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 paper, 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 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, one-class support vector domain description, and deep neural networks 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, multi-information, negentropy, principal component analysis, 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.

Manuscript received April 3, 2010; revised December 13, 2010; accepted December 15, 2010. This work was supported in part by the Project CICYT-FEDER TEC2009-13696, Project AYA2008-05965-C04-03, and Project CSD2007-00018, and the Grant BES-2007-16125.

The authors are with the Image Processing Laboratory, Universitat de València, Paterna 46980, Spain (e-mail: valero.laparra@uv.es; gustavo.camps@uv.es; jesus.malo@uv.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNN.2011.2106511

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 discrete cosine transform in many natural signals such as speech, images and video, assumes a Gaussian source [3], [7]. More recently, linear independent component analysis (ICA), which 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 dependences 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]. Nonlinear methods using nonexplicit 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 paper, 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, it 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 paper, 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 irrespective of the selected class of rotations.

The remainder of this paper is organized as follows. In 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 and differentiable, and converges for a wide class of orthonormal transforms, even including random rotations (RND). 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 $\mathbf{y}$, $\mathcal{G} : \mathbb{R}^d \to \mathbb{R}^d$, 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 *negentropy*,[1] $J(\mathbf{x}) = \mathrm{D_{KL}}\left(p(\mathbf{x})|\mathcal{N}(\mathbf{0}, \mathbf{I})\right)$. Negentropy can be decomposed as the sum of two nonnegative 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 [26, Eq. (5)], by considering as contrast PDF $\prod_i q_i(x_i) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The multi-information is [26]

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

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

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

Given a data distribution from the unknown PDF, in general both I and J_m will be nonzero. 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, Ψ. This is easy as will be shown in the next section.
   However, this is an ambitious approach, since looking for independent components is a nontrivial (intrinsically multivariate and nonlinear) problem. According to this, linear ICA techniques will not succeed in completely removing the multi-information, and thus a nonlinear postprocessing is required.
2) *Reducing J_m*: As stated above, this univariate problem is easy to solve by using the appropriate Ψ. 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 PP methods focused on looking for interesting projections [16], [17]. Since the core of these methods is looking for meaningful projections

[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.

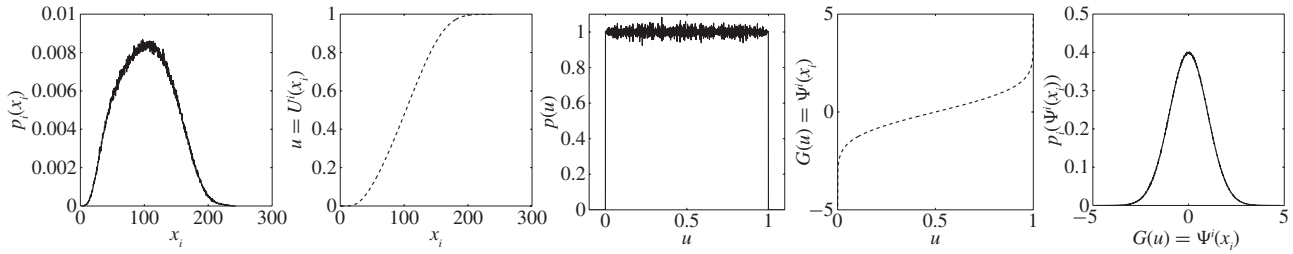

Fig. 1. Example of marginal Gaussianization in some particular dimension i. From left to right: marginal PDF of x_i, uniformization transform $u = U^i(x_i)$, PDF of the uniformized variable $p(u)$, Gaussianization transform $G(u)$, and PDF of the Gaussianized variable $p_i(\Psi^i(x_i))$.

