on EKLD of the sampling distributions is the same as that of the KLD. We find this result to hold in general, particularly in image applications such as those presented in Section 5, though for brevity we do not present more illustrations here. We also note that this procedure is analogous to the classical hypothesis testing routines of basic inference on sampling distributions. Nonetheless, more theoretical examination of the sensitivity of our image

Table 1. Four discrimination cases and the corresponding Kullback-Leibler Distance

case	f	g_1	g_2	condition
(1)	Uniform(0, 1)	Uniform(0, a)	Uniform(0, b)	$1 < a < b$
(2)	exponential(1)	exponential(a)	exponential(b)	$a < b$
(3)	Uniform(0, 1)	exponential(a)	exponential(b)	$a < b$
(4)	exponential(1)	Uniform(0, a)	Uniform(0, b)	$a < b$

Table 2. Four theoretical formula of Kullback-Leibler Distances

case	KLD	condition	KLD theoretical value
(1)	$J(U(0,a), U(0,b))$	$a < b$	$\log \frac{b}{a}$
(2)	$J(\exp(a), \exp(b))$	$a < b$	$\log \frac{b}{a} + \frac{a}{b} - 1$
(3)	$J(U(0,a), \exp(b))$	$a < b$	$\log \frac{b}{a} + \frac{1}{2}\frac{a}{b}$
(4)	$J(\exp(a), U(0,b))$	$a < b$	$(1 - e^{-b/a})[\log \frac{b}{a} - 1] + (\frac{b}{a}e^{-\frac{b}{a}} - 1)$

analysis methods to EKLD comparisons through image sampling distributions is an item of future research.

4.2. The discrimination based on the EKLD

Given f and g, two density functions, the KLD $J(f, g)$ represents a measure of distance between them. Numerically, given samples drawn from f and g, the EKLD $J_E(\hat{f}, \hat{g})$ of the sample mean distribution is an approximation of $J(f, g)$. To demonstrate the discrimination of distributions via EKLD, we consider the following problem. If the KLD between the sample mean distribution of $f(x)$ and that of $g_1(x)$ is smaller than the corresponding KLD between $f(x)$ and $g_2(x)$, can we conclude that the true KLD between $f(x)$ and $g_1(x)$ is smaller than that between $f(x)$ and $g_2(x)$? Let us consider the following examples to investigate this idea.

We wish to discriminate between distributions g_1 and g_2 with respect to f in the four cases in Table 4.2. Table 4.2 presents the theoretical KLD values in each case. Suppose f is "closer" to g_1 than to g_2 in terms of the KLD. We will study whether EKLD can correctly discriminate between g_1 and g_2.

To be concrete, we consider the distributions listed in Table 4.2. Table 4.2 also presents the KLD values for each of the corresponding comparisons of f with g_1 and f with g_2. EKLD evaluates the distance between the sample mean distributions of the distributions of interest. Figure 2 presents the EKLD comparisons for each of the four cases in Table 4.2. It is clear that, from Figure 2, the larger sample size is, the KLD of the sample mean from the "large" distance distribution is larger than the one from the "small" distance distribution with some minor exception points. That is,

Table 3. Four theoretical Kullback-Leibler Distances

case	$J(f,\ g_1)$	$J(f,\ g_2)$
(1)	$J(U(0,1), U(0,10)) = 2.303$	$J(U(0,1), U(0,100)) = 4.605$
(2)	$J(\exp(1), \exp(10)) = 1.403$	$J(\exp(1), \exp(100)) = 4.596$
(3)	$J(U(0,1), \exp(10)) = 2.353$	$J(U(0,1), \exp(100)) = 3.606$
(4)	$J(\exp(1), U(0,10)) = 0.303$	$J(\exp(1), U(0,100)) = 2.605$

EKLD can be used well to discriminate between g_1 and g_2.

4.3. Calculation of the differential entropy

In this section, we illustrate computation of the differential entropy using our Edgeworth approximation. We use this application to compare our approach to the estimated entropy of (Hall and Morton, 1993) based on density estimation. Note that the differential entropy

$$S(f) = -\int f(\mathbf{u}) \log f(\mathbf{u})\, d\mathbf{u}$$

may be written in terms of the neg-entropy

$$S(f) = S(\phi_f) - J(f, \phi_f)\,. \tag{11}$$

Differential entropy calculations via density estimation are computationally slow due to the choice of 'bandwidth' and kernel functions (Joe, 1989; Hall, 1987). Furthermore, density estimation is not applicable to evaluate differential entropy for problems in dimensions greater than three. Differential entropy computations via the Edgeworth expansion of the neg-entropy do not suffer from these shortfalls. Furthermore, the order of Edgeworth approximation is $O(n^{-3/2})$, while the density estimation approximation is of order $O(n^{-1/2})$. This difference in order explains the difference in absolute error between the two techniques. In the remainder of this section, we highlight these difference through a number of numerical studies.

Table 5.2 presents $S(f)$ when $f = \phi_Z$, the standard normal distribution in one dimension (the theoretical value is 1.42) using EKLD and density estimation for samples of size $n = 100$, 200, 300, 400, and 500. Note that even in this simple one dimensional example, the approximate value of $S(\phi)$ by EKLD is more accurate than that by density estimation.

Tables 5.2 - 5.2 illustrate situations in which density estimation can not be used to evaluate differential entropy, but the Edgeworth expanded neg-entropy is not only feasible, but produces excellent approximations. Table 5.2 presents $S(f)$ when f is a bivariate Gaussian distribution with three different dispersions. Table 5.2 displays $S(f)$ when f is a three-dimensional Gaussian distributions with two different dispersions. Table 5.2 shows the 4-dimensional, 5-dimensional, and

8-dimensional numerical results of $S(f)$ when $f = \phi_{\mathbf{Z}}$, $\mathbf{Z}$ denotes a standard normal random vector. Here, in Table 5.2 - 5.2, we use the Edgeworth expansion with order $O(n^{-3/2})$.

