Iterative Gaussianization: from ICA to Random Rotations

Valero Laparra, Gustavo Camps-Valls, *Senior Member, IEEE* and Jesús Malo

***Abstract*—Most signal processing problems involve the challenging task of multidimensional probability density function (PDF) estimation. In this work, we propose a solution to this problem by using a family of Rotation-based Iterative Gaussianization (RBIG) transforms. The general framework consists of the sequential application of a univariate marginal Gaussianization transform followed by an orthonormal transform. The proposed procedure looks for differentiable transforms to a known PDF so that the unknown PDF can be estimated at any point of the original domain. In particular, we aim at a zero mean unit covariance Gaussian for convenience.**

RBIG is formally similar to classical iterative Projection Pursuit (PP) algorithms. However, we show that, unlike in PP methods, the particular class of rotations used has no special qualitative relevance in this context, since looking for *interestingness* is not a critical issue for PDF estimation. The key difference is that our approach focuses on the univariate part (marginal Gaussianization) of the problem rather than on the multivariate part (rotation). This difference implies that one may select the most convenient rotation suited to each practical application.

The differentiability, invertibility and convergence of RBIG are theoretically and experimentally analyzed. Relation to other methods, such as Radial Gaussianization (RG), one-class support vector domain description (SVDD), and deep neural networks (DNN) is also pointed out. The practical performance of RBIG is successfully illustrated in a number of multidimensional problems such as image synthesis, classification, denoising, and multi-information estimation.

***Index Terms*—Gaussianization, Independent Component Analysis (ICA), Principal Component Analysis (PCA), Negentropy, Multi-information, Probability Density Estimation, Projection Pursuit.**

I. INTRODUCTION

MANY signal processing problems such as coding, restoration, classification, regression or synthesis greatly depend on an appropriate description of the underlying probability density function (PDF) [1]–[5]. However, density estimation is a challenging problem when dealing with high-dimensional signals because direct sampling of the input space is not an easy task due to the curse of dimensionality [6]. As a result, specific problem-oriented PDF models are typically developed to be used in the Bayesian framework.

The conventional approach is to transform data into a domain where *interesting* features can be easily (i.e. marginally) characterized. In that case, one can apply well-known marginal techniques to each feature independently and then obtain a description of the multidimensional PDF. The most popular approaches rely on linear models and statistical independence. However, they are usually too restrictive to describe general data distributions. For instance, principal component analysis (PCA) [7], that reduces to DCT in many natural signals such as speech, images and video, assumes a Gaussian source [3], [7]. More recently, linear ICA, that reduces to wavelets in natural signals, assumes that observations come from the linear combination of independent non-Gaussian sources [8]. In general, these assumptions may not be completely correct, and residual dependencies still remain after the linear transform that looks for independence. As a result, a number of problem-oriented approaches have been developed in the last decade to either describe or remove the relations remaining in these linear domains. For example, parametric models based on joint statistics of wavelet coefficients have been successfully proposed for texture analysis and synthesis [5], image coding [9] or image denoising [10]. Non-linear methods using non-explicit statistical models have been also proposed to this end in the denoising context [11], [12] and in the coding context [13], [14]. In function approximation and classification problems, a common approach is to first linearly transform the data, e.g. with the most relevant eigenvectors from PCA, and then applying nonlinear methods such as artificial neural networks or support vector machines in the reduced dimensionality space [3], [4], [7].

Identifying the *meaningful* transform for an easier PDF description in the transformed domain strongly depends on the problem at hand. In this work we circumvent this constraint by looking for a transform such that the transformed PDF is known. Even in the case that this transform is qualitatively *meaningless*, being differentiable, allows us to estimate the PDF in the original domain. Accordingly, in the proposed context, the role (*meaningfulness*) of the transform is not that relevant. Actually, as we will see, an infinite family of transforms may be suitable to this end, so one has the freedom to choose the most convenient one.

In this work, we propose to use a unit covariance Gaussian as target PDF in the transformed domain and iterative transforms based on arbitrary rotations. We do so because the match between spherical symmetry and rotations makes it possible to define a cost function (negentropy) with nice theoretical properties. The properties of negentropy allow us to show that one Gaussianization transform is always found no matter the selected class of rotations.

The remainder of the paper is organized as follows. In

Manuscript received April 2010;
This work was partially supported by projects CICYT-FEDER TEC2009-13696, AYA2008-05965-C04-03 and CSD2007-00018, and grant BES-2007-16125.
Authors are with Image Processing Laboratory (IPL) at the Universitat de València, Spain. Address: C/ Catedrático Escardino, 46980 - Paterna (València) Spain. E-mail: {lapeva,gcamps,jmalo}@uv.es, http://ipl.uv.es

Section II we present the underlying idea that motivates the proposed approach to Gaussianization. In Section III, we give the formal definition of the Rotation-based Iterative Gaussianization (RBIG), and show that the scheme is invertible, differentiable and it converges for a wide class of orthonormal transforms, even including random rotations. Section IV discusses the similarities and differences of the proposed method and Projection Pursuit (PP) [15]–[18]. Links to other techniques (such as single-step Gaussianization transforms [19], [20], one-class support vector domain descriptions (SVDD) [21], and deep neural network architectures [22]) are also explored. Section V shows the experimental results. First, we experimentally show that the proposed scheme converges to an appropriate Gaussianization transform for a wide class of rotations. Then, we illustrate the usefulness of the method in a number of high-dimensional problems involving PDF estimation: image synthesis, classification, denoising and multi-information estimation. In all cases, RBIG is compared to related methods in each particular application. Finally, Section VI draws the conclusions of the work.

II. Motivation

This section considers a solution to the PDF estimation problem by using a differentiable transform to a domain with known PDF. In this setting, different approaches can be adopted which will motivate the proposed method.

Let $\mathbf{x}$ be a d-dimensional random variable with (unknown) PDF, $p_{\mathbf{x}}(\mathbf{x})$. Given some bijective, differentiable transform of $\mathbf{x}$ into $\mathbf{y}$, $\mathcal{G}:\mathbb{R}^d\rightarrow\mathbb{R}^d$, such that $\mathbf{y}=\mathcal{G}(\mathbf{x})$, the PDFs in the original and the transformed domains are related by [23]:

$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{y}}(\mathcal{G}(\mathbf{x}))\left|\frac{d\mathcal{G}(\mathbf{x})}{d\mathbf{x}}\right| = p_{\mathbf{y}}(\mathcal{G}(\mathbf{x}))|\nabla_{\mathbf{x}}\mathcal{G}(\mathbf{x})|, \quad (1)$$

where $|\nabla_{\mathbf{x}}\mathcal{G}|$ is the determinant of the Jacobian matrix. Therefore, the unknown PDF in the original domain can be estimated from a transform of known Jacobian leading to an appropriate (known or straightforward to compute) target PDF, $p_{\mathbf{y}}(\mathbf{y})$.

One could certainly try to figure out direct (or even closed form) procedures to transform particular PDF classes into a target PDF [19], [20]. However, in order to deal with any possible PDF, iterative methods seem to be a more reasonable approach. In this case, the initial data distribution should be iteratively transformed in such a way that the target PDF is progressively approached in each iteration.

The appropriate transform in each iteration would be the one that maximizes a similarity measure between PDFs. A sensible cost function here is the Kullback-Leibler divergence (KLD) between PDFs. In order to apply well-known properties of this measure [24], [25], it is convenient to choose a unit covariance Gaussian as target PDF, $p_{\mathbf{y}}(\mathbf{y})=\mathcal{N}(\mathbf{0},\mathbf{I})$. With this choice, the cost function describing the divergence between the current data, $\mathbf{x}$, and the unit covariance Gaussian is the hereafter called negentropy[1], $J(\mathbf{x}) = \mathrm{D}_{\mathrm{KL}}\left(p(\mathbf{x})|\mathcal{N}(\mathbf{0},\mathbf{I})\right)$. Negentropy can be decomposed as the sum of two non-negative quantities, the multi-information and the marginal negentropy:

$$J(\mathbf{x}) = I(\mathbf{x}) + J_m(\mathbf{x}). \quad (2)$$

This can be readily derived from Eq. (5) in [25], by considering as contrast PDF $\prod_i q_i(x_i) = \mathcal{N}(\mathbf{0},\mathbf{I})$. The multi-information is [26]:

$$I(\mathbf{x}) = \mathrm{D}_{\mathrm{KL}}(p(\mathbf{x})|\textstyle\prod_i p_i(x_i)) \quad (3)$$

Multi-information measures statistical dependence, and it is zero if and only if the different components of $\mathbf{x}$ are independent. The marginal negentropy is defined as:

$$J_m(\mathbf{x}) = \sum_{i=1}^{d}\mathrm{D}_{\mathrm{KL}}\left(p_i(x_i)|\mathcal{N}(0,1)\right) \quad (4)$$

Given a data distribution from the unknown PDF, in general both I and J_m will be non-zero. The decomposition in (2) suggests two alternative approaches to reduce J:

1) *Reducing I*: This implies looking for interesting (independent) components. If one is able to obtain $I=0$, then $J=J_m\geq 0$, and this reduces to solving a marginal problem. Marginal negentropy can be set to zero with the appropriate set of dimension-wise Gaussianization transforms, $\mathbf{\Psi}$. This is easy as will be shown in the next section.
   However, this is an ambitious approach since looking for independent components is a non-trivial (intrinsically multivariate and nonlinear) problem. According to this, linear ICA techniques will not succeed in completely removing the multi-information, and thus a nonlinear post-processing is required.
2) *Reducing J_m*: As stated above, this univariate problem is easy to solve by using the appropriate $\mathbf{\Psi}$. Note that I will remain constant since it is invariant under dimension-wise transforms [26]. In this way, one ensures that the cost function is reduced by J_m. Then, a further processing has to be taken in order to come back to a situation in which one may have the opportunity to remove J_m again. This additional transform may consist of applying a rotation $\mathbf{R}$ to the data, as will be shown in the next section.