(usually ICA algorithms), they suffer from a big computational complexity, for example, robust ICA algorithms such as RADICAL [27] would lead to extremely slow Guassanization algorithms, whereas relatively more convenient alternatives such as FastICA [28] may not converge in all cases. This may explain why, so far, Gaussianization techniques have been applied just to low-dimensional (audio) signals in either simple contexts based on pointwise nonlinearities [29], [30], or after *ad hoc* speech-oriented feature extraction steps [31]. In this paper, 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

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 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 intuition 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 (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 nonlinear 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, (1). Invertibility guarantees that the transform is bijective, which is a necessary condition to apply (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, i.e., $\mathbf{\Psi}^i_{(k)}$, 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

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

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

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

1-D 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, $\mathbf{R}^{-1}_{(k)} = \mathbf{R}^{\top}_{(k)}$. 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, i.e., 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 (7), each element in $\nabla_{\mathbf{x}^{(k)}} \mathbf{\Psi}_{(k)}$ can be easily computed by applying the chain rule on u defined in (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 of the selected rotations $\mathbf{R}_{(k)}$.

C. Convergence Properties

Here we prove two general properties of random variables that 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

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

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

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

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

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

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

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

■

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

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

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

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

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

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

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

■

The above properties suggest the convergence of the proposed Gaussianization method. Property 3.1 (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 (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 following 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. 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

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	×	√	Max ΔJ	$\mathcal{O}(2md(d+1)n)$
PCA	√	√	ΔJ = 2nd order	$\mathcal{O}(d^2(d+1)n)$
RND	√	√	$\Delta J \geq 0$	$\mathcal{O}(d^3)$

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

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 (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 suboptimal 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 $\Psi_{(k+1)}(\mathbf{x}^{(k)}) = \mathbf{x}^{(k)}$, the next PCA, i.e., $\mathbf{R}_{(k+1)}$, 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 and 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 RND 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 RND, 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 suboptimal, yet nonnegative ($\Delta J \geq 0$), i.e., 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.

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 PP techniques [15], [16] is looking for interesting linear projections according to some projection index measuring interestingness, and then 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]

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

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, 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

[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.

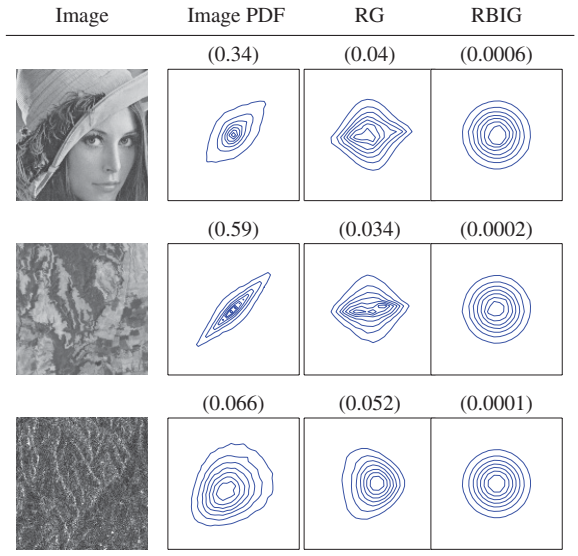

Fig. 2. Gaussianization of pairs of neighbor pixels from different images with RG and RBIG: natural image (top row), remote sensing landsat channel in the optical range (middle row), and intensity of a ERS2 synthetic aperture radar (SAR) image (bottom row). Contour plots show the PDFs in the corresponding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

"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] and related models [40] share this symmetry). Radial Gaussianization (RG) was specifically developed to deal with these particular kinds 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.

Fig. 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 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 and has 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 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, nonlinear 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 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 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

TABLE II
AVERAGE (± 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

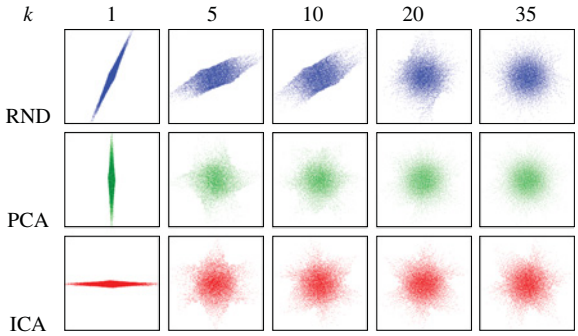

Fig. 3. Scatter plots of 2-D data in different iterations for the considered rotation matrices: (top) RND, (middle) PCA, and (bottom) ICA.

