UNIVERSITAT DE VALÈNCIA

Departaments d'Informàtica i Matemàtica Aplicada.

Learning efficient image representations: Connections between statistics and neuroscience

MAY, 2011

AUTHOR:

Valero Laparra Pérez-Muelas

ADVISORS:

DR. Jesús Malo López
DR. Gustavo Camps i Valls

Contents

Summary

This thesis summarizes different works developed in the framework of analyzing the relation between image processing, statistics and neuroscience. These relations are analyzed from the *efficient coding hypothesis* point of view (H. Barlow [1961] and Attneave [1954]). This hypothesis suggests that the human visual system has been adapted during the ages in order to process the *visual information* in an efficient way, i.e. taking advantage of the statistical regularities of the visual world. Under this classical idea different works in different directions are developed.

One direction is analyzing the statistical properties of a revisited, extended and fitted classical model of the human visual system. No statistical information is used in the model. Results show that this model obtains a representation with good statistical properties, which is a new evidence in favor of the *efficient coding hypothesis*. From the statistical point of view, different methods are proposed and optimized using natural images. The models obtained using these statistical methods show similar behavior to the human visual system, both in the spatial and color dimensions, which are also new evidences of the *efficient coding hypothesis*. Applications in image processing are an important part of the Thesis. Statistical and neuroscience based methods are employed to develop a wide set of image processing algorithms. Results of these methods in denoising, classification, synthesis and quality assessment are comparable to some of the most successful current methods.

Resumen

Esta Tesis resume diferentes trabajos realizados bajo el marco de el análisis de las relaciones entre el procesado de imágenes, la estadística y la neurociencia. Estas relaciones son analizadas desde el punto de vista de la *hipótesis de la codificación eficiente* (H. Barlow [1961] y Attneave [1954]). Dicha hipótesis sugiere que el sistema visual humano se ha ido adaptando durante los años para poder procesar la *información visual* de forma eficiente, es decir, para aprovechar las regularidades estadísticas del mundo visual. Con esta idea de fondo se han realizado trabajos en diferentes direcciones.

Una dirección ha sido analizar las propiedades estadísticas del sistema visual humano. Para ello se ha usado un modelo clásico el cual se ha revisado, extendido y ajustado. Nótese que el modelo no hace uso de la estadística en ningún momento. Los resultados muestran que este modelo obtiene una representación con buenas propiedades estadísticas para las imágenes naturales, lo cual es una nueva evidencia en favor de la *hipótesis de la codificación eficiente*. Desde el punto de vista estadístico, se han propuesto diferentes métodos y se han optimizado utilizando datos de imágenes naturales. Estos modelos estadísticos aprenden un comportamiento similar al del sistema visual humano, tanto en las dimensiones espaciales como en las dimensiones de color. Esto también supone una evidencia en favor de la *hipótesis de la codificación eficiente*. Una parte importante de la Tesis es el empleo de estos métodos, tanto los estadísticos como los basados en neurociencia, para desarrollar distintas aplicaciones de procesado de imágenes. Por ejemplo, en aplicaciones de restauración, clasificación, síntesis y calidad de imagen se obtienen resultados similares a algunos de los mejores métodos actuales.

Chapter 1

Introduction

THIS Thesis is a compendium of works that, from different points of view, focuses on studying the relation between statistics and perception in the Human Visual System (HVS). This relation makes both statistics and perception a suitable criteria for developing image processing applications. This idea is synthesized in Fig. 1.1.

This relation has constituted a fruitful field in order to understand how the brain is designed,that is, to answer the fundamental question of *"What is the goal of the brain?"* The brain mechanisms have been adapted during the ages in order to process *natural data*. Even rejecting the plausible evolution hypothesis [Darwin, 1859] one should agree with the next two facts. On the one hand, the brain has to work under some restrictions (size, energy, time...), and on the other hand, it processes a huge amount of information efficiently. Both things together suggest that the brain has evolved in order to be as optimal as possible.

Following this idea, H. Barlow [1961] and Attneave [1954] stated the so-called *redundancy reduction hypothesis*, opening a new direction for understanding how the brain works. This hypothesis interprets the optimality of the brain in statistical terms, by looking for a representation where the redundant information is discarded. This hypothesis has been modeled during the last decades and renamed as *efficient coding hypothesis*, see [H. B. Barlow, 2001] for a nice review. Personally, I prefer the second name because it does not make any assumption about the procedure used by the brain (redundancy reduction) but stems for the goal followed (optimal coding). Using the idea of representing efficiently the visual data, this Thesis proposes and analyzes different models and tries to extract information about the HVS behavior.

Understanding the HVS is a challenging task since it is the human perception mechanism that collects and processes most amount of data. Imagine that you are a regular human that wakes up in the morning to go to work. One may accept that at the early few seconds, the brain is only performing essential functions. The amount of raw information collected by the visual system during the first minute is around 5308 terabytes. This rough estimation comes from assuming 80 photoreceptors/deg within a field of view of $90^o \times$

160º, 120 spectral data *per* photoreceptor (5 nm of resolution in the range of [400-700] nm), 100 images/sec (critical fusion rate about 50Hz), 8 bytes/sample and 2 eyes. Therefore, the human visual system should be adapted to process this huge amount of information in order to process *the important information* only. Note that, if the HVS was not optimized in order to efficiently process the regularities of the environment, our brain would collapse within seconds.

Of course, the main question to solve is *What does important information mean?* In order to simplify this question, here we only focus in the early stages of the HVS. Therefore, images will be seen as textures, and no qualitative structure will be taken into account. This point of view will allow us to ignore the high level information like for instance *familiar shapes* (that for sure is used).

In order to avoid processing unnecessary information, the brain should take into account the statistical regularities of images. Note that, any *machine* that collects and processes data should be constrained in the same way. Taking into account regularities is necessary because, only if one knows the position of all the particles in the universe and all the laws governing the interactions between them, one would be able to interpret the world as a deterministic system, like in Laplace Demon's [Laplace, 1814]. Therefore, humans (and any other *machine*) should interpret the world statistically, taking into account what is *normal to happen*, i.e. what are the probabilities of the possible events.

In this direction, information theory [Shannon, 1948] has been used in order to obtain details about how the brain works. The relation between statistics and information is straightforward. Intuitively, one can measure the amount of information of an event by relating it with the inverse of the probability that this event occurs, i.e. the more probable is an event the less information will give to us. Barlow proposed to use this mathematical treatment of information as a tool to understand the brain.

Taking all the above things into account, one may think of designing an optimal system from a statistical point of view, and then exploring its similarity with the brain. A fruitful framework in computational neuroscience consists on trying to understand how the brain works by analyzing natural data. Specifically, *natural image statistics* has been used as a tool to understand how the HVS processes visual data. This direction goes from statistics to neuroscience (Fig. 1.1).

Note that an important part of simulating the behavior of the HVS implies to understand what are the restrictions that the HVS has to deal with, and moreover, understanding how to implement them mathematically. This makes this issue directly dependent on accounting for a previous knowledge about the behavior of the HVS.

Reverse engineering may help us to learn the behavior of a system by analyzing its structure, function and operation. Therefore, taking the brain as case study may help us to learn how to design optimal systems. We have seen that the brain is able to process a huge amount of data. Of course, the capacity of collecting and storing data, and the

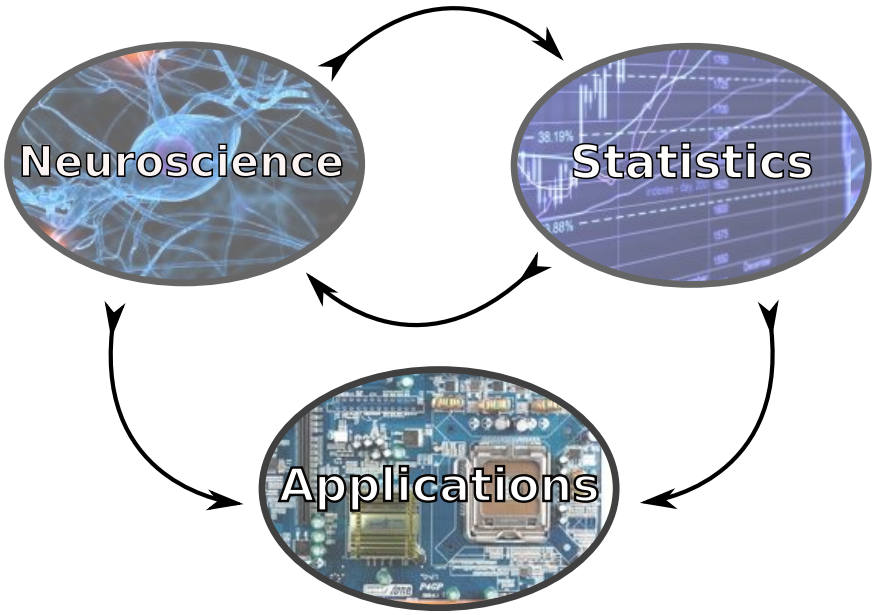


Figure 1.1: Road map of this Thesis.

computational capabilities of current computers is bigger than the human brain. However, the ability of the brain to infer information from data is bigger than any *intelligent system* designed by the humans so far. Therefore we can take advantage of having an *inference machine* and use reverse engineering in order to learn how to design optimal systems. This is represented by the arrow that goes from neuroscience to statistics (Fig. 1.1).

Both aspects, the neuroscience point of view and the statistical point of view have direct consequences when implementing image processing applications. Whether if we want to solve a task for which humans are prepared (e.g. object classification) or not (e.g. measuring the amount of gamma radiation), having a probabilistic description of the possible event will help us to select an optimal solution, arrow from *statistics* to *applications* in figure 1.1. Moreover, the HVS is a very good tool in order to assess the result of some image processing algorithms, arrow from *neuroscience* to *applications* in figure 1.1. For instance, the best way to evaluate the performance of denoising algorithms is by visual inspection.

1.1 Human Visual System as a reference

You do not really understand something unless you can explain it to your grandmother (and she understands it). This quote (usually attributed to Albert Einstein) could be used in order to test how much we know about how the brain works. Unfortunately, not everyone has a grandmother to teach [1]. Nowadays, we can change this quote by *You do not really understand something unless you can program it in Matlab (and without bugs)*. Therefore, we might say that we understand how the HVS works when we have a successful computational

[1] Pilar Celda and Isabel Marín, in memoriam.

model. This model should agree with the psychophysical and physiological knowledge. If one would have the (insane) idea of using this knowledge to implement a computational model, one should face many problems. The main problem is that psychophysical results make reference to the behavior of the whole system, i.e. humans can not *feel* when a single neuron is active. Therefore, these results can be used only to implement a model of the whole brain. Another problem is the huge amount of diverse measures, in different experiments, with different kind of uncertainties. Even if all these things were solved, one would always be able to find works stating opposite conclusions [Lehrer, 2010]. From the physiological point of view, gathering measures is even more complicated, and even much more acquire *awake* measures (which for studying the HVS would be a quite important requirement). These problems (along with the current emphimpossibility of communication between engineers and psychophysicists/neuro-physiologists) makes the task of implementing a perfect computational model of the HVS a 'to do work' for the next generations.

Nevertheless, a number of people is working towards implementing computational models of some specific tasks of the brain. Although these models have a lot of limitations, they are still useful tools in order to check the current knowledge of the brain's behavior. Moreover, these models, to a greater or lesser extent, can be checked in efficiency terms and can be also used to improve some engineering applications. Specifically, the HVS models can be used in order to figure out facts about the statistics of natural images, and more importantly, to understand the restrictions and the goals of the HVS.

Chapter 2 analyzes the statistical properties of a computational model of the early stages of the HVS, which is physiologically inspired and psychophysically fitted. We will show theoretically and quantitatively how this model obtains a representation of visual data with good statistically properties.

1.2 Statistics as a tool to optimize a system

From the statistical point of view, designing a system as efficient as the HVS is a challenging task. When designing a system to infer information from data, *decision theory* shows that two ingredients are needed: the probability density function (PDF) of the data and the cost function associated to the possible events, i.e. *Bayes risk* [Bernardo & Smith, 1994]. In the attempts to explain statistically the HVS, the cost function is sometimes assumed to be Euclidean or even ignored. Although looking only to the PDF can give us clues about the behavior of the HVS, it is worth keeping in mind that the cost function is an important part of the decision theory. Even neglecting the cost function issue, obtaining a plausible statistical explanation of the visual system is a challenging task, since there are some open issues:

■ What is the goal of the Human Visual System?

Looking at the HVS as a statistically optimized system raises a lot of questions. The main one is that it is not yet clear what is the criterion to optimize. Of course, one may think of maximizing the amount of extracted information from the world, as a desirable goal. Trends in this direction have tried to find a transformation that obtains a representation of the data with independent components. Even though this is a desirable representation for many reasons, it is not a mandatory one in reality. Moreover, sometimes, it is incompatible with representing the information with the minimum possible error (which could be also a desirable situation). Therefore, although independence is a useful way to obtain a probabilistic description of the data, maybe it is not the goal in the HVS. Anyway, the HVS should employ some strategy to exploit the statistical regularities of the visual world, which bring us directly to the next open issues.

■ What is "the visual world"?

When the idea is to extract information from some measurements, the quality and representativity of the acquired data is as important as the method to infer information. In image statistics, *natural* data are typically used, which is very often reduced to forest images [2]. Using this kind of images assumes that the HVS has been adapted during many years and only lately man made things are natural in our environment (e.g. first building constructions date around 5000 years ago). However, the learning process in the HVS also involves the first months of life, and one can argue that these months are essential in defining the final behavior of the HVS. Moreover, the adaptation capacity to each specific scene should be taken into account. Therefore, we are going to take a loose definition, *natural data* is what involves the surroundings of the HVS nowadays.

Other issue is related to what kind of information should contain the visual data properly to train the system. For instance: Is there any sense in using images with complicated information (like an human body) in order to train an early-stage vision model? Is it going to take advantage of the structure? Unrepresentative images will only bias the results of the model and the expectation of the scientist.

■ Restrictions of the Human Visual System

There are a lot of restrictions when thinking in the brain as a system. The most obvious is the size: a limited number of neurons are available. Knowing the exact amount of neurons dedicated to each specific task in the brain would give us a very useful

[2] i.e. The van Hateren image database [Hateren & Schaaf, 1998] is the most used in natural image statistics works

information[3]. Another important restrictions are related to the capacity of a single neuron of processing data, the amount of neural noise and the speed of the different neurons. A lot of research has been done is this sense. However statistical methods rarely try to include them.

■ Computational resources

The optimization of the vision system has been carried out during 2700 million years. This gives an idea of its complexity[4]. Moreover, optimization of the HVS in specific human takes months, which implies a huge amount of data, approximately 1.500 million of natural images with high resolution. In an idyllic case, the statistical methods applied to extracting HVS features usually employ 60000 images[5].

■ Statistical methods

Advances in computing machines have allowed us to start thinking in methods to extract statistical information from data that were unimaginable only few years ago, or even using methods *imagined* years ago but with unbearable computational complexity. This implies that, in the last years, many of statistical methods have been developed (and which is better, some of them are useful). Almost all of the machine learning algorithms have been applied in image processing problems. A number of them could be interpreted in order to extract information of how the HVS should work. However, none of the statistical methods applied to reproduce the behavior of the whole brain has been successful yet.

One special mention should be made about the called *curse of dimensionality*. It establishes that the amount of data necesary to estimating a PDF increases exponentially with the data dimension. This problem makes that the data-driven statistical methods always obtain partial and biased solutions.

In Chapter 4, three different works that involve extracting HVS features from image data by using statistical methods are presented. In section 4.1 a method to design a sensor system with tunable metric is proposed and used to explain statistically some abilities of the color mechanisms. In section 4.2 a maximum likelihood method is used to obtain some features of the V1 region of the brain. In section 4.3 the ability to explain the shape of the linear filters using the independence assumption is evaluated over texture images.

[3]The proportion of neurons in the lateral geniculate nucleus and in V1 is around 1:1000 [H. B. Barlow, 2001], which suggests that the redundancy reduction hypothesis does not apply between these stages.

[4]This affirmation takes the evolution theory as a fact. However, from a creationist point of view, the HVS is made *in the image and likeness* as a superior being, which is also unattainable from a computational point of view.

[5]Most of the natural image databases contains less than 10000 images.

1.3 Neuroscience and statistics for image processing

An indirect way to test HVS based models or statistical image models is by applying them in image processing problems. Moreover, proposing useful image processing algorithms could be an endless way to earn money for a scientist, or at least, it can be a way to bring back to the society the inversion made in him/her. Both concepts analyzed here, *perception* and *statistics*, can be used to inspire image processing algorithms.

On the one hand, a correct HVS model should be useful in man-oriented tasks. For instance, they could help humans in some tasks, such as image evaluation or object recognition. In section 3.1, the ability of the HVS model is employed by using it to evaluate the quality of images.

On the other hand, image statistics is also important when designing image processing algorithms. One only has to take a look at the formulation for the optimal design of the main image processing tasks: quantization [Gersho & Gray, 1992], denoising [Portilla et al., 2003], classification [R. Duda & Hart, 1973], information Theory [Cover & Tomas, 1991] and synthesisBernardo & Smith [1994]. In all of them the PDF of the data is involved.

In fact, most of these tasks are based on the Bayes'risk rule. In section 5.2 we revise the capability of the projection pursuit method to obtain a description of multidimensional PDFs. A computationally convenient extension of projection pursuit Friedman & Tukey [1974] is presented and evaluated in a variety of image processing problems. However, as also stated above, obtaining a plausible estimation of the true PDF for explaining multidimensional data is complicated. A lot of methods are based on using the statistical regularities of data but without estimating a PDF explicitly. One of the most popular machines are kernel methods. In section 5.1 we explore the capability of a kernel-based method, the support vector regression, in taking advantage of statistical image information. This capability is evaluated by using the model in image denoising.

1.4 Thesis organization

This Thesis is organized following the scheme in Fig. 1.1. The arrows coming out from *neuroscience* are based in a formulation of a classical model of the HVS (until V1). The model is revisited, expanded and psychophysically fitted. Section 2 correponds to the arrow that arrives to *statistics* and involves the results obtained in [Malo & Laparra, 2010a]. The statistical properties of this model are analyzed: approximate PDF factorization and substantial mutual information reduction. Note that no statistical information is used to fit the V1 model, and hence these results are a complementary evidence in favor of the efficient coding hypothesis. Another related work is Malo & Laparra [2010b].

Section 3 corresponds to the arrow that arrives to *applications* and involves the results obtained in [Laparra, Marí, & Malo, 2010]. In this work, the computational HVS model is

applied as a quality image metric. Experiments on a number of databases including a wide range of distortions show that this model is fairly competitive with newer approaches, robust, and easy to interpret in linear terms.

The arrows coming out from *statistics* are basically learning methods applied to explain aspects of the HVS or to develop image processing tools. Section 4 corresponds to the arrow that arrives to *neuroscience* and involves three works: [Laparra, Jiménez, et al., 2011a], [Laparra, Gutman, et al., 2011] and [Laparra & Bethge, 2011].

The work [Laparra, Jiménez, et al., 2011a] is reported in section 4.1. In this section a method to design a set of sensors with two main features is proposed: (i) the shape of the sensors is able to be non-linear, and (ii) the metric of the sensors is tunable for different criteria. This method is applied over natural colors obtaining similar behavior as the HVS color mechanisms: the system is nonlinear and adaptive to changing environments. The reported adaptation under D65 and A illuminations has been reproduced by gathering a *new database* of colorimetrically calibrated images of natural objects under these illuminants, thus overcoming the limitations of existing databases. Moreover, the obtained results suggest that color perception at this low abstraction level may be guided by an error minimization strategy [D. MacLeod & Twer, 2003] rather than by the information maximization principle [Laughlin, 1983]. Another related works are [Laparra & Malo, 2008a,b; Laparra, Tuia, et al., 2011].

The work [Laparra, Gutman, et al., 2011] is exposed in section 4.2. This section proposes an extension of the complex Independent Components Analysis (ICA) method applied to natural images. We show that linear complex-valued ICA learns complex cell properties from Fourier-transformed natural images, i.e. two Gabor-like filters with quadrature-phase relationships. Conventional methods for complex-valued ICA assume that the phases of the output signals follow uniform distributions. We relax this assumption by modeling the phase information of the output sources in the complex-valued ICA estimation. The resulting model of phases shows that the distributions are often far from uniform, and the shapes of the Gabor filters are also changed.

The work [Laparra & Bethge, 2011] is exposed in section 4.3. It consists of measuring the amount of mutual information reduced by using different linear representations. Results stress the idea that the shape of the filters in V1 could not be probably dued to an independence goal of this stage in the brain. Also, it can be seen that the ability on redundancy reduction of the linear transforms is very image-dependent and therefore adaptation of the filters in V1 to different environments is necessary for efficient coding.

The arrow that comes out from *statistics* and arrives to *applications* summarizes the results published in [Laparra, Gutiérrez, et al., 2010] and [Laparra, Camps-Valls, & Malo, 2011]. The first work reveals the ability of kernel methods in including statistical information. Specifically we use support vector regression (SVR) in the wavelet domain to impose natural image features to noisy images. The specific signal relations are obtained

from mutual information measures computed on a representative image database. Results under several noise levels and noise sources show that: (1) the proposed method outperforms conventional wavelet methods that assume coefficient independence, and (2) it performs similarly to state-of-the-art methods that do explicitly include these relations. Therefore, the proposed machine learning approach can be seen as a more flexible alternative to the explicit description of wavelet coefficient relations. Another related works are [Armengot et al., 2010; Camps-Valls et al., 2011, 2010; Laparra et al., 2008]. The second work presents a method to estimate multidimensional PDFs, based on projection pursuit techniques. 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. It is shown that, unlike in projection pursuit, 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 differentiability, invertibility and convergence of the method are theoretically and experimentally studied. Also, the practical performance is illustrated in a number of multidimensional problems such as image synthesis, classification, denoising, and multi-information estimation. Another related works are [Laparra et al., 2009; Laparra et al., 2009].

Chapter 6 summarizes the general conclusions and the lessons learned during this Thesis.

Introducción (Castellano)

ESTA Tesis es un compendio de trabajos que, desde diferentes puntos de vista, se centra en una idea: la relación entre estadística y la percepción del Sistema Visual Humano (de aquí en adelante SVH). Dicha relación convierte a ambos, estadística y percepción, en criterios útiles para el desarrollo de aplicaciones en procesado de imágenes. La idea básica de la Tesis está sintetizada en la figura 1.2.

El estudio de esta relación ha sido un campo muy fructífero a la hora de entender cómo está diseñado el cerebro, es decir, a la hora de responder a la pregunta: *¿Cómo funciona el cerebro?* Los mecanismos del cerebro han sido adaptados a lo largo de los años para procesar *datos naturales*. Incluso si ignorásemos la más que plausible teoría de la evolución [Darwin, 1859], los siguientes dos hechos son irrefutables. Por un lado, el cerebro debe trabajar bajo muchas restricciones (tamaño, energía, tiempo...), mientras que por otro lado, procesa una cantidad ingente de información de forma eficiente. Ambos hechos sugieren que el cerebro debe haber evolucionado para ser lo más óptimo posible.

A partir de esta idea, H. Barlow [1961] y Attneave [1954] empezaron la llamada *hipótesis de la reducción de redundancia*, abriendo una nueva dirección para entender cómo funciona el cerebro. Esta hipótesis interpreta la optimalidad del cerebro en términos estadísticos, buscando una representación donde la información redundante es desechada. Dicha hipótesis ha sido modelada a lo largo de los últimos años y renombrada cómo *hipótesis de la codificación eficiente*, en [H. B. Barlow, 2001] se puede encontrar un buen resumen sobre el tema. Personalmente, yo prefiero el segundo nombre puesto que no hace ninguna asunción sobre el procedimiento empleado por el cerebro (reducción de redundancias) sino que describe el fin buscado (codificación óptima). Empleando la idea de representar eficientemente la información visual, esta Tesis propone y analiza diferentes modelos, e intenta extraer información sobre cuál es el funcionamiento del SVH.