The relevant difference between the approaches is that, in the first one, the important part is looking for the interesting representation (multivariate problem), while in the second approach the important part is the univariate Gaussianization. In this second case, the class of rotations has no special qualitative relevance: in fact, marginal Gaussianization is the only part reducing the cost function.

The first approach is the underlying idea in Projection Pursuit methods focused on looking for interesting projections [16], [17]. Since the core of these methods is looking for meaningful projections (usually ICA algorithms), they suffer from a big computational complexity: for example, robust ICA algorithms such as RADICAL [27] would lead to extremely slow Gaussianization algorithms whereas relatively more convenient alternatives such as FastICA [28] may not converge in all cases. This may explain why, so far, Gaussianization

[1] This usage of the term negentropy slightly differs from the usual definition [24] where negentropy is taken to be KLD between $p_{\mathbf{x}}(\mathbf{x})$ and a multivariate Gaussian of the same mean and covariance. However, note that this difference has no consequence assuming the appropriate input data standardization (zero mean and unit covariance), which can be done without loss of generality.

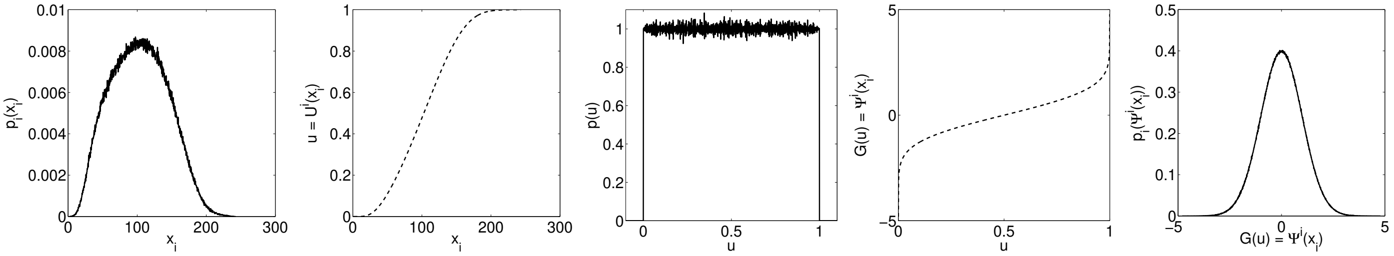

Fig. 1. Example of marginal Gaussianization in some particular dimension i. From left to right: marginal PDF of x_i, uniformization transform $u = U^i(x_i)$, PDF of the uniformized variable $p(u)$, Gaussianization transform $G(u)$, and PDF of the Gaussianized variable $p_i(\Psi^i(x_i))$.

techniques have been applied just to low-dimensional (audio) signals in either simple contexts based on point-wise non-linearities [29], [30], or after *ad hoc* speech-oriented feature extraction steps [31]. In this work, we propose following the simpler second approach using the most computationally convenient rotation. Intentionally, we do not pay attention to the meaningfulness of the rotations.

III. Rotation-based Iterative Gaussianization (RBIG)

This section first introduces the basic formulation of the proposed method, and then analyzes the properties of differentiability, invertibility, and convergence. Finally, we discuss on the role of the rotation matrix used in the scheme.

A. Iterative Gaussianization based on arbitrary rotations

According to the above reasoning, we propose the following class of Rotation-based Iterative Guassianization (RBIG) algorithms: given a d-dimensional random variable $\mathbf{x}^{(0)}$, following an unknown PDF, $p(\mathbf{x}^{(0)})$, in each iteration k, a two-step processing is performed:

$$\mathcal{G} : \mathbf{x}^{(k+1)} = \mathbf{R}_{(k)} \cdot \mathbf{\Psi}_{(k)}(\mathbf{x}^{(k)}) \tag{5}$$

where $\mathbf{\Psi}_{(k)}$ is the marginal Gaussianization of each dimension of $\mathbf{x}^{(k)}$ for the corresponding iteration, and $\mathbf{R}_{(k)}$ is a generic rotation matrix for the marginally Gaussianized variable $\mathbf{\Psi}_{(k)}(\mathbf{x}^{(k)})$.

The freedom in choosing the rotations is consistent with the intuitive fact that there is an infinite number of ways to twist a PDF in order to turn it into a unit covariance Gaussian. In principle, any of these choices is equally useful for our purpose, i.e. estimating the PDF in the original domain using Eq. (1). Note that when using different rotations, the qualitative meaning of the same region of the corresponding Gaussianized domain will be different. As a result, in order to work in the Gaussianized domain, one has to take into account the value of the point-dependent Jacobian. Incidentally, this is also the case in the PP approach, and more generally, in any non-linear approach. However, the interpretation of the Gaussianized domain is not an issue when working in the original domain. Finally, it is important to note that the method just depends on univariate (marginal) PDF estimations. Therefore, it does not suffer from the curse of dimensionality.

B. Invertibility and differentiation

The considered class of Gaussianization transforms is *differentiable* and *invertible*. Differentiability, allows us to estimate the PDF in the original domain from the Jacobian of the transform in each point, cf. Eq. (1). Invertibility guarantees that the transform is bijective which is a necessary condition to apply Eq. (1). Additionally, it is convenient for generating samples in the original domain by sampling the Gaussianized domain.

Before getting into the details, we take a closer look at the basic tool of marginal Gaussianization. Marginal Gaussianization in each dimension i and each iteration k, $\mathbf{\Psi}^i_{(k)}$, can be decomposed into two equalization transforms: (1) marginal uniformization, $U^i_{(k)}$, based on the cumulative density function of the marginal PDF, and (2) Gaussianization of a uniform variable, $G(u)$, based on the inverse of the cumulative density function of a univariate Gaussian: $\mathbf{\Psi}^i_{(k)} = G \odot U^i_{(k)}$, where:

$$u = U^i_{(k)}(x_i^{(k)}) \quad = \quad \int_{-\infty}^{x_i^{(k)}} p_i(x_i'^{(k)})\, dx_i'^{(k)} \tag{6}$$

$$G^{-1}(x_i) \quad = \quad \int_{-\infty}^{x_i} g(x_i')\, dx_i' \tag{7}$$

and $g(x_i)$ is just a univariate Gaussian. Figure 1 shows an example of the marginal Gaussianization of a one-dimensional variable x_i.

One dimensional density estimation is an issue by itself, and it has been widely studied [4], [32]. The selection of the most convenient density estimation procedure depends on the particular problem and, of course, the univariate Gaussianization step in the proposed algorithm could benefit from the extensive literature on the issue. In our case, we take a practical approach and no particular model is assumed for the marginal variables to keep the method as general as possible. Accordingly, the univariate Gaussianization transforms are computed from the cumulative histograms. Of course, alternative analytical approximations could be introduced at the cost of making the model more rigid. On the positive side, parametric models may imply better data regularization and avoid overfitting. However, exploring the effect of alternative density estimators will not be analyzed here.

Let us consider now the issue of invertibility. By simple manipulation of (5), it can be shown that the inverse transform is given by:

$$\mathcal{G}^{-1} : \mathbf{x}^{(k)} = \mathbf{\Psi}^{-1}_{(k)}(\mathbf{R}^{\top}_{(k)} \cdot \mathbf{x}_{(k+1)}). \tag{8}$$

The rotation $\mathbf{R}_{(k)}$ is not a problem for invertibility since the inverse is just the transpose, $\mathbf{R}_{(k)}^{-1} = \mathbf{R}_{(k)}^{\top}$. However, the key to ensure transform inversion is the invertibility of $\mathbf{\Psi}_{(k)}$. This is trivially ensured when the support of each marginal PDF is connected, that is, there are no holes (zero probability regions) in the support. In this way all the marginal CDFs are strictly monotonic and hence invertible. Note that the existence of holes in the support of the joint PDF is not a problem as long as it gives rise to marginal PDFs with a connected support. Problems in inversion will appear only when the joint PDF gives rise to clusters that are so distant that their projections onto the axes do not overlap. However, in such a situation, it may make more qualitative sense to consider that distinct clusters come from different sources and learn each one with a different Gaussianization transform.