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, the 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: i.e., orthonormal ICA, PCA, and RND. Synthetic data of varying dimensions ($d = 2, \ldots, 16$) was generated by first sampling from a uniform distribution (UU) 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, (13). A total of 10000 samples was used for the methods, and we show average and standard deviation results for five independent random realizations.

2-D scatter plots in Fig. 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. The 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, which is a univariate population estimate since (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.

The bottom plots in Fig. 4 give the result of the multivariate Gaussianity test in [41], 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 fewer iterations to converge, but it is closely followed by PCA; 3) RND 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 [41] 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 that matters, but the time required to complete this path. Table II compares the number of iterations for appropriate convergence and the CPU time of five 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 itself is an iterative procedure 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 tradeoff

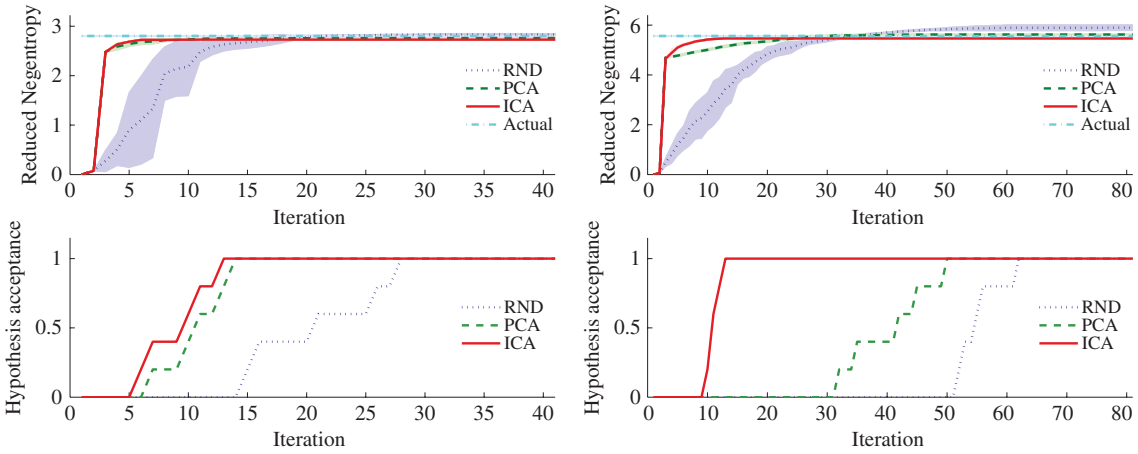

Fig. 4. Cumulative negentropy reduction (top) and multivariate Gaussian significance test (bottom) for each iteration in 2-D (left) and 4-D (right) synthetic problem. Average and standard deviation results from five realizations are shown.

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 [41]. 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.

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 (2). Essentially, one learns the sequence of transforms to Gaussianize a given dataset, and the I estimate reduces to computing 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: UU, GG, and a marginally composed exponential and Gaussian distribution (EG). An arbitrary rotation was applied in each case to obtain nonzero multi-information. In all cases, we used 10000 samples and repeated the experiments for 10 realizations. Two kinds of experiments were performed.

1) A 2-D experiment, where RBIG results can be compared to the results of naive (histogram-based) mutual information estimates (NE) and to previously reported 2-D estimates such as the Rudy estimate (RE) [42] (see Table III).
2) 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 2-D experiment to standard approaches. The ground-truth (GT) result is also given for comparison purposes.