Entender cómo funciona el SVH es una tarea complicada puesto que es el sistema de percepción humana que recoge y procesa mayor cantidad de información. Pongamos por ejemplo que el lector es un simple humano que debe levantarse por la mañana e ir a trabajar. Por supuesto, se puede asumir que a primera hora de la mañana el cerebro está realizando sólo tareas esenciales. La cantidad estimada de información recogida por el sistema visual durante el primer minuto de la mañana está alrededor de 5.308 Terabytes.

Esta simple estimación asume 80 fotorreceptores/grado dentro de un campo de visión de $90^o \times 160^o$, 120 datos espectrales por fotorreceptor (5 nm de resolución en el rango de [400-700] nm), 100 imágenes/seg (ratio crítico de fusión de 50 Hz), 8 bytes/muestra y 2 ojos. Por tanto, el SVH debe estar adaptado para procesar esta gran cantidad de información de forma que emplee *sólo la información importante*. Nótese que, si el SVH no estuviera optimizado para procesar eficientemente las regularidades del entorno, nuestro cerebro se colapsaría en segundos.

Por supuesto, la principal cuestión es *¿Qué información es importante?* Para simplificar esta cuestión, en esta Tesis nos centraremos en los primeros estadios del SVH. Por tanto, las imágenes deberán ser vistas cómo texturas sin hacer caso a estructuras cuálitativas. Este punto de vista nos permitirá ignorar información de alto nivel (la cuál es ciertamente usada por nuestro sistema visual) cómo por ejemplo *formas familiares*.

Para evitar procesar información innecesaria, el cerebro debe tener en cuenta las regularidades estadísticas. Nótese que, cuálquier *máquina* que recoja y procese datos debe estar diseñada con el mismo criterio. Tener en cuenta las regularidades estadísticas es necesario puesto que, sólo sabiendo la posición de todas las partículas del universo y todas las leyes que gobiernan sus interacciones, seríamos capaces de ver el mundo cómo un problema determinista, cómo en el demonio de Laplace [Laplace, 1814]. Por tanto, los humanos (y cuálquier otra *máquina*) debe interpretar el mundo de forma estadística, teniendo en cuenta *qué es normal*, es decir que probabilidades tienen los posibles eventos.

En este sentido, la teoría de la información [Shannon, 1948] ha sido usada para obtener detalles de cómo funciona el cerebro. Estadística e información están directamente relacionadas. Intuitivamente, se puede medir la cantidad de información que proporciona un evento relacionándola con la probabilidad inversa de que pase, es decir, cuanto más probable es un evento menos información nos proporciona. Barlow propuso utilizar este tratamiento matemático de la información cómo herramienta para entender el cerebro.

Teniendo en cuenta todo lo anterior, se podría pensar en diseñar un sistema que procesase información de forma óptima desde un punto de vista estadístico, y explorar sus similitudes con el cerebro. Este es un campo fructífero en neurociencia computacional: en él se intenta averiguar cómo funciona el cerebro analizando datos extraídos de la naturaleza. Específicamente, *la estadística de las imágenes naturales* ha sido usada cómo herramienta para entender cómo el SVH procesa la información visual. Esta es la dirección que va de la estadística a la neurociencia en la figura 1.2.

Nótese que una parte importante al tratar de simular el comportamiento del SVH implica entender las restricciones que este tiene, y además, entender cómo implementarlas matemáticamente. Para ello debemos tener un conocimiento previo sobre el comportamiento del SVH.

La ingeniería inversa se basa en aprender el comportamiento de un sistema a través de analizar su estructura, su funcionamiento y su forma de operar. Por tanto, si tomamos

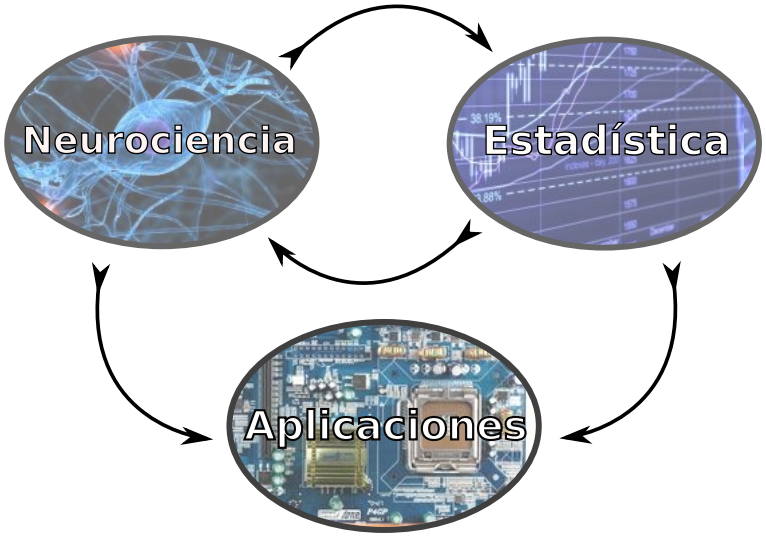


Figure 1.2: Esquema básico de la Tesis.

el cerebro cómo ejemplo podemos aprender cómo diseñar sistemas óptimos. Hemos visto que el cerebro es capaz de procesar una gran cantidad de información. Por supuesto, la capacidad de recoger y almacenar datos, y la capacidad de cálculo de los ordenadores hoy en día es mayor que la de un cerebro humano. Sin embargo, la habilidad que tiene el cerebro para inferir información a partir de datos es mayor que la de cuálquier *sistema inteligente* diseñado por el ser humano. Por tanto, podemos aprovechar que tenemos una *máquina de inferir información* y usar la ingeniería inversa para poder aprender cómo diseñar sistemas óptimos. Esta dirección es la flecha que va desde neurociencia a estadística en la figura 1.2.

Ambos aspectos, la neurociencia y la estadística, tienen consecuencias directas a la hora de implementar aplicaciones de procesado de imágenes. Tanto si queremos resolver tareas para las que los humanos están preparados (tipo reconocimiento de objetos) o no (tipo medir la cantidad de radiación gamma), tener una descripción probabilística de los posibles eventos nos ayudaría a seleccionar la opción óptima, flecha que va desde *estadística* a *aplicaciones* en la figura 1.2. Además, el SVH es una muy buena herramienta para poder mejorar los resultados de los algoritmos de procesado de imágenes, flecha que va desde *neurociencia* a *aplicaciones* en la figura 1.2. Por ejemplo, la mejor manera de evaluar algoritmos de eliminación de ruido es mediante inspección ocular.

1.5 Sistema Visual Humano cómo referencia

No entiendes realmente algo hasta que eres capaz de explicárselo a tu abuela (y lo entiende). Esta cita (habitualmente atribuida a Albert Einstein) podría ser usada para testear cuánto sabemos sobre el funcionamiento del cerebro. Desgraciadamente, no todo el mundo tiene una

abuela a la que explicar algo [6]. Hoy en día podríamos cambiar esta frase por: *No entiendes algo hasta que eres capaz de programarlo en Matlab (y compila sin errores)*. Por tanto, podremos decir que entendemos cómo funciona el SVH cuándo tengamos un modelo computacional que reproduzca su funcionamiento. Dicho modelo debería reproducir los comportamientos observados tanto psicofísicos cómo fisiológicos.

Si uno tuviera la (insana) idea de usar el conocimiento actual para implementar un modelo computacional, se encontraría con muchos problemas. El principal sería que los resultados psicofísicos hacen referencia al comportamiento de todo el sistema al mismo tiempo, es decir, los humanos no podemos *sentir* cuándo una neurona está activa. Por tanto, estos resultados pueden ser solamente usados para implementar un modelo completo del cerebro. Otro problema es la gran cantidad de medidas distintas, en diferentes experimentos, y con diferentes tipos de errores. Incluso si resolviésemos estos problemas, seríamos capaces de encontrar estudios que concluyesen comportamientos distintos [Lehrer, 2010]. Desde el punto de vista fisiológico, tomar medidas es incluso más complicado, y mucho más medidas con *muestras* [7] despiertas (lo cuál sería importante para estudiar el comportamiento del SVH). Estos problemas (junto a la imposibilidad de comunicación entre ingenieros y psicofísicos/fisiólogos) dejan para generaciones futuras la tarea de implementar un modelo computacional perfecto del SVH.

Sin embargo, una gran cantidad de personas está trabajando hoy en día en implementar modelos computacionales que reproduzcan tareas específicas del cerebro. Aunque estos modelos tienen una gran cantidad de limitaciones, son herramientas útiles para evaluar cuál es nuestro conocimiento actual del funcionamiento del cerebro. Además, estos modelos, en mayor o menor medida, pueden ser analizados en términos de eficiencia y ser utilizados para mejorar aplicaciones de ingeniería. Específicamente, los modelos del SVH pueden ser usados para encontrar nuevas características estadísticas de las imágenes naturales, y más importante, para entender las restricciones y la finalidad del SVH.

El capítulo 2 analiza las propiedades estadísticas de un modelo computacional de los primeros estadios del SVH, el cuál está inspirado en datos psicofísicos y cuyos parámetros son ajustados con datos fisiológicos. Se mostrará de forma teórica y de forma práctica cómo este modelo obtiene una representación de los datos visuales con buenas características estadísticas.

1.6 Estadística cómo herramienta para optimizar un sistema

Desde un punto de vista estadístico, diseñar un sistema tan eficiente cómo el SVH es una tarea compleja. La *teoría de la decisión* nos enseña que existen dos ingredientes básicos a la hora de diseñar un sistema de inferencia: la función de densidad de probabilidad PDF

[6] En memoria de Pilar Celda e Isabel Marín.

[7] *Muestra* es un eufemismo para designar a los animales de laboratorio

(por sus siglas en inglés) de los datos y la función de coste asociada a los posibles eventos. Estos son los elementos de la función de riesgo de Bayes [Bernardo & Smith, 1994]. Normalmente, los métodos usados para explicar el SVH de forma estadística asumen una función de coste Euclídea o incluso la ignoran. Aunque una estimación correcta de la PDF de los datos visuales puede ofrecernos muchos detalles sobre cómo funciona el SVH, es necesario recordar que la función de coste es una parte fundamental para realizar inferencia. Incluso ignorando la función de coste, obtener una explicación estadística plausible del SVH es muy complicado, puesto que existen muchos problemas que resolver, cómo revisamos a continuación.

■ ¿Cuál es el objetivo del Sistema Visual Humano?

Ver el SVH cómo un sistema optimizado estadísticamente abre muchas cuestiones. La principal es que no está claro qué debe optimizar. Por supuesto, maximizar la cantidad de información extraída, podría ser un objetivo deseable. Intentos en esta dirección pretenden encontrar una transformación que obtenga una representación de los datos donde los componentes de los datos transformados sean independientes entre si. Aunque esta sería una representación deseable por muchas razones, no es un requisito imprescindible. Además, en algunos casos, este tipo de representaciones no son compatibles con el hecho de representar la información con el mínimo error de reconstrucción (la cuál es también una situación deseable). Por tanto, aunque la independencia es una opción útil [8] para obtener una descripción estadística de los datos, es posible que no sea el objetivo del SVH. De todos modos, el SVH debe emplear una estrategia para explotar las regularidades estadísticas del mundo visual. Lo cuál nos lleva directamente a la siguiente cuestión.

■ ¿Qué se pude entender por "mundo visual"?

Cuando se intenta extraer información de medidas, la calidad y representatividad de los datos adquiridos es igual de importante que el método para inferir información. En el estudio de la estadística de imágenes normalmente se usan datos *naturales*, los cuáles se reducen a imágenes de bosques [9]. El uso de este tipo de imágenes asume que el SVH se ha adaptado durante millones de años de los cuáles sólo últimamente existen cosas no naturales, hechas por el hombre (por ejemplo las primeras edificaciones datan de hace 5000 años). Sin embargo, el proceso de aprendizaje del SVH se desarrolla durante los primeros meses de vida, y se podría argumentar que estos meses son esenciales a la hora de definir el comportamiento final del SVH. Además, la capacidad de adaptación a cada situación específica también debería ser tenida en cuenta. Por tanto, durante la Tesis tomaremos una definición amplia de *datos naturales*: son los que nos envuelven hoy en día.

[8] En este caso se hace referencia a independencia estadística no espacial.

[9] La base de datos de imágenes naturales más usada es la de [Hateren & Schaaf, 1998].

Otro problema relacionado es qué clase de información deben contener los datos visuales utilizados para entrenar los sistemas. Es decir, ¿tiene algún sentido usar imágenes con información complicada (por ejemplo formas humanas) para entrenar modelos de las primeras etapas del SVH? ¿Va a utilizar el modelo dicha estructura? Imágenes no representativas únicamente desviarán los resultados del modelos y las expectativas del científico.

■ Restricciones del Sistema Visual Humano

Cuando se piensa en el cerebro cómo un sistema existe una gran cantidad de restricciones. La más obvia es el tamaño, es decir el limitado número de neuronas. Saber la cantidad exacta de neuronas dedicadas a cada tarea específica sería una información muy útil [10]. Otras restricciones importantes son: la capacidad de procesado de una neurona, la cantidad de ruido neuronal y la velocidad de las diferentes neuronas. Aunque se ha realizado mucha investigación en este sentido, los métodos estadísticos rara vez introducen estas restricciones.

■ Recursos computacionales

La optimización del sistema visual se ha realizado durante 2.700 millones de años. Lo cuál da una idea de su complejidad [11]. Además, la optimización específica del SVH en cada ser humano es realizada durante meses, lo cuál implica una gran cantidad de datos, alrededor de 1.500 millones de imágenes de alta resolución. En el mejor de los casos, los métodos estadísticos utilizados para extraer características del SVH son optimizados utilizando 60.000 imágenes de baja resolución cómo mucho[12].

■ Métodos estadísticos

La capacidad de computación de las máquinas actuales nos ha permitido pensar en métodos estadísticos de extracción de información inimaginables hace unos años, o incluso usar métodos *imaginados* hace años pero con una complejidad computacional inabordable hasta ahora. Esto implica que, en los últimos años, ha aparecido una gran variedad de métodos estadísticos (y lo que es incluso mejor, algunos son útiles). Casi todos los algoritmos de máquinas de aprendizaje han sido aplicados en problemas de procesado de imágenes. Muchos de ellos podrían ser interpretados para poder extraer información de

[10] La proporción de neuronas entre el núcleo geniculado lateral y la corteza visual primaria V1 es de 1:1000 [H. B. Barlow, 2001], lo cuál sugiere que la hipótesis de reducción de redundancia no se aplica a esta etapa.

[11] Esta afirmación implica la aceptación de la teoría de la evolución cómo un hecho. Sin embargo, desde un punto de vista creacionista el SVH esta diseñado *a imagen y semejanza* de un ser superior, lo cuál es también inabordable desde un punto de vista computacional.

[12] La mayoría de las bases de datos contienen menos de 10.000 imágenes de baja resolución.

cómo debería funcionar el SVH. Sin embargo, ninguno de los métodos estadísticos aplicados hasta la fecha es capaz de reproducir completamente el funcionamiento del cerebro.

Dentro de los problemas de los métodos estadísticos merece una mención especial *la maldición de la dimensionalidad*. Esta establece que la cantidad de datos necesarios para estimar una PDF plausible incrementa exponencialmente con la dimensión. Este problema genera que los métodos estadísticos dependientes de los datos siempre obtienen una solución sesgada y parcial.

En el capítulo 4, se presentan tres trabajos diferentes que implican extracción de características del SVH usando métodos estadísticos. En la sección 4.1, se presenta un método para diseñar estadísticamente un sistema de sensores con distintas funciones objetivo y se entrena para extraer algunas características de los mecanismos de color del SVH. En la sección 4.2 se propone un método basado en máxima verosimilitud para obtener características de el área visual primaria V1 del SVH. En la sección 4.3 se evalúa la habilidad de los filtros lineales para conseguir una representación donde los componentes de texturas naturales son independientes.

1.7 Neurociencia y estadística para el procesado de imágenes

Un modo indirecto de testear los modelos tanto estadísticos cómo del SVH es aplicando estos en problemas de procesado de imágenes. Además, proponer algoritmos útiles podría ser una fuente inagotable de ingresos para los científicos, o al menos, una manera de devolver a la sociedad la inversión hecha en él/ella. Los dos conceptos analizados aquí, *neurociencia* y *estadística*, pueden ser usados para inspirar dichos algoritmos. Por un lado, un modelo apropiado del SVH podría ser útil en muchas tareas sustituyendo al ser humano, por ejemplo para el reconocimiento atomático de objetos. En la sección 3.1 la habilidad de un modelo del SVH es utilizada para evaluar la calidad de imágenes. Por otro lado, una descripción la estadística de las imágenes es importante en el diseño de de algoritmos de procesado de imágenes. Esto salta a la vista al observar la formulación para el diseño óptimo de algunas tareas de procesado de imagen: cuantización [Gersho & Gray, 1992], limpieza de ruido [Portilla et al., 2003], clasificación [R. Duda & Hart, 1973], teoría de la información [Cover & Tomas, 1991] y síntesis [Bernardo & Smith, 1994]. En todas ellas es necesaria una estimación de la PDF. De hecho, la mayoría de estas tareas están basadas en la fórmula del riesgo de Bayes. En la sección 5.2 se utilizan dichas fórmulas a través de la estimación de PDFs multidimensionales estimadas mediante un método de Projection Pursuit. Concretamente, se presenta una extensión del método que lo hace computacionalmente útil y se evalúa en diferentes problemas de procesado de imágenes. Sin embargo, cómo se ha dicho antes, obtener una PDF que explique correctamente datos multidimensionales es complicado. Muchos métodos están basados en el uso de las regularidades estadísticas de los datos pero sin estimar implícitamente la PDF, unos de los mas popu-

lares son los métodos kernel. En la sección 5.1 se explora la capacidad de uno de estos métodos, las Máquinas de Vectores Soporte para Regresión, para utilizar información estadística de las imágenes. Esta capacidad es evaluada usando el método en una aplicación de limpieza de ruido.

1.8 Organización de la Tesis

La Tesis está organizada siguiendo el esquema de la figura 1.2. Las flechas que salen de *neurociencia* están basadas en la formulación del modelo clásico del SVH (hasta V1). El modelo se revisa, se expande y se ajusta psicofísicamente. La flecha que llega a *estadística* se desarrolla en el capítulo 2 y hace referencia a los resultados obtenidos en [Malo & Laparra, 2010a]. En este trabajo se analizan las propiedades estadísticas del modelo: factorización aproximada de la PDF de las imágenes naturales y reducción sustancial de la información mutua. Nótese que no se ha utilizado información estadística en el modelo. Estos resultados son una evidencia complementaria a favor de la hipótesis de la codificación eficiente.

La parte relacionada con la flecha que llega a *aplicaciones* se expone en el capítulo 3, y hace referencia a los resultados obtenidos en [Laparra, Marí, & Malo, 2010]. En este trabajo el modelo es aplicado en evaluación de calidad de imágenes. Experimentos sobre un amplio número de bases de datos que incluyen un amplio rango de distorsiones muestran que este modelo obtiene resultados similares a las últimas aproximaciones propuestas en el campo, y es fácil de interpretar en términos lineales.

Las flechas que salen de *estadística* hacen referencia básicamente a métodos de aprendizaje aplicados para, o bien explicar aspectos del SVH o para obtener herramientas de procesado de imágenes. La flecha que llega a *neurociencia* se desarrolla en el capítulo 4 y hace referencia a tres trabajos: [Laparra, Jiménez, et al., 2011a], [Laparra, Gutman, et al., 2011] y [Laparra & Bethge, 2011]. El primer trabajo, [Laparra, Jiménez, et al., 2011a], se desarrolla en la sección 4.1 y propone un método de diseño de sistema de sensores con dos características principales: (i) la forma de los sensores puede ser no lineal, y (ii) la métrica de los sensores pude ser seleccionada con distintos criterios. Este método se aplica sobre colores de imágenes naturales obteniendo un sistema de sensores con un comportamiento similar a los mecanismos de color del SVH: el sistema es no lineal y adaptativo a los cambios en el entorno. La capacidad de adaptación se evalúa utilizando una base de datos creada especialmente para el problema, donde de muestran distintos tipos de objetos bajo los iluminantes D65 y A, y por tanto se evitan las limitaciones de las bases de datos existentes. Los resultados obtenidos sugieren que la percepción de color en este nivel de abstracción está posiblemente determinada por una estrategia de minimización del error [D. MacLeod & Twer, 2003] en lugar de una estrategia de maximización de la información [Laughlin, 1983]. Otros trabajos realizados en el marco de la Tesis relacionados son [Laparra & Malo, 2008a,b; Laparra, Tuia, et al., 2011].

El segundo trabajo, [Laparra, Gutman, et al., 2011], está expuesto en 4.2 y propone una extensión del Análisis de Componentes Independientes ICA (por sus siglas en inglés) Complejo aplicado a imágenes naturales. Se muestra que el ICA complejo lineal aprende características de las células complejas del área visual primaria V1, obteniendo filtros de Gabor en cuadratura de fase. Los métodos convencionales que realizan un ICA complejo asumen distribuciones de fase uniforme en la señal de salida. En este trabajo se suaviza esta asunción modelando la distribución de fase de las salidas. El modelo resultante muestra que las distribuciones son muchas veces no uniformes, y las formas de los filtros de Gabor también cambian. El trabajo [Laparra & Bethge, 2011] se expone en 4.3. En él se mide la cantidad de reducción de redundancia que se puede obtener mediante diferentes transformaciones lineales. Los resultados subrayan la idea de que la forma de los filtros en V1 es posible que no se deba a un objetivo de independencia en esta etapa del cerebro. También se puede observar que la habilidad de las distintas transformaciones lineales depende de las características de las imágenes a tratar, y por tanto, la adaptación de estos filtros a diferentes situaciones es necesaria para una codificación eficiente de los datos.

Los resultados referentes a la flecha que sale de *estadística* y llega a *aplicaciones* se presentan en el capítulo 5. Estos resultados han sido publicados en [Laparra, Gutiérrez, et al., 2010] y [Laparra, Camps-Valls, & Malo, 2011]. El primer trabajo muestra la habilidad de los métodos kernel para incluir información estadística. Específicamente se realiza regresión utilizando máquinas de vectores soporte en el dominio wavelet para imponer características de imágenes naturales a imágenes ruidosas. Las relaciones son obtenidas de medidas de información mutua realizadas sobre una base de datos de imágenes representativa. Los resultados muestran que: (1) el método propuesto mejora métodos convencionales que asumen independencia entre coeficientes y (2) obtiene resultados similares a métodos bien establecidos. Por tanto, la aproximación propuesta puede ser vista cómo una alternativa más flexible a las que usan una descripción explícita de las relaciones de los coeficientes wavelets. Otros trabajos realizados en el marco de la Tesis relacionados con métodos kernel son [Armengot et al., 2010; Camps-Valls et al., 2011, 2010; Laparra et al., 2008]. El segundo trabajo presenta un método para estimar PDFs multidimensionales, basado en técnicas projection pursuit Friedman & Tukey [1974]. El marco general consiste en la aplicación secuencial de transformaciones de Gausianización marginales, seguidas de una transformación ortonormal. El procedimiento propuesto transforma los datos a una distribución Gaussiana multidimensional, por tanto el valor de la PDF en el dominio original puede ser estimado en cuálquier punto. Se muestra que, al contrario que en el projection pursuit básico, la clase particular de rotaciones usada no tiene un impacto relevante sobre los resultados cuálitativos en el resultado, puesto que la búsqueda de proyecciones interesantes no es crítica para la estimación de la PDF. La diferenciabilidad, invertibilidad y convergencia del método son analizadas teóricamente y experimentalmente. El uso práctico del método se ilustra en diferentes problemas de procesado de imágenes: síntesis,

clasificación, limpieza de ruido y estimación de la información mutua. Otros trabajos relacionados, realizados en el marco de la Tesis son [Laparra et al., 2009; Laparra et al., 2009].