The Jacobian of the series of K iterations is just the product of the corresponding Jacobian in each iteration:

$$\nabla_{\mathbf{x}}\mathcal{G} = \prod_{k=1}^{K} \mathbf{R}_{(k)} \cdot \nabla_{\mathbf{x}^{(k)}} \mathbf{\Psi}_{(k)} \tag{9}$$

Marginal Gaussianization, $\mathbf{\Psi}_{(k)}$, is a dimension-wise transform, whose Jacobian is the diagonal matrix,

$$\nabla_{\mathbf{x}^{(k)}} \mathbf{\Psi}_{(k)} = \begin{pmatrix} \frac{\partial \mathbf{\Psi}^1_{(k)}}{\partial x_1^{(k)}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\partial \mathbf{\Psi}^d_{(k)}}{\partial x_d^{(k)}} \end{pmatrix} \tag{10}$$

According to the two equalization steps in each marginal Gaussianization, Eq. (7), each element in $\nabla_{\mathbf{x}^{(k)}} \mathbf{\Psi}_{(k)}$ can be easily computed by applying the chain rule on u defined in Eq. (6):

$$\begin{aligned} \frac{\partial \mathbf{\Psi}^i_{(k)}}{\partial x_i^{(k)}} &= \frac{\partial \mathcal{G}}{\partial u}\frac{\partial u}{\partial x_i^{(k)}} = \left(\frac{\partial \mathcal{G}^{-1}}{\partial x_i}\right)^{-1} p_i(x_i^{(k)}) \\ &= g(\Psi^i_{(k)}(x_i^{(k)}))^{-1} p_i(x_i^{(k)}) \end{aligned} \tag{11}$$

Again, the differentiable nature of the considered Gaussianization is independent from the selected rotations $\mathbf{R}_{(k)}$.

C. Convergence properties

Here we prove two general properties of random variables, which are useful in the contexts of PDF description and redundancy reduction.

Property 3.1 (Negentropy reduction): Marginal Gaussianization reduces the negentropy and this is not modified by any posterior rotation:

$$\Delta J = J(\mathbf{x}) - J(\mathbf{R}\mathbf{\Psi}(\mathbf{x})) \geq 0, \forall \mathbf{R} \tag{12}$$

Proof: Using Eq. (2), the negentropy reduction due to marginal Gaussianization followed by a rotation is:

$$\Delta J = J(\mathbf{x}) - J(\mathbf{R}\mathbf{\Psi}(\mathbf{x})) = J(\mathbf{x}) - J(\mathbf{\Psi}(\mathbf{x}))$$

since $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is rotation invariant. Therefore,

$$\Delta J = I(\mathbf{x}) + J_m(\mathbf{x}) - I(\mathbf{\Psi}(\mathbf{x})) - J_m(\mathbf{\Psi}(\mathbf{x}))$$

Since the multi-information is invariant under dimension-wise transforms [26] (such as $\mathbf{\Psi}$), and the marginal negentropy of a marginally Gaussianized variable is zero,

$$\Delta J = J_m(\mathbf{x}) \geq 0, \forall \mathbf{R}$$

■

Property 3.2 (Redundancy reduction): Given a marginally Gaussianized variable, $\mathbf{\Psi}(\mathbf{x})$, any rotation reduces the redundancy among coefficients,

$$\Delta I = I(\mathbf{\Psi}(\mathbf{x})) - I(\mathbf{R}\mathbf{\Psi}(\mathbf{x})) \geq 0, \forall \mathbf{R} \tag{13}$$

Note that this property also implies that the combination of marginal Gaussianization and rotation gives rise to redundancy reduction since $I(\mathbf{\Psi}(\mathbf{x})) = I(\mathbf{x})$.

Proof: Using Eq. (2) on both $I(\mathbf{\Psi}(\mathbf{x}))$ and $I(\mathbf{R}\mathbf{\Psi}(\mathbf{x}))$, the redundancy reduction is:

$$\Delta I = J(\mathbf{\Psi}(\mathbf{x})) - J_m(\mathbf{\Psi}(\mathbf{x})) - J(\mathbf{R}\mathbf{\Psi}(\mathbf{x})) + J_m(\mathbf{R}\mathbf{\Psi}(\mathbf{x})).$$

Since negentropy is rotation invariant and the marginal negentropy of a marginally Gaussianized variable is zero,

$$\Delta I = J_m(\mathbf{R}\mathbf{\Psi}(\mathbf{x})) \geq 0, \forall \mathbf{R}$$

■

The above properties suggest the convergence of the proposed Gaussianization method. Property 3.1 (Eq. (12)) ensures that the distance between the PDF of the transformed variable to a zero mean unit covariance multivariate Gaussian is reduced in each iteration. Property 3.2 (Eq. (13)) ensures that redundancy among coefficients is also reduced after each iteration. According to this the distance to a Gaussian will decay to zero for a wide class of rotations.

D. On the rotation matrices

Admissible rotations are those that change the situation after marginal Gaussianization in such a way that J_m is increased. Using different rotation matrices gives rise to different properties of the algorithm.

The above Properties 3.1 and 3.2 provide some intuition on the suitable class of rotations. By using (12) and (13) in the sequence (5), one readily obtains the relations:

$$\Delta J_{(k)} = J_m(\mathbf{x}^{(k)}) = \Delta I_{(k-1)}, \tag{14}$$

and thus, interestingly, the amount of negentropy reduction (the convergence rate) at some iteration k will be determined by the amount of redundancy reduction obtained in the previous iteration, $k-1$. Since dependence can be analyzed in terms of correlation and non-Gaussianity [25], the intuitive candidates for $\mathbf{R}$ include orthonormal ICA, hereafter simply referred to as ICA, which maximizes the redundancy reduction; and PCA, which removes correlation. Random rotations (RND) will be considered here as an extreme case to point out that looking for interesting projections is not critical to achieve convergence. Note that other rotations are possible, for instance, a quite sensible choice would be randomly selecting projections that uniformly recover the surface of an hypersphere [33]. Other possibilities include extension to complex variables [34].

As an illustration, Table I summarizes the main characteristics of the method when using ICA, PCA and RND. The table analyzes the closed-form nature of each rotation, the theoretical convergence of the method, the convergence rate (negentropy reduction *per* iteration), and the computational cost of each rotation. Section V-A is devoted to the experimental confirmation of the reported characteristics of convergence presented here.

Using ICA guarantees the theoretical convergence of the Gaussianization process since it seeks for the maximally non-Gaussian marginal PDFs. Therefore, the negentropy reduction ΔJ (Eq. (12)) is always strictly positive, except for the case that the Gaussian PDF has been finally achieved. This is consistent with previously reported results [17]. Moreover, the convergence rate is optimal for ICA since it gives rise to the maximum $J_m(\mathbf{x})$ (indicated in Table I with 'Max ΔJ'). However, the main problem of using ICA as the rotation matrix is that it has no closed-form solution, so ICA algorithms typically resort to iterative procedures with either difficulties in convergence or high computational load.

Using PCA leads to sub-optimal convergence rate because it removes second-order redundancy (indicated in Table I with 'ΔJ = 2nd order'), but it does not maximize the marginal non-Gaussianity $J_m(\mathbf{x})$. Using PCA guarantees the convergence for every input PDF except for one singular case: consider a variable $\mathbf{x}^{(k)}$ which is not Gaussian but all its marginal PDFs are univariate Gaussian and with a unit covariance matrix. In this case, $\Delta J_{(k+1)} = J_m(\mathbf{x}^{(k)}) = 0$, i.e. no approximation to the Gaussian in negentropy terms is obtained in the next iteration. Besides, since $\mathbf{\Psi}_{(k+1)}(\mathbf{x}^{(k)}) = \mathbf{x}^{(k)}$, the next PCA, $\mathbf{R}_{(k+1)}$, will be the identity matrix, thus $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)}$: as a result, the algorithm may get stuck into a negentropy local minimum. In our experience, this undesired effect never happened in real datasets. On the other hand, advantages of using PCA is that the solution is closed-form, very fast, and even though the convergence rate is lower than for ICA, the solution is achieved in a fraction of the time.

Using RND transforms guarantees the theoretical convergence of the method since random rotations ensure that, even in the above considered singular case, the algorithm will not be stuck into this particular non-Gaussian solution. On the contrary, if the achieved marginal non-Gaussianity is zero after an infinite number of random rotations, it is because the desired Gaussian solution has been finally achieved (Cramer-Wold Theorem [38]). In practice, the above property of RND can be used as a way to check convergence when using other rotations (e.g. PCA): when the zero marginal non-Gaussianity situation is achieved, a useful safety check consists of including RND-based iterations. In the RND case, the convergence rate is clearly sub-optimal, yet non-negative ($\Delta J \geq 0$): the amount of negentropy reduction may take any value between zero and the maximum achieved by ICA. However, the method is much faster in practice: even though it may take more iterations to converge, the cost of each transform does not depend on the number of samples. The rotation matrix can be computed by fast orthonormalization techniques [37]. In this case, the computation time of the rotation is negligible compared to that of the marginal Gaussianization.

TABLE I
PROPERTIES OF THE GAUSSIANIZATION METHOD FOR DIFFERENT ROTATIONS (SEE COMMENTS IN THE TEXT).

Rotation	Closed -form	Theoretical convergence	Convergence rate	CPU cost† [35]–[37]
ICA	$\times$	$\surd$	Max ΔJ	$\mathcal{O}(2md(d+1)n)$
PCA	$\surd$	$\surd$	ΔJ = 2nd order	$\mathcal{O}(d^2(d+1)n)$
RND	$\surd$	$\surd$	$\Delta J \geq 0$	$\mathcal{O}(d^3)$

† Computational cost considers n samples of dimension d. The cost for the ICA transform is that of FastICA running m iterations.