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

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

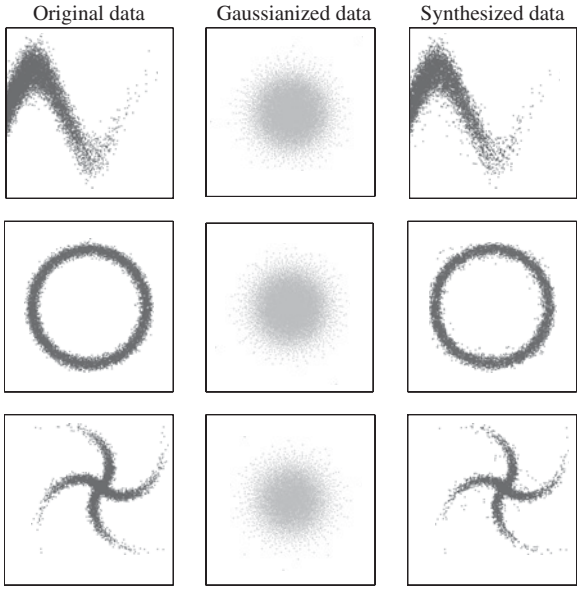

Fig. 5. Toy data examples synthesized using RBIG.

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.

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

Dim. d	EG		GG		UU	
	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

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

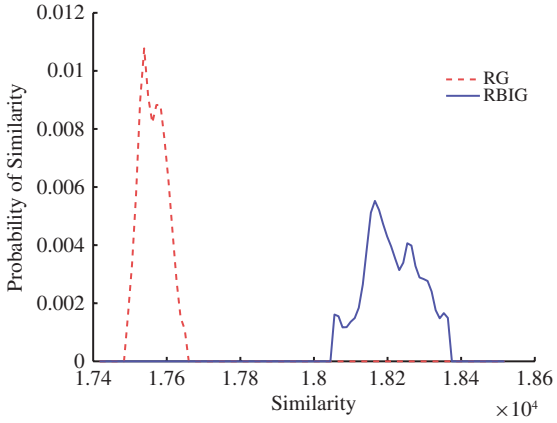

Fig. 7. Histogram of the similarity (inner product) between the distribution of original and synthesized face images for 300 realizations. For reference, the average image energy (average inner product) in the original set is 1.81×10^4.

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: Fig. 5 shows examples of 2-D 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 the synthesized data approximately follow the original PDF.

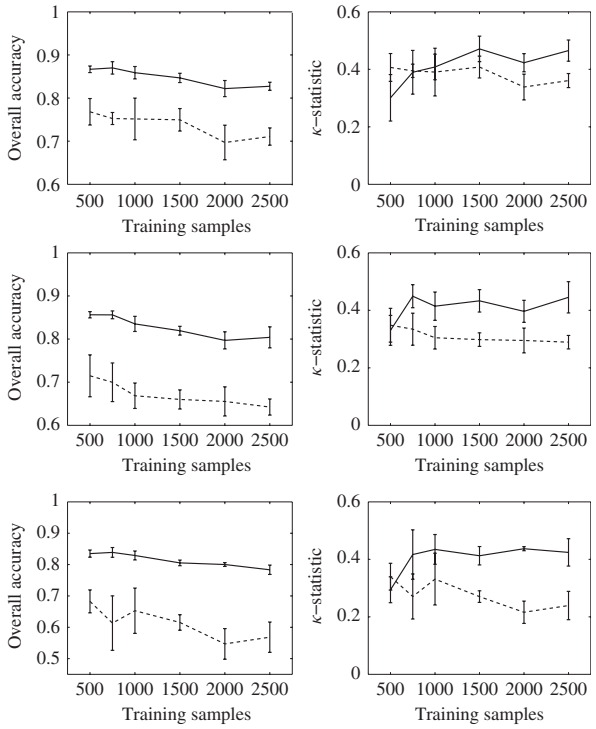

Fig. 8. Overall accuracy (left) and kappa statistic, κ (right) for RBIG (solid line) and SVDD (dashed line) in different scenes: Naples 1995 (top), Naples 1999 (center) and Rome 1995 (bottom).