5. Image Analysis

5.1. LSDB from the local basis dictionaries

Recent advances in imaging technology produce a large quantity of images over almost a continuous spatial spectrum as well as resolution. Image modeling is essential for the description and characterization of image features, large scale computations using images, and image compression. The most difficult problem in image modeling is the 'curse of dimensionality'. In particular, reliable estimates of probability density functions of high dimensional data, such as images, from a finite number of samples are hard to obtain in general. It is thus of paramount importance to extract relevant features from the images, reduce the dimensionality of the problem, and simplify the model by assuming statistical relationship among these features.

Image features are defined as the expansion coefficients of an image relative to some basis. The Karhunen-Loève Basis provides a decorrelated coordinate system. Saito (Saito, 1994) developed and considered a local basis library to extract features from images for classification and regression. The basis library consists of a collection of *local basis dictionaries* such as wavelet packets, local cosine/sine bases, or local Fourier bases. Each dictionary consists of a redundant number of the basis vectors with the specific characters in scale, position, and frequency. These basis vectors are organized as a quadtree in a hierarchical manner ranging from very localized spikes to global oscillations with different frequencies.

Image modeling techniques using the feature extractors have been proposed by various group of scientists. Saito (Saito, 1998, 2001) developed an algorithm to find the *least statistically-dependent basis* (LSDB) by quickly selecting from the local basis library a basis that is "closest" to the statistical independence in the sense of relative entropy. He used the differential entropy $S(f_{X_i})$ of each coordinate estimated by the method of density estimation as the selection criterion of LSDB:

$$B_{\mathrm{LSDB}} = \arg\min_{B \in \mathcal{D}} \sum_{i=1}^{n} S(f_{X_i}) \,,$$

where $\mathcal{D}$ is a basis dictionary. Based on the relationship of differential entropy $S(f)$ and neg-entropy $J(f, \phi)$

$$J(f, \phi) = S(\phi) - S(f) \,,$$

we can rewrite the selection criterion as the form

$$B_{\mathrm{LSDB}} = \arg\min_{B \in \mathcal{D}} \sum_{i=1}^{n} \left(S(\phi_{X_i}) - J(f_{X_i}, \phi_{X_i}) \right)$$

where neg-entropy $J(f_{X_i}, \phi_{X_i})$ can be estimated by the method of Edgeworth expansion.

To demonstrate the comparison between the LSDB selected by the method of density estimation and Edgeworth expansion, we use the data set of face images, 'Rogues Gallery Problem'. This dataset consists of digitized pictures of faces of 143 people, provided by Prof. L. Sirovich of Mount Sinai School of Medicine via Prof. M. V. Wickerhauser of Washington University. We randomly selected 72 faces to be the training dataset. Each image is of dimension 128×128. Figure 3 (a) is the average face of the training set. We want to choose the LSDB partition pattern (segments) from the local dictionary to investigate features of the face: profile, eyebrow, eye, nose, and mouth. We will use both the methods of density estimation and Edgeworth expansion to choose the LSDB of the face from the local cosine dictionary.

Figure 3 (b)–(d) show the partition patterns of LSDB selected from the local cosine dictionary by using density estimation and Edgeworth expansion with order $O(n^{-3/2})$ and $O(n^{-2})$. Here, there are 103 LSDB segments generated by the method of density estimation (Figure 3(b)); 142 LSDB segments are chosen by the method of Edgeworth expansion up to the order 1.5 ((Figure 3(c)); and 190 LSDB segments are chosen by the method of Edgeworth expansion up to the order 2 (Figure 3(d)). We observe that as the order of the Edgeworth expansion increases, the LSDB tries to split the image into finer segments. In particular, the LSDB segments in Figure 3(d) using the Edgeworth expansion up to the order 2, catches finer features around the eye area than Figure 3(b), which was selected by density estimation. Furthermore, the LSDB chosen by the Edgeworth expansion up to order 1.5 describes the same features as that by density estimation. On the other hand, the LSDB chosen by the Edgeworth expansion up to order 2 describes the facial features exactly, such as the eyebrow, which is not characterized by either density estimation nor the Edgeworth expansion with order 1.5. Therefore, we may conclude that the method of Edgeworth expansion up to $O(n^{-2})$ provides the best LSDB describing more facial features than the others.

5.2. Image Synthesis Validation via Edgeworth KLD

Given n samples of a dependent random vector $\mathbf{X}$ (image) of p dimensions ($n > p$), a synthesis can be obtained by several simulation methods. It is then of critical importance to know the accuracy of the methods in a quantitative manner. That is, how to judge the closeness of these syntheses to the original image?

In this section, we will show how to use the EKLD to compare three methods for image synthesis obtained by the method of PCA, ICA and INGA. The Principal Components Analysis (PCA), under a strict assumption of normality, transforms the sample set orthogonally to independent Gaussian distributed random variables. The Independent Component Analysis (ICA), under the assumption that the sample set is a linear mixture of the independent source, is a linear process which tries

to transform the sample set to independent components. The Iterative Nonlinear Gaussianization Algorithm (INGA) — an extension to PCA. While PCA merely transforms a set of correlated random variables into a set of uncorrelated random variables, INGA nonlinearly transforms them to the standard multivariate Gaussian variables in an attempt to minimize the statistical dependence among the transformed coordinates, at a similar computational cost to PCA. The difference between INGA and ICA lies in two aspects, although both seek statistically-independent coordinate systems. First, INGA seeks a *nonlinear* transform whereas ICA seeks a linear one. Second, the motivation of INGA is really *resampling* and *simulation* rather than blind source separation and blind deconvolution. Each of these simulation methods consist of forward and backward processes. See (Lin; et. al., 2001) for more details of these three processes in the context of image synthesis.

To validate these synthesis procedures, we compare the EKLD of the sample mean distribution for each simulation method. The fact is that the smaller EKLD is, the closer (or more similar) the simulated samples are to the originals.

Our model validation procedure may be summarized as follows.

(a) `Simulate` the stochastic process of interest with INGA and generate 100 datasets each of which contains 100 simulated samples.

(b) `Compute` the EKLD in (8) between the original distribution and the simulated distribution using the sample means and the sample covariance matrices computed from the original samples and a simulated dataset containing 100 simulated samples. Perform this EKLD computation for each dataset. This results in 100 EKLDs.