IV. Relation to other methods

In this section we discuss the relation of RBIG to previously reported Gaussianization methods, including iterative PP-like techniques [15]–[17] and direct approaches suited for particular PDFs [19], [20], [39]. Additionally, relations to other machine learning tools, such as Support Vector Domain Description [21] and deep neural networks [22] are also considered.

A. Iterative Projection Pursuit Gaussianization

As stated above, the aim of Projection Pursuit (PP) techniques [15], [16] is looking for interesting linear projections according to some projection index measuring interestingness, and *after*, this interestingness is captured by removing it through the appropriate marginal equalization, thus making a step from structure to disorder. When interestingness or structure is defined by departure from disorder, non-Gaussianity or negentropy, PP naturally leads to iterative application of non-orthogonal ICA transforms followed by marginal Gaussianization, as in [17]:

$$\mathcal{G} : \mathbf{x}^{(k+1)} = \mathbf{\Psi}_{(k)}(\mathbf{R}^{\mathrm{ICA}} \cdot \mathbf{x}^{(k)}) \quad (15)$$

As stated in Section II, this is *Approach 1* to the Gaussian goal. Unlike PP, RBIG aims at the Gaussian goal following *Approach 2*. The differences between (15) and (5) (reverse order between the multivariate and the univariate transforms) suggest the different qualitative weight given to each counterpart. While PP gives rise to an *ordered* transition from structure to disorder[2], RBIG follows a *disordered* transition to disorder.

B. Direct (single-iteration) Gaussianization algorithms

Direct (non-iterative) Gaussianization approaches are possible if the method has to be applied to restricted classes of PDFs, for example: (1) PDFs that can be marginally Gaussianized in the *appropriate axes* [19], or (2) elliptically symmetric PDFs so that the final Gaussian can be achieved by equalizing the length (norm) of the whitened samples [20], [39].

The method proposed in [19] is useful when combined with tools that can identify marginally Gaussianizable components,

[2] In PP the structure of the unknown PDF in the input domain is progressively removed in each iteration starting from the most relevant projection and continuing by the second one, and so on, until total disorder (Gaussianity) is achieved.

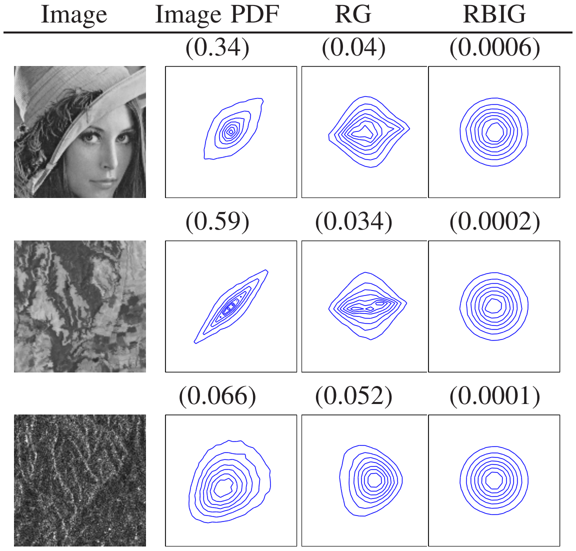

Fig. 2. Gaussianization of pairs of neighbor pixels from different images with RG and RBIG: natural image (top row), remote sensing Landsat channel in the optical range (middle row) and intensity of a ERS2 synthetic aperture radar (SAR) image (bottom row). Contour plots show the PDFs in the corresponding domains. The estimated mutual information (in bits) is given in parenthesis.

somewhat related to ICA transforms. Nevertheless, the use of alternative transformations is still an open issue. Erdogmus et al. proposed PCA, vector quantization or clustering as alternatives to ICA in order to find the most potentially 'Gaussianizable' components. In this sense, the method could be seen as a particular case of PP in that it only uses one iteration: first finding the most appropriate representation and then using marginal Gaussianization.

Elliptically symmetric PDFs constitute a relevant class of PDFs in image processing applications since this kind of functions is an accurate model of natural images (e.g. Gaussian Scale Mixtures [10] share this symmetry). Radial Gaussianization (RG) was specifically developed to deal with these particular kind of models [20]. This transform consists of a nonlinear function that acts radially, equalizing the histogram of the magnitude (energy) of the data to obtain the histogram of the magnitude of a Gaussian. Other methods have exploited this kind of transformation to generalize it to L_p symmetric distributions [39]. Obviously, elliptical symmetry is a fair assumption for natural images, but it may not be appropriate for other problems. Even in the image context, particular images may not strictly follow distributions with elliptical symmetry, therefore if RG-like transforms are applied to these images, they will give rise to non-Gaussianized data.

Figure 2 shows this effect in three types of acquired images: (1) a standard grayscale image, i.e. a typical example of a natural photographic image, (2) a band (in the visible range) of a remote sensing multispectral image acquired by the Landsat sensor, and (3) a ERS2 synthetic aperture radar (SAR) intensity image for the same scene (of course out of the visible range). In these illustrative examples, RG and RBIG were trained with the data distribution of pairs of neighbor pixels for each image, and RBIG was implemented using PCA rotations according to the results in Section V-A. Both RG and RBIG strongly reduce the mutual-information of pairs of neighbor pixels (see the mutual information values, in bits), but it is noticeable that RG is more effective, higher I reduction, in the natural image cases (photographic and visible channel images), in which the assumption of elliptically symmetric PDF is more reliable. However, it obviously fails when considering non-natural (radar) images, far from the visible range (I is not significantly reduced). The proposed method is more robust to these changes in the underlying PDF because no assumption is made.

C. Relation to Support Vector Domain Description

The Support Vector Domain Description (SVDD) is a one-class classification method that finds a minimum volume sphere in a kernel feature space that contains $1-\nu$ fraction of the *target* training samples [21]. The method tries to find the transformation (implicit in the kernel function) that maps the *target* data into a hypersphere. The proposed RBIG method and the SVDD method are conceptually similar due to their *apparent* geometrical similarity. However, RBIG and SVDD represent two different approaches to the one-class classification problem: PDF estimation versus separation boundary estimation. RBIG for one-class problems may be naively seen as if test samples were transformed and classified as *target* if lying inside the sphere containing $1-\nu$ fraction of the learned Gaussian distribution. According to this interpretation, both methods reduce to the computation of spherical boundaries in different feature spaces. However, this is not true in the RBIG case: note that the value of the RBIG Jacobian is not the same at every location in the Gaussianized domain. Therefore, the optimal boundary to reject a ν fraction of the training data is not necessarily a sphere in the Gaussianized domain. In the case of the SVDD, though, by using an isotropic RBF kernel, all directions in the kernel feature spaces are treated in the same way.

D. Relation to Deep Neural Networks

RBIG is essentially an iterated sequence of two operations: non-linear dimension-wise squashing functions and linear transforms. Intuitively, these are the same processing blocks used in a feedforward neural network (linear transform plus sigmoid-shaped function in each hidden layer). Therefore, one could see each iteration as one hidden layer processing of the data, and thus argue that complex (highly non-Gaussian) tasks should require more hidden layers (iterations). This view is in line with the field of *deep learning* in neural networks [22], which consists of learning a model with several layers of nonlinear mappings. The field is very active nowadays because some tasks are highly nonlinear and require accurate design of processing steps of different complexity. Note, that it may appear counterintuitive the fact that full Gaussianization of a dataset is eventually achieved with a large enough number of iterations, thus leading to overfitting in the case of a neural network with such number of layers. Nevertheless, note that capacity control also applies in RBIG: we have observed that early-stopping criteria must be applied to allow good generalization properties. In this setting, one can see early

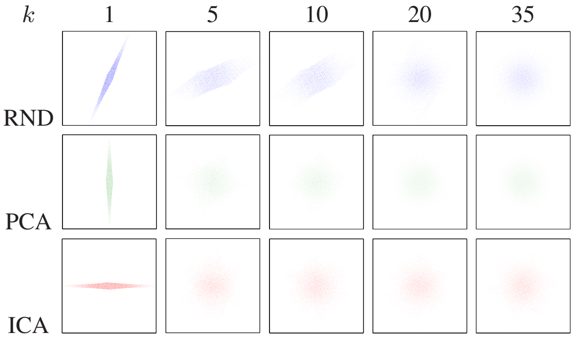

Fig. 3. Scatter plots of a 2D data in different iterations for the considered rotation matrices: RND (top), PCA (middle) and ICA (bottom).

TABLE II
AVERAGE ($\pm$ STD. DEV.) CONVERGENCE RESULTS.