2) Face Synthesis: In this experiment, 2500 face images were extracted from [43], 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. Fig. 6 shows illustrative examples of original and synthesized faces with RG and RBIG.

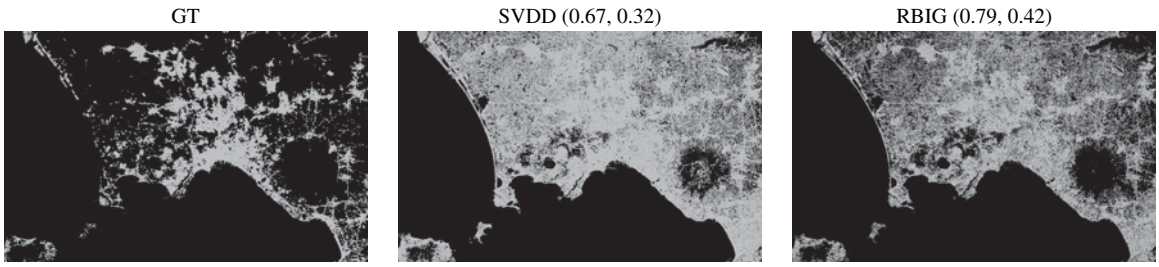

Fig. 9. GT and classification maps obtained with SVDD and RBIG for the Naples 1995 scene. The white points represent urban area and the black points represent non-urban area. The corresponding overall accuracy and κ-statistic are given in parenthesis.

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 (see Fig. 7). 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 a one-class classifier. Its performance is illustrated in the challenging problem of detecting urban areas from multispectral and SAR images. The GT data for the images used in this section were collected in the urban expansion monitoring ESA-ESRIN DUP project[3] [44]. 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 area discrimination [44]. 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 threefold cross validation in the training set, optimizing the κ statistic [45]. Training sets of different size of the target class were used in the range $[500, 2500]$. We assumed a scarce knowledge of the nontarget 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.

Fig. 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 that of the proposed method for small 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 nontarget class.

Fig. 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 nontarget data gives rise to too broad a 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], [46]–[48]