(c) `Display` the distribution of these 100 EKLDs by its 90% confidence interval (c.i.) . The smaller EKLD is, the closer (or more similar) the simulated samples are to the originals.

We illustrate our methods on the so-called cigar, spike, and eye data sets (for details see Lin; et. al., 2001) and compare the syntheses produced by INGA, PCA and ICA with Monte Carlo estimates of the EKLD sampling distribution.

(a) Cigar data

Figure 4 shows the original cigar sample and syntheses obtained by INGA, PCA, and ICA. By visual inspection, we note that the INGA synthesis is better than the others, but we desire a quantitative comparison of the three syntheses. Table 5.2, row 1 displays 90% c.i. of the EKLD for the INGA, PCA, and ICA syntheses. We note that INGA outperforms PCA and ICA on average, though the difference between the three algorithms is not statistically significant.

(b) Spike process data

Figure 5 shows the original two-dimensional spike data set and syntheses obtained by INGA, PCA, and ICA. By visual inspection, we note that the INGA synthesis is better than that of

PCA and ICA. Table 5.2, row 2 shows 90% c.i. of the EKLD for the INGA, PCA, and ICA syntheses. Again, on average, INGA is superior to PCA and ICA though the difference is not statistically significant.

(c) Eye data

As a higher dimensional comparison of the three synthesis algorithms, we use 25-dimensional extracted eye images. Figure 6 shows the original data and syntheses by INGA, PCA, and ICA. It is very difficult to visually compare the syntheses here. A quantitative comparison through the EKLD is thus crucial towards evaluating these synthesis procedures. Table 5.2, row 3 shows 90% c.i. of the EKLD for the INGA, PCA, and ICA syntheses. We may conclude that the INGA synthesis is significantly superior to that of PCA and ICA. Furthermore, ICA does not significantly improve upon PCA.

Acknowledgment

This work was partially supported by grants from NSF-EPA DMS-99-78321 (JJL, NS, RAL), NSF DMS-99-73032 (NS), and ONR YIP N00014-00-1-046 (NS).

Appendix A: Covariant and Contravariant System

To define the covariant and contravariant system more precisely, we start with a vector $\mathbf{x}$ with m components $x^1, x^2, \ldots, x^m$. We define u as a d-dimensional array whose elements are functions of the components of $\mathbf{x}$, taken d at a time. We write $u = u^{i_1 i_2 \cdots i_d} = (x^{i_1}, x^{i_2}, \ldots, x^{i_d})^T$ where the d components need not be distinct and T denotes the transposition. Consider the transformation $y = g(x)$ from $x^1, \ldots, x^m$ to new variables $y^1, \ldots, y^m$ and let $c_i^r \equiv c_i^r(x) = \frac{\partial y^r}{\partial x^i}$ having full rank for all x. If $\bar{u}$, the value of u for the transformed variables $y^r, r = 1, 2, \ldots, m$, satisfies

$$\bar{u}^{r_1 r_2 \ldots r_d} = c_{i_1}^{r_1} c_{i_2}^{r_2} \cdots c_{i_d}^{r_d} u^{i_1 i_2 \ldots i_d}$$

then u is said to be a *contravariant tensor*. On the other hand, if u is a *covariant tensor*, we write $u = u_{i_1 i_2 \cdots i_d}$ and the transformation law for covariant tensor is

$$\bar{u}_{r_1 r_2 \ldots r_d} = d_{r_1}^{i_1} d_{r_2}^{i_2} \ldots d_{r_d}^{i_d} \, u_{i_1 i_2 \ldots i_d}$$

where $d_r^i = \frac{\partial x^i}{\partial y^r}$, the matrix inverse of c_i^r, satisfies the relationship $c_i^r d_r^j = \delta_i^j = c_r^j d_i^r$.

Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be independent and identically distributed m-dimensional random vectors. Denote the components of each random vector by $\mathbf{X} = (X^1, \cdots, X^m)$, with mean $\mu = (\mu^1, \ldots, \mu^m)$ and moments

$$\kappa^{i_1 \ldots i_v} = E(X^{i_1} - \mu^{i_1}) \ldots (X^{i_v} - \mu^{i_v}),$$

where $1 \leq i_k \leq m$, $1 \leq k \leq m$. The cumulants of $\mathbf{X}$ are the coefficients of the cumulant generating function

$$\kappa_{\mathbf{X}}(\mathbf{t}) = \log(M_{\mathbf{X}}(\mathbf{t})) = \sum_{i_1,\ldots,i_v=0}^{\infty} \frac{1}{i_1!\ldots i_v!}\kappa^{i_1,\ldots,i_v} t_{i_1}\ldots t_{i_v}\,,$$

where $M_{\mathbf{X}}$ is the moment generating function of $\mathbf{X}$. Here, $\kappa^{i_1,\ldots,i_v}$ is called the vth cumulant of $\mathbf{X}$. The following are the relationships between moments and cumulants.

$$\begin{aligned}
\kappa^{ij} &= \kappa^{i,j} + \kappa^i\kappa^j \\
\kappa^{ijk} &= \kappa^{i,j,k} + (\kappa^i\kappa^{j,k} + \kappa^j\kappa^{i,k} + \kappa^k\kappa^{i,j}) + \kappa^i\kappa^j\kappa^k \\
&= \kappa^{i,j,k} + \kappa^i\kappa^{j,k}[3] + \kappa^i\kappa^j\kappa^k \\
\kappa^{ijkl} &= \kappa^{i,j,k,l} + \kappa^i\kappa^{j,k,l}[4] + \kappa^{i,j}\kappa^{k,l}[3] + \kappa^i\kappa^j\kappa^{k,l}[6] + \kappa^i\kappa^j\kappa^k\kappa^l\,,
\end{aligned} \tag{12}$$

where $\kappa^i\kappa^{j,k}[3]$ is the sum over the three partitions of three indices. The following is a complete list of the 15 partitions of four items, one column for each of the five types (McCullagh, 1987)