	RND		PCA		ICA	
Dim.	iterations	time [s]	iterations	time [s]	iterations	time [s]
2	14 ± 3	0.01 ± 0.01	7 ± 3	0.005 ± 0.002	3 ± 1	6 ± 5
4	44 ± 6	0.06 ± 0.01	33 ± 6	0.05 ± 0.01	11 ± 1	564 ± 223
6	68 ± 7	0.17 ± 0.01	43 ± 12	0.1 ± 0	11 ± 2	966 ± 373
8	92 ± 4	0.3 ± 0.1	54 ± 23	0.2 ± 0	16 ± 1	1905 ± 534
10	106 ± 10	0.4 ± 0	58 ± 25	0.3 ± 0.1	19 ± 1	2774 ± 775
12	118 ± 10	0.5 ± 0.2	44 ± 5	0.2 ± 0.1	21 ± 2	3619 ± 323
14	130 ± 8	0.7 ± 0.1	52 ± 21	0.4 ± 0.1	19 ± 1	4296 ± 328
16	139 ± 10	0.7 ± 0	73 ± 36	0.4 ± 0.2	22 ± 1	4603 ± 932

stopping in the Gaussianization method as a form of model regularization. This is certainly an interesting research line to be pursued in the future.

Finally, we would like to note that it does not escape our notice that the exploitation of the RBIG framework in the previous contexts might eventually be helpful in designing new algorithms or helping understanding them from different theoretical perspectives. This is of course out of the scope of this paper.

V. Experimental Results

This section shows the capabilities of the proposed RBIG methods in some illustrative examples. We start by experimentally analyzing the convergence of the method depending on the rotation matrix in a controlled toy dataset, and give useful criteria for early-stopping. Then, method's performance is illustrated for mutual information estimation, image synthesis, classification and denoising. In each application, results are compared to standard methods in the particular field. A documented Matlab implementation is available at http://www.uv.es/vista/vistavalencia/RBIG.htm.

A. Method convergence and early-stopping

The RBIG method is analyzed here in terms of convergence rate and computational cost for different rotations: orthonormal ICA, PCA and RND. Synthetic data of varying dimensions ($d = 2, \ldots, 16$) was generated by first sampling from a uniform distribution hypercube and then applying a rotation transform. This way we can compute the ground-truth negentropies of the initial distributions, and estimate the reduction in negentropies in every iteration by estimating the difference in marginal negentropies, cf. Eq. (13). A total of 10000 samples was used for the methods, and we show average and standard deviation results for 5 independent random realizations.

Two-dimensional scatter plots in Figure 3 qualitatively show that different rotation matrices give rise to different solutions in each iteration but, after a sufficient number of iterations, all of them transform the data into a Gaussian independently of the rotation matrix.

RBIG convergence rates are illustrated in Fig. 4. Top plots show the negentropy reduction for the different rotations as a function of the number of iterations and data dimension. We also give the actual negentropy estimated from the samples, is an univariate population estimate since Eq. (12) can be used. Successful convergence is obtained when the accumulated reduction in negentropy tends to the actual negentropy value (cyan line). Discrepancies are due to the accumulation of computational errors in the negentropy reduction estimation in each iteration.

Bottom plots in Fig. 4 give the result of the multivariate Gaussianity test in [40]: when the outcome of the test is 1, it means accepting the hypothesis of multidimensional Gaussianity. Several conclusions can be extracted: (1) the method converges to a multivariate Gaussian independently of the rotation matrix; (2) ICA requires a less number of iterations to converge, but it is closely followed by PCA; (3) random rotations take a higher number of iterations to converge and show high-variance in the earlier iterations; and (4) convergence in cumulative negentropy is consistent with the parametric estimator in [40] which, in turn, confirms the analysis in Table I.

Despite the previous conclusions, and as pointed out before, in practical applications, it is not the length of the path to the Gaussian goal what matters, but the time required to complete this path. Table II compares the number of iterations for appropriate convergence and the CPU time of 5 realizations of RBIG with different matrix rotations (RND, PCA and ICA) in several dimensions. While, in general, CPU time results are obviously implementation dependent, note that results in Table II are fairly consistent with the computational burden per iteration shown in Table I since each ICA computation is an iterative procedure itself which needs m iterations.

The use of ICA rotations critically increases the convergence time. This effect is more noticeable as the dimension increases, thus making the use of ICA computationally unfeasible when the number of dimensions is moderate or high. The use of PCA in RBIG is consequently a good trade-off between Gaussianization error and computational cost if the number of iterations is properly chosen. An early-stopping criterion could be based on the evolution of the cumulative negentropy reduction, or of a multivariate test of Gaussianity such as the one used here [40]. Both are sensible strategies for early-stopping. According to the observed performance, we restrict ourselves to the use of PCA as the rotation matrix in the experiments hereafter. Note that by using PCA, the algorithm might not converge in a singular situation (see Section III-D). However, we checked that such singular situation never happened by jointly using both criteria in each iteration.

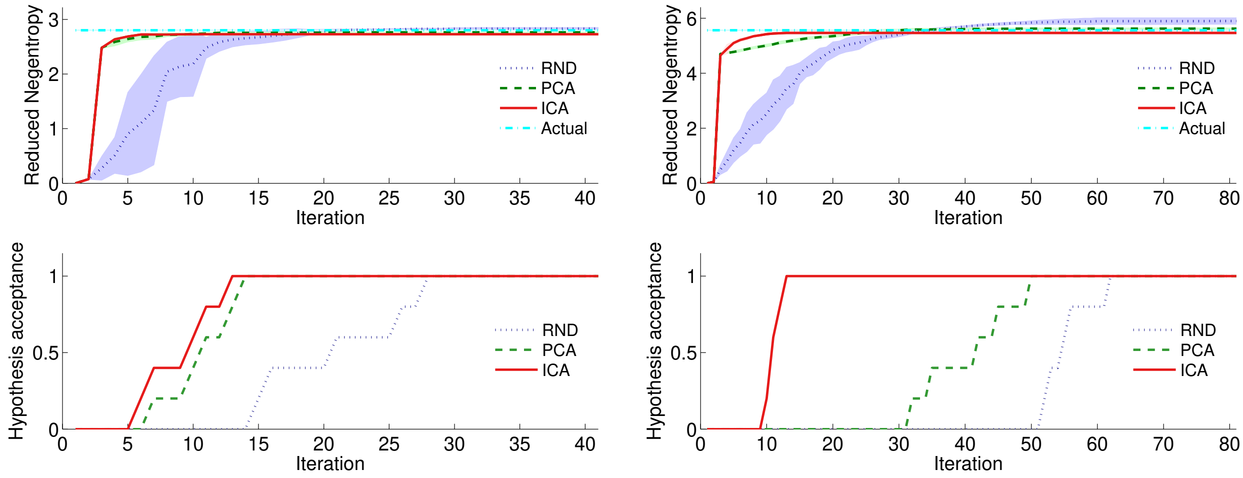

Fig. 4. Cumulative negentropy reduction (top) and multivariate Gaussian significance test (bottom) for each iteration in 2D (left) and 4D (right) dimensional synthetic problem. Average and standard deviation results from 5 realizations is shown.

TABLE III
AVERAGE ($\pm$ STD. DEV.) MULTI-INFORMATION (IN BITS) FOR THE DIFFERENT ESTIMATORS IN 2D PROBLEMS.

DIST	EG	GG	UU
RBIG	0.49 ± 0.01	1.38 ± 0.004	0.36 ± 0.03
NE	0.35 ± 0.02	1.35 ± 0.006	0.39 ± 0.002
RE	0.32 ± 0.01	1.29 ± 0.004	0.30 ± 0.002
Actual	0.51	1.38	0.45

TABLE IV
MULTI-INFORMATION (IN BITS) WITH RBIG IN DIFFERENT d-DIMENSIONAL PROBLEMS.

Dim.	EG		GG		UU	
d	RBIG	Actual	RBIG	Actual	RBIG	Actual
3	1.12 ± 0.03	1.07	1.91 ± 0.01	1.9	1.6 ± 0.1	1.6
4	5 ± 0.1	5.04	1.88 ± 0.02	1.86	2.2 ± 0.1	2.2
5	4.7 ± 0.1	4.82	1.77 ± 0.02	1.75	2.7 ± 0.1	2.73
6	7.8 ± 0.1	7.9	2.11 ± 0.01	2.08	3.5 ± 0.1	3.72
7	6.2 ± 0.1	6.33	2.68 ± 0.03	2.65	3.6 ± 0.1	3.92
8	8.1 ± 0.1	8.19	2.72 ± 0.02	2.68	4.1 ± 0.1	4.29
9	9.5 ± 0.1	9.6	3.22 ± 0.02	3.18	5.3 ± 0.1	5.69
10	12.7 ± 0.1	13.3	3.45 ± 0.03	3.4	5.8 ± 0.2	6.24

B. Multi-information Estimation

As previously shown, RBIG can be used to estimate the negentropy, and therefore could be used to compute multi-information (I) of high dimensional data (Eq. (2)). Essentially, one learns the sequence of transforms to Gaussianize a given dataset, and the I estimate reduces to compute the cumulative ΔI since, at convergence, full independence is supposedly achieved. We illustrate the ability of RBIG in this context by estimating multi-information in three different synthetic distributions with known I: uniform distribution (UU), Gaussian distribution (GG), and a marginally composed exponential and Gaussian distribution (EG). An arbitrary rotation was applied in each case to obtain non-zero multi-information. In all cases, we used $10,000$ samples and repeated the experiments for 10 realizations. Two kinds of experiments were performed:

- A 2D experiment, where RBIG results can be compared to the results of naive (histogram-based) mutual information estimates (NE), and to previously reported 2D estimates such as the Rudy estimate (RE) [41] (see Table III).
- A set of d-dimensional experiments, where RBIG results are compared to actual values (see Table IV).