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

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

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

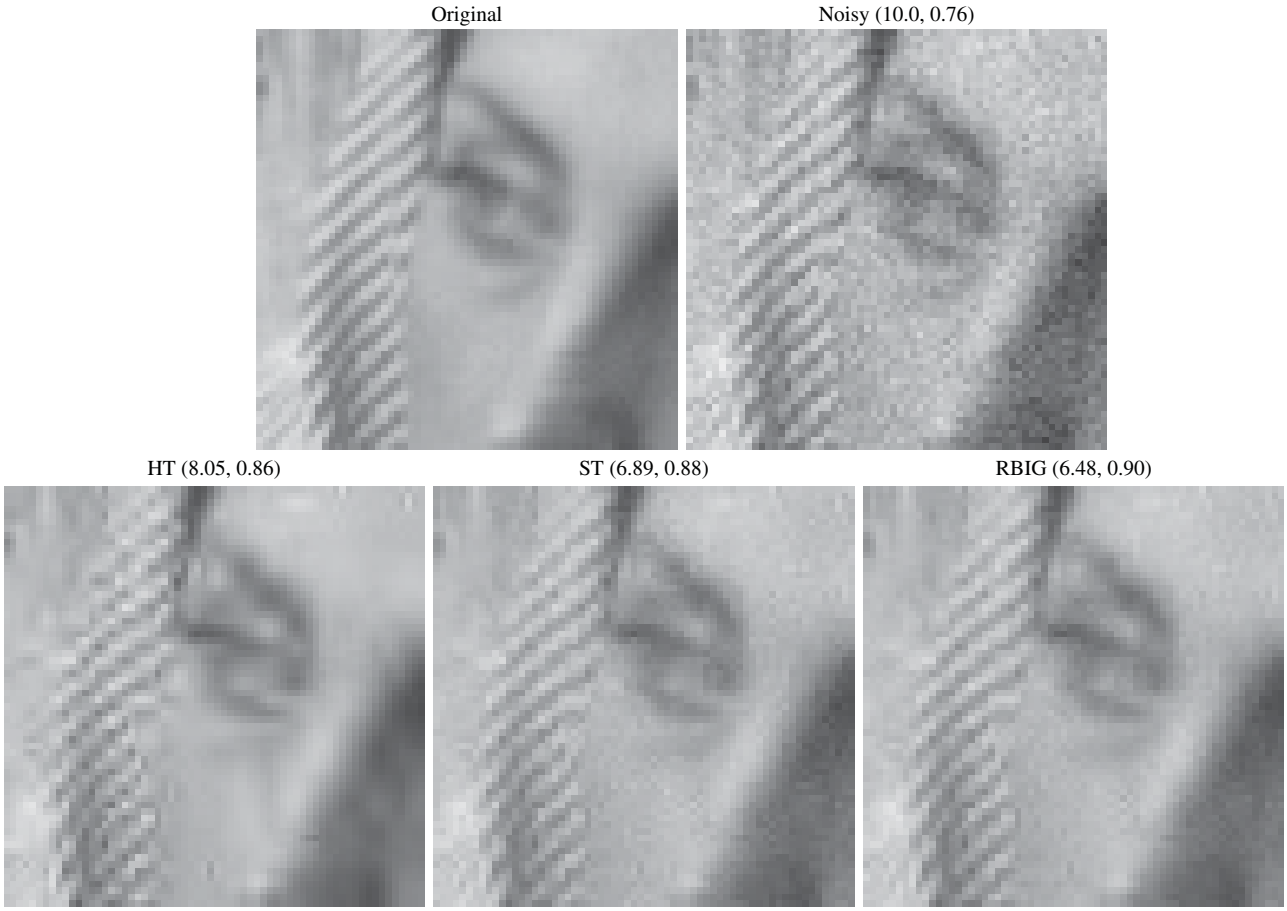

Fig. 10. Original noisy (noise variance $\sigma_n^2 = 100$) and restored "Barbara" images. The root-mean-square error (RMSE) and the perceptually meaningful structural similarity index (SSIM) [49] are given in parentheses.

methods. The conventional approach consists of using parametric models to be plugged into (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 the orthonormal QMF wavelet domain with four frequency scales [50], and 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], [51]. In order to evaluate (16), we sampled the posterior PDF at 8000 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 [52]. Note that the classical hard-thresholding (HT) and soft-thresholding (ST) results [46] 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 [47].

Fig. 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 SSIM [49]. Note that the RBIG method obtains better results (numerically and visually) than the classical methods due to the more accurate PDF estimation.

VI. Conclusion

In this paper, we proposed an alternative solution to the PDF estimation problem by using a family of 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 showed that a wide class of orthonormal transforms (including trivial RND) is well suited to Gaussianization purposes. The

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

freedom to choose the most convenient rotation is the difference with formally similar techniques, such as PP, which is focused on looking for interesting projections (which is an intrinsically more difficult problem). In this way, here we proposed 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 establishing a convenient family of rotations for a good Jacobian or convenient criteria to ensure the generalization ability, are subjects for future research.

VII. Acknowledgment

The authors would like to thank M. Bethge for his constructive criticism of this paper and E. Simoncelli for the estimulating discussion on "meaningful versus meaningless transforms."

References

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

[2] M. Banham and A. Katsaggelos, "Digital image restoration," *IEEE Signal Process. Mag.*, vol. 14, no. 2, pp. 24–41, Mar. 1997.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, Nov. 2000.

[4] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*. New York: Springer-Verlag, Aug. 2003.

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

[6] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization* (Probability and Statistics). New York: Wiley, Sep. 1992.