$ijkl$	$i\|jkl$	$ij\|kl$	$i\|j\|kl$	$i\|j\|k\|l$
	$j\|ikl$	$ik\|jl$	$i\|k\|jl$	
	$k\|ijl$	$il\|jk$	$i\|l\|jk$	
	$l\|ijk$		$j\|k\|il$	
			$j\|l\|ik$	
			$k\|l\|ij$	

Let $S_n = \sum_{i=1}^n \mathbf{X}_i$ and $\mathbf{Z} = (S_n - \mu)/\sqrt{n}$ be the sum and standardized sum of random vectors $\mathbf{X}_1, \cdots, \mathbf{X}_n$ such that the cumulant $\kappa^{i_1,\ldots,i_v}$ of $\mathbf{Z}$ is of the order $n^{1-\frac{v}{2}}$. Then the Edgeworth expansion of $p_{\mathbf{Z}}$ up to order five about its best normal approximate is given by (Barndorff-Nielsen and Cox, 1989; Kendall and Stuart, 1977)

$$\begin{aligned}
& p_{\mathbf{Z}}(\mathbf{z};\kappa) \\
= \; & \phi_m(\mathbf{z};\kappa) \quad \left[1 + \frac{1}{3!}\kappa^{i,j,k}h_{ijk}(\mathbf{z};\kappa) + \frac{1}{4!}\kappa^{i,j,k,l}h_{ijkl}(\mathbf{z};\kappa) + \frac{10}{6!}\kappa^{i,j,k}\kappa^{l,p,q}h_{ijklpq}(\mathbf{z};\kappa)\right] \\
& + \quad O(n^{-\frac{3}{2}})
\end{aligned} \tag{13}$$

where

$$\phi_m(\mathbf{z};\kappa) = (2\pi)^{-m/2}\{\det(\kappa)\}^{-1/2}\exp(-\frac{1}{2}\kappa_{i,j}z^iz^j)\,,$$

denotes the m-dimensional multivariate normal distribution of zero mean and covariance matrix $\kappa = [\kappa^{i,j}]$, with $\kappa^{i,j} = E(Z^iZ^j)$ and $[k_{i,j}]$ represents κ^{-1}. The covariant Hermite polynomial $h_{i_1\ldots i_v}$ is defined as

$$\phi_m(\mathbf{x};\kappa)\, h_{i_1\ldots i_v}(\mathbf{x};\kappa) \;=\; (-1)^v\partial_{i_1}\ldots\partial_{i_v}\,\phi_m(\mathbf{x};\kappa)\,,$$

where $\partial_i = \partial/\partial x^i$. For the later use, the contravariant Hermite polynomial $h^{i_1\cdots i_v}$ is defined as

$$\phi_m(\mathbf{x};\kappa)\, h^{i_1\ldots i_v}(\mathbf{x};\kappa) \;=\; (-1)^v \partial^{i_1}\ldots\partial^{i_v}\,\phi_m(\mathbf{x};\kappa)\,,$$

with $\partial^i = \kappa^{i,j}\partial_j$. The first four covariant and contravariant Hermite polynomials are

$$\begin{aligned}
h_i &= x_i\,, & h^i &= x^i\,,\\
h_{ij} &= x_i x_j - \kappa_{i,j}\,, & h^{ij} &= x^i x^j - \kappa^{i,j}\,,\\
h_{ijk} &= x_i x_j x_k - \kappa_{i,j} x_k [3]\,, & h^{ijk} &= x^i x^j x^k - \kappa^{i,j} x^k [3]\,,\\
h_{ijkl} &= x_i x_j x_k x_l - \kappa_{i,j} x_k x_l [6] + \kappa_{i,j}\kappa_{k,l}[3]\,, & h^{ijkl} &= x^i x^j x^k x^l - \kappa^{i,j} x^k x^l [6] + \kappa^{i,j}\kappa^{k,l}[3]\,,
\end{aligned}$$

where the new notation x_i is defined as $x_i = \kappa_{i,j} x^j$.

Appendix B : Properties of Hermite polynomials

The expansion (7) may be simplified via certain properties of the Hermite polynomials (Skovgaard, 1981). First recall

$$h^{i_1\ldots i_v}(\mathbf{x};\kappa) \;=\; \kappa^{i_1,j_1}\ldots\kappa^{i_v,j_v} h_{j_1\ldots j_v}(\mathbf{x};\kappa)\,.$$

If the components of $\mathbf{X}$ are uncorrelated and of unit variance, then $\kappa^{i,i} = \kappa_{i,i} = 1$, $\kappa^{i,j} = \kappa_{i,j} = 0$. The covariant-contravariant Hermite polynomials for the multivariate distribution of $\mathbf{X}$ is then formed by taking all possible products of the Hermite polynomials (McCullagh, 1987):

$$\begin{aligned}
h_{i\ldots i}(\mathbf{x}) &= h^{i\ldots i}(\mathbf{x}) = H_v(x^i), \quad i\ldots i \text{ denotes } v \text{ repetitions}\\
h_{i\ldots ij\ldots j}(\mathbf{x}) &= h^{i\ldots ij\ldots j}(\mathbf{x}) = H_{v-t}(x^i)H_t(x^j), \quad i\ldots i \text{ denotes } v-t \text{ repetitions}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j\ldots j \text{ denotes } t \text{ repetitions}\\
h_{i_1\ldots i_v}(\mathbf{x}) &= h^{i_1\ldots i_v}(\mathbf{x}) = H_1(x^{i_1})\ldots H_1(x^{i_v}).
\end{aligned}$$