Table III shows the results (in bits) for the mutual information estimation in the 2D experiment to standard approaches. The ground-truth result is also given for comparison purposes.

For Gaussian and exponential-Gaussian data distributions, RBIG outperforms the rest of methods, but when data are marginally uniform, NE yields better estimates. Table IV extends the previous results to multidimensional cases, and compares RBIG to the actual I. Good results are obtained in all cases. Absolute errors slightly increase with data dimensionality.

C. Data Synthesis

RBIG obtains an invertible Gaussianization transform that can be used to generate (or synthesize) samples. The approach is simple: the transform $\mathcal{G}$ is *learned* from the available training data, and then synthesized samples are obtained from random Gaussian samples in the transformed domain inverted back to the original domain using $\mathcal{G}^{-1}$. Two examples are given here to illustrate the capabilities of the method.

1) Toy data: Figure 5 shows examples of 2D non-Gaussian distributions (left column) transformed into a Gaussian (center column). The right column was obtained sampling data from a zero mean unit covariance Gaussian and inverting back the transform. This example visually illustrates that synthesized data approximately follow the original PDF.

2) Face synthesis: In this experiment, $2,500$ face images were extracted from [42], eye-centered, cropped to have the same dimensions, mean and variance adjusted, and resized to 17×15 pixels. Images were then reshaped to 255-dimensional vectors, and Gaussianized with RG and RBIG. Figure 6 shows illustrative examples of original and synthesized faces with RG and RBIG.

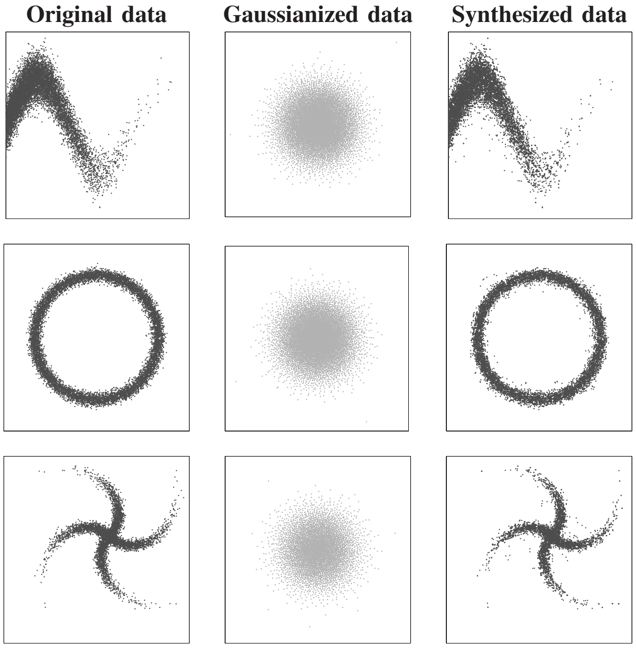

Fig. 5. Toy data examples synthesized using RBIG.

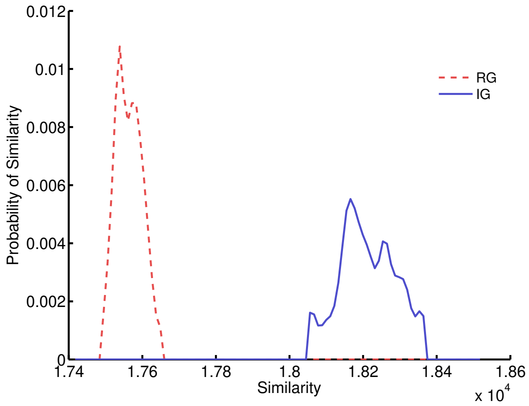

Fig. 7. Histogram of the similarity (inner product) between the distribution of original and synthesized face images for 300 realizations. For reference, the average image energy (average inner product) in the original set is $1.81 \cdot 10^4$.

Note that both methods achieve good visual qualitative performance. In order to assess performance quantitatively, we compared 200 actual and synthesized images using the inner product as a measure of local similarity. We averaged this similarity measure over 300 realizations and show the histograms for RG and RBIG. Results suggest that the distribution of the samples generated with RBIG is more realistic (similar to the original dataset) than the obtained with RG.

D. One-class Classification

In this experiment, we assess the performance of the RBIG method as one-class classifier. Performace is illustrated in the challenging problem of detecting urban areas from multispectral and SAR images. The ground-truth data for the images used in this section were collected in the Urban Expansion Monitoring (UrbEx) ESA-ESRIN DUP project[3] [43]. The considered test sites were the cities of Rome and Naples, Italy, for two acquisitions dates (1995 and 1999). The available features were the seven Landsat bands, two SAR backscattering intensities (0–35 days), and the SAR interferometric coherence. We also used a spatial version of the coherence specially designed to increase the urban areas discrimination [43]. After this preprocessing, all features were stacked at a pixel level, and each feature was standardized.

We compared the RBIG classifier based on the estimated PDF for urban areas with the SVDD classifier [21]. We used the RBF kernel for the SVDD whose width was varied in the range $\sigma \in [10^{-2}, \ldots, 10^2]$. The fraction rejection parameter was varied in $\nu \in [10^{-2}, 0.5]$ for both methods. The optimal parameters were selected through 3-fold cross-validation in the training set optimizing the κ statistic [44]. Training sets of different size for the target class were used in the range $[500, 2500]$. We assumed a scarce knowledge of the non-target class: 10 outlier examples were used in all cases. The test set was constituted by $10,000$ pixels of each considered image. Training and test samples were randomly taken from the whole spatial extent of each image. The experiment was repeated for 10 different random realizations in the three considered test sites.

Figure 8 shows the estimated κ statistic and the overall accuracy (OA) in the test set achieved by SVDD and RBIG in the three images. The κ scores are relatively small because samples were taken from a large spatial area thus giving rise to a challenging problem due to the variance of the spectral signatures. Results show that SVDD behavior is similar to the proposed method for small size training sets. This is because more target samples are needed by the RBIG for an accurate PDF estimation. However, for moderate and large training sets the proposed method substantially outperforms SVDD. Note that training size requirements of RBIG are not too demanding: using 750 samples in a 10-dimensional problem is enough for RBIG to outperform SVDD when very little is known about the non-target class.

Figure 9 shows the classification maps for the representative Naples95 scene for SVDD and RBIG. Note that RBIG better rejects the ‘non-urban’ areas (in black). This may be because SVDD training with few non-target data gives rise to a too broad boundary. As a result, too many pixels are identified as belonging to the target class (in white). Another relevant observation is the noise in neighboring pixels, which may come from the fact that no spatial information was used. This problem could be easily alleviated by imposing some post-classification smoothness constraint or by incorporating spatial texture features.

E. Image Denoising

Image denoising tackles the problem of estimating the underlying image, $\mathbf{x}$, from a noisy observation, $\mathbf{x}_n$, assuming an additive degradation model: $\mathbf{x}_n = \mathbf{x} + \mathbf{n}$. Many methods have exploited the Bayesian framework to this end [10], [45]–

[3] http://dup.esrin.esa.int/ionia/projects/summaryp30.asp

Fig. 6. Example of real (top) and synthesized faces with RG (middle) and RBIG (bottom).

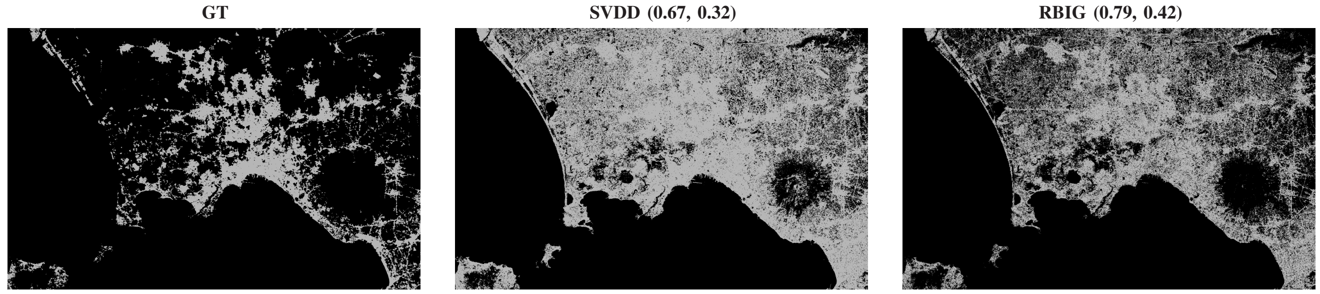

Fig. 9. Ground-truth (GT) and classification maps obtained with SVDD and RBIG for the Naples 1995 scene. The white points represent urban area and the black points represent non-urban area. The corresponding overall accuracy and κ-statistic are given in parenthesis.