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

[8] A. Hyvärinen, "Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation," *Neural Comput.*, vol. 11, no. 7, pp. 1739–1768, 1999.

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

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

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

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

[13] J. Malo, I. Epifanio, R. Navarro, and E. P. Simoncelli, "Nonlinear image representation for efficient perceptual coding," *IEEE Trans. Image Process.*, vol. 15, no. 1, pp. 68–80, Jan. 2006.

[14] G. Camps-Valls, J. Gutiérrez, G. Gómez-Pérez, and J. Malo, "On the suitable domain for SVM training in image coding," *J. Mach. Learn. Res.*, vol. 9, pp. 49–66, Jan. 2008.

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

[16] P. J. Huber, "Projection pursuit," *Ann. Stat.*, vol. 13, no. 2, pp. 435–475, 1985.

[17] S. S. Chen and R. A. Gopinath, "Gaussianization," in *Proc. Neural Inf. Process. Syst.*, 2000, pp. 423–429.

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

[19] D. Erdogmus, R. Jenssen, Y. Rao, and J. C. Principe, "Gaussianization: An efficient multivariate density estimation technique for statistical signal processing," *J. VLSI Signal Process.*, vol. 45, nos. 1–2, pp. 67–83, 2006.

[20] S. Lyu and E. P. Simoncelli, "Nonlinear extraction of independent components of natural images using radial Gaussianization," *Neural Comput.*, vol. 21, no. 6, pp. 1485–1519, Jun. 2009.

[21] D. M. J. Tax and R. P. W. Duin, "Support vector domain description," *Pattern Recognit. Lett.*, vol. 20, nos. 11–13, pp. 1191–1199, Nov. 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, Jul. 2006.

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

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

[25] J.-F. Cardoso, "Dependence, correlation and Gaussianity in independent component analysis," *J. Mach. Learn. Res.*, vol. 4, nos. 7–8, pp. 1177–1203, 2003.

[26] M. Studeny and J. Vejnarova, *The Multi-Information Function as a Tool for Measuring Stochastic Dependence*. Norwell, MA: Kluwer, Jan. 1998, pp. 261–298.

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

[28] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, May 1999.

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

[30] S. Squartini, A. Bastari, and F. Piazza, "A practical approach based on Gaussianization for post-nonlinear underdetermined BSS," in *Proc. Int. Conf. Commun. Circuits Syst.*, Guilin, China, Jun. 2006, pp. 528–532.

[31] B. Xiang, U. V. Chaudhari, J. Navratil, G. N. Ramaswamy, and R. A. Gopinath, "Short-time Gaussianization for robust speaker verification," in *Proc. Int. Conf. Acoustic Speech Signal Process.*, 2002, pp. 681–684.

[32] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London, U.K.: Chapman & Hall, 1986.

[33] C. A. León, J.-C. Massé, and L.-P. Rivest, "A statistical model for random rotations," *J. Multivar. Anal.*, vol. 97, no. 2, pp. 412–430, Feb. 2006.

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

[35] V. Zarzoso, P. Comon, and M. Kallel, "How fast is fastICA," in *Proc. 14th Eur. Signal Process. Conf.*, Florence, Italy, Sep. 2006, pp. 1–5.

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

[37] G. H. Golub and C. F. Van Loan, *Matrix Computations* (Mathematical Sciences), 3rd ed. Baltimore, MD: The Johns Hopkins Univ. Press, Oct. 1996.

[38] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1, 3rd ed. New York: Wiley, Jan. 1968.

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

[40] J. Malo, and V. Laparra, "Psychophysically tuned divisive normalization factorizes the PDF of natural images," *Neural Comput.*, vol. 22, no. 12, pp. 3179–3206, Dec. 2010.

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

[42] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," *Signal Process.*, vol. 16, no. 3, pp. 233–248, Mar. 1989.