Second, recall the useful orthogonality properties (**?**) in the Hermite polynomials

$$\begin{aligned}
\int \phi(x)H_p(x)H_q(x)\,dx &= p!\,\delta_{pq}\\
\int \phi(x)H_3^2(x)H_4(x)\,dx &= 3!^3\,,\\
\int \phi(x)H_3^2(x)H_6(x)\,dx &= 6!\\
\int \phi(x)H_3^3(x)\,dx &= 0\\
\int \phi(x)H_3^4(x)\,dx &= 93\cdot 3!^2\,,
\end{aligned}$$

where $H_k(x)$ is the standard kth order Hermite polynomial. In the case of two dimensions ($m = 2$) with uncorrelated components of unit variance,

$$p_{\mathbf{Z}}(\mathbf{z};\kappa) \;=\; \phi_2(\mathbf{z};\kappa)\;[1 + v_1(\mathbf{z};\kappa) + v_2(\mathbf{z};\kappa) + v_3(\mathbf{z};\kappa)] \;+ o(n^{-1})$$

where

$$
\begin{aligned}
v_1(\mathbf{z};\kappa) &= \kappa^{1,1,1}h_{111}(\mathbf{z}) + 3\kappa^{1,1,2}h_{112}(\mathbf{z}) + 3\kappa^{1,2,2}h_{122}(\mathbf{z}) + \kappa^{2,2,2}h_{222}(\mathbf{z})\,,\\
v_2(\mathbf{z};\kappa) &= \kappa^{1,1,1,1}h_{1111}(\mathbf{z}) + 4\kappa^{1,1,1,2}h_{1112}(\mathbf{z})\\
&+ 6\kappa^{1,1,2,2}h_{1122}(\mathbf{z}) + 4\kappa^{1,2,2,2}h_{1222}(\mathbf{z}) + \kappa^{2,2,2,2}h_{2222}(\mathbf{z})\,,\\
v_3(\mathbf{z};\kappa) &= \kappa^{1,1,1}\kappa^{1,1,1}h_{111111}(\mathbf{z}) + 6\kappa^{1,1,1}\kappa^{1,1,2}h_{111112}(\mathbf{z})\\
&+ 15\kappa^{1,1,1}\kappa^{1,2,2}h_{111122}(\mathbf{z}) + 20\kappa^{1,1,1}\kappa^{2,2,2}h_{111222}(\mathbf{z})\\
&+ 15\kappa^{1,1,2}\kappa^{2,2,2}h_{112222}(\mathbf{z}) + 6\kappa^{1,2,2}\kappa^{2,2,2}h_{122222}(\mathbf{z})\\
&+ \kappa^{2,2,2}\kappa^{2,2,2}h_{222222}(\mathbf{z})
\end{aligned}
$$

and

$$
\begin{aligned}
h_{111}(\mathbf{z}) = h^{111}(\mathbf{z}) &= H_3(z^1) & h_{112}(\mathbf{z}) = h^{112}(\mathbf{z}) &= H_2(z^1)H_1(z^2)\\
h_{122}(\mathbf{z}) = h^{122}(\mathbf{z}) &= H_1(z^1)H_2(z^2) & h_{222}(\mathbf{z}) = h^{222}(\mathbf{z}) &= H_3(z^2)\\
h_{1111}(\mathbf{z}) = h^{1111}(\mathbf{z}) &= H_4(z^1) & h_{1112}(\mathbf{z}) = h^{1112}(\mathbf{z}) &= H_3(z^1)H_1(z^2)\\
h_{1122}(\mathbf{z}) = h^{1122}(\mathbf{z}) &= H_2(z^1)H_2(z^2) & h_{1222}(\mathbf{z}) = h^{1222}(\mathbf{z}) &= H_1(z^1)H_3(z^2)\\
h_{111111}(\mathbf{z}) = h^{111111}(\mathbf{z}) &= H_6(z^1) & h_{111112}(\mathbf{z}) = h^{111112}(\mathbf{z}) &= H_5(z^1)H_1(z^2)\\
h_{111122}(\mathbf{z}) = h^{111122}(\mathbf{z}) &= H_4(z^1)H_2(z^2) & h_{111222}(\mathbf{z}) = h^{111222}(\mathbf{z}) &= H_3(z^1)H_3(z^2)\\
h_{112222}(\mathbf{z}) = h^{112222}(\mathbf{z}) &= H_2(z^1)H_4(z^2) & h_{122222}(\mathbf{z}) = h^{122222}(\mathbf{z}) &= H_1(z^1)H_5(z^2)\\
h_{222222}(\mathbf{z}) = h^{222222}(\mathbf{z}) &= H_6(z^2)\,.
\end{aligned}
$$

The correlation term $v_1(\mathbf{z};\kappa)$, $v_2(\mathbf{z};\kappa)$, and $v_3(\mathbf{z};\kappa)$ will reduce to

$$
\begin{aligned}
v_1(\mathbf{z};\kappa) &= \kappa^{1,1,1}H_3(z^1) + 3\kappa^{1,1,2}H_2(z^1)H_1(z^2) + 3\kappa^{1,2,2}H_1(z^1)H_2(z^2) + \kappa^{2,2,2}H_3(z^2)\,,\\
v_2(\mathbf{x};\kappa) &= \kappa^{1,1,1,1}H_4(z^1) + 4\kappa^{1,1,1,2}H_3(z^1)H_1(z^2)\\
&+ 6\kappa^{1,1,2,2}H_2(z^1)H_2(z^2) + 4\kappa^{1,2,2,2}H_1(z^1)H_3(z^2) + \kappa^{2,2,2,2}H_4(z^2)\,,\\
v_3(\mathbf{z};\kappa) &= \kappa^{1,1,1}\kappa^{1,1,1}H_6(z^1) + 6\kappa^{1,1,1}\kappa^{1,1,2}H_5(z^1)H_1(z^2)\\
&+ 15\kappa^{1,1,1}\kappa^{1,2,2}H_4(z^1)H_2(z^2) + 20\kappa^{1,1,1}\kappa^{2,2,2}H_3(z^1)H_3(z^2)\\
&+ 15\kappa^{1,1,2}\kappa^{2,2,2}H_2(z^1)H_4(z^2) + 6\kappa^{1,2,2}\kappa^{2,2,2}H_1(z^1)H_5(z^2)\\
&+ \kappa^{2,2,2}\kappa^{2,2,2}H_6(z^2)\,.
\end{aligned}
$$

Table 4. Numerical results of $S(\phi_X)$ (1.42, theoretical value).