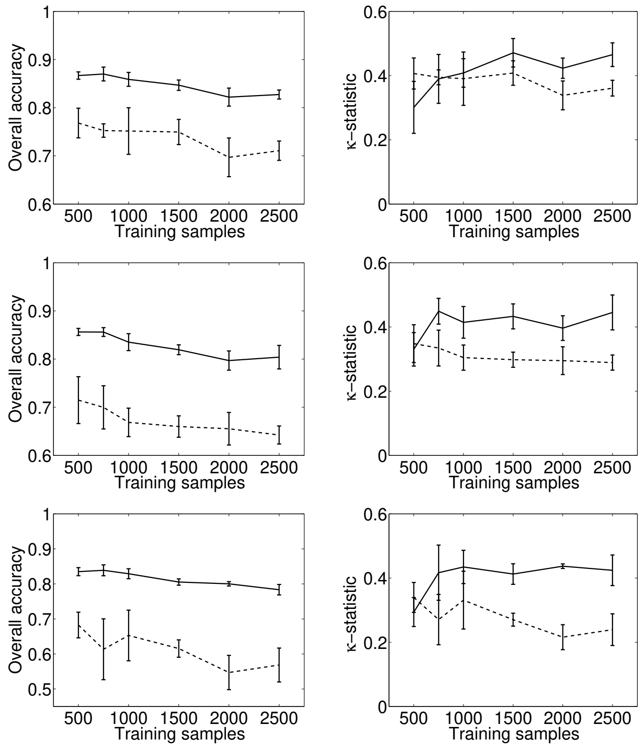

Fig. 8. Overall accuracy (left) and kappa statistic, κ (right) for RBIG (solid line) and SVDD (dashed line) in different scenes: Naples 1995 (top), Naples 1999 (center) and Rome 1995 (bottom).

[47]:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}^*}{\operatorname{argmin}} \left\{ \int \mathcal{L}(\mathbf{x}, \mathbf{x}^*) p(\mathbf{x}|\mathbf{x}_n) d\mathbf{x} \right\}, \tag{16}$$

where $\mathbf{x}^*$ is the candidate image, $\mathcal{L}(\mathbf{x}, \mathbf{x}^*)$ is the cost function, and $p(\mathbf{x}|\mathbf{x}_n)$ is the posterior probability of the original sample $\mathbf{x}$ given the noisy sample $\mathbf{x}_n$. This last term plays an important role since it can be decomposed (using the Bayes rule) as

$$p(\mathbf{x}|\mathbf{x}_n) = Z^{-1} p(\mathbf{x}_n|\mathbf{x}) p(\mathbf{x}), \tag{17}$$

where Z^{-1} is a normalization term, $p(\mathbf{x}_n|\mathbf{x})$ is the noise model (probability of the noisy sample given the original one), and $p(\mathbf{x})$ is the prior (marginal) sample model.

Note that, in this framework, the inclusion of a feasible image model, $p(\mathbf{x})$, is critical in order to obtain a good estimation of the original image. Images are multidimensional signals whose PDF $p(\mathbf{x})$ is hard to estimate with traditional methods. The conventional approach consists of using parametric models to be plugged into Eq. (17) in such a way that the problem can be solved analytically. However, mathematical convenience leads to the use of too rigid image models. Here we use RBIG in order to estimate the probability model of natural images $p(\mathbf{x})$.

In this illustrative example, we use the L_2-norm as cost function, $\mathcal{L}(\mathbf{x}, \mathbf{x}^*) = ||\mathbf{x} - \mathbf{x}^*||_2$, and an additive Gaussian noise model, $p(\mathbf{x}_n|\mathbf{x}) = \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I})$. We estimated $p(\mathbf{x})$ using 100 achromatic images of size 256×256 extracted from the McGill Calibrated Colour Image Database[4]. To do this, images were transformed using orthonormal QMF wavelet domain with four frequency scales [48], an then each subband was converted to patches in order to obtain different PDF models for each subband according to well-known properties of natural images in wavelet domains [12], [49]. In order to evaluate Eq. (16), we sampled the posterior PDF at $8,000$

[4]http://tabby.vision.mcgill.ca/

points from the neighborhood of each wavelet coefficient by generating samples with the PDF of the noise model ($p(\mathbf{x}_n|\mathbf{x})$), and evaluated the probability for each sample with the PDF obtained in the training step $p(\mathbf{x})$. The estimated coefficient $\hat{\mathbf{x}}$ is obtained as the expected value over the 8000 samples of the posterior PDF. Obtaining the expected value is equivalent to using the L_2 norm [50]. Note that the classical hard-thresholding (HT) and soft-thresholding (ST) results [45] are a useful reference since they can be interpreted as solutions to the same problem with a marginal Laplacian image model and L_1 and L_2 norms respectively [46].

Figure 10 shows the denoising results for the 'Barbara' image corrupted with Gaussian noise of $\sigma_n^2 = 100$ using marginal models (HT and ST), and using a RBIG as the PDF estimator. Accuracy of the results is measured in Euclidean terms (RMSE), and using a perceptually meaningful image quality metric such as the Structural Similarity Index (SSIM) [51]. Note that RBIG method obtains better results (numerically and visually) than the classical methods due to the more accurate PDF estimation.

VI. Conclusions

In this work, we proposed an alternative solution to the PDF estimation problem by using a family of Rotation-based Iterative Gaussianization (RBIG) transforms. The proposed procedure looks for differentiable transforms to a Gaussian so that the unknown PDF can be computed at any point of the original domain using the Jacobian of the transform.

The RBIG transform consists of the iterative application of univariate marginal Gaussianization followed by a rotation. We show that a wide class of orthonormal transforms (including trivial random rotations) is well suited to Gaussianization purposes. The freedom to choose the most convenient rotation is the difference with formally similar techniques, such as Projection Pursuit, focused on looking for interesting projections (which is an intrinsically more difficult problem). In this way, here we propose to shift the focus from ICA to a wider class of rotations since interesting projections as found by ICA are not critical to solve the PDF estimation problem in the original domain. The suitability of multiple rotations to solve the PDF estimation problem may help to revive the interest of classical iterative Gaussianization in practical applications. As an illustration, we showed promising results in a number of multidimensional problems such as image synthesis, classification, denoising, and multi-information estimation.

Particular issues in each of the possible applications, such as stablishing a convenient family of rotations for a good Jacobian or convenient criteria to ensure the generalization ability, are a matter for future research.

References

[1] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Press, Boston, 1992.

[2] M. Banham and A. Katsaggelos, "Digital image restoration," *IEEE Sig.Proc.Mag.*, vol. 14, pp. 24–41, 1997.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*, Wiley-Interscience, 2 edition, November 2000.

[4] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*, Springer, August 2003.

[5] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int. J. Comp. Vis.*, vol. 40, no. 1, pp. 49–71, 2000.

[6] David W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization (Wiley Series in Probability and Statistics)*, Wiley-Interscience, September 1992.

[7] I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, 1986.

[8] A. Hyvärinen, "Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation," *Neur. Comp.*, pp. 1739–1768, 1999.

[9] R.W. Buccigrossi and E.P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Tr. Im. Proc.*, vol. 8, no. 12, pp. 1688–1701, 1999.

[10] J Portilla, V Strela, M Wainwright, and E P Simoncelli, "Image denoising using a scale mixture of Gaussians in the wavelet domain," *IEEE Trans Image Processing*, vol. 12, no. 11, pp. 1338–1351, November 2003.

[11] J. Gutiérrez, F. Ferri, and J. Malo, "Regularization operators for natural images based on nonlinear perception models," *IEEE Tr. Im. Proc.*, vol. 15, no. 1, pp. 189–200, 2006.

[12] V. Laparra, J. Gutiérrez, G. Camps-Valls, and J. Malo, "Image denoising with kernels based on natural image relations," *Journal of Machine Learning Research*, vol. 11, pp. 873–903, Feb. 2010.

[13] J. Malo, I. Epifanio, R. Navarro, and E. Simoncelli, "Non-linear image representation for efficient perceptual coding," *IEEE Trans. Im. Proc.*, vol. 15, no. 1, pp. 68–80, 2006.

[14] G. Camps-Valls, J. Gutiérrez, G. Gómez, and J. Malo, "On the suitable domain for SVM training in image coding," *Journal of Machine Learning Research*, vol. 9, pp. 49–66, 2008.

[15] J.H. Friedman and J.W. Tukey, "A projection pursuit algorithm for exploratory data analysis," *IEEE Trans. Comp.*, vol. C-23, no. 9, pp. 881–890, 1974.

[16] P. J. Huber, "Projection pursuit," *The Annals of Statistics*, vol. 13, no. 2, pp. 435–475, 1985.

[17] S.S Chen and R.A. Gopinath, "Gaussianization," in *NIPS*, 2000, pp. 423–429.

[18] E. Rodriguez-Martinez, J.Y. Goulermas, Tingting Mu, and J.F. Ralph, "Automatic induction of projection pursuit indices," *Neural Networks, IEEE Transactions on*, vol. 21, no. 8, pp. 1281 –1295, 2010.

[19] D. Erdogmus, R. Jenssen, Y. Rao, and J. Principe, "Gaussianization: An efficient multivariate density estimation technique for statistical signal processing," *J. VLSI Sig. Proc.*, vol. 45(17), pp. 67–83, 2006.

[20] S Lyu and E P Simoncelli, "Nonlinear extraction of 'independent components' of natural images using Radial Gaussianization," *Neur. Comp.*, 2009, Accepted for publ., 10/08.