El capítulo 6 resume las conclusiones generales aprendidas durante el proceso de la realización de la Tesis.

Chapter 2

From Neuroscience to Statistics

2.1 Statistical Properties of Divisive Normalization Model

Horace Barlow suggested that functional properties of biological vision sensors should be matched to the signal statistics faced by these sensors [H. Barlow, 1961]. The conventional approach to confirm the plausibility of such efficient coding hypothesis goes *from image statistics to perception*.

Over the last decades a number of evidences in the above *conventional* direction have been reported. First, the shape of the linear receptive fields in V1 was derived using different network architectures and learning algorithms to optimize different statistical criteria such as energy minimization, enforcing decorrelation of the outputs or maximizing the mutual information between input and output: for instance, in [Linsker, 1986; T. Sanger, 1989; T. D. Sanger, 1990] low-pass filtered random noise was used as a rough model for natural images to feed the networks, while [Foldiak, 1989] focused on information transmission. Then, more attention was devoted to statistical independence beyond decorrelation. When higher order moments are considered in natural images (using linear ICA), sets of localized and oriented edge detectors are found [Bell & Sejnowski, 1997; Hateren & Schaaf, 1998; Olshausen & Field, 1996]. Another linear feature of perception explained from the spectrum of natural images and maximization of signal to noise ratio is the spatial frequency sensitivity [Van Hateren, 1992, 1993]. Van Hateren works also explain a global non-linear dependence on the luminance in accordance with Weber's law.

More recently, attention has shifted from the linear receptive fields and the luminance non-linearity to the specific non-linearities of V1 cells, namely surround effects and contrast adaptation or gain control. In this case, parametric models using divisive normalization [Schwartz & Simoncelli, 2001] or other specific non-linearities [Kayser et al., 2003] have been fitted using image statistics and efficient coding arguments. Feedback and feedforward connections in hierarchical networks have been used to reproduce surround inhibition [Rao & Ballard, 1999]. Non-parametric approaches, such as non-linear ICA used

in [Malo & Gutiérrez, 2006], exemplifies the *image statistics to perception* way of reasoning since the right non-linearities directly emerge from the images using a not perceptually inspired functional form.

However, despite the above evidences, nowadays there is a productive debate about the generality of the efficient coding hypothesis, or the strict applicability of redundancy reduction arguments [H. B. Barlow, 2001; E. Simoncelli, 2003]. In this debate, two complementary lines of research are possible:

- The conventional direction, *from image statistics to perception*, as described above.
- The reverse direction, i.e. *from perception to image statistics*. This approach starts from the response of real neurons at different stages along the visual pathway, or equivalently from the response of a psychophysical model. When such a perception system is stimulated with natural images it is possible to obtain statistical measurements about the transmitted signal at different processing stages. The eventually good statistical behavior of the perceptual responses at a certain stage (e.g. independence) suggests that the efficient coding hypothesis is correct, since the brain is reducing the redundancy in the signal along the visual pathway, even though no statistical information was used in computing these responses (direct recordings or perceptually transformed signals).

In this work we take the second approach:

We show that the psychophysical divisive normalization masking model has appealing statistical properties (e.g. factorization of the PDF of natural images) even though no statistical information is used to fit the model. Therefore, this work can be seen as the *reverse approach* version of [Malo & Gutiérrez, 2006; Schwartz & Simoncelli, 2001], thus providing an original evidence in favor of the efficient coding hypothesis.

The structure is as follows. In section 2.1.1 we review the standard non-linear model of the V1 visual cortex and propose a new (indirect) psychophysical procedure to set its parameters. In our case, the model parameters are obtained to predict perceived distortions on a large subjectively rated database. Details on the parameters setting will be latter presented in chapter 3, section 3.1.2. Chapter 3 also shows that the proposed model works better than state-of-the-art image quality metrics. Section 2.1.2 analytically shows how the proposed perception model may factorize a plausible PDF for natural images (which captures local image dependencies). Section 2.1.3 empirically shows the good statistical behavior of the perceptual model when confronted to natural images: the non-linear part of the V1 model strongly reduces the mutual information between coefficients of the previous linear stage and approximately achieves the predicted component independence, thus confirming the match between the psychophysical model and the image statistics. Section 2.1.4 shows that the fitted model qualitatively reproduces traditional psychophysics (frequency sensitivity and masking). Finally, section 2.2 draws the conclusions of the work.

2.1.1 The Divisive Normalization V1 model

The perceptual image representation considered here is based on the standard psychophysical and physiological model that describes the early visual processing up to the V1 cortex. The linear part of the model describes the shape of the receptive fields as linear edge detectors tuned to different scales [J. Daugman, 1980; A. Watson, 1983, 1987], and accounts for the threshold contrast sensitivity [Campbell & Robson, 1968; Malo, Pons, Felipe, & Artigas, 1997; Mullen, 1985]. The non-linear part of the model accounts for the non-linearities related to contrast masking [Carandini & Heeger, 1994; Carandini et al., 1997; Foley, 1994; Heeger, 1992; A. Watson & Solomon, 1997]. In this model, the input image, $\mathbf{x} = (x_1, \cdots, x_N)$, is first analyzed by a set of wavelet-like linear sensors, $\mathbf{T}_{ij}$, that provide a scale and orientation decomposition of the image [J. Daugman, 1980; A. Watson, 1983, 1987]. The linear sensors have a frequency dependent linear gain according to the Contrast Sensitivity Function (CSF), $\mathbf{S}_{ii}$, [Campbell & Robson, 1968; Malo, Pons, Felipe, & Artigas, 1997; Mullen, 1985]. The weighted response of these sensors is non-linearly transformed according to the divisive normalization gain control, $\mathbf{R}$ [Carandini & Heeger, 1994; Carandini et al., 1997; Foley, 1994; Heeger, 1992; A. Watson & Solomon, 1997]:

$$\mathbf{x} \xrightarrow{\mathbf{T}} \mathbf{w} \xrightarrow{\mathbf{S}} \mathbf{w}' \xrightarrow{\mathbf{R}} \mathbf{r} \tag{2.1}$$

In this scheme, the set of local-frequency analyzers (matrix $\mathbf{T}$) and the slopes of their responses (matrix $\mathbf{S}$) constitute the linear part of the model. The diagonal in $\mathbf{S}$, is described by a function that depends on the scale, $e = 1, 2, 3, 4$, (e ranges from fine to coarse), may depend on the orientation, $o = 1, 2, 3$, (the o values stand for horizontal, diagonal and vertical), but it is constant for every spatial position, $\mathbf{p}$:

$$S_i = S_{(e,o,\mathbf{p})} = A_o \cdot \exp\left(-\frac{(4-e)^\theta}{s_o^\theta}\right) \tag{2.2}$$

where A_o is the maximum gain for the considered orientation, s_o controls the bandwidth of the frequency response, and θ determines the sharpness of the decay with spatial frequency. The rows of the matrix $\mathbf{T}$ contain the linear receptive fields of V1 neurons. In this model we used an orthogonal 4-scales QMF wavelet transform[1] [E. Simoncelli & Adelson, 1990] to model such receptive fields. $\mathbf{S}$ is a diagonal matrix containing the linear gains to model the CSF. Finally, $\mathbf{R}$ is the divisive normalization response which describes the non-linear behavior:

$$\mathbf{R}(\mathbf{w}')_i = r_i = \text{sign}(w_i') \frac{|S_{ii} \cdot w_i|^\gamma}{\beta_i^\gamma + \sum_{k=1}^n H_{ik} |S_{kk} \cdot w_k|^\gamma} \tag{2.3}$$

where $\mathbf{H}$ is a kernel matrix that controls how the responses of neighboring linear sensors, k, affect the non-linear response of sensor i. The constants β_i determine the minimum contrast for significant response saturation.

[1] http://www.cns.nyu.edu/∼lcv/software.php

Even though in the original use of Divisive Normalization for image quality purposes [Teo & Heeger, 1994] the interaction kernel weights every sensor in a certain neighborhood in the same way, here we use the Gaussian interaction kernel proposed by Watson and Solomon [A. Watson & Solomon, 1997], which has been successfully used in block-frequency domains [Camps-Valls et al., 2008; Epifanio et al., 2003; Gutiérrez et al., 2006; Malo et al., 2006], and in steerable wavelet domains [Laparra, Gutiérrez, et al., 2010]. In the orthogonal wavelet domain this reduces to:

$$H_{ik} = H_{(e,o,\mathbf{p}),(e',o',\mathbf{p}')} = K \cdot \exp\left(-\left(\frac{(e-e')^2}{\sigma_e^2} + \frac{(o-o')^2}{\sigma_o^2} + \frac{(\mathbf{p}-\mathbf{p}')^2}{\sigma_p^2}\right)\right) \tag{2.4}$$

where $(e, o, \mathbf{p})$ and $(e', o', \mathbf{p}')$ refer to the scale, orientation and spatial position meaning of the wavelet coefficients i and k respectively, and K is a normalization factor to ensure $\sum_k H_{ik} = 1$.

In our implementation of the model we set the profile of the regularizing constants β_i according to the standard deviation of each subband of the wavelet coefficients of natural images in the selected wavelet representation. This is consistent with the interpretation of the values β_i as priors of the amplitude of the coefficients [Schwartz & Simoncelli, 2001]. This profile (computed from 100 images of a calibrated image data base [Olmos & Kingdom, 2004]) is further multiplied by a constant b to be set in the optimization process. Section 3.1.2 gives further details on the parametrization and the optimization process.

The color version of the V1 response model involves the same functional form of spatial transforms described above applied to the image channels in an opponent color space [Martínez-Uriegas, 1997]. In particular, we used the standard YUV (luminance, yellow-blue, red-green) representation [Pratt, 1991]. According to the well known differences in frequency sensitivity in the opponent channels [Mullen, 1985], we will allow for different matrices $\mathbf{S}$ in each channel. We will assume the same behavior for the other spatial transforms since the non-linear behavior of the chromatic channels is similar to the achromatic non-linearities [Martínez-Uriegas, 1997].

The natural way to set the parameters of the model is empirical: by fitting low-level perception data, either physiological recordings [Heeger, 1992] or threshold psychophysics [A. Watson & Solomon, 1997]. This low-level approach is not straightforward because the experimental literature is often interested in a subset of the parameters, and a variety of experimental settings is used (e.g. different stimuli, different contrast definitions, etc.). As a result, it is not easy to unify the wide range of data into a common computational framework. Alternative (theoretical) approaches involve using image statistics and the efficient coding hypothesis to derive the parameters [Malo & Gutiérrez, 2006; Olshausen & Field, 1996; Schwartz & Simoncelli, 2001]. Obviously, this is not an option in our case since our aim is assessing the statistical efficiency of a non-statistically optimized model.

Instead, in this work we used an empirical but *indirect* approach: we set the parameters of the model to reproduce experimental (but higher-level) visual results such as image

quality assessment as in [A. Watson & J.Malo, 2002]. In particular, we optimized the V1 model to obtain an image distortion metric that maximizes the correlation with the subjective ratings of a subset of the LIVE Quality Assessment Database[2] [Sheikh, Sabir, & Bovik, 2006]. Section 3.1.2 gives further details on the parametrization and the optimization process.

Figure 2.1 shows the optimal values for the linear gains $\mathbf{S}$, the saturation constants, β^γ, and the interaction kernel H. Note that the interaction kernel is (1) convolutional (i.e., each coefficient is normalized by other nearby coefficients) and (2) for each coefficient, neighbors are taken only from the orientation bands at the same scale. The optimal value for the excitation and inhibition exponent was $\gamma = 1.7$. An implementation of the proposed model is available on-line[3].

Section 2.1.4 shows that the obtained model simultaneously accounts for a wide variety of suprathreshold distortions as well as for the basic trends of threshold psychophysics (e.g. frequency sensitivity and contrast masking).

2.1.2 PDF factorization through V1 Divisive Normalization

In this section, we assume a plausible joint PDF model for natural images in the wavelet domain and we show that this PDF is factorized by a divisive normalization transform, given that some conditions apply. The analytical results shown here predict quite characteristic marginal PDFs in the transformed domain. In section 2.1.3 we will empirically check the predictions made here by applying the normalization model proposed above to a set of natural images.

Image model

It is widely known that natural images display a quite characteristic behavior in the wavelet domain: on the one hand, they show heavy-tailed marginal PDFs, $P_{w_i'}(w_i')$ (see Fig. 2.2), and, on the other hand, the variance of one particular coefficient is related to the variance of the neighbors. These relations are easy to see by looking at the so called bow-tie plot: the conditional probability of a coefficient given the values of its neighbors, $P(w_j'|w_i')$, normalized by the maximum of the function for each value of w_i' (see Fig. 2.2). In this representation tilting of the conditional density suggests that the coefficients are correlated, but more importantly, it can be seen that the variance of one coefficient strongly depends on the variance of the neighbor. These observations on the marginal and conditional PDFs have been used to propose leptokurtotic functions to model the marginal PDFs [Hyvärinen, 1999b; E. Simoncelli, 1997; E. P. Simoncelli, 1999] and models of the

[2] http://live.ece.utexas.edu/research/quality/

[3] http://www.uv.es/vista/vistavalencia/standard_V1_model/

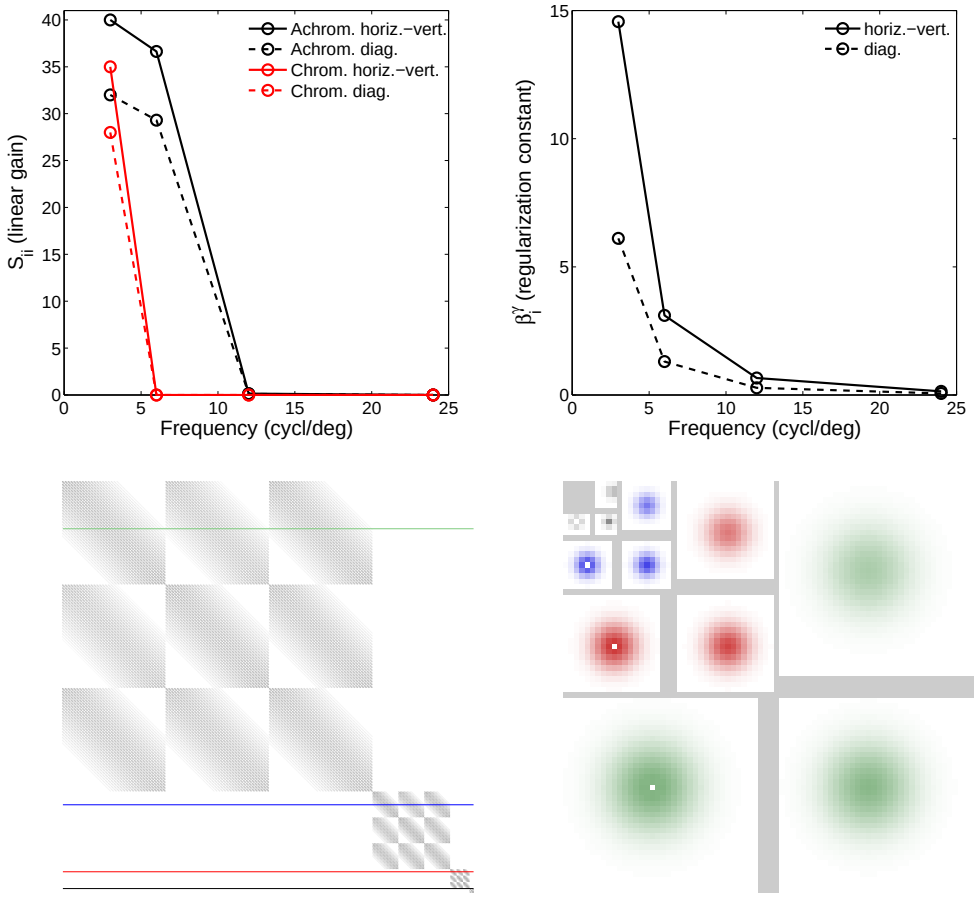


Figure 2.1: *Linear gains S (top left), saturation constants β^γ (top right), and kernel H (bottom left). The particular structure of the interaction kernel comes from the particular arrangement of wavelet coefficients used in the transform [E. Simoncelli & Adelson, 1990]. The bottom right figure shows the individual rows highlighted in different colors in the kernel figure. Each row corresponds to the particular coefficients in white in the bottom right figure. The different shades of color represent the interaction intensity with the spatial and orientation neighbors. In this example we assumed 72×72 discrete images sampled at 64 cycles per degree. According to this, the spatial extent of the subbands is 1.125 degrees.*

conditional PDFs in which the variance of one coefficient depends on the variance of the neighbors [Buccigrossi & Simoncelli, 1999; Schwartz & Simoncelli, 2001].

Inspired on these conditional models, we propose the following joint PDF (for the N-dimensional vectors $\mathbf{w}'$), in which, each element of the diagonal matrix, Σ, depends on the neighbors:

$$P_{\mathbf{w}'}(\mathbf{w}') = \frac{1}{Z} \frac{1}{|\Sigma(\mathbf{w}')|^{1/2}} e^{-\frac{1}{2}\mathbf{w}'^T \cdot \Sigma(\mathbf{w}')^{-1} \cdot \mathbf{w}'} \tag{2.5}$$

where

$$\Sigma_{ii}(\mathbf{w}') = (\beta_i^\gamma + \sum_j H_{ij} \cdot |w_j'|^\gamma)^{\frac{2}{\gamma}}, \tag{2.6}$$

and Z is simply a normalization constant.

The diagonal matrix $\Sigma(\mathbf{w}')$ can be thought as playing similar role as the covariance matrix in a regular Gaussian PDF. However, note that $\Sigma(\mathbf{w}')$ is point dependent (i.e. it is not a covariance matrix), and even though it is diagonal, it introduces relations among the energies of neighbor coefficients (see eq. (2.6)). Therefore, this joint PDF *is not* Gaussian, and the coefficients of the wavelet transform are not independent since the joint PDF, $P_{\mathbf{w}'}(\mathbf{w}')$, cannot be factorized by its marginal PDFs, $P_{w_i'}(w_i')$.

The proposed PDF is inspired by the models used in [Buccigrossi & Simoncelli, 1999; Schwartz & Simoncelli, 2001] since it tries to describe the relations among neighbor coefficients in wavelet domains using linear combinations of them. The differences include (1) the specific exponent, a sort of norm, γ, applied to the coefficients of the wavelet transform used in the linear combination (whether you consider amplitudes, $\gamma = 1$, as in [Buccigrossi & Simoncelli, 1999]; energy, $\gamma = 2$, as in [Schwartz & Simoncelli, 2001]; or some generic γ, here); and (2) the fact that here we are proposing a joint PDF model while in those cases the model was conditional.

A 2D example using the above joint PDF illustrates its suitability to capture the reported marginal and conditional behavior of wavelet coefficients: see the predictions shown in Fig. 2.2.

V1 normalized components are approximately independent

Here we compute the PDF of the natural images in the divisive normalized representation assuming (1) the above image model, and (2) the match between the parameters of the V1 representation and the parameters of the image model. Specifically, the match between the denominator in the perceptual response (eq. 2.3) and the matrix Σ in the image model (eq. 2.6).

We will use the fact that given the PDF of a random variable, $\mathbf{w}'$, and some transform, $\mathbf{r} = \mathbf{R}(\mathbf{w}')$, the PDF of the transformed variable can be computed by [Stark & Woods, 1994],

$$P_{\mathbf{r}}(\mathbf{r}) = P_{\mathbf{w}'}(\mathbf{R}^{-1}(\mathbf{r})) \cdot |\nabla_{\mathbf{r}} \mathbf{R}^{-1}| \tag{2.7}$$

Considering that the divisive normalization (in vector notation) can be expressed as

$$\mathbf{r} = sign(\mathbf{w}')\, \Sigma(\mathbf{w}')^{-\frac{\gamma}{2}} \cdot |\mathbf{w}'|^\gamma \tag{2.8}$$

where $|\cdot|^\gamma$ is an element-wise exponentiation, the inverse $\mathbf{R}^{-1}$ can be obtained from one of these (equivalent) expressions [Malo et al., 2006]:

$$|\mathbf{w}'|^\gamma = (\mathbf{I} - \mathbf{D}_{|\mathbf{r}|}\mathbf{H})^{-\mathbf{1}} \cdot \mathbf{D}_{\mathbf{fl}} \cdot |\mathbf{r}| \tag{2.9}$$

$$\mathbf{w}' = sign(\mathbf{r})\Sigma(\mathbf{w}')^{\frac{1}{2}} \cdot |\mathbf{r}|^{\frac{1}{\gamma}} \tag{2.10}$$

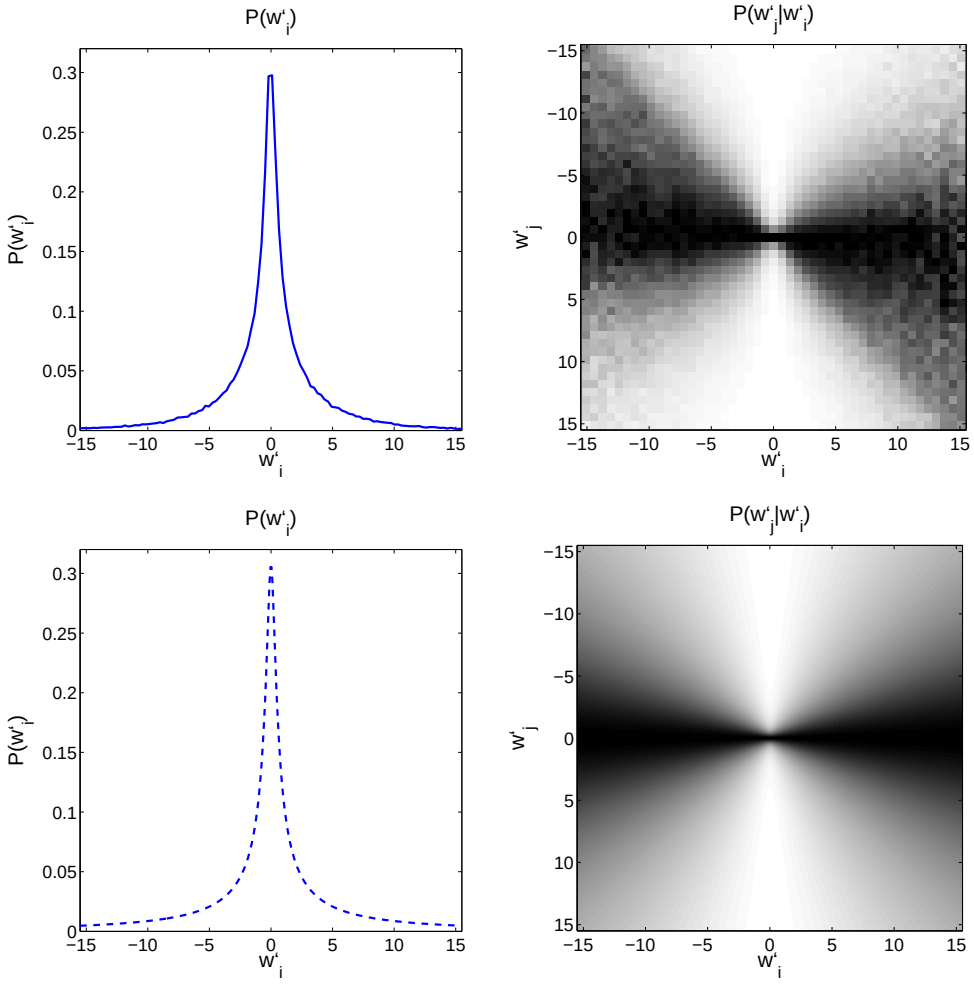


Figure 2.2: *Top: empirical behavior of wavelet coefficients of natural images (marginal PDF -left- and conditional PDF -right-). Darker values indicate higher probability. Bottom: simulated behavior according to the proposed model. In this 2D experiment we considered two coefficients of the second scale of* $\mathbf{w}'$ *(computed for 10000 images of the database [Olmos & Kingdom, 2004], using* $3 \cdot 10^6$ *samples). We used* $S_{ii} = 0.14$, $\beta_i = 0.4$, $H_{ii} = 0.7$ *and* $H_{ij} = 0.3$ *and* $\gamma = 1.7$, *according to the psychophysically fitted model.*

where $\mathbf{D_v}$ are diagonal matrices with the vector $\mathbf{v}$ in the diagonal. Plugging $\mathbf{w}'$, eq. 2.10, into the image model we obtain,

$$P_{\mathbf{w}'}(\mathbf{R}^{-1}(\mathbf{r})) = \frac{1}{Z} \frac{1}{|\Sigma(\mathbf{w}')|^{1/2}} e^{-\frac{1}{2}(|\mathbf{r}|^{1/\gamma})^T \cdot I \cdot (|\mathbf{r}|^{1/\gamma})} \tag{2.11}$$