n	by Edgeworth/abs.err	by density estimation/abs.err
100	1.439/0.019	1.389/0.031
200	1.425/0.005	1.392/0.028
300	1.424/0.004	1.404/0.016
400	1.424/0.004	1.413/0.007
500	1.422/0.002	1.425/0.005

Table 5. 2-dim numerical results of $S(\phi_X)$.

covariance	$cov = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$cov = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}$	$cov = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$
true vale	2.8379	2.7907	2.3271
n	by Edgeworth/abs.err	by Edgeworth/abs.err	by Edgeworth/abs.err
100	2.5642/0.2737	2.6181/0.1726	2.4041/0.0769
200	2.7545/0.0833	2.6751/0.1156	2.2617/0.0653
300	2.7874/0.0505	2.8281/0.0374	2.2739/0.0530
400	2.8090/0.0288	2.7820/0.0086	2.3442/0.0172
500	2.8529/0.0149	2.7897/0.0009	2.3169/0.0101

Table 6. 3-dim numerical results of $S(\phi_X)$.

covariance	$cov = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$cov = \begin{pmatrix} 1 & 0.8 & 0.6 \\ 0.8 & 1 & 0.4 \\ 0.6 & 0.4 & 1 \end{pmatrix}$
true vale	4.2568	3.5087
n	by Edgeworth/abs.err	by Edgeworth/abs.err
100	4.0904/0.1663	3.2448/0.2639
200	4.3208/0.0640	3.4428/0.0659
300	4.1976/0.0591	3.4757/0.0330
400	4.2800/0.0232	3.4737/0.0350
500	4.2730/0.0162	3.4899/0.0187

Table 7. 4-dim, 5-dim, 8-dim numerical results of $S(\phi_{\mathbf{Z}})$ with identity covariance.

dimension	4-dim	5-dim	8-dim
true vale	5.6757	7.0946	11.35151
n	by Edgeworth/abs.err	by Edgeworth/abs.err	by Edgeworth/abs.err
100	5.3947/0.2810	6.7851/0.3095	11.0698/0.2816
200	5.4864/0.1893	7.3411/0.2464	11.1345/0.2169
300	5.5043/0.1714	6.9629/0.1317	11.1246/0.2268
400	5.5904/0.0852	6.9771/0.1175	11.1223/0.2291
500	5.7089/0.0331	7.0451/0.0495	11.2801/0.0713

Table 8. The 90% confidence interval for the EKLD of the INGA, PCA and ICA.

example	INGA	PCA	ICA	Conclusion
2-dim cigar (Figure 4)	(0.1237) 0.1847 (0.7051)	(0.1920) 1.0892 (2.8520)	(0.1047) 0.9823 (3.8764)	INGA produces the best syntheses.
2-dim spike (Figure 5)	(0.0976) 0.1628 (0.7512)	(0.5361) 1.9211 (3.7647)	(0.6301) 1.1541 (4.2062)	INGA produces the best syntheses.
25-dim eye (Figure 6)	(2.8531) 3.1781 (3.8114)	(10.0919) 19.694 (28.743)	(12.713) 20.789 (30.418)	INGA produces the best syntheses. Here, the original eye dataset has been compressed from 144-dim to 25-dim via PCA.

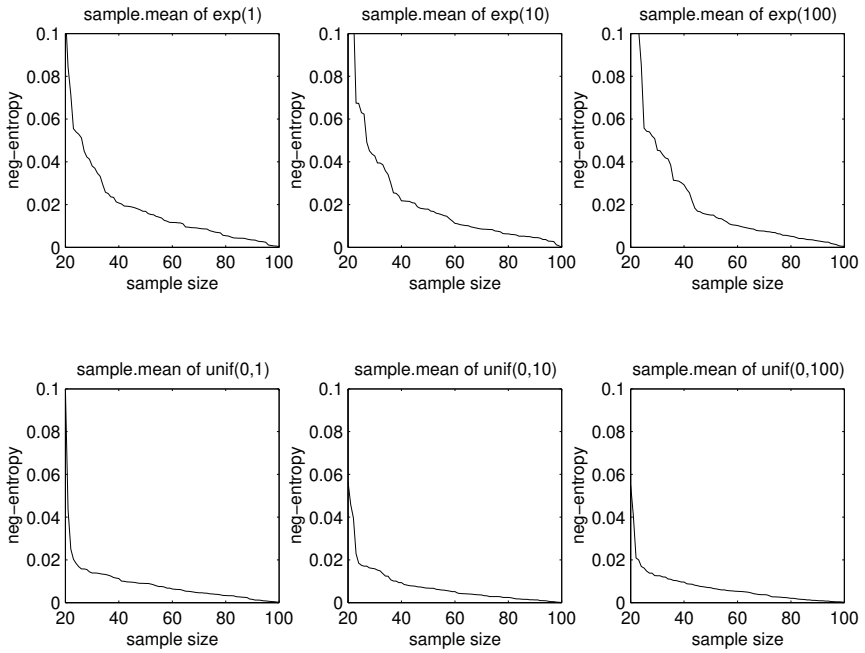

Fig. 1. neg-entropy of the sample mean from different distributions.

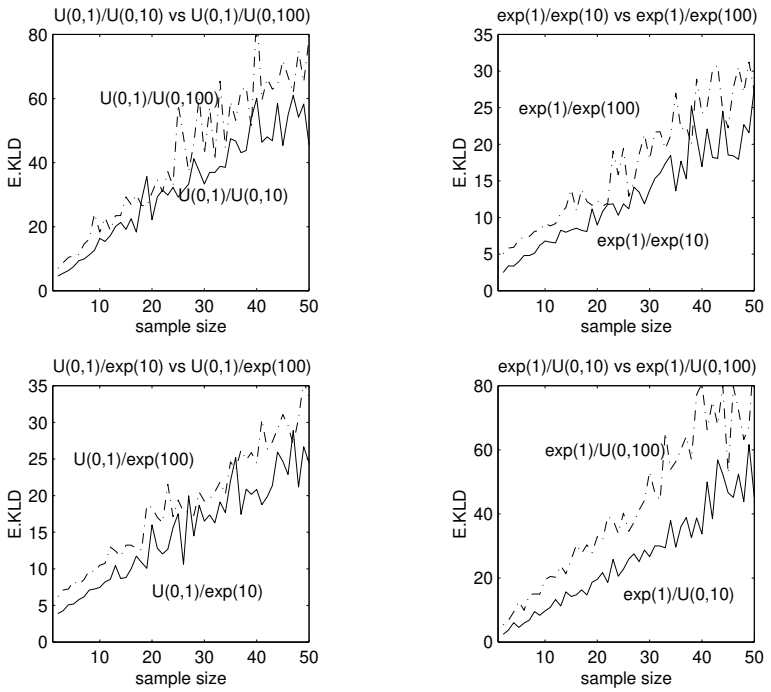

Fig. 2. EKLD between the sample mean from “large” and “small” distance.