[43] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[44] L. Gómez-Chova, D. Fernández-Prieto, J. Calpe, E. Soria, J. Vila, and G. Camps-Valls, "Urban monitoring using multitemporal SAR and multispectral data," *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 234–243, Mar. 2006,

[45] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Measure.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.
[46] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Stat. Assoc.*, vol. 90, no. 432, pp. 1200–1224, Dec. 1995.
[47] E. P. Simoncelli, "Bayesian denoising of visual images in the wavelet domain," in *Bayesian Inference in Wavelet Based Models*, vol. 141, P. Müller and B. Vidakovic, Eds. New York: Springer-Verlag, Jun. 1999, ch. 18, pp. 292–308.
[48] M. A. T. Figueiredo and R. D. Nowak, "Wavelet-based image estimation: An empirical Bayes approach using Jeffrey's noninformative prior," *IEEE Trans. Image Process.*, vol. 10, no. 9, pp. 1322–1331, Sep. 2001.
[49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[50] E. P. Simoncelli and E. H. Adelson, "Subband transforms," in *Subband Image Coding*, J. Wodds, Ed. Norwell, MA: Kluwer, 1990, ch. 4, pp. 143–192.
[51] J. Liu and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Trans. Image Process.*, vol. 10, no. 11, pp. 1647–1658, Nov. 2001.
[52] J. M. 51 and A. F. M. Smith, "Bayesian theory," *Measure. Sci. Technol.*, vol. 12, no. 2, p. 221, 2001.

Valero Laparra was born in Valencia, Spain, in 1983. He received the B.Sc. degree in telecommunications engineering, the M.Sc. degrees in electronics engineering and mathematics, in 2005, 2007, and 2010, respectively. He is currently pursuing the Ph.D. degree at the Image Processing Laboratory, University of Valencia, Spain.

His current research interests include information processing in the brain and its application to engineering problems.

Gustavo Camps-Valls (M'04–SM'07) was born in 1972. He received the B.Sc. degree in physics and electronics engineering in 1996 and 1998, respectively, and the Ph.D. degree in physics in 2002, all from the Universitat de València, Spain.

He is currently an Associate Professor in the Department of Electronics Engineering and member of the Image Processing Laboratory, Universitat de València. He was a Visiting Professor at the Remote Sensing Laboratory, University of Trento, Trento, Italy, and the Max Planck Institute for Biological Cybernetics, Tübingen, Germany. He has published 70 research papers in journals and more than 100 papers in conference proceedings, and has written several book chapters. He is the editor of the books *Kernel Methods in Bioengineering, Signal and Image Processing* (Hershey, PA: IGI, 2007) and *Kernel Methods for Remote Sensing Data Analysis* (New York: Wiley, 2009). He serves as a referee for many international journals and conferences. His current research interests include development of machine learning algorithms for signal and image processing.

Dr. Camps-Valls has been a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society since 2009, and an Associate Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, Improvement Science Research Network's *Signal Processing Journal*, and the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.

Jesús Malo was born in 1970. He received the M.Sc. and Ph.D. degrees in physics from the Universitat de València, Spain, in 1995 and 1999, respectively.

He worked as a Fulbright Post-Doctoral Fellow at the Vision Group, National Aeronautics and Space Administration Ames Research Center, Mountain View, CA, in 2000 and 2001 (with A. B. Watson), and the Laboratory of Computational Vision, Center for Neural Science, New York University, New York (with E. P. Simoncelli). Currently, he is with the Image Processing Laboratory, Universitat de València. His current research interests include models of low-level human vision, their relations with information theory, and their applications to image processing and vision science experimentation, Fourier, MATLAB, Equipo Crónica, Jim Jarmusch, Jordi Savall, Pixies, Manara, la Bola de Cristal, Faemino y Cansado, and beauty in general.

Dr. Malo is a member of the Asociación de Mujeres Investigadoras y Tecnólogas, and an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING. He was the recipient of the Vistakon European Research Award in 1994.