Taking derivatives on the inverse, eq. 2.9, the determinant of the Jacobian is:

$$|\nabla_{\mathbf{r}} \mathbf{R}^{-1}| = det\left(\frac{1}{\gamma}\Sigma(\mathbf{w}')^{1/2} \cdot D_{|\mathbf{r}|^{\frac{1}{\gamma}-1}} \cdot \left(I + \underbrace{D_{\beta^{-\gamma}} \cdot H \cdot (I - D_{|\mathbf{r}|}H)^{-1} \cdot D_{\beta^{\gamma}} \cdot D_{|\mathbf{r}|}}_{M(\mathbf{r})}\right)\right)$$

$$|\nabla_{\mathbf{r}}\mathbf{R}^{-1}| = det\left(\frac{1}{\gamma}\Sigma(\mathbf{w}')^{1/2}\cdot D_{|\mathbf{r}|^{\frac{1}{\gamma}-1}}\cdot(I+M(\mathbf{r}))\right)$$

$$|\nabla_{\mathbf{r}}\mathbf{R}^{-1}| = |\Sigma(\mathbf{w}')|^{1/2}\cdot\prod_{i=1}^{N}\frac{1}{\gamma}|r_i|^{\frac{1}{\gamma}-1}det(I+M(\mathbf{r}))^{\frac{1}{N}}$$

Since $det(I+M(\mathbf{r}))^{\frac{1}{N}}\approx 1$ in natural images[4], it follows,

$$|\nabla_{\mathbf{r}}\mathbf{R}^{-1}| \approx |\Sigma(\mathbf{w}')|^{1/2}\cdot\prod_{i=1}^{N}\frac{1}{\gamma}|r_i|^{\frac{1}{\gamma}-1} \tag{2.12}$$

Therefore, from Eqs. (2.7), (2.11) and (2.12), it follows that the joint PDF of the normalized signal is just the product of N functions that depend solely on r_i:

$$P_{\mathbf{r}}(\mathbf{r}) \approx \prod_{i=1}^{N}\frac{1}{\gamma\, Z^{1/N}}\,|r_i|^{\frac{1}{\gamma}-1}\,e^{-\frac{|r_i|^{2/\gamma}}{2}} = \prod_{i=1}^{N}P_{r_i}(r_i) \tag{2.13}$$

i.e., we have factorized the joint PDF into its marginal PDFs.

Note that our proof of factorization holds whenever the normalization transform is performed using parameters that are matched to those of the image probability model. Nevertheless, the shape of the marginal densities of the normalized coefficients *d*o depend on those parameters. Figure 2.3 illustrates this fact by showing marginal PDF's for different values of γ.

In particular, if the appropriate value were $\gamma = 1$, the transform would give rise to Gaussian marginal PDFs thus becoming similar to Radial Gaussianization transforms as suggested in [Lyu & Simoncelli, 2009]. However, note that different values of γ in the transform would imply a better (or worse) match between the denominator of the normalization and the matrix Σ of the image model. This match is required to achieve the factorization result in Eq. (2.13).

2.1.3 Statistical results

This section assesses the statistical independence performance of the psychophysically fitted V1 representation (i.e. the validity of Eq. 2.13) by (1) analyzing the marginal and conditional probabilities of the transformed coefficients, and (2) by mutual information measures. To do so, 10000 natural image patches of size 72×72 from the McGill database [Olmos & Kingdom, 2004] were considered and transformed to the linear V1 representation (the wavelet domain), and to the non-linear V1 representation.

[4]We found that the average value and standard deviation of this determinant on 10000 images taken from McGill calibrated image dataset [Olmos & Kingdom, 2004] is: $\langle det(I+M(\mathbf{r}))^{\frac{1}{N}}\rangle = 1.013\pm 0.003$.

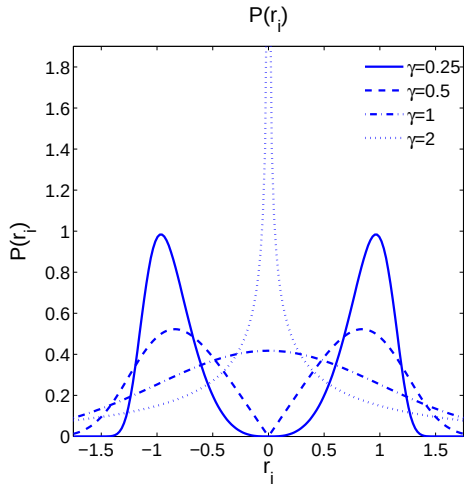


Figure 2.3: *Family of marginal PDFs of the normalized coefficients r_i as a function of γ.*

Marginal and conditional PDFs

Figure 2.4 shows the experimental and the predicted marginal and conditional PDFs in the normalized domain. These results correspond to two spatial neighbors of the second scale and horizontal orientation ($3 \cdot 10^6$ 2D samples). Similar results are obtained for other subbands. For the sake of illustration, in the case of the marginal PDFs, we show the results for different values of the exponent γ in the transform: the psychophysically optimal value, $\gamma = 1.7$, and other values, $\gamma = 0.5$ and $\gamma = 0.25$, due to the characteristic bimodal shape of the predicted marginal PDFs in those cases (see Fig. 2.3).

Bimodal results are obtained in the marginal PDFs for the (psychophysically non-optimal) values of γ as predicted by the theory. However, note that the agreement with the theoretical prediction is much better for the psychophysically optimal exponent, thus indicating the match of the psychophysical vision model to image statistics.

The result for the conditional probability shows that the vision model substantially reduces the redundancy among neighbor coefficients with regard to the linear wavelet representation: note that the bow-tie has practically disappeared (compare with the equivalent result in Fig. 2.2), in close agreement with the theoretical prediction.

Mutual Information results

Mutual information (MI) between pairs of neighbor coefficients of image samples in the spatial domain, in the linear V1 image representation (wavelet domain), and in the V1 non-linear representation were computed. The eventual reduction of MI values would point out the redundancy reduction along the visual pathway. In order to assess the magnitude of the achieved reductions we also include the results of two non-linear statistically-based techniques designed to give rise to independent components in images: Radial Gaussian-

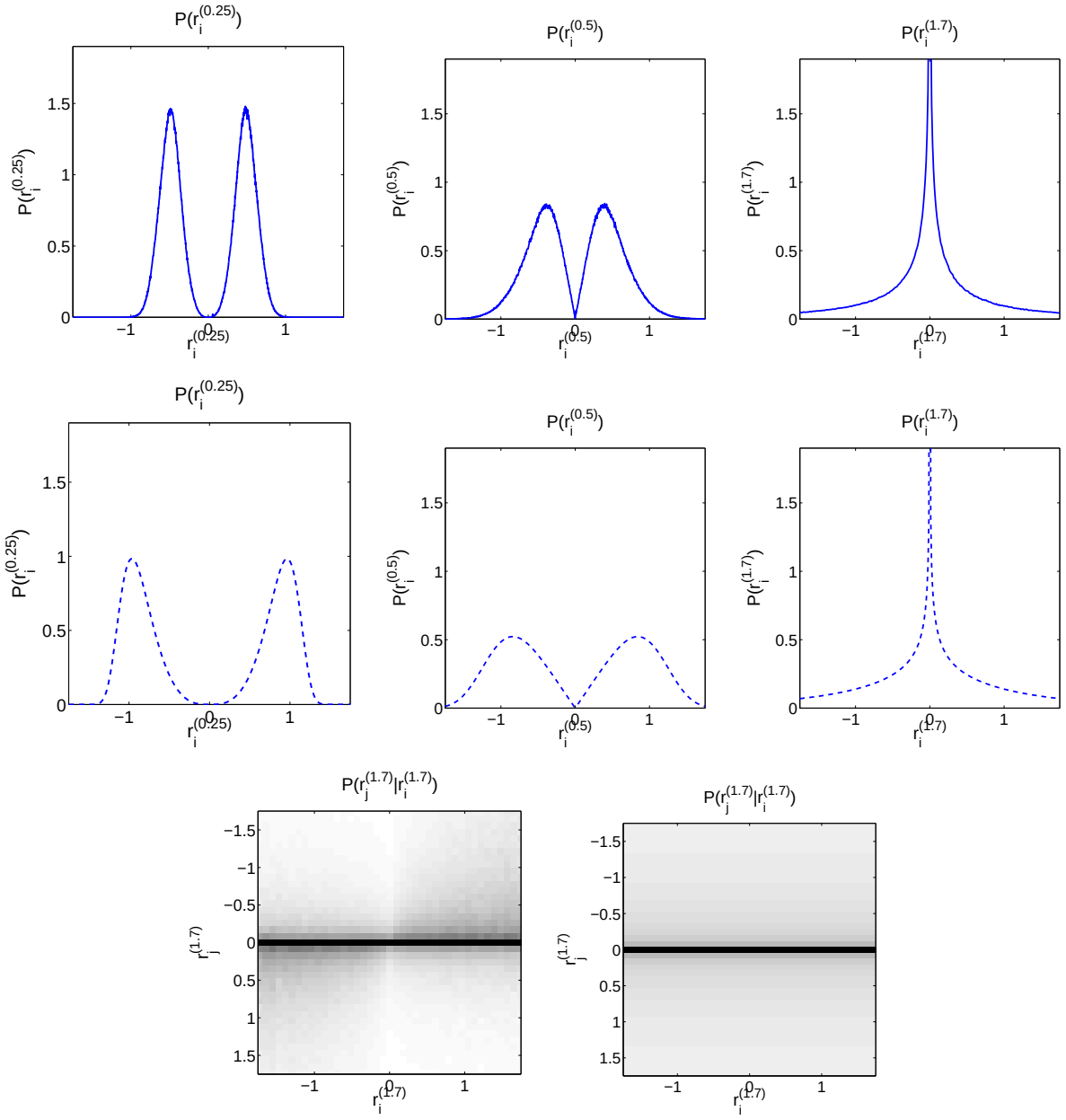


Figure 2.4: *Marginal and conditional PDFs in the response domain. The first row shows the experimental marginal PDF of the responses for illustrative values of the exponent $\gamma = 0.25, 0.5, 1.7$. The second row shows the corresponding predictions according to the theoretical results in section 2.1.2. The third row shows the experimental (left) and the theoretical (right) conditional distributions for pairs of coefficients of the psychophysically optimal V1 model ($\gamma = 1.7$).*

ization using L_2 norm as in [Lyu & Simoncelli, 2009], and L_p norm as in [Eichhorn et al., 2009]. These transforms start from a whitened linear representation of image vectors fol-

lowed by a univariate (radial) non-linear transform tuned to obtain Gaussian distribution of the L_2 or L_p lengths of the vectors.

In order to make the comparison easier, we used the same initial linear stage (wavelets) in those non-linear transforms. It is true that orthogonal wavelets may not be the best linear transform to achieve independence, but it is important to stress that (1) the selected linear stage is not critical for the final independence results obtained by using Radial Gaussianization techniques as pointed out in [Eichhorn et al., 2009], and (2) the aim of this work is not looking for the ultimate transform to achieve independence, but to show that the brain substantially reduces redundancy through the gain control non-linearity. The second non-linear stage in these illustrative Gaussianization techniques was performed by equalizing the L_2 and L_p lengths respectively, as done in [Lyu & Simoncelli, 2009]. In our simulations we used $p = 1.2$ in the L_p norm according to the results in [Eichhorn et al., 2009]. This is the optimal norm for ICA, while other linear representations are optimal for exponents in the range [1.2, 2]. As stated above, choosing different linear representations with norm exponents in the cited range gives rise to similar independence results [Eichhorn et al., 2009].

Section 2.1.3 gives the details on the used MI estimator and its errors: it shows that the errors are small compared to the MI differences presented in this section, thus ensuring the significance of the differences.

We performed two experiments:

1. The first one tries to obtain a rough estimate of the global redundancy reduction ability of the linear (wavelet) and the non-linear (divisive normalization) stages of the V1 model. In this experiment we computed the MI among one coefficient and all the other coefficients (both in the spatial domain and in the local frequency domains, also including Radial Gaussianization using both L_2 and L_p).

2. The second experiment consists of a more accurate analysis of the different possible relations in the local frequency representations, $\mathbf{w}$, $\mathbf{r}$, and Radial Gaussianization using L_2 and L_p: (1) *intra-band*, measuring the MI of one coefficient with its $9 \times 9 - 1$ neighbors of the same subband, (2) *inter-orientation*, measuring the MI of one coefficient with its corresponding 5×5 spatial neighbors in a subband of the same scale but different orientation, and (3) *inter-scale*, measuring the MI of one coefficient in a coarser scale with its 2×2 sons in the corresponding finer scale.

Figure 2.5 shows representative results of the first experiment. In each MI computation 10^4 2D samples from the McGill database [Olmos & Kingdom, 2004] were used. The MI values in the spatial domain monotonically decrease with distance, as previously reported in [Lyu & Simoncelli, 2009]. The MI values among neighbors in the local frequency domains decrease as the distance in space, orientation and scale increases. The behavior is similar for coefficients of other scales and orientations.

As expected, the statistically tuned Gaussianization techniques obtain quite good independence results on the considered data set. Interestingly, the psychophysically tuned transform (that uses no statistical optimization at all) obtains very similar results in redundancy reduction. These results show that about 86% of the average MI in the spatial domain is reduced by the linear wavelet transform, while the non-linear psychophysical transform further reduces an additional 82% of the remaining MI in the linear wavelet domain. As a consequence, the non-linear V1 representation reduces about 98% of the average MI in the spatial domain, which is comparable to the reductions achieved by the statistically tuned Radial Gaussianization techniques using L_2 norm (99.2%) and L_p norm (99.5%).

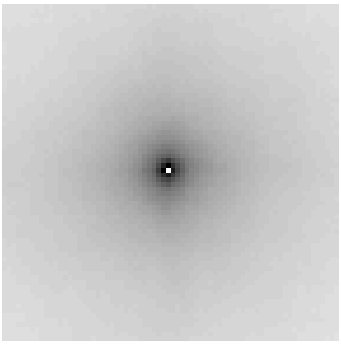
Figure 2.6 shows a representative subset of the results of the second experiment: intra-band and inter-orientation MI values for the different orientations of the second scale, and inter-scale MI values for parents of the third scale and the corresponding sons of the second scale. Overlapping blocks of the different subbands were used to obtain more samples for a reliable MI estimation. The intra-scale, inter-orientation and inter-scale results were computed using $0.8 \cdot 10^6$, $1.3 \cdot 10^6$, and $0.7 \cdot 10^6$ 2D samples respectively.

Again, the results show that the statistically tuned Radial Gaussianization transforms substantially reduce the redundancy among the different neighbor coefficients with regard to the linear wavelet representation. The psychophysically optimal divisive normalized representation (second column) achieves very similar results. This means that the redundancy removal obtained through the psychophysical transform is significant. This is consistent with the removal of the bow-tie relations in the conditional probability plots (Figure 2.4).

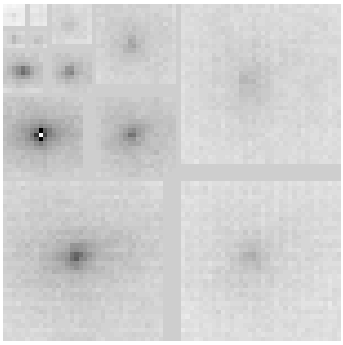
Moreover, it is interesting to note the similarity between the exponents to be used in the L_p norm in [Eichhorn et al., 2009], and the psychophysical value for γ in the V1 normalization (which normalizes each wavelet coefficient by a sort of γ norm of its neighbors). In the first case, the value for better independence is in the range [1.2, 2]. In the psychophysically tuned V1 transform, $\gamma = 1.7$.

The fact that relatively more redundancy is reduced in the intra-band and inter-orientation cases may be due to the quantization of the masking kernel which was necessary for practical computational reasons (see comment in section 3.1.2). The quantization of the kernel in divisive normalization removes inter-scale interactions so the normalization is not as effective in that situation. Note that the optimal kernel in figure 2.1 does not reflect inter-scale interactions.

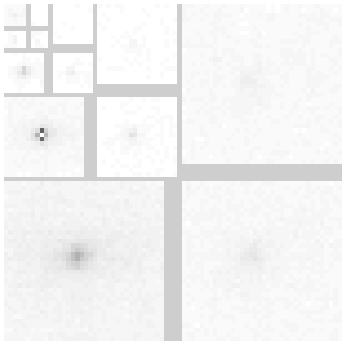
Summarizing, the agreement between the experimental and the predicted marginal and conditional PDFs of $\mathbf{r}$, and the substantial MI reduction with regard to the linear wavelet domain confirm the theoretical result in eq. (2.13): the psychophysically optimal divisive normalization is well matched to image statistics and approximately factorizes the PDF of



$MI_{\mathbf{x}}$: *0.38*, [0.21, 1.70] bits



$MI_{\mathbf{w}}$: *0.052*, [10^{-3}, 0.27] bits



$MI_{\mathbf{r}}$: *0.009*, [10^{-4}, 0.15] bits

$MI_{RG\,L_2}$: *0.003*, [$6 \cdot 10^{-5}$, 0.06] bits

$MI_{RG\,L_p}$: *0.002*, [10^{-5}, 0.05] bits

Figure 2.5: *MI results between one coefficient (the one in white) and its neighbors in the spatial domain (top) the linear V1 response, wavelet domain (middle left); the non-linear V1 response domain (middle right); the Radial Gaussianization using L_2 norm (bottom left); and the Radial Gaussianization using L_p norm (bottom right). The numbers in each case represent the average and the range of MI values found in bits. The top figure is scaled so that the black and white correspond to the maximum MI value in the spatial domain, 1.70 bits, and 0 bits respectively. All the other figures are scaled with regard to the maximum MI value in the wavelet domain: white and black correspond to the limits of the range [0, 0.27] bits.*

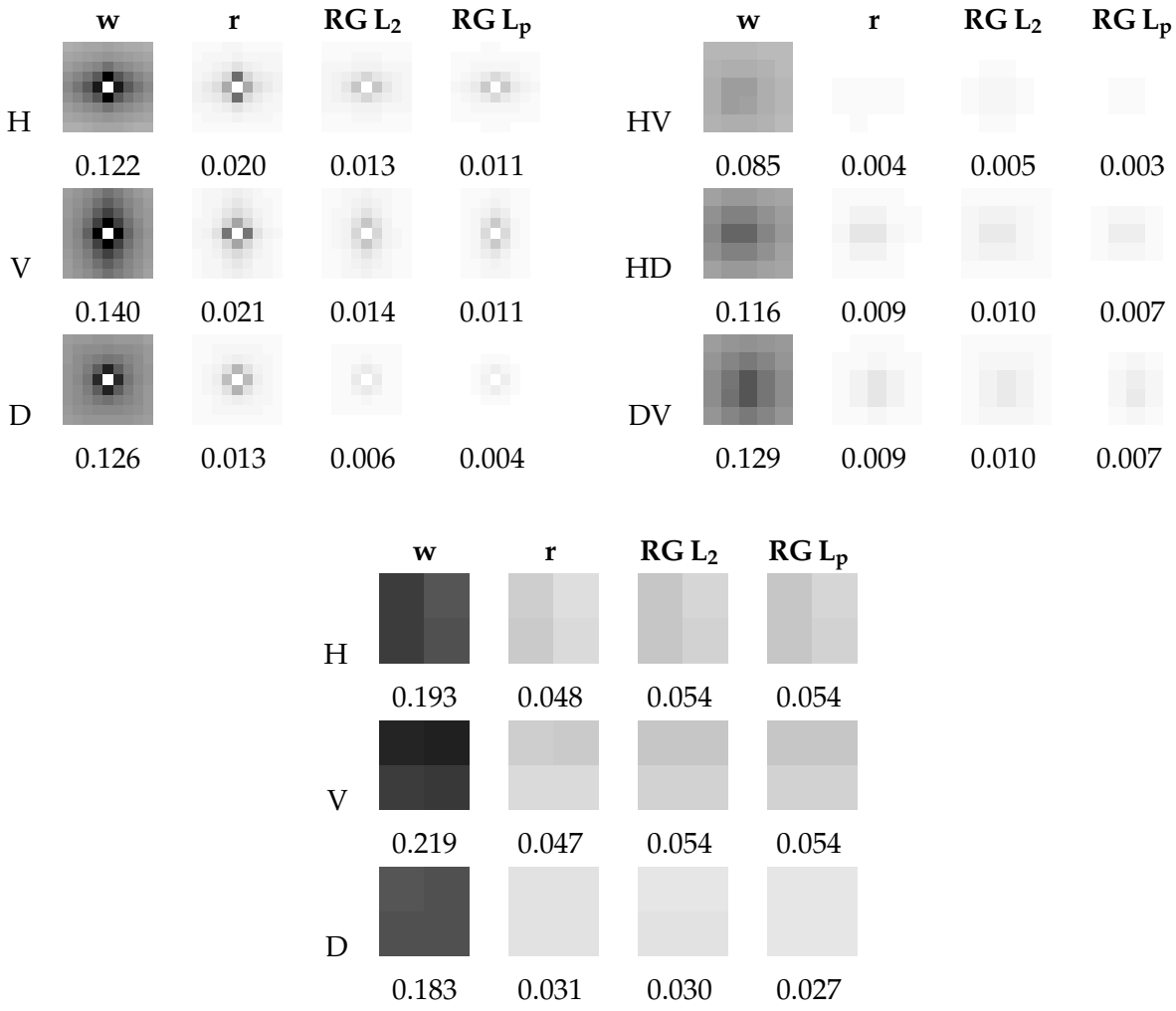


Figure 2.6: *MI (in bits) between pairs of coefficients in the linear V1 representation (wavelet,* $\mathbf{w}$*) and in the non-linear V1 representation (normalized response,* $\mathbf{r}$*). The last two columns in each panel show the results of Radial Gaussianization techniques using* L_2 *norm and* L_p *norm respectively. The top left panel shows intra-band relations within 2nd scale subbands of different orientation. The top right panel shows inter-orientation relations for the 2nd scale coefficients. The bottom panel shows inter-scale relations of coefficients of the 3rd scale with their sons in the 2nd scale for different orientations. All images are scaled so that back and white correspond to the maximum MI value and 0 bits respectively. The numbers represent the average MI value (in bits) in each image.*

natural images.