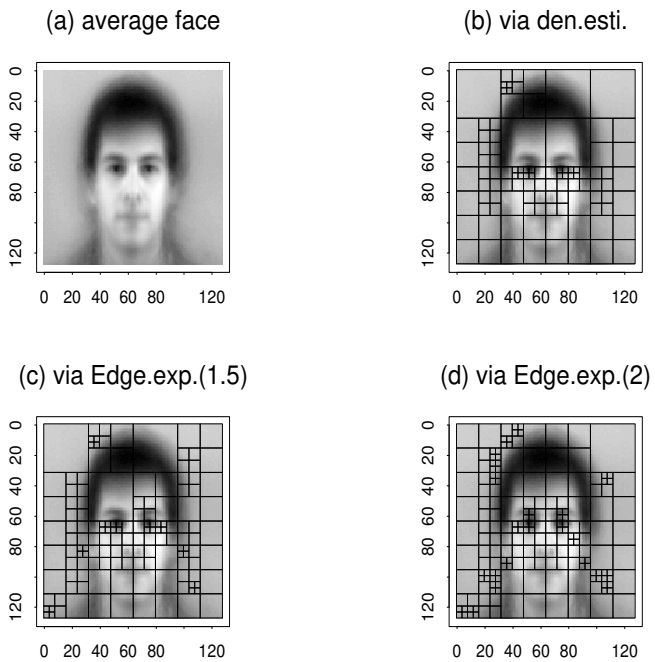

Fig. 3. Comparison of LSDB chosen by using density estimation and Edgeworth Expansion with order $O(n^{-1.5})$ and $O(n^{-2})$.

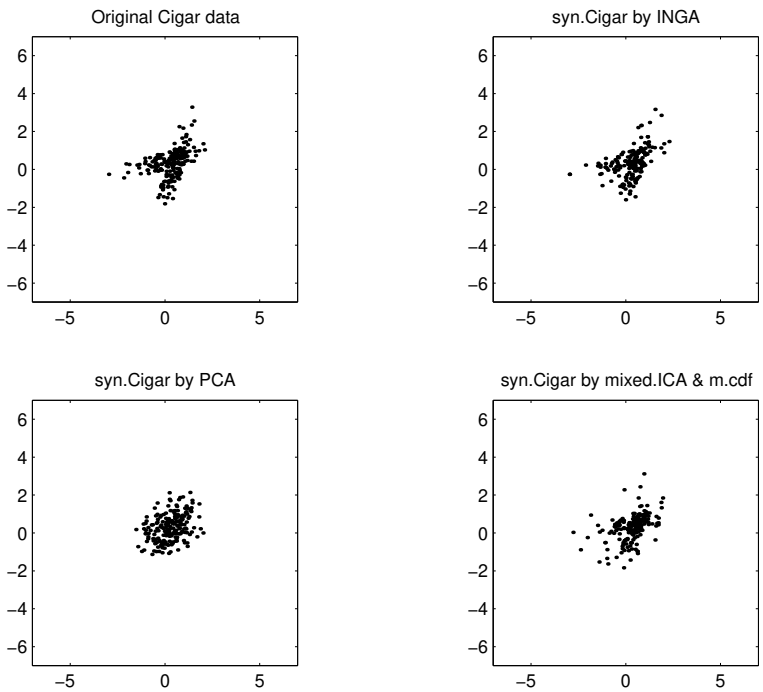

Fig. 4. Resampling of the "cigar" data by INGA, PCA and ICA. The first row left to right: original samples; resamples by INGA; The second row left to right: resamples by PCA and ICA.

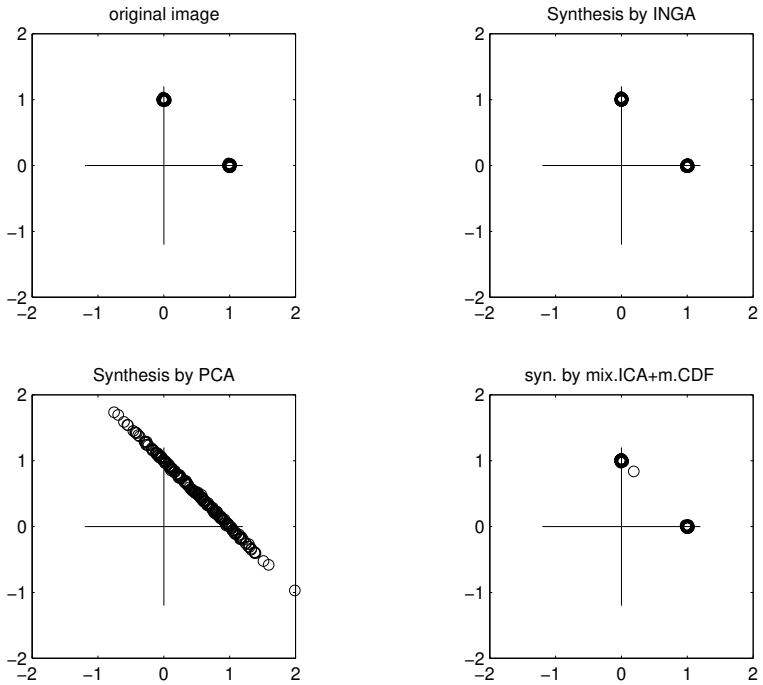

Fig. 5. The spike process simulation ($p = 2$). The first row left to right: original samples; resamples by INGA; The second row left to right: resamples by PCA and ICA.