[21] David Tax and Robert P.W. Duin, "Support vector domain description," *Pattern Recognition Letters*, vol. 20, pp. 1191–1199, 1999.

[22] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, July 2006.

[23] H. Stark and J.W. Woods, *Probability, Random Processes and Estimation Theory for Engineers*, Prentice-Hall, Englewood Cliffs, New Jersey, 1986.

[24] P. Comon, "Independent component analysis: A new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[25] Jean-François Cardoso, "Dependence, correlation and gaussianity in independent component analysis," *J. Mach. Learn. Res.*, vol. 4, pp. 1177–1203, 2003.

[26] M. Studeny and J. Vejnarova, *The Multi-information function as a tool for measuring stochastic dependence*, pp. 261–298, Kluwer, January 1998.

[27] E.G. Learned and J.W. Fisher, "ICA using spacings estimates of entropy," *J. Mach.Learn.Res.*, vol. 4, pp. 1271–1295, 2003.

[28] A. Hyvärinen, "Fast and robust fixed-point algorithms for ICA," *IEEE Trans. Neur. Net.*, vol. 10, pp. 626–634, 1999.

[29] K. Zhang and L.W. Chan, "Extended gaussianization method for blind separation of post-nonlinear mixtures," *Neur. Comp.*, vol. 17, no. 2, pp. 425–452, 2005.

[30] S. Squartini, A. Bastari, and F. Piazza, "A practical approach based on Gaussianization for post-nonlinear underdetermined BSS," in *2006 ICCCSP*, Orlando, Florida, 2006.

[31] B. Xiang, U.V. Chaudhari, G.N. Ramaswamy, and R.A. Gopinath, "Short-time gaussianization for robust speaker verification," in *IEEE ICASSP 02'*, Orlando, Florida, 2002.

[32] B. W. Silverman, *Density estimation for statistics and data analysis*, Chapman & Hall/CRC Monographs on Statistics and Applied Probability, 1986.

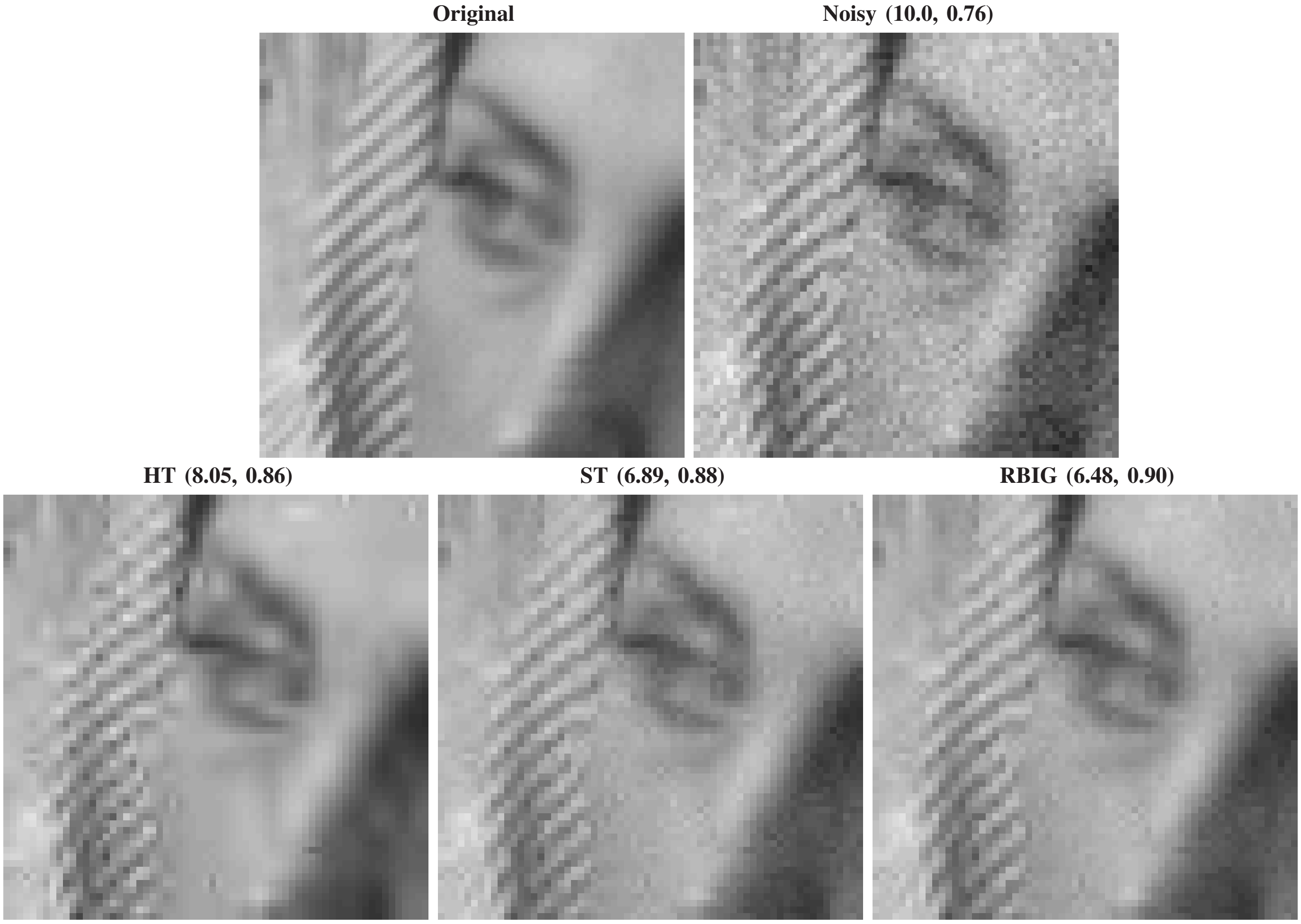

Fig. 10. Original, noisy (noise variance $\sigma_n^2 = 100$) and restored 'Barbara' images. The root-mean-square-error (RMSE) and the perceptually meaningful Structural Similarity Measure (SSIM) [51] are given in parentheses.

[33] Carlos A. León, Jean-Claude Massé, and Louis-Paul Rivest, "A statistical model for random rotations," *J. Multivar. Anal.*, vol. 97, no. 2, pp. 412–430, 2006.

[34] M. Novey and T. Adali, "Complex ICA by negentropy maximization," *Neural Networks, IEEE Transactions on*, vol. 19, no. 4, pp. 596 –609, 2008.

[35] Vicente Zarzoso, Pierre Comon, and Mariem Kallel, "How fast is fastica," in *In: Proc. EUSIPCO-2006, XIV European Signal Processing Conference*, 2006.

[36] A. Sharma and K. Paliwal, "Fast principal component analysis using fixed-point algorithm," *Pattern Recognition Letters*, vol. 28, no. 10, pp. 1151–1155, July 2007.

[37] Gene H. Golub and Charles F. Van Loan, *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)(3rd Edition)*, The Johns Hopkins University Press, 3rd edition, October 1996.

[38] W. Feller, *An Introduction to Probability Theory and Its Applications, Vol. 1*, Wiley, 3 edition, January 1968.

[39] J. Eichhorn, F. Sinz, and M. Bethge, "Natural image coding in v1: How much use is orientation selectivity?," *PLoS Comput Biol*, vol. 5, no. 4, pp. e1000336+, April 2009.

[40] Gábor J. Székely and Maria L. Rizzo, "A new test for multivariate normality," *J. Multivar. Anal.*, vol. 93, no. 1, pp. 58–80, 2005.

[41] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," *Signal Processing*, vol. 16, no. 3, pp. 233–246, 1989.

[42] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 643–660, 2001.

[43] L. Gómez-Chova, D. Fernández-Prieto, J. Calpe, E. Soria, J. Vila-Francés, and G. Camps-Valls, "Urban monitoring using multitemporal SAR and multispectral data," *Pattern Recognition Letters*, vol. 27, no. 4, pp. 234–243, Mar 2006, 3rd Pattern Recognition in Remote Sensing Workshop, Kingston Upon Thames, ENGLAND, AUG 27, 2004.

[44] Jacob Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, April 1960.

[45] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage.," *J. Am. Stat. Assoc.*, vol. 90, pp. 432, 1995.

[46] E P Simoncelli, "Bayesian denoising of visual images in the wavelet domain," in *Bayesian Inference in Wavelet Based Models*, P Müller and B Vidakovic, Eds., pp. 291–308. Springer-Verlag, New York, Spring 1999.

[47] Robert D. Nowak Mario A. T. Figueiredo, "Wavelet-based image estimation: an empirical bayes approach using jeffrey's noninformative prior.," *IEEE Transactions on Image Processing*, vol. 10, 2001.

[48] E.P. Simoncelli and E.H. Adelson, *Subband Image Coding*, chapter Subband Transforms, pp. 143–192, Kluwer Academic Publishers, Norwell, MA, 1990.

[49] Juan Liu and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Transactions on Image Processing*, vol. 10, pp. 1647–1658, 2001.

[50] José M Bernardo and Adrian F M Smith, "Bayesian theory," *Measurement Science and Technology*, vol. 12, no. 2, pp. 221, 2001.

[51] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.