Measuring Mutual Information

Mutual information (MI) between two random variables is defined as the difference between the sum of marginal entropies and the joint entropy [Cover & Tomas, 1991]:

$$MI(v_1, v_2) = h(v_1) + h(v_2) - h(v_1, v_2) \tag{2.14}$$

Since MI is invariant under point-wise transforms [Cover & Tomas, 1991], our MI estimator first equalizes the marginal PDF of each coefficient to obtain uniform densities in the range $[0, 1]$. Then, the joint entropy is computed by using the 2D histogram and the Miller-Madow correction [Miller, 1955]. In our implementation, the total number of bins in the 2D histogram was set to the square root of the number of available samples. In our case, the marginal entropies are zero due to the uniformization step. Therefore the MI is equal to minus the joint entropy.

In order to assess the accuracy of the above estimator we tested it for two particular densities of known MI: (1) Gaussian densities, whose MI can be computed in closed-form [Cover & Tomas, 1991], and (2) the image model in the wavelet domain, Eq. 2.5, whose MI can be obtained by numerical integration of the joint PDF.

In Table 2.1 we show the mean and the standard deviation of the percentage of error for 2D PDFs of different MI as a function of the number of samples used in the estimation. The explored range of MI values is $[0.01, 0.32]$ bits, and the number of samples is in the range $[10^4, 10^6]$. These error percentages have been obtained with 100 different realizations for each sample size. These results ensure that the estimation error is always below the MI differences shown in section 2.1.3.

2.1.4 Reproducing low-level and high-level psychophysics

In this section we show that the model optimized to account for (high-level) image quality opinion also accounts for the fundamental trends of (low-level) threshold psychophysics.

Figures 2.7-2.9 show the results of three experiments: (1) reproduction of subjectively rated distortion, (2) reproduction of frequency-dependent threshold contrast sensitivity, and (3) reproduction of contrast masking non-linearities.

In the first experiment, the generalization ability and robustness of the model to account for a wide variety of suprathreshold distortions is assessed by checking its performance on a more general image quality database (with more images and distortions of different nature than the used for fitting the model). Here we applied the model (optimized for 83 images) to predict distortions on the whole LIVE database (779 distorted images), plus on the whole TID database [Ponomarenko et al., 2008]. The extension to the TID database is

Gaussian PDFs						
	Number of samples ($\times 10^4$)					
MI	1	2.5	6.3	15.8	39.8	100
0.01	14 ± 18	9 ± 10	5 ± 6	3 ± 3	2 ± 2	1.3 ± 1.5
0.04	7 ± 7	4 ± 4	4 ± 3	2 ± 2	2.1 ± 1.3	1.6 ± 0.7
0.09	8 ± 5	5 ± 3	4 ± 2	2.8 ± 1.3	2.4 ± 0.8	1.7 ± 0.5
0.14	8 ± 4	5 ± 2	4 ± 1.5	3.3 ± 1.0	2.5 ± 0.6	1.9 ± 0.4
0.19	8 ± 3	6 ± 2	5 ± 1.5	3.5 ± 0.9	2.6 ± 0.6	1.9 ± 0.3
0.24	8 ± 3	6 ± 2	5 ± 1.2	3.3 ± 0.7	2.7 ± 0.5	2.0 ± 0.3
0.28	8 ± 2	6 ± 1.8	5 ± 0.9	3.7 ± 0.7	2.6 ± 0.4	2.0 ± 0.3
0.31	8 ± 2	6 ± 1.6	5 ± 1.1	3.7 ± 0.6	2.8 ± 0.4	2.0 ± 0.2
0.32	9 ± 2	6 ± 1.6	5 ± 1.0	3.7 ± 0.6	2.8 ± 0.4	2.1 ± 0.2
Image PDF model in the wavelet domain (Section 2.1.2)						
	Number of samples ($\times 10^4$)					
MI	1	2.5	6.3	15.8	39.8	100
0.21	8 ± 3	5 ± 2	2.9 ± 1.4	1.5 ± 0.6	1.5 ± 0.6	1.0 ± 0.3

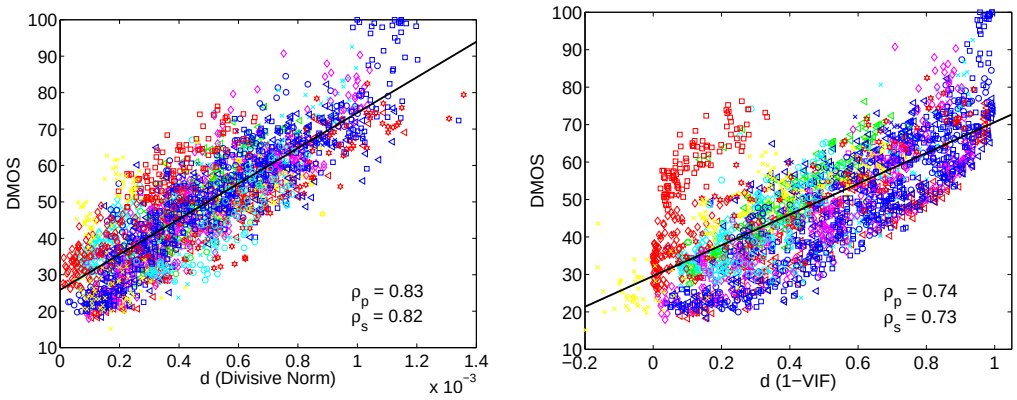
Table 2.1: Relative error (in %) of the mutual information estimator on Gaussian densities and on the proposed image model in the wavelet domain.

challenging since it does not only contain different images, but more importantly, it includes 12 kinds of distortion not included in the LIVE database. The model was finally applied to 2479 distorted images. The performance of the proposed model (figure 2.7.a) can be compared to the performance of the state-of-the-art Visual Information Fidelity index (VIF) [Sheikh & Bovik, 2006] of the same authors as the LIVE database (figure 2.7.b). Note that the VIF metric fails to account for some of the distortions in the TID database (represented by different symbols/colors in the plots) while the proposed V1 image representation model obtains significantly better correlation when considering a wide range of distortions (see the Pearson and Spearman correlation coefficients at the plots). More details on the performance of the proposed model as image quality metric are given in section 3.1. The Matlab implementation of the metric is available on-line[5].

The second experiment shows how the model accounts for the threshold frequency sensitivity. Here, the response of the model to a given incremental pattern (target), $\Delta\mathbf{x}$, seen on top of a background, $\mathbf{x}$, is computed as the perceptual distance $d(\mathbf{x}, \mathbf{x} + \Delta\mathbf{x})$. The CSF can be simulated by computing the above distances between sinusoids with fixed contrast, but different frequencies and orientations, and a uniform gray background. Figures 2.8.a

[5] http://www.uv.es/vista/vistavalencia/div_norm_metric/

○ Gauss. in color channels	× Spatially correl. noise	◇ Masked noise
□ High Freq.	★ Impulse noise	◁ Color Quantization
× Denois. artifacts	★ JPG transm. errors	◁ J2K transm. errors (LIVE)
○ Non-ecc patttern	□ Local block-wise	◇ Mean shift
× Contrast scale	□ J2K compression (LIVE)	◇ JPG compression (LIVE)
◁ Additive Gaussian (LIVE)	○ Gaussian blur (LIVE)	◁ Fast fading



(a) (b)

Figure 2.7: *Reproduction of high-level perception results. The figures show the correlation among the predicted distortion, d, and the observers opinion, DMOS, for the distance in the proposed V1 image representation (a), and the state-of-the-art VIF metric (b). The different symbols in the plot and legend represent images with distortions of different nature. For details on the different distortions see [Ponomarenko et al., 2008; Sheikh, Sabir, & Bovik, 2006].*

and 2.8.b compare the result of this simulation for achromatic sinusoids in a wide range of spatial frequencies with the corresponding achromatic CSF of the Standard Spatial Observer [A. Watson & Ramirez, 2000]. Note that the model approximately reproduces the band pass behavior, the overall bandwidth, and the oblique effect.

The third experiment simulates contrast masking results. In order to do so, the contrast of a Gabor patch is increased on top of different backgrounds (sinusoids with different contrasts and orientations). As widely known [Foley, 1994; A. Watson & Solomon, 1997], the visibility of the target increases quickly for low contrast targets, while remains more stable for higher contrast targets, thus revealing a non-linear response. Moreover, the visibility of the target is reduced as the contrast of the background is increased. This effect is bigger when the the background has the same orientation as the target.

Figures 2.9.a and 2.9.b show the response curves of the model to vertical targets for the different background sets: vertical (left) and horizontal (right). The model response to the target is a saturating non-linearity when the target is shown on top of no back-

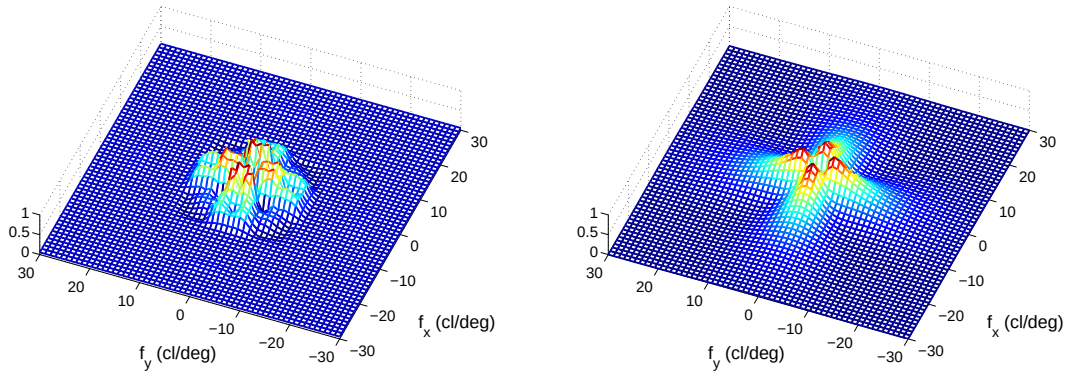


Figure 2.8: *Reproduction of (low-level) frequency-dependent sensitivity. In the plots, the achromatic CSF as predicted by the proposed V1 model (a) is compared to the Standard Spatial Observer CSF (b).*

ground (auto-masking). The model predicts the reduction of the response when the target is shown on top of a background (cross-masking). The reduction increases with the contrast of the mask. Moreover, note that the reduction in visibility is bigger for backgrounds of the same nature (vertical target and vertical background). Therefore, the behavior of the model with the proposed parameters is compatible with the low-level behavior of human observers reported elsewhere [A. Watson & Solomon, 1997].

Figures 2.9.c and 2.9.d show contrast incremental thresholds ΔC for non-zero mask contrast as a function of the test contrast. These plots have been obtained from the previous response curves (with $C_{mask} = 0.1$) looking for the amount of contrast deviation needed to obtain a constant increment in the response (or distance). The left plot corresponds to target and background of the same frequency and orientation while the right plot corresponds to the orthogonal orientation situation. In both cases the thresholds increase with contrast (as expected from saturating responses). However, when target and background have the same orientation the sensitivity is reduced (thresholds increase faster). Figures 2.9.e and 2.9.f show equivalent experimental results by Foley [Foley, 1994] explicitly reproduced from [A. Watson & Solomon, 1997], which display the same behavior.

To summarize, the results in this section show that the divisive normalization model optimized to reproduce high level distortions (such as those in the LIVE database) can simultaneously reproduce the basic features of low-level psychophysics (e.g. frequency sensitivity and contrast masking), while being robust enough to account for a wider range of suprathreshold distortion data (TID database).

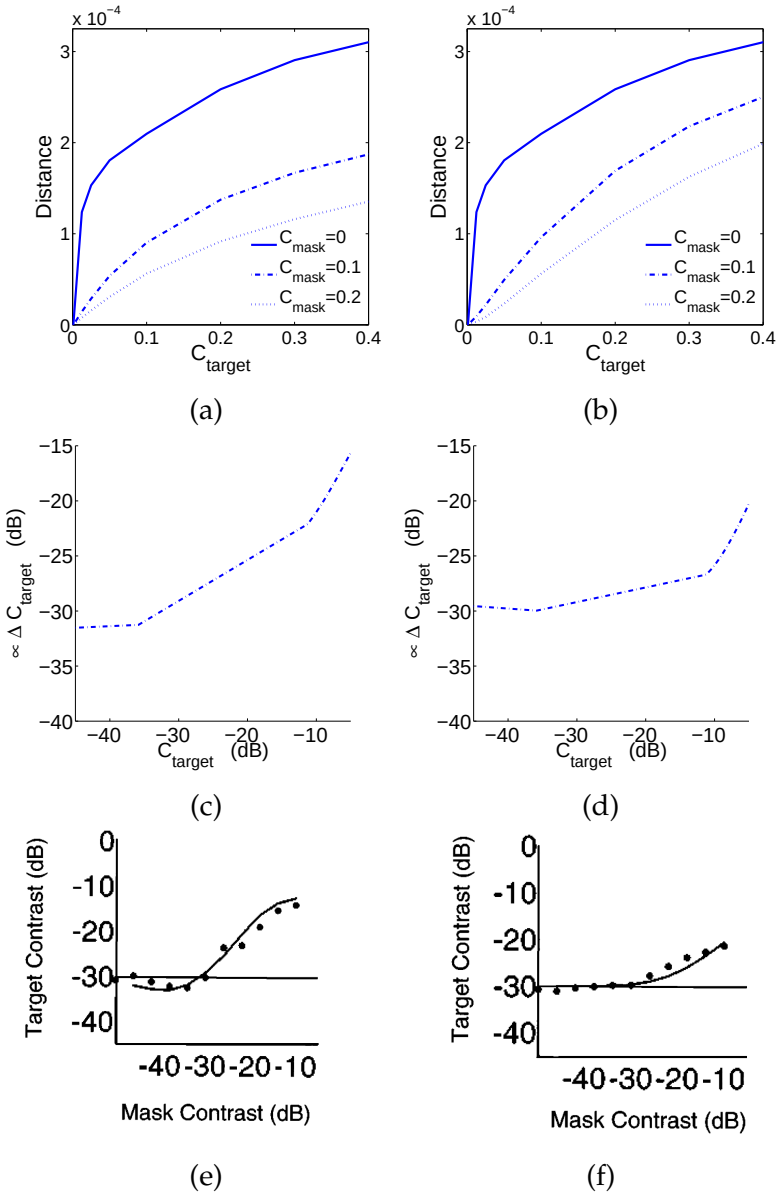


Figure 2.9: *Reproduction of (low-level) masking non-linearities and contrast incremental thresholds. Top row non-linear response: (a) response to Gabor targets of increasing contrast seen on top of sinusoids of the same frequency and orientation, and (b) equivalent responses on top of orthogonal sinusoids. Middle row: contrast incremental thresholds ΔC as a function of the test contrast when mask and test have the same orientation (c) and orthogonal orientations (d). Bottom row: equivalent ΔC psychophysical data by Foley [Foley, 1994], as reported in [A. Watson & Solomon, 1997]. In the middle and bottom rows contrast is expressed in dB: $C_{dB} = 20 \log_{10} C$.*

2.2 Chapter conclusions

Here we showed that the non-linear stage of the standard V1 cortex model optimized to reproduce image quality psychophysics substantially increases the independence of the image coefficients obtained in the linear stage. Theoretical results (confirmed by experiments) show that the psychophysically tuned V1 model approximately factorizes a plausible joint PDF for natural images in the wavelet domain: bow-tie dependencies are almost removed and redundancy among coefficients is substantially reduced.

Therefore, the results presented here confirm the efficient coding hypothesis in a novel direction: *from perception to image statistics*. These results complement the standard approach to validate the hypothesis (e.g. *from image statistics to perception*) taken in [Kayser et al., 2003; Malo & Gutiérrez, 2006; Olshausen & Field, 1996; Schwartz & Simoncelli, 2001].

It is true that redundancy reduction is not the only goal in early visual processing [H. B. Barlow, 2001], but the results presented here suggest that this initial set of perceptual transforms performs a sort of non-linear independent components extraction.

Further work should address additional issues such as (1) internal noise, and (2) redundancy of the sign (or phase) of the wavelet coefficients. On the one hand, it is worth noting that efficient coding is not equivalent to redundancy reduction, except in the noise-free case [Dayan & Abbott, 2001; E. Simoncelli & Olshausen, 2001]. That is the case here since the assumed V1 model is deterministic. In the more general stochastic case, the factorization of the PDF is an idealized goal, and may need refinement under conditions of response noise (surely present in real neurons). On the other hand, sign or phase information is not taken into account in the image model since the PDF is symmetric around the origin, and signs (and their eventual relations) are not modeled in any way. The proposed divisive normalization model does not take this issue into account either since it acts on the amplitude of the wavelet coefficients. A separate or complementary model for the signs of image coefficients is needed. Extensions of the perception model could be fitted by using the specific distortions in subjectively rated image databases consisting of phase alteration (e.g. fast fading or JPEG2000 transmission errors).

Chapter 3

From Neuroscience to Applications

3.1 Divisive Normalization model as image quality metric

Reproducing subjective opinion of image distortion has two broad applications: in *engineering*, image quality metrics may replace the (time consuming) human evaluation to assess the results of the algorithms, and in *vision science*, image quality results may provide insight on the way the brain processes visual information.

Nowadays there is a fruitful debate about the right approach in the image quality assessment problem. The image quality metrics have been classified according to the following broad taxonomy [Wang & Bovik, 2009]: (1) error visibility techniques based on human visual system (HVS) models, (2) structural similarity techniques, and (3) information theoretic techniques.

The classical *error visibility* approach to simulate human judgement naturally tried to include empirical aspects of the HVS in the image metric [Ahumada, 1993; A. B. Watson, 1993]. Basic features taken into account include decomposition in orientation and scale channels [Lubin, 1993], contrast sensitivity [Ahumada, 1993; Nill, 1985; Saghri et al., 1989; X. Zhang & Wandell, 1996], and contrast masking non-linearities through simple pointwise models [Barten, 1990; Daly, 1990; Malo, Pons, & Artigas, 1997], the more general Divisive Normalization [Epifanio et al., 2003; Malo et al., 2006; Teo & Heeger, 1994], or equivalently, the non-uniform nature of just noticeable differences (JND) [Chandler & Hemami, 2007]. The final distance measure is typically obtained from a certain summation norm of the difference vector in the internal image representation domain (Minkowski pooling) [Ahumada, 1993; A. Watson & Solomon, 1997].

Recently, alternatives to the above empirical approach have been proposed: *structural similarity* methods [Wang et al., 2004a,b; Wang & Simoncelli, 2005b] and *information theoretic* methods [Sheikh & Bovik, 2006; Sheikh et al., 2005]. The common ground of these new techniques rely on the relation between image statistics and the behavior of the visual system [H. Barlow, 1961; Schwartz & Simoncelli, 2001]: since the organization and

non-linearities of visual sensors seem to emerge from image statistics [Malo & Gutiérrez, 2006; Olshausen & Field, 1996], it is sensible to assess the image distortion by measuring the departure of the corrupted image from the average behavior of natural images.

The *structural similarity* approach quantifies visual quality by comparing three statistical measures in the original and distorted images: mean (related to luminance), variance (related to contrast), and cross-correlation (related to structure). The aim of using a characterization of the structure is achieving invariance under small changes in the image [Wang & Simoncelli, 2005a]. This general concept has been applied both in the spatial domain (SSIM) [Wang et al., 2004a] and in multi-scale image representations (MSSIM) [Wang et al., 2004b] and (CW-SSIM) [Wang & Simoncelli, 2005b].

The *information theoretic* approach (VIF [Sheikh & Bovik, 2006]) quantifies the similarity by comparing the information that could ideally be extracted by the brain from the distorted and the original image, respectively. The authors assume a certain image source model and characterize the HVS as a simple channel that introduces additive noise in the wavelet domain. The amount of information that can be extracted from the original signal from the perceived images is modeled by the mutual information between the output of the above simplified HVS model and the original image.

Despite some of the new approaches claim to be a new philosophy [Wang et al., 2004a], a number of qualitative relations have been pointed out among the newer approaches and Divisive Normalization masking models [Seshadrinathan & Bovik, 2008; Sheikh & Bovik, 2006; Sheikh et al., 2005].

However, no explicit comparison has been done with metrics based on updated versions of the Divisive Normalization error visibility. Moreover, the new approaches criticize the classical error visibility approach in many ways:

- Suprathreshold problem. Since the empirical HVS models are based on near threshold measurements using too simple (academic) stimuli, it is argued that there is no guarantee that these models are applicable to suprathreshold distortions on complex natural images [Sheikh & Bovik, 2006; Wang et al., 2004a].

- Geometric limitations of error visibility techniques. In [Wang et al., 2004a], the authors criticize linear and point-wise non-linear HVS models because they give rise to too rigid discrimination regions, while stressing the flexibility of structural measures. Nevertheless, the authors (qualitatively) recognize that general Divisive Normalization models (including inter-coefficient masking) may induce a richer geometric behavior.

- Minkowski pooling assumes statistical independence among error coefficients. It has been argued that this is not an appropriate summation strategy in linear domains where there are statistical relations among coefficients [Wang et al., 2004a]. This criticism is certainly appropriate for linear (CSF-based) HVS models. Again,

this would not be the case for image representations with reduced relations among coefficients.

These un-addressed criticisms, and the fact that the new approaches are easy to use in a number of engineering applications [Wang & Bovik, 2009], have popularized the idea of their superiority over the error visibility approach.

The aim of this work is to provide new results in favor of the classical error visibility approach by showing that the above criticisms do not apply to the Divisive Normalization masking models, and by showing that what will be referred to as Divisive Normalization metric (originally proposed as image quality measure in [Teo & Heeger, 1994]) can be easily adapted to be competitive with the new approaches. This is an additional evidence to confirm the link among the different strategies, and suggests that, despite the criticisms, Divisive Normalization masking models should still be considered in the image quality discussion.

The structure is as follows. In Section 3.1.1 we review and generalize the Divisive Normalization masking model, and we show that the resulting metric successfully addresses the criticisms against the error visibility approach. Finally we show the relation of the proposed metric to other error visibility techniques. In Section 3.1.5 we compare the performance of the proposed metric to structural similarity techniques SSIM [Wang et al., 2004a] and MSSIM [Wang et al., 2004b], and information theoretic techniques VIF [Sheikh & Bovik, 2006]. An extensive comparison is made according to standard procedures in a number of recently available subjectively rated databases including a total of 2173 distorted images and 25 kinds of distortion. Finally in Section 3.2 we draw the conclusions of the work and discuss additional issues that may improve the Divisive Normalization performance.

3.1.1 The Divisive Normalization model as metric

In this section we first review the associated error visibility metric to the Divisive Normalization model exposed in section 2.1.1. In subsection 3.1.2 we describe the procedure to set the parameters of the model. Previously (in sections 2.1.4 to 2.1.3), we have addressed two of the main criticisms made against error visibility techniques: we showed (1) the ability to simultaneously reproduce high level and low level distortion data, and (2) the statistical independence effect that justifies uniform Minkowski pooling. Here also we are going to adress the third criticism: (3) the geometric richness of the model (Sec. 3.1.3). Finally in subsection 3.1.4 we show how the proposed metric relates to other error visibility metrics.

Given an input image, $\mathbf{x}$, and its distorted version, $\mathbf{x}' = \mathbf{x} + \Delta\mathbf{x}$, the above model provides two response vectors, $\mathbf{r}$, and $\mathbf{r}' = \mathbf{r} + \Delta\mathbf{r}$. The perceived distortion can be obtained through the appropriate pooling of the one dimensional deviations in the vector $\Delta\mathbf{r}$. Non-quadratic pooling norms have been reported [Ahumada, 1993; A. Watson & J.Malo, 2002;

A. Watson & Solomon, 1997]. Moreover, different summation exponents, for the pooling across spatial position, q_p, and frequency, q_f, may be used:

$$d_{pf}(\mathbf{x}, \mathbf{x}') = \frac{1}{n} \left[\sum_{\mathbf{f}} \left[[\sum_{\mathbf{p}} \Delta r_{\mathbf{fp}}^{q_p}]^{\frac{1}{q_p}} \right]^{q_f} \right]^{\frac{1}{q_f}} \tag{3.1}$$

$$d_{fp}(\mathbf{x}, \mathbf{x}') = \frac{1}{n} \left[\sum_{\mathbf{p}} \left[[\sum_{\mathbf{f}} \Delta r_{\mathbf{fp}}^{q_f}]^{\frac{1}{q_f}} \right]^{q_p} \right]^{\frac{1}{q_p}} \tag{3.2}$$

where $\mathbf{f} \equiv \{e, o\}$. In this general case, the order in which dimensions are pooled matters. Pooling across space and frequency is not commutative unless both pooling exponents are the same. In particular, Teo and Heeger proposed to compute the perceived distortion as the Euclidean norm of the difference vector (quadratic Minkowski pooling exponent $q_p = q_f = 2$).

According to the well known differences in frequency sensitivity in the achromatic and chromatic channels [Mullen, 1985], we will allow for different matrices $\mathbf{S}$ in the YUV channels. In particular, we will allow for different gains (A_{oY}, $A_{oU} = A_{oV}$) and different bandwidths (s_{oY}, $s_{oU} = s_{oV}$). We will assume the same behavior for the other spatial transforms since the non-linear behavior of the chromatic channels is similar to the achromatic non-linearities [Martínez-Uriegas, 1997].

3.1.2 Setting model parameters

In the original work introducing the metric based on Divisive Normalization [Teo & Heeger, 1994] and in the sequels [Epifanio et al., 2003; Malo et al., 2006] the parameters were inspired in psychophysical facts. In general there are three basic strategies to obtain the parameters of the model:

- The *direct empirical approach* implies fitting the parameters to reproduce direct low-level perception data such as physiological recordings on V1 neurons (as in [Heeger, 1992]), or psychophysical measurements of contrast incremental thresholds (as in [A. Watson & Solomon, 1997]). Since the realization of direct experiments is beyond the scope of this work, this low-level empirical approach is not straightforward because the physiological and psychophysical literature is often interested in a subset of the parameters, and a variety of experimental settings is used in these restricted experiments (e.g. different selected stimuli, different contrast units...). As a result, it is not easy to unify the wide range of experimental results into a common computational framework.

- The *indirect empirical approach* implies fitting the parameters of the model to reproduce higher level visual tasks such as image quality assessment: for instance, in

[A. Watson & J.Malo, 2002] the authors fitted the parameters of the Standard Spatial Observer to the VQEG subjectively rated data.

- The *statistically-based approach* assumes that the goal of the different signal transforms is to increase the independence among the coefficients of the image representation [H. Barlow, 1961; Malo & Gutiérrez, 2006; Olshausen & Field, 1996]. In this case, the parameters of the model may be optimized to maximize some statistical independence measure as in [Schwartz & Simoncelli, 2001].

In this work we take the second approach: we fitted the parameters of the Divisive Normalization metric to maximize the Pearson correlation with the subjective ratings of a subset of the LIVE Quality Assessment Database [Sheikh, Z.Wang, et al., 2006]. In order to point out the generalization ability of the proposed metric, we optimized the Divisive Normalization model just for 3 of the 27 images in the database (*house, sailing2* and *womanhat*) that represents about 10% of the available data. In Section 3.1.5 we not only test the behavior of the model in the whole dataset but also in other databases not including LIVE distortions (TID [Ponomarenko et al., 2008], IVC [Le Callet & Autrusseau, 2005], and Cornell [Chandler & Hemami, 2007] [1]). By using this testing strategy, we address one of the criticisms to the error visibility techniques: the model is applicable to a variety of new supra-threshold distortions, while still reproducing the low-level psychophysical results (as shown in Section 2.1.4).

Assuming the same behavior in the horizontal and vertical directions ($o = 1, 3$), and assuming that the oblique effect in the frequency sensitivity [A. Watson & Ramirez, 2000] is described by a single attenuation of the gain in the diagonal direction (i.e. $A_2 = d \cdot A_1$ in every chromatic channel), the model described so far has 13 free parameters:

$$\Omega \equiv \{ A_{1Y}, d, A_{1UV}, s_Y, s_{UV}, \theta, \gamma, b, \sigma_e, \sigma_o, \sigma_{\mathbf{p}}, q_s, q_f \}. \tag{3.3}$$

In order to simplify the optimization process, we did not explore all the dimensions of the parameter space at the same time, but optimized the parameters using a three stages procedure obtaining local optima in restricted subspaces. We first obtained the basic parameters of the model by neglecting the chromatic channels, the oblique effect and the non-quadratic summation, i.e. using $A_{1UV} = 0$, $d = 1$, and $q_s = q_f = 2$, thus reducing the dimensions of the parameter space to 8, $\Omega_1 \equiv \{ A_Y, s_Y, \theta, \gamma, b, \sigma_e, \sigma_o, \sigma_{\mathbf{p}} \}$. Afterwards, we checked the eventual improvements obtained from the previous (local) optimal configuration by considering the chromatic channels and allowing different values for the sensitivity in the diagonal direction, $\Omega_2 \equiv \{ A_{UV}, s_{UV}, d \}$. Finally, different summation exponents for the spatial and frequency pooling (in both possible orders) were considered $\Omega_3 \equiv \{ q_s, q_f \}$.

[1] Available at: `http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html`.

The only computational inconvenience of the proposed metric is the size of the kernel H. In order to circumvent this problem, two approximations were necessary:

- Kernel thresholding and quantization. The Gaussian interaction matrices were converted to sparse matrices by eliminating those elements below a given threshold, that in our experiments was set to $1/500$ of the maximum in each interaction neighborhood. Once the best Gaussian kernel was obtained, their size was further reduced by quantizing it using 6 bits. No appreciable reduction of the performance was introduced by this quantization, while extremely reducing the storage requirements.

- Limitation of the image size. The LIVE database include images of size 512×768. This size implies a huge kernel. Since the computation and storage of a number of non-quantized kernels is necessary for the optimization process, we decided to restrict ourselves to work with cropped versions of the images in the database. The cropped versions of the images were obtained by selecting the 256×256 area around the most salient point of each (original) image for 10 observers. The most salient point was estimated as the average of the points selected by the observers. This approximation is relevant just in the optimization process. Actually, the resulting Divisive Normalization is used for images of any size by applying it first to each 256×256 block of the image and then by merging the result of each block into a single pyramid.

The parameter ranges were set starting from an initial guess obtained from the low-level psychophysical behavior [A. Watson & Solomon, 1997] and previous use of similar models in image processing applications [Camps-Valls et al., 2008; Epifanio et al., 2003; Gutiérrez et al., 2006; Malo et al., 2006]. The explored ranges for the parameters and the optimal values found are shown in Table 3.1. The optimal pooling strategy found in our experiment was Eq. 3.2: first sum over subbands and then over spatial positions. Figure 2.1 shows the shape of the linear gains $\mathbf{S}$, the regularization constants β^γ and the interaction kernel H when using the optimal parameters. The structure of the interaction kernel comes from the particular arrangement of wavelet coefficients used in the transform [E. Simoncelli & Adelson, 1990].

3.1.3 Geometry of the Divisive Normalized domain

Assuming a quadratic pooling in the distance computation, a number of analytical results can be obtained that show the appealing geometric behavior of the proposed metric. This behavior still holds for non quadratic schemes.

In the quadratic summation case, the Euclidean metric, I, in the Divisive Normalization domain may be interpreted as using non-Euclidean (Riemannian) metrics, M, in other image representation domains [Epifanio et al., 2003; Malo et al., 2006]. The metric matrix,

Table 3.1: Parameter space, optimal values found, and improvement of the Pearson correlation in the progressive stages of the optimization.

Parameter	Range	Optimal	Correlation
A_Y	$30,\ldots,60$	**40**	
s_Y	$0.25,\ldots,3$	**1.5**	
θ	$2,\ldots,8$	**6**	
γ	$0.5,\ldots,3$	**1.7**	
b	$0.5,\ldots,8$	**2**	
σ_e	$0.15,\ldots,3$	**0.25**	
σ_o	$0.15,\ldots,3$	**3**	
$\sigma_{\mathbf{p}}$	$0.03,\ldots,0.4$	**0.25** (in deg)	$\rho_p = 0.916$
A_{UV}	$30,\ldots,40$	**35**	
s_{UV}	$0.25,\ldots,1.5$	**0.5**	
d	$0.6,\ldots,1.4$	**0.8**	$\rho_p = 0.922$
q_p	$0.5,\ldots,6$	**2.2**	
q_f	$0.5,\ldots,6$	**4.5**	$\rho_p = 0.931$

M, is a quadratic form that determines the size and shape (orientation) of the ellipsoidal discrimination regions in the corresponding image representation domain. The diagonal or non-diagonal nature of the metric determines whether the discrimination regions are oriented along the axes of the representation. The magnitude of the metric elements determines the size of the discrimination regions.

In particular, in the spatial, the wavelet, and the normalized representations, we have:

$$\begin{aligned} d(\mathbf{x}, \mathbf{x}+\Delta\mathbf{x})^2 &= \Delta\mathbf{x}^T \cdot M(\mathbf{x}) \cdot \Delta\mathbf{x} = \\ &= \Delta\mathbf{w}^T \cdot M(\mathbf{w}) \cdot \Delta\mathbf{w} = \Delta\mathbf{r}^T \cdot I \cdot \Delta\mathbf{r} \end{aligned} \tag{3.4}$$

Since the sequence of transforms are differentiable, a small distortion $\Delta\mathbf{r}$ may be written as:

$$\Delta\mathbf{r} = \nabla\mathbf{R}(\mathbf{w}') \cdot \mathbf{S} \cdot \mathbf{T} \cdot \Delta\mathbf{x} \tag{3.5}$$

Therefore (from Eqs. 3.4 and 3.5), the expression of the metrics in the spatial and the wavelet domain are:

$$M(\mathbf{x}) = \mathbf{T}^T \cdot \mathbf{S} \cdot \nabla\mathbf{R}(\mathbf{w}')^T \cdot \nabla\mathbf{R}(\mathbf{w}') \cdot \mathbf{S} \cdot \mathbf{T} \tag{3.6}$$

$$M(\mathbf{w}) = \mathbf{S} \cdot \nabla\mathbf{R}(\mathbf{w}')^T \cdot \nabla\mathbf{R}(\mathbf{w}') \cdot \mathbf{S} \tag{3.7}$$

According to the above expressions, the metric in the spatial and wavelet domains criti-

cally depends on the Jacobian of the Divisive Normalization, which is:

$$\nabla \mathbf{R}(\mathbf{w}')_{ij} = \frac{\partial \mathbf{R}_i}{\partial w'_j} = \tag{3.8}$$
$$= \gamma \left(\frac{|w'_i|^{\gamma-1}}{\beta_i + \sum_k H_{ik}|w'_k|^{\gamma}} \cdot \delta_{ij} - \frac{|w'_i|^{\gamma}|w'_j|^{\gamma-1}}{(\beta_i + \sum_k H_{ik}|w'_k|^{\gamma})^2} \cdot H_{ij} \right)$$

A number of interesting geometrical conclusions can be obtained from the above expressions:

- Linear image spaces are not perceptually Euclidean since the distortion metric is image dependent. As one could expect from contrast masking, the non-linear nature of the Divisive Normalization transform implies that the visibility of a given distortion $\Delta\mathbf{x}$ depends on the background image $\mathbf{x}$.

- Discrimination regions increase with the contrast of the image. Note that the elements of the Jacobian $\nabla\mathbf{R}$ (Eq. 3.8) decrease as the magnitude of the wavelet coefficients (or contrast of the image components) increases. The reduction of the sensitivity is bigger in high activity regions where a number of linear sensors $|w'_k|$ have non-zero values in the denominators of Eq. 3.8.

- Discrimination regions are not aligned with the axes of the wavelet representation. Note that the Jacobian has a positive diagonal contribution (proportional to δ_{ij}) and a negative non-diagonal contribution due to the kernel, H_{ij}, and depending on w'_i and w'_j with $i \neq j$. This coupling implies that the discrimination ellipsoids are not oriented along the axes of the wavelet representation. Since the Jacobian is input dependent, it can not be strictly diagonalized in any linear representation.

The above considerations on the metric, $M(\mathbf{w})$, analytically demonstrate that the appealing geometric behavior of structural similarity techniques (as in Fig. 4 in [Wang et al., 2004a]) can be shared by error visibility techniques when considering non-linearities including relations among wavelet coefficients (e.g. Divisive Normalization).

Note also that the above considerations (that show that the geometric criticism does not apply to the Divisive Normalization metric) still hold even though non-quadratic schemes are considered. In that general case the shape of the discrimination regions will not be ellipsoidal, but still its size and orientation will be determined by $\nabla\mathbf{R}(\mathbf{w}') \cdot \mathbf{S} \cdot \mathbf{T}$ or $\nabla\mathbf{R}(\mathbf{w}') \cdot \mathbf{S}$.

3.1.4 Relations to other *error visibility* metrics

The proposed model can reproduce Just Noticeable Differences (JNDs), which is a key factor in other recent error visibility metrics [Chandler & Hemami, 2007]. JNDs of a certain target can be computed from the inverse of the slope of the corresponding non-linear response.

On the other hand, if the proposed model is simplified to be completely linear by setting, $\nabla \mathbf{R} = I$, the proposed metric reduces to $M(\mathbf{w}) = \mathbf{S}^2$. In this case, the distortion is just the sum of differences in the transform domain weighted by the contrast sensitivity values (as in [Nill, 1985]): $d(\mathbf{x}, \mathbf{x}') = \left(\sum_i S_i^2 \Delta w_i^2\right)^{\frac{1}{2}}$.

If the proposed model is simplified to be point-wise non-linear by neglecting the non-diagonal elements in $\nabla \mathbf{R}$, a contrast dependent behavior (smaller sensitivity for higher contrasts) is achieved as in [Barten, 1990; Daly, 1990; Malo, Pons, & Artigas, 1997].

3.1.5 Metric results

In this section we compare the performance of the proposed Divisive Normalization metric (code available at[2]) with structural similarity metrics (SSIM [Wang et al., 2004a] and MSSIM [Wang et al., 2004b]), and information theoretic measures (VIF [Sheikh & Bovik, 2006]) on on-line available subjectively rated databases (LIVE [Sheikh, Sabir, & Bovik, 2006; Sheikh, Z.Wang, et al., 2006], TID [Ponomarenko et al., 2008], IVC [Le Callet & Autrusseau, 2005], Cornell[3]). Note that more recent structural measures on wavelet domains (such as CW-SSIM [Wang & Simoncelli, 2005b]) are designed to take into account phase distortions (translations and rotations). For registered images, as is the case in the available databases, the results of CW-SSIM basically reduce to the results of previously reported structural measures[4].

On-line available implementations from the authors were used in each case (SSIM[5], VIF and MSSIM[6]). In the SSIM case, the two available implementations were used: the standard one (`ssim_index.m`), and a posterior recommended modification (`ssim.m`) that subsamples the images to look for the best scale to apply SSIM. This will be referred as SSIM_{sub} in the experiments. SSIM results will not be shown since they are always worse than those obtained with SSIM_{sub}. The whole set of results is available at [7]. In every case, we used the RGB to Luminance conversion recommended by the authors. In the experiments we also include the Euclidean measure RMSE for illustration purposes.

The experiments will be analyzed in two parts: (1) LIVE database, and (2) additional databases with different distortions.

This distinction comes from the fact that even though a small subset of images of the LIVE database was used to derive the parameters of the Divisive Normalization model, all the five distortions in the LIVE database were used. One could argue that using LIVE to check the performance of the model is not fair since it learnt the distortions. According to this, we will show the results on the whole LIVE database for illustrative purposes,

[2]Available at: `http://www.uv.es/vista/vistavalencia/div_norm_metric/`

[3]Available at: `http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html`

[4]Personal communication by Z. Wang.

[5]SSIM available at: `http://www.ece.uwaterloo.ca/~z70wang/research/ssim/`.

[6]MSSIM and VIF available at: `http://live.ece.utexas.edu/research/quality/`.

[7]`http://www.uv.es/vista/vistavalencia/div_norm_metric/div_norm.html`.

but more interestingly, we will check the generalization ability of the model using data of other subjectively rated databases corresponding to distortions *not included* in the LIVE database.

The good performance of the proposed metric on the new data can not come from over fitting a particular database, but from the fact that it accurately models human perception. A different indication of this accuracy is that even though the model was set using suprathreshold data, it also reproduces the basic trends of threshold psychophysics (frequency sensitivity and contrast masking, as shown in Figs. 2.8 and 2.7).

■ Accuracy of a metric: correlations and calibration functions

Representing the ground truth subjective distortions (referred to as DMOS) as a function of the distances, d, computed by some metric leads to a scatter plot. Ideally, the data in this scatter plot should follow a straight line thus showing a perfect correlation among the computed distances and the subjective ratings. In real situations the data depart from this ideal behavior.

From the *engineering* point of view, any monotonic (not necessarily linear) relation between d and DMOS is good enough provided that the calibration function, DMOS $= f(d)$, is known by the metric user. According to this, non-parametric rank order correlation measures (such as the Spearman correlation) or prediction errors using standard non-linear calibration functions have been used to measure the accuracy of the distortion metrics [Group, 2008; Sheikh, Sabir, & Bovik, 2006]. Rank order correlations have been criticized for a number of reasons [Sheikh, Sabir, & Bovik, 2006]: they do not take into account the magnitude of the deviation from the predicted behavior and, as a result, it is difficult to obtain useful confidence intervals to discriminate between metrics. Therefore, even though the Spearman correlation is usually given for illustrative purposes, F-test on the quotient of the sum of squared prediction errors and standard non-linear calibration functions are usually preferred [Sheikh, Sabir, & Bovik, 2006], and has been extensively used [Chandler & Hemami, 2007; Group, 2008; Sheikh & Bovik, 2006; Sheikh, Sabir, & Bovik, 2006; A. Watson & J.Malo, 2002].

However, from the *vision science* point of view, systematic deviations from the linear prediction suggest a failure (or limitation) of the underlying model: residual non-linearities should be avoided by including the appropriate (perceptually meaningful) correction in the model, instead of using an *ad-hoc* calibration afterwards. Besides, since distortion metrics are commonly used without reference to such calibration functions [8], the unexperienced user may (erroneously) interpret the metric results in a linear way.

In the experiments below we analyze the results of the considered metrics by using the

[8] The software implementations [Chandler & Hemami, 2007; Sheikh & Bovik, 2006; Wang et al., 2004a,b] do not come with this non-linearity.

standard F-test [Sheikh, Sabir, & Bovik, 2006] along with the intuitive linear calibration and the previously reported non-linear calibration functions [Chandler & Hemami, 2007; Group, 2008; Sheikh & Bovik, 2006]. Even though we feel that linear calibration is the most intuitive scale for the final user and the most challenging situation for a model intended to reproduce human perception, we will see that the basic message (the proposed error visibility metric is competitive with the newer techniques) is independent from the calibration measure. This is good since F-test may be criticized as well because it depends on an arguable choice of the calibration function. For illustration purposes we will also include the (linear) Pearson's correlation and the Spearman's correlation. Note that the Pearson's correlation on the raw data as done here conveys the same kind of information as the F-test when using a linear calibration function. The difference is that the F-test is useful to establish confidence levels in the results so that it is easy to assess when the differences in prediction errors (or Pearson correlation) are statistically significant.

■ Performance of the metrics

In this section we show the scatter plots, the correlations, the fitted calibration functions and the F-test results for (1) the LIVE database, and (2) additional databases (TID, IVC, and Cornell) excluding LIVE-like distortions. Note that distortions in Cornell database are different since it consists of achromatic images only.

As stated above, we used the linear calibration and three additional non-linear calibration functions used in the literature: a 4 parameter logistic [Chandler & Hemami, 2007], a 5 parameter logistic [Sheikh & Bovik, 2006; Sheikh, Sabir, & Bovik, 2006], and a 4th order polynomial [Group, 2008]. In every case, the calibration functions were fitted using the Nelder-Mead simplex search method [Lagarias et al., 1998] with equivalent initial guesses (according to the corresponding ranges of the distances).

Provided that the prediction errors of the metrics m_i and m_j are independent and Gaussian, the F-test gives the probability that the sum of squared errors of metric i, ε_i^2, is smaller than the corresponding value of metric j, ε_j^2. This probability, $P(\varepsilon_i^2 < \varepsilon_j^2)$, can be used to assess if metric i is better than metric j. The F-test has been applied to compare among the previously reported metrics [Chandler & Hemami, 2007; Sheikh, Sabir, & Bovik, 2006]. Here we apply the same standard procedure. In the case of the proposed Divisive Normalization metric the correlation between its residuals and the residuals of the other metrics is similar or smaller than the equivalent results among the other (previously compared) metrics. Therefore, the independence condition holds as accurately as in previously reported comparisons. Unless explicitly stated, residuals can be taken as Gaussian according to previously used kurtosis-based criteria [Chandler & Hemami, 2007; Sheikh, Sabir, & Bovik, 2006]. Therefore, the Gaussianity condition holds as accurately as in previously reported comparisons.

None of the available image quality databases used an experimental procedure similar to A. Watson & Kreslake [2001] (which gives rise to subjective ratings in meaningful JND units). The differences in the experimental procedures imply that the available results are not ready to be merged into a single database. Nevertheless, the different DMOS data were linearly scaled to fall within the range of the LIVE database for visualization purposes. In this work we used the DMOS scores that come with the on-line file `databaserelease2.zip`, as used in [Sheikh & Bovik, 2006]. This convenient linear DMOS scaling is not a problem since (1) a separate analysis for each database is done, and (2) it does not modify the correlation results (either Pearson or Spearman), nor the F-test results (since the scaling is taken into account in fitting the corresponding calibration functions and it cancels out in the quotient of squared errors).

Figures 3.1-3.4 show the scatter plots and the fitted functions for the considered metrics in the considered situations (1) LIVE, Fig. 3.1; and (2) TID, Fig. 3.2, IVC, Fig. 3.3, Cornell, Fig. 3.4. Each distortion is identified by a different symbol/color combination. The details on these distortions can be found in the corresponding references. In every case increasing functions were obtained by linearly turning similarity measures, s, into distortion measures, d (as indicated in the plots).

Note that non-linear fitting functions may be unreliable: too flexible fitting functions (such as the 4th order polynomial and the 5 parameter sigmoid) may give rise to non-monotonic behavior. The behavior of these functions strongly depends on the considered data, thus suggesting that it may not account for more general data.

In tables 3.2-3.5 we show the results of the F-test for the quotients of the sum of residuals of the considered metrics in the two considered situations: (1) LIVE, table 3.2; and (2) TID, table 3.3, IVC, table 3.4, Cornell, table 3.5. In these tables, highlighted cells in a row mean that the model in the row is better than the model in the column at 90% confidence level.

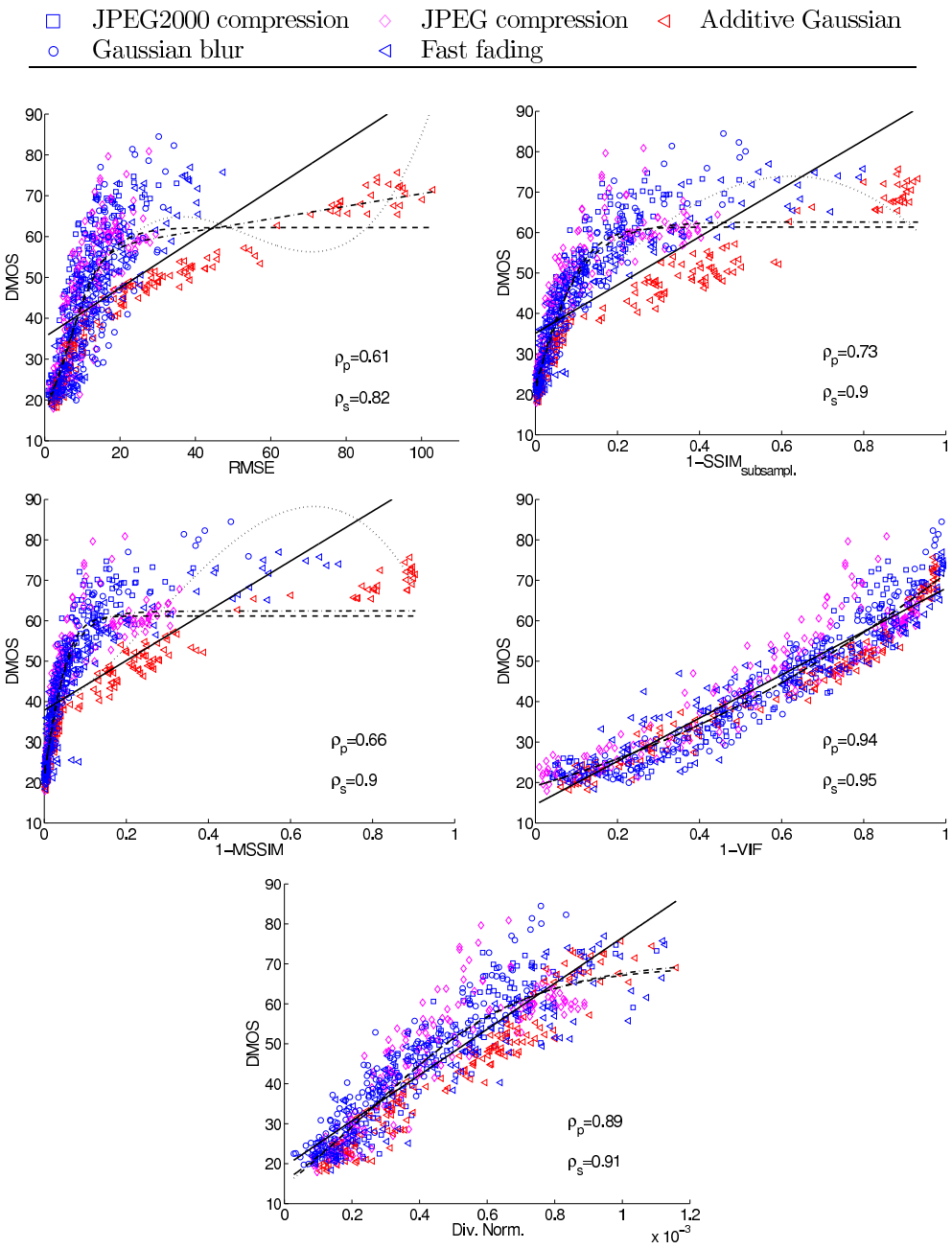


Figure 3.1: Scatter plots, fitted functions and correlation coefficients for the considered metrics on the LIVE database. The legend shows the symbols representing each distortion in the LIVE database. The solid line represents the linear fitting. The dashed line represents the 4 parameter sigmoid function used in [Chandler & Hemami, 2007], the dash-dot line represents the 5 parameter sigmoid used in [Sheikh & Bovik, 2006; Sheikh, Sabir, & Bovik, 2006]. The dotted line stands for the 4th order polynomial used in [Group, 2008].

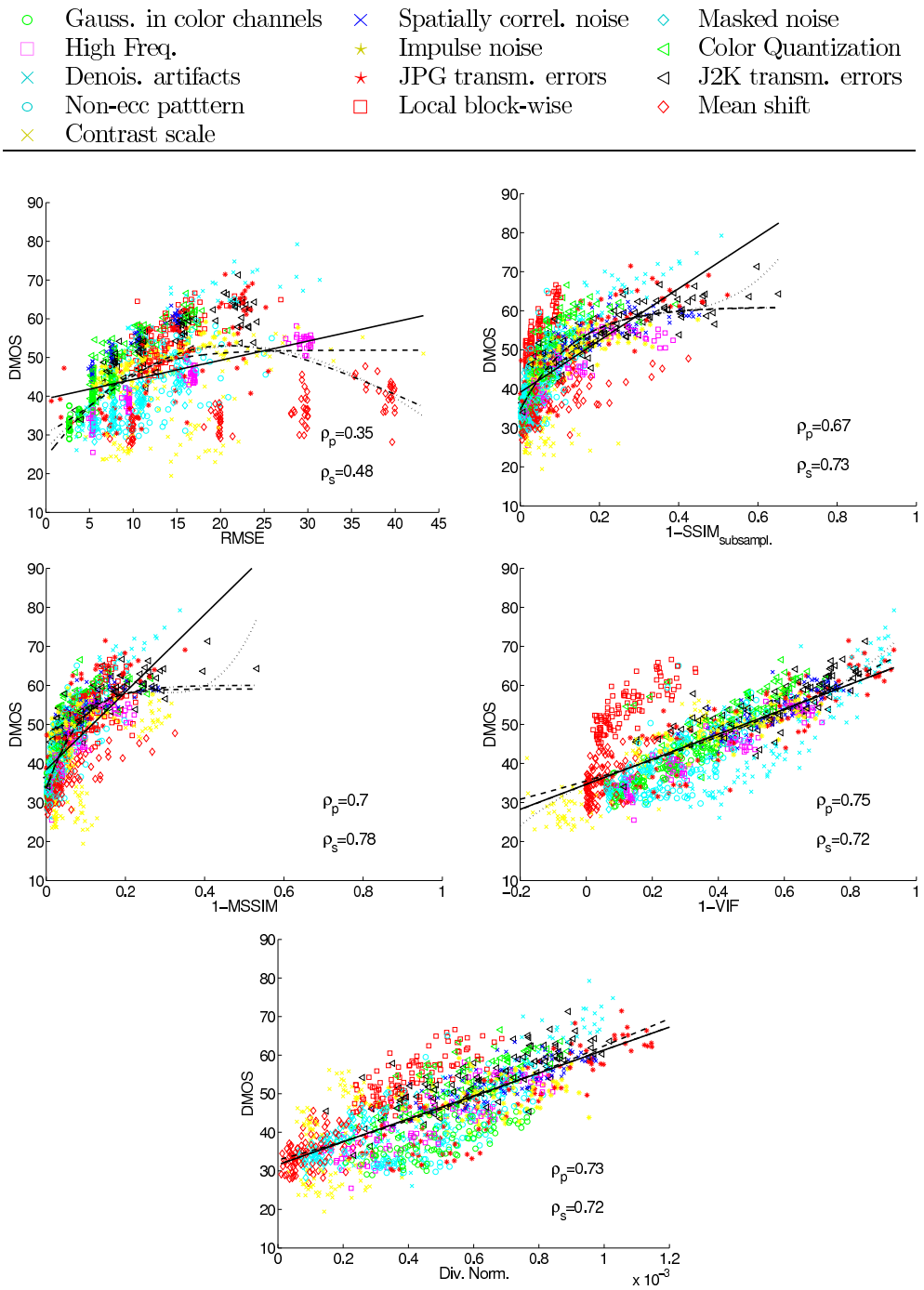


Figure 3.2: Scatter plots, fitted functions and correlation coefficients for the considered metrics on the TID database (excluding LIVE-like distortions). The legend represents the symbols corresponding to the distortions which are not present in the LIVE database. Line styles for the calibration functions have the same meaning as in figure 3.1.

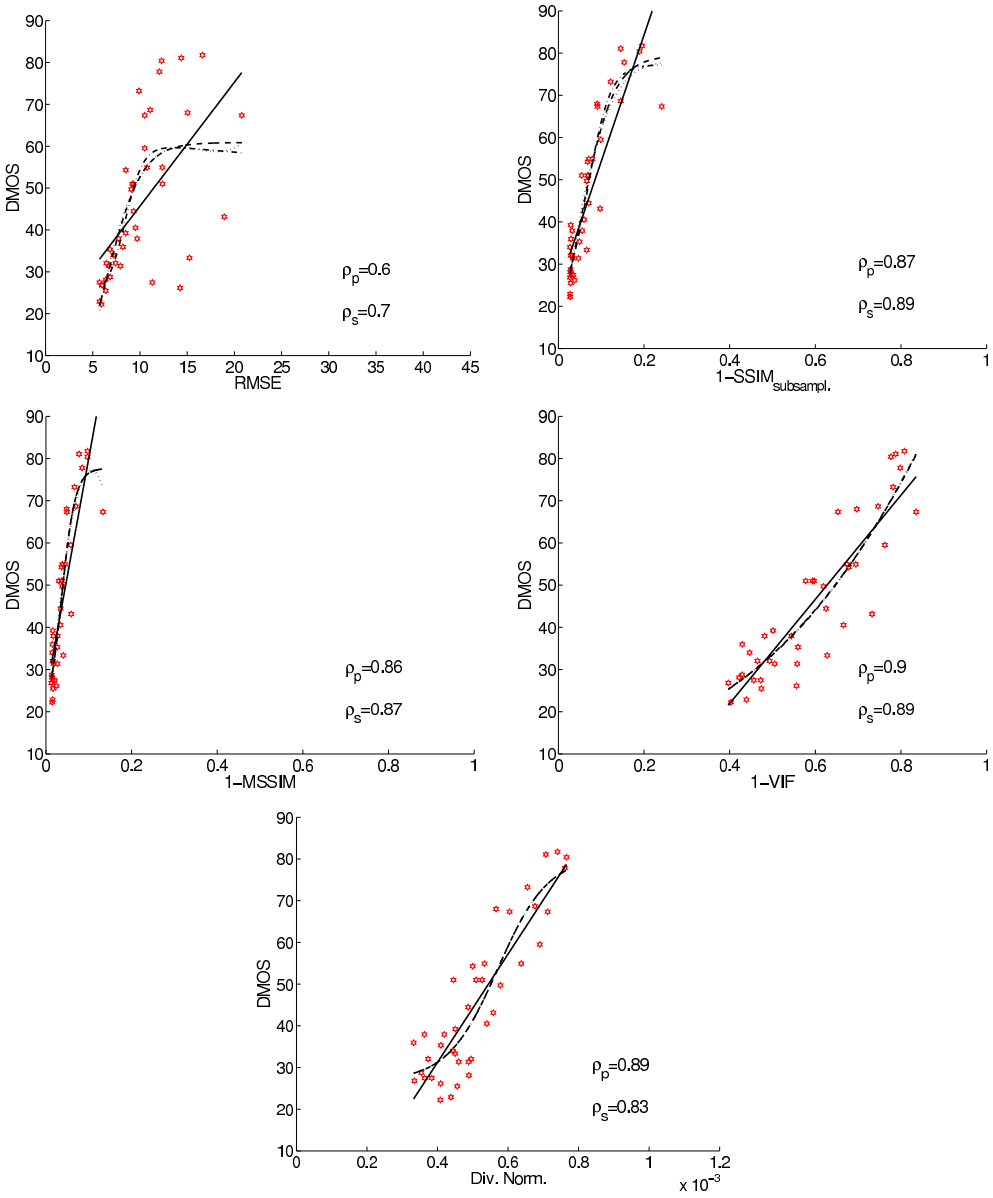


Figure 3.3: Scatter plots, fitted functions and correlation coefficients for the considered metrics on the IVC database (excluding LIVE-like distortions). The only non-LIVE distortion in the IVC database is what they call LAR distortion (see [Le Callet & Autrusseau, 2005] for details). Line styles for the calibration functions have the same meaning as in figure 3.1.

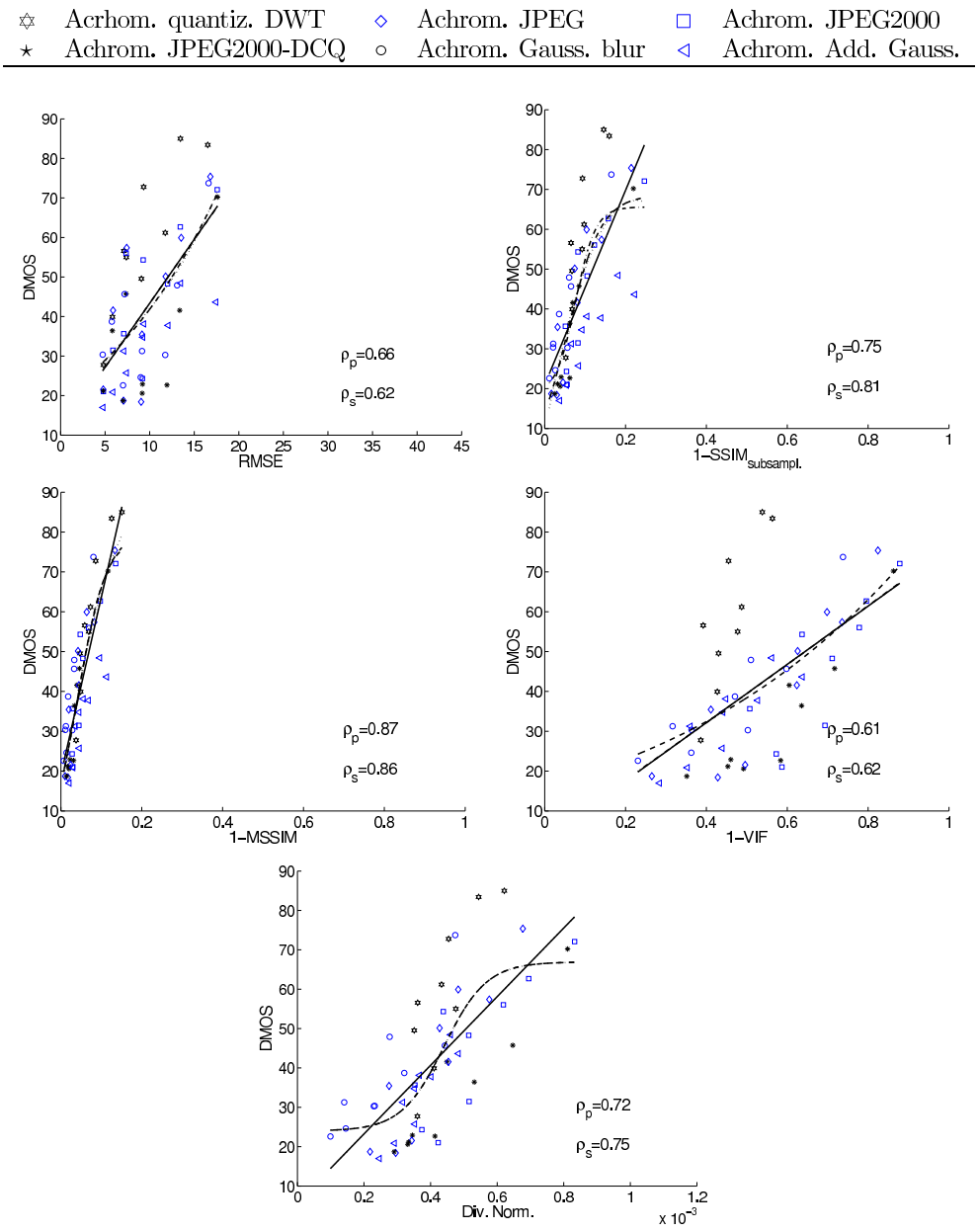


Figure 3.4: Scatter plots, fitted functions and correlation coefficients for the considered metrics on the Cornell database. The legend represents the symbols corresponding to the distortions which are not present in the LIVE database (no Cornell distortion is present in LIVE since Cornell is an achromatic database). Line styles for the calibration functions have the same meaning as in figure 3.1.

Table 3.2: Quality of metrics on the LIVE database (F-test): probability that the model in the row is better than the model in the column for the linear and several non-linear fits. Highlighted cells mean that model in the row is better than the model in the column at 90% confidence level. The models highlighted with * have non-Gaussian residuals, so the result is not strictly correct.

$P(\varepsilon^2_{row} < \varepsilon^2_{col})$ Linear Fit

	RMSE	$SSIM_{sub.}$	MSSIM	VIF	DN
RMSE:	-	0.00	0.08	0.00	0.00
$SSIM_{sub.}$:	1.00	-	1.00	0.00	0.00
MSSIM:	0.92	0.00	-	0.00	0.00
VIF:	1.00	1.00	1.00	-	1.00
DN:	1.00	1.00	1.00	0.00	-

$P(\varepsilon^2_{row} < \varepsilon^2_{col})$ 4 parameter Sigmoid Fit

RMSE:	-	0.00	0.00	0.00	0.00
$SSIM_{sub.}$:	1.00	-	0.64	0.00	0.09
MSSIM:	1.00	0.36	-	0.00	0.04
VIF*:	1.00	1.00	1.00	-	1.00
DN:	1.00	0.91	0.96	0.00	-

$P(\varepsilon^2_{row} < \varepsilon^2_{col})$ 5 parameter Sigmoid Fit

RMSE:	-	0.00	0.00	0.00	0.00
$SSIM_{sub.}$:	1.00	-	0.58	0.00	0.14
MSSIM:	1.00	0.42	-	0.00	0.10
VIF*:	1.00	1.00	1.00	-	1.00
DN:	1.00	0.86	0.90	0.00	-

$P(\varepsilon^2_{row} < \varepsilon^2_{col})$ 4th order polynomial Fit

RMSE:	-	0.12	1.00	0.00	0.00
$SSIM_{sub.}$:	0.88	-	1.00	0.00	0.00
MSSIM:	0.00	0.00	-	0.00	0.00
VIF:	1.00	1.00	1.00	-	1.00
DN:	1.00	1.00	1.00	0.00	-

Table 3.3: Quality of metrics on the TID database (excluding LIVE-like distortions). See caption of table 3.2 for details.

$P(\varepsilon^2_{row} < \varepsilon^2_{col})$ Linear Fit

	RMSE	$SSIM_{sub.}$	MSSIM	VIF*	DN
RMSE:	-	0.00	0.00	0.00	0.00
$SSIM_{sub.}$:	1.00	-	0.12	0.00	0.00
MSSIM:	1.00	0.88	-	0.01	0.03
VIF*:	1.00	1.00	0.99	-	0.75
DN:	1.00	1.00	0.97	0.25	-

$P(\varepsilon^2_{row} < \varepsilon^2_{col})$ 4 parameter Sigmoid Fit

RMSE:	-	0.00	0.00	0.00	0.00
$SSIM_{sub.}$:	1.00	-	0.01	0.10	0.29
MSSIM*:	1.00	0.99	-	0.86	0.97
VIF*:	1.00	0.90	0.14	-	0.77
DN:	1.00	0.71	0.03	0.23	-

$P(\varepsilon^2_{row} < \varepsilon^2_{col})$ 5 parameter Sigmoid Fit

RMSE:	-	0.00	0.00	0.00	0.00
$SSIM_{sub.}$:	1.00	-	0.01	0.13	0.32
MSSIM*:	1.00	0.99	-	0.91	0.98
VIF*:	1.00	0.87	0.09	-	0.75
DN:	1.00	0.68	0.02	0.25	-

$P(\varepsilon^2_{row} < \varepsilon^2_{col})$ 4th order polynomial Fit

RMSE:	-	0.00	0.00	0.00	0.00
$SSIM_{sub.}$:	1.00	-	0.01	0.05	0.31
MSSIM*:	1.00	0.99	-	0.79	0.97
VIF:	1.00	0.95	0.21	-	0.86
DN:	1.00	0.69	0.03	0.14	-

Table 3.4: Quality of metrics on the IVC database (excluding LIVE-like distortions). See caption of table 3.2 for details.

$P(\varepsilon^2_{row} < \varepsilon^2_{col})$ Linear Fit

	RMSE	$SSIM^{\star}_{sub.}$	MSSIM*	VIF	DN
RMSE:	-	0.00	0.00	0.00	0.00
$SSIM^{\star}_{sub.}$:	1.00	-	0.62	0.28	0.34
MSSIM*:	1.00	0.38	-	0.19	0.23
VIF:	1.00	0.72	0.81	-	0.56
DN:	1.00	0.66	0.77	0.44	-

$P(\varepsilon^2_{row} < \varepsilon^2_{col})$ 4 parameter Sigmoid Fit

RMSE:	-	0.00	0.00	0.00	0.00
$SSIM_{sub.}$:	1.00	-	0.72	0.86	0.86
MSSIM*:	1.00	0.28	-	0.69	0.68
VIF:	1.00	0.14	0.31	-	0.50
DN:	1.00	0.14	0.32	0.50	-

$P(\varepsilon^2_{row} < \varepsilon^2_{col})$ 5 parameter Sigmoid Fit

RMSE:	-	0.00	0.00	0.00	0.00
$SSIM_{sub.}$:	1.00	-	0.75	0.87	0.87
MSSIM*:	1.00	0.25	-	0.68	0.68
VIF:	1.00	0.13	0.32	-	0.50
DN:	1.00	0.13	0.32	0.50	-

$P(\varepsilon^2_{row} < \varepsilon^2_{col})$ 4th order polynomial Fit

RMSE:	-	0.00	0.00	0.00	0.01
$SSIM_{sub.}$:	1.00	-	0.64	0.81	0.89
MSSIM*:	1.00	0.36	-	0.70	0.82
VIF:	1.00	0.19	0.30	-	0.65
DN:	0.99	0.10	0.18	0.35	-

Table 3.5: Quality of metrics on the (achromatic) Cornell database. See caption of table 3.2 for details.

$P(\varepsilon^2_{row} < \varepsilon^2_{col})$ Linear Fit

	RMSE	$SSIM_{sub.}$	MSSIM	VIF*	DN
RMSE:	-	0.17	0.00	0.63	0.27
$SSIM_{sub.}$:	0.83	-	0.03	0.90	0.63
MSSIM:	1.00	0.97	-	1.00	0.99
VIF*:	0.37	0.10	0.00	-	0.17
DN:	0.73	0.37	0.01	0.83	-

$P(\varepsilon^2_{row} < \varepsilon^2_{col})$ 4 parameter Sigmoid Fit

RMSE:	-	0.07	0.00	0.64	0.18
$SSIM_{sub.}$:	0.93	-	0.06	0.97	0.72
MSSIM:	1.00	0.94	-	1.00	0.98
VIF*:	0.36	0.03	0.00	-	0.10
DN:	0.82	0.28	0.02	0.90	-

$P(\varepsilon^2_{row} < \varepsilon^2_{col})$ 5 parameter Sigmoid Fit

RMSE:	-	0.05	0.00	0.63	0.17
$SSIM_{sub.}$:	0.95	-	0.08	0.97	0.74
MSSIM:	1.00	0.92	-	1.00	0.98
VIF*:	0.37	0.03	0.00	-	0.10
DN:	0.83	0.26	0.02	0.90	-

$P(\varepsilon^2_{row} < \varepsilon^2_{col})$ 4th order polynomial Fit

RMSE:	-	0.08	0.00	0.64	0.28
$SSIM_{sub.}$:	0.92	-	0.06	0.96	0.80
MSSIM:	1.00	0.94	-	1.00	0.99
VIF*:	0.36	0.04	0.00	-	0.17
DN:	0.72	0.20	0.01	0.83	-

3.1.6 Discussion

In the LIVE case, VIF is the best performing metric. The proposed Divisive Normalization metric is the second best and shows a significantly better performance than structural methods. This reveals that the proposed model can adequately account for the whole database even though its parameters were set by using the 10% of the data (and cropped images). This good performance is independent from the fitting function.

When considering a wider range of distortions (TID and IVC), and using the most challenging linear fit, no algorithm outperforms the proposed Divisive Normalization metric. In the (small) Cornell database, MSSIM is the only metric that significantly outperforms the proposed metric. However, note that the proposed metric significantly outperforms MSSIM in the (bigger) LIVE case no matter the calibration function.

To summarize, the proposed metric performs quite well in the LIVE database (5 distortions) and successfully generalizes to a wide range of distortions (e.g. 20 new distortions in the TID, IVC and Cornell databases). This suggests that the parameters found are perceptually meaningful thus giving rise to a robust metric. In most of the cases the proposed metric is statistically indistinguishable from structural and information theoretic methods. In some particular cases, it is outperformed by VIF (as in LIVE) or by MSSIM (as in Cornell), but it is important to note that, conversely, it significantly outperforms MSSIM in LIVE, and works better than VIF in Cornell (at 80% confidence level). The above is true for all the considered calibration functions.

As a result, the proposed error visibility metric based on Divisive Normalization seems to be competitive with structural and information theoretic metrics. It is quite robust and easy to interpret in linear terms. This is consistent with the fact that the criticisms made to the error visibility techniques do not apply to the Divisive Normalization metric as shown in sections 2.1.4, 3.1.3 and 2.1.3.

3.2 Chapter conclusions

In this work, the classical Divisive Normalization metric [Teo & Heeger, 1994] was revisited to address the criticisms raised against error visibility techniques. It was straightforwardly fitted by using a small subset of the subjectively rated LIVE database, and proved to generalize quite well for the whole database as well as for more general databases including distortions of different nature (e.g. TID, IVC, Cornell) .

We showed that the three basic criticisms made against error visibility techniques do not apply to the Divisive Normalization metric: (1) even though the Divisive Normalization is inspired in low-level (threshold) psychophysical and physiological data, it can account for higher-level (suprathreshold) distortions while approximately reproducing the frequency sensitivity and masking results. (2) It was shown that the Divisive Normalization repre-

sentation reduces the statistical relations among the image coefficients, thus justifying the use of uniform Minkowski summation strategies in the normalized domain. (3) It was analytically shown that the Divisive Normalization has a rich geometric behavior, so it is not a singular feature of structural similarity metrics.

The experiments show that the proposed metric is competitive with structural and information theoretic metrics, it performs consistently when facing a wide range of distortions, and it is easy to interpret in linear terms. These results suggest that the classical error visibility approach based on gain control models should still be considered in the image quality discussion.

In fact, the proposed Divisive Normalization framework can still be improved in many ways. The linear chromatic and spatial transforms can be improved by (1) using non-linear color representations to account for the chromatic adaptation ability of human observers [M. Fairchild, 1997], and (2) better wavelet transforms may be used for a better simulation of V1 receptive fields (e.g. steerable pyramids [E. Simoncelli et al., 1992]). Useful Divisive Normalization transforms for image enhancement have already been proposed on steerable pyramids [Lyu & Simoncelli, 2009]. Different wavelet basis (as in CW-SSIM [Wang & Simoncelli, 2005b]) could be used to introduce translation and rotation invariance. Better (non-linear) color representations can be useful to assess changes in average luminance or in the spectral radiance (i.e. including color constancy). Linear models may overestimate the effect of such distortions. The proposed non-linear transform can also be generalized since masking interactions among sensors of different chromatic channels may occur [Gegenfurtner & Kiper, 1992], but they were not considered here in order to keep the interaction kernel small. Summation over the color dimension can be generalized as well by including different summation exponents on the opponent channels. Another issue to be explored is the role of the low-frequency residual which was neglected in this work. Weber-law like non-linearities should be used in this case (in agreement with non-linear color appearance models) together with and appropriate relative weight between the low-pass and the higher frequency subbands. From a more general point of view, the proposed model may be complemented by bottom-up techniques for saliency prediction based on the V1 image representation [Zhaoping, 2006]. Finally, better optimization techniques instead of the reported exhaustive search of the parameter subspaces may be used in order to obtain a more accurate estimation of the optimal parameters with a reduced computational burden.

Chapter 4

From Statistics to Neuroscience

4.1 Color vision mechanisms from Sequential Principal Curves Analysis

Human color vision is mediated by opponent mechanisms, achromatic and chromatic, with two fundamental properties [Abrams et al., 2007; M. Fairchild, 2005; Hillis & Brainard, 2005]: (i) their response is nonlinear given some fixed observation conditions (e.g. illuminant, spatial context, etc.); and (ii) they are able to compensate changes in the observation conditions to keep the perception (color) of the objects constant despite the changes in the input. On the one hand, the nonlinear response of the mechanisms is revealed by the non-uniform nature of discrimination thresholds throughout the tristimulus space [Cole et al., 1990; Krauskopf & Gegenfurtner, 1992; Romero et al., 1993; Wyszecki & Stiles, 1982]. On the other hand, the adaptation ability of these mechanisms is revealed by asymmetric color matching experiments [Breneman, 1987; M. Luo et al., 1991; M. Luo & Rhodes, 1999; M. A. Webster & Mollon, 1991]: different physical inputs give rise to the same perception (corresponding stimuli) for equivalent shifts in the context.

The standard empirical models of color vision assume that the system is formed by three linear photoreceptors sensitive to long, medium and short wavelengths (LMS). These models try to reproduce the above mentioned effects using three basic ingredients as reported in [Abrams et al., 2007; M. Fairchild, 2005; Hillis & Brainard, 2005]: (i) context dependent weighting of the sensitivity of LMS mechanisms, also known as Von Kries normalization, (ii) linear transform to an opponent color space, and (iii) nonlinear saturation of the achromatic and the opponent-chromatic responses.

Following the classical suggestion by Barlow on the relation between image statistics and neural behavior [H. Barlow, 1961; H. B. Barlow, 2001], a large body of literature argues that mechanisms underlying the perception of object colors are organized according to the statistical regularities of the signals confronted by the sensory systems. However, very often, the proposed statistical approaches deal with color discrimination and color

constancy in a *separated way*: they are not able to address both adaptation and nonlinearities jointly.

On the one hand, linear approaches based on decorrelation (linear principal component analysis, PCA) and higher order redundancy removal (linear independent component analysis, ICA) explain the existence of spectrally opponent chromatic channels with the right spatial sensitivity. The seminal work of Buchsbaum & Gottschalk [1983] derives opponent channels from PCA applied to LMS signals subject to a rough model of natural radiances (white noise). Atick et al. [Atick, 1992] derive the achromatic and chromatic Contrast Sensitivity Functions using decorrelation arguments constrained with error minimization and an idealized model of spatio-spectral radiances. Ruderman et al. [D. Ruderman & Chiao, 1998] obtain Fourier-like spatio-chromatic opponent sensors using PCA on the LMS signals obtained from real reflectance measurements. Wachtler et al. [Wachtler et al., 2001] apply linear ICA pixel-wise and patchwise on real hyperspectral photographic images and obtain better coding results than using PCA. Doi et al. [Doi et al., 2003] use PCA and ICA to derive spatio-chromatic properties of the lateral geniculate nucleus (LGN) and the primary visual cortex (V1). Even though the above techniques do not explicitly address adaptation, Webster and Mollon [M. Webster & Mollon, 1997] show that the mean shift (or chromatic adaptation), and the covariance shift (or contrast adaptation) can be roughly reproduced by dimension-wise normalization of the LMS responses, and PCA followed by whitening using the set of adaptation colors under different illuminants. Their work combines an extension of the measurements reported in [M. A. Webster & Mollon, 1991] (which used single-color adaptation) with the decorrelation-oriented explanation given by Atick et al. [Atick et al., 1993]. The obvious problem with linear methods is that they cannot explain non-uniform discrimination, i.e. the nonlinear response.

On the other hand, color discrimination and the associated nonlinearities have been statistically addressed from a different point of view. In this case, the key is the consideration of the limited resolution of any physical sensor and its optimal design to deal with non-uniformly distributed signals. Two different criteria have been proposed in this context. First, Laughlin [Laughlin, 1983] argued that limited resolution mechanisms should be designed to maximize the information transfer (*infomax*). In noise-free scenarios, the *infomax* principle leads to component independence and nonlinear responses related to the marginal probability density functions (PDFs) [Bell & Sejnowski, 1995]. Second, MacLeod et al. proposed that, in order to minimize the representation error in the presence of neural noise, the response of the color sensors should be related to some power of the marginal PDFs in the color opponent directions [D. MacLeod & Twer, 2003; D. A. MacLeod, 2003; Twer & MacLeod, 2001]. The same reasoning has been applied to explain physiological nonlinearities at the LGN [Goda et al., 2009]. Unfortunately, in these studies, an explicit multidimensional data-driven algorithm to get the optimal set of sensors remained un-

addressed: in their experiments they just showed marginal PDFs in predefined linear axes [Goda et al., 2009; D. MacLeod & Twer, 2003; D. A. MacLeod, 2003; Twer & MacLeod, 2001]. And more importantly, no attempt has been made to explain adaptation using *infomax* or *error minimization*. Again, just one of the aspects, the nonlinear behavior, was addressed by these techniques.

Among the statistically inspired approaches, a remarkable exception to the separate consideration of color discrimination and adaptation is the work by Abrams et al. [2007], where the authors investigate whether optimizing the nonlinearity is compatible with optimal color constancy or adaptation. The problem in this case is that the presented model already has the appropriate *parametric formulation* adopted from the empirical models [Hillis & Brainard, 2005]. Therefore, it may be argued that, even though the model was statistically fitted, the expected behavior is somehow imposed in advance through the use of a convenient functional form. According to this, it is not shown that the behavior emerges directly from data, but the fact that the selected functional model may be simultaneously optimal in color discrimination and adaptation.

Finally, we should note an important shortcoming present across the literature: many of the above statistical studies rely on *simplified databases*. In particular, most of the works dealing with adaptation to changes in the illuminant usually assume that the input radiance is just the product of the spectral reflectance and the illuminant radiance: they assume flat Lambertian surfaces. Therefore, relevant nonlinear phenomena in the image formation process, such as mutual illumination in rough surfaces, are neglected.

In this work, we address the *simultaneous* explanation of (i) the nonlinear behavior of achromatic and chromatic mechanisms in a fixed adaptation state; and (ii) the change of such behavior, i.e. adaptation, under the change of observation conditions. This is done by proposing a single non-parametric method based on Principal Curves (PCs) [Delicado, 2001; Hastie & Stuetzle, 1989]: the Sequential Principal Curves Analysis (SPCA). The method exploits the flexibility of PCs to find adaptive (eventually non-linear) sensors. Moreover, SPCA is equipped with tunable local metric so that the proposed analysis may follow either the *infomax* or the *error minimization* principles. Given the fact that psychophysical adaptation data are given under D65 and A illuminations [M. Luo et al., 1991; M. Luo & Rhodes, 1999], in this work, the statistical analysis is made on a *new database* consisting of colorimetrically calibrated images of natural objects under these calibrated illuminations. SPCA reproduces the psychophysical behavior on color discrimination thresholds, discount of the illuminant and corresponding pairs in asymmetric color matching. These color vision properties are demonstrated to emerge directly from realistic data regularities without assuming any *a priori* functional form. Moreover, the results suggest that color perception at this low abstraction level may follow an error minimization strategy, as suggested by MacLeod [D. MacLeod & Twer, 2003; D. A. MacLeod, 2003; Twer & MacLeod, 2001], instead of the information maximization principle suggested in

Laughlin [Laughlin, 1983].

The remainder is organized as follows. Section 4.1.1 motivates the statistical study of color vision by reviewing (i) the basic features of color PDFs and their changes, and (ii) the non-linear and adaptive nature of color vision mechanisms. Section 4.1.2 reviews the results on *infomax* and *error minimization* in *unsupervised manifold learning*. This motivates our computational approach, which is presented in Section 4.1.3: the Sequential Principal Curves Analysis (SPCA) with local metric. Section 4.1.4 describes the design of the experiments: the database of calibrated natural color images used and how the simulation of color vision phenomena from SPCA is carried out. Section 4.1.5 shows how the proposed statistical technique simultaneously reproduces experimental data on color discrimination and adaptation, compares its performance with empirical color appearance models (CIELab, LLab, CIECAM), and discusses the biological implications of the results.

4.1.1 Facts on color PDFs and color mechanisms behavior

This section motivates the problem of characterizing nonlinearities and adaptation of the color vision mechanisms. First, we review the special characteristics of the color manifolds. The considered perception phenomena are reviewed through two sets of results: (i) the nonlinear behavior of the achromatic and the chromatic opponent mechanisms [Cole et al., 1990; Krauskopf & Gegenfurtner, 1992; Romero et al., 1993; Wyszecki & Stiles, 1982]; and (ii) the ability to compensate changes in spectral illumination according to experiments on corresponding colors [Breneman, 1987; M. Luo et al., 1991; M. Luo & Rhodes, 1999].

Non-uniformities and shifts in color manifolds

Observation of the natural world gives rise to measurements that typically live in low dimensional manifolds. The shape of these manifolds and its PDF depend on the interesting features of the underlying phenomenon. For example, in vision, it is known that the spectral reflectance of objects is intrinsically low dimensional [Maloney, 1986]. Moreover, the physical constraints on the reflectance as well as the geometry of the surfaces give rise to particular statistics of the tristimulus values [Koenderink, 2010; Motoyoshi et al., 2007]. However, these distributions are also modified by additional, eventually non-interesting, causes [Funt & Drew, 1993]: both mutual illumination in surfaces of complex geometry and changes in the spectral illumination and its geometry introduce nonlinear changes in the PDF of the tristimulus values that are difficult to characterize. The examples in Fig. 4.1 show real (top) and synthetic (bottom) examples of the nonlinear nature of changes in color distributions, which clearly cannot be compensated by linear transforms.

Top panel in Figure 4.1 illustrates the basic features of the tristimulus PDF in natural scenes and its changes:

- The distribution is remarkably non-uniform, i.e. densely populated around the color direction determined by the illuminant, and of decreasing density for higher saturation values (see LMS distribution).
- It displays a strong correlation between the LMS values.
- Different illumination geometry gives rise to different data distribution along the principal axes. In particular, the highlighted regions differ in the illumination angle: the scene illuminated with D65 (A) is relatively more populated in the high (low) luminance region.

The change in orientation of the PDF due to the change in the spectral radiance of the illuminant can be approximately compensated by a rotation, and the change in total radiance by a suitable scaling [Atick et al., 1993; M. Webster & Mollon, 1997]. However, the classical *PCA plus whitening* linear transform is not able to compensate the uneven data distribution along the principal axes. In the example, black dots represent the prediction of the data under D65 from A data using PCA+whitening. In this case, the transformed data are still relatively more concentrated in the low luminance region. Nonlinear transforms either in D65 or A data would be required to equalize the distributions along the individual subspaces. The effect of such equalization would remove the shaded regions that remain in the linearly compensated image.

The inability of linear transforms can be better stressed in a synthetic example with controlled reflectance, surface and illumination geometries. The bottom panel in Fig. 4.1 shows a triangularly undulated surface illuminated with D65 and A radiances from different angles (11^o and 22^o). In this synthetic case, a one-to-one correspondence between the colors can be established so the linear transform can be optimized to minimize the least squares error, which is not possible in real examples. The result of such optimal linear transform (not restricted to be based on rotations) is represented by the black dots and by the reconstructed image. In this case, the best possible linear transform is unable to compensate for the color changes induced by the change in spectral radiance, surface and illumination geometries: it can shift the yellowish colors to the region of the gray-blueish colors but, as a byproduct, the low luminance colors are markedly desaturated. In this example, the nonlinear change in the PDF comes from the differences in the surface and illumination geometries. This stresses the fact that general changes occurring in colors of natural surfaces are nonlinear.

Nonlinear behavior of achromatic and opponent chromatic mechanisms

The sensitivity of some underlying sensory mechanism is related to its discrimination ability according to the classical Fechner's hypothesis used in psychophysics [Hillis & Brainard, 2005; Laming, 1997]: given a unidimensional measure x and an eventually nonlinear trans-

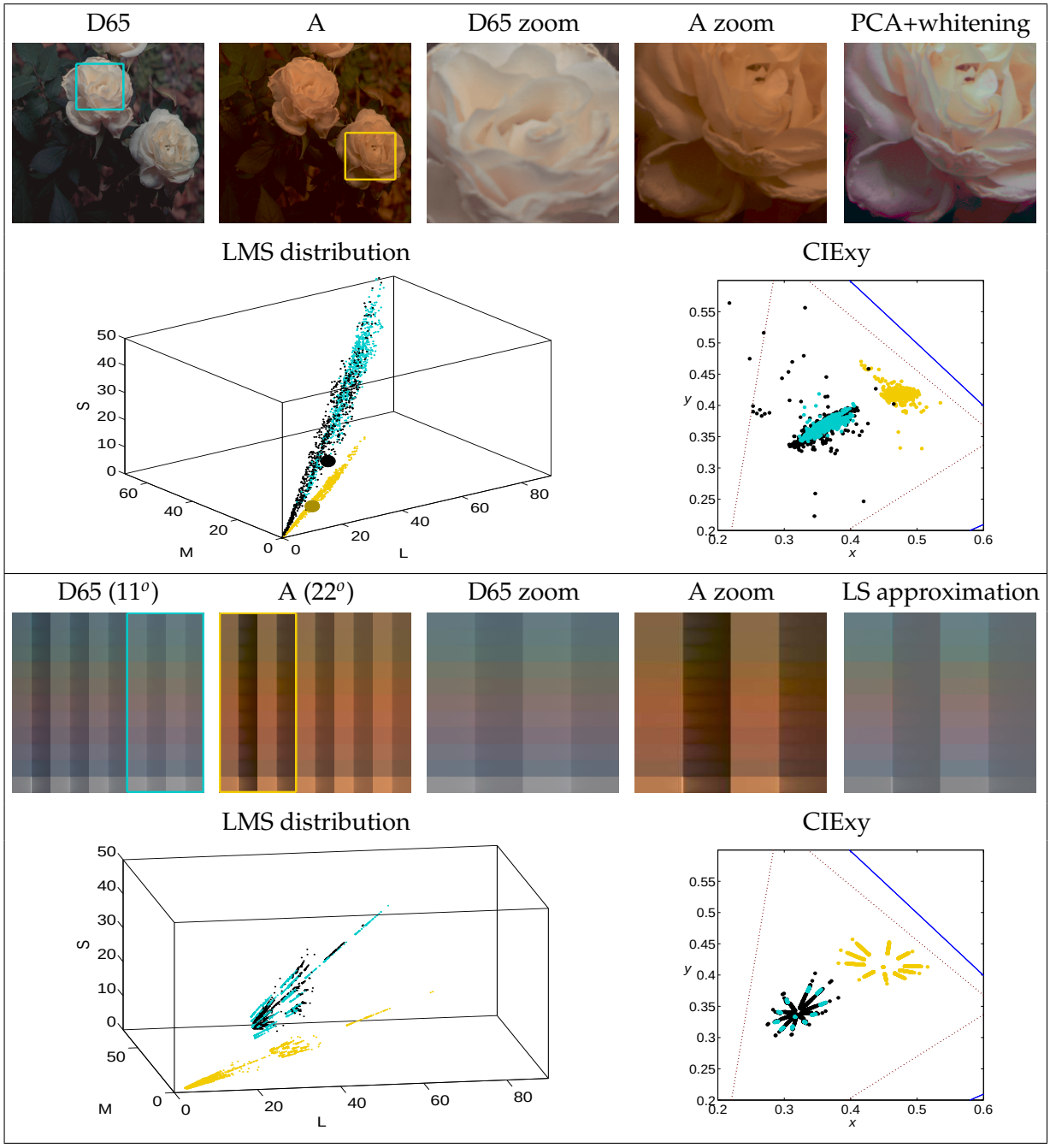


Figure 4.1: Illustration of nonlinearities and shifts in color manifolds. Images, tristimulus distributions in LMS and CIExy chromatic coordinates are shown for real (top panel) and synthetic (bottom panel) examples. **Top:** Flower images under diffuse D65 (blue dots) and A (yellow dots) illuminant. The synthesized colors (black dots and right image) are obtained from linear manifold matching by PCA plus whitening. A particular color (large yellow dot) and its linearly obtained corresponding pair (large black dot) have been highlighted for convenient comparison with equivalent illuminant compensation results that will be shown in Section 4.1.4. **Bottom:** Undulated surface under Lambertian D65 and A illuminants tilted 11^o and 22^o, respectively. The synthesized image is obtained by fitting the best least squares linear regression (black dots).

form R, the bigger the incremental threshold (or just noticeable difference), $\Delta x(x)$, the smaller the slope (or sensitivity) of the underlying mechanism,

$$\frac{dR(x)}{dx} \propto \frac{1}{\Delta x(x)}. \tag{4.1}$$

Therefore, the response can be estimated from the experimental incremental thresholds by

$$R(x) = R(x^o) + \beta \int_{x^o}^{x} \frac{1}{\Delta x(x')} dx', \tag{4.2}$$

where β is an irrelevant scaling factor. As a result, if the incremental thresholds are not constant over the stimulus range, the response of the underlying mechanism is nonlinear.

Figure 4.2 (top panel curves) shows the experimental behavior of the achromatic (A), red-green (T), and yellow blue (D) mechanisms. Top left plot shows the experimental increase of the luminance thresholds (data derived from Fig. 7.10.1 in [Wyszecki & Stiles, 1982], page 569). Luminance has been expressed in terms of the Ingling and Tsou color space [Ingling & Tsou, 1977] for appropriate comparison to the theoretical predictions below. The increase in thresholds is classically known as the Weber's law [Wyszecki & Stiles, 1982]. Middle left figure shows the saturating nonlinearity of the underlying brightness perception mechanism using Eq. (4.2). Top center and top right plots show the V-shaped curves of the color incremental thresholds for red-green and yellow-blue stimuli (replotted from Figs. 10 and 11 in [Krauskopf & Gegenfurtner, 1992]). Axes in the figures have been expressed in absolute T and D units by re-scaling the original data (given in threshold relative units) using threshold values for appropriate comparison to the theoretical predictions. Krauskopf and Gegenfurtner [Krauskopf & Gegenfurtner, 1992] used two adaptation conditions: (i) white adaptation point (at the origin in the T and D axes); and, in the case of the T discrimination, they also used (ii) a reddish adaptation point (about 2.5 in the rescaled T axis). Middle central and right plots show the corresponding response curves using Eq. (4.2).

The data show two interesting features of T and D color perception mechanisms:

- The discrimination is optimal (minimum threshold) at the chromaticity of the adaptation point, and then the sensitivity decreases as one departs from it. This gives rise to the saturating sigmoidal response curves with maximum slope at the adaptation point.
- A change in the adaptation state implies a shift in the response curve in order to set the maximum sensitivity region at the new adaptation point.

Adaptation and corresponding pairs

Corresponding colors are two sets of tristimulus values that give rise to the same perceived color when one sample is observed under two different light sources [Breneman,