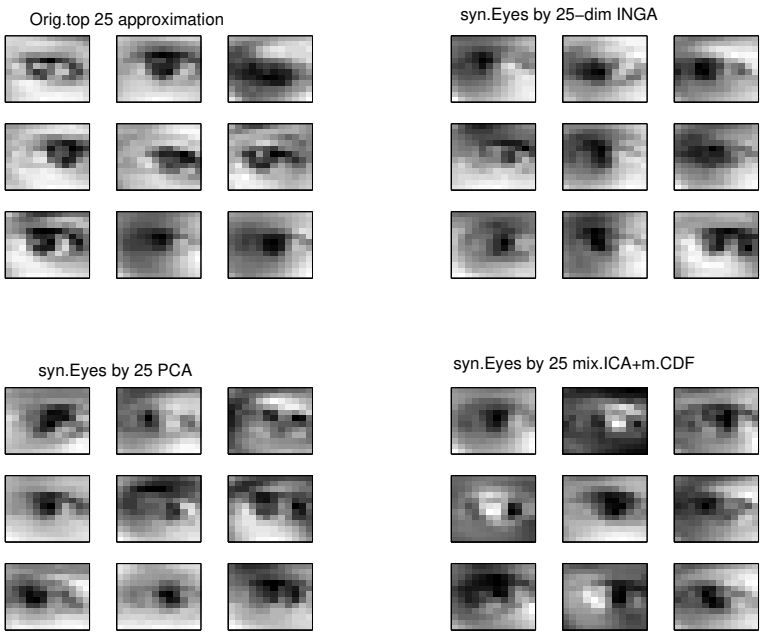

Fig. 6. Simulations of the eye image database by INGA and PCA. The first row left to right: original samples; resamples by INGA; The second row left to right: resamples by PCA and ICA.

References

Abramowitz, M and Stegun, I.A. (1972). *Handbook of Mathematical Functions*, Dover, New York.

Anscombe, F.J. (1961). Examination of residuals. *Proc. 4th Berkeley Symp.*, **1**, 1–36.

Barndorff-Nielsen, O.E. and Cox, D.R. (1989). *Inference and Asymptotics.* Chapman and Hall, London.

Bickel, P.J. (1978). Using residuals robustly I: Tests for heteroscedasticity non-linearity, *Ann. Statist.*, **6**, 266–291.

Brillinger, D.R. (1994). Some basic aspects and uses of higher-order spectra. *Signal Processing*, **36**, 239–249.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, **36**, 287–314.

Cross, G. and Jain, A. (1983). Markov random field texture models. *IEEE Trans. Pattern Anal. Machine Intell.*, **5**, 25–39.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, **6**, 721–741.

Hall, P. (1987). On Kullback-Leibler loss and density estimation. *Ann. Statist.*, **15**, 1491–1519.

Hall, P. and Morton, S.C. (1993). On the estimation of entropy. *Ann. Inst. Statist. Math.*, **45**, 69–88.

Hinkley, D.V. (1985). Transformation diagnostics for linear models. *Biometrika*, **72**, 487–496.

Joe, H. (1989). Estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Statist. Math.*, **41**, 683–697.

Jones, M.C. and Sibson, R. (1987). What is projection pursuit?, *J. Royal Statist. Soc. London, Ser. A*, **150**, 1–36.

Jutten, C. and Herault, J. (1991). Blind separation of sources, Part I: An adaptive algorithm based on neuromimatic architecture. *Signal Processing*, **24**, 1–10.

Kendall, M. and Stuart, A. (1977). *The Advanced Theory of Statistics*, Vol. 1. Oxford Univ. Press, New York.

Kullback, S. (1959). *Information Theory and Statistics*, Wiley, New York, Republished by Dover in 1997.

Lin, J.-J., Saito, N., and Levine, R. A. (2001). An iterative nonlinear Gaussianization algorithm for Image Simulation and Synthesis, submitted for publication.

McCullagh, P. and Pregibon, D. (1987). K-statistics and dispersion effects in regression. *Ann. Statist.* **15**, 202–219.

McCullagh, P. (1987). *Tensor Methods in Statistics.* Chapman and Hall, London.

Popat, K. and Picard, R.W. (1997) Cluster-based probability model and its application to image and texture processing. *IEEE Trans. Image Proc.* **6**, 268–284.

Portilla, J. and Simoncelli, E.P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *Intern. J. Comput. Vision*, **40**, 49–71.

Simoncelli, E.P. (1997). Statistical models for images: compression, restoration and synthesis. *31st Asilomar conference on Signals Systems, and Computers*, Pacific Grove, CA. Nov. 2–5.

Saito, N. (1994). *Local Feature Extraction and Its Applications Using a Library of Bases.* PhD thesis, Department of Mathematics, Yale University, New Haven, CT 06520 USA.

Saito, N. (1998). Least statistically-dependent basis and its application to image modeling. In *Wavelet Applications in Signal and Image Processing VI*, Eds. A.F. Laine, M.A. Unser, and A. Aldroubi, Proc. SPIE 3458, 24-38.

Saito, N. (2001). Image approximation and modeling via least statistically-dependent bases. *Pattern Recognition*, **34**, 159–178.

Skovgaard, I.M. (1981). Transformation of an Edgeworth expansion by a sequence of smooth functions. *Scand. J. Statist.*, **8**, 207–217.

Watanabe, S. (1965). Karhunen-Loève expansion and factor analysis: Theoretical remarks and applications. In *Trans. 4th Prague Conf. Inform. Theory, Statist. Decision Functions, Random Processes*, (Prague), Publishing House of the Czechoslovak Academy of Sciences, 635–660.

Zhu, S.C., Wu, Y. and Mumford, D. (1998). FRAME: Filters, Random fields And Maximum Entropy–Toward a unified theory for texture modeling. *Intern. J. Comput. Vision* **27**, 1–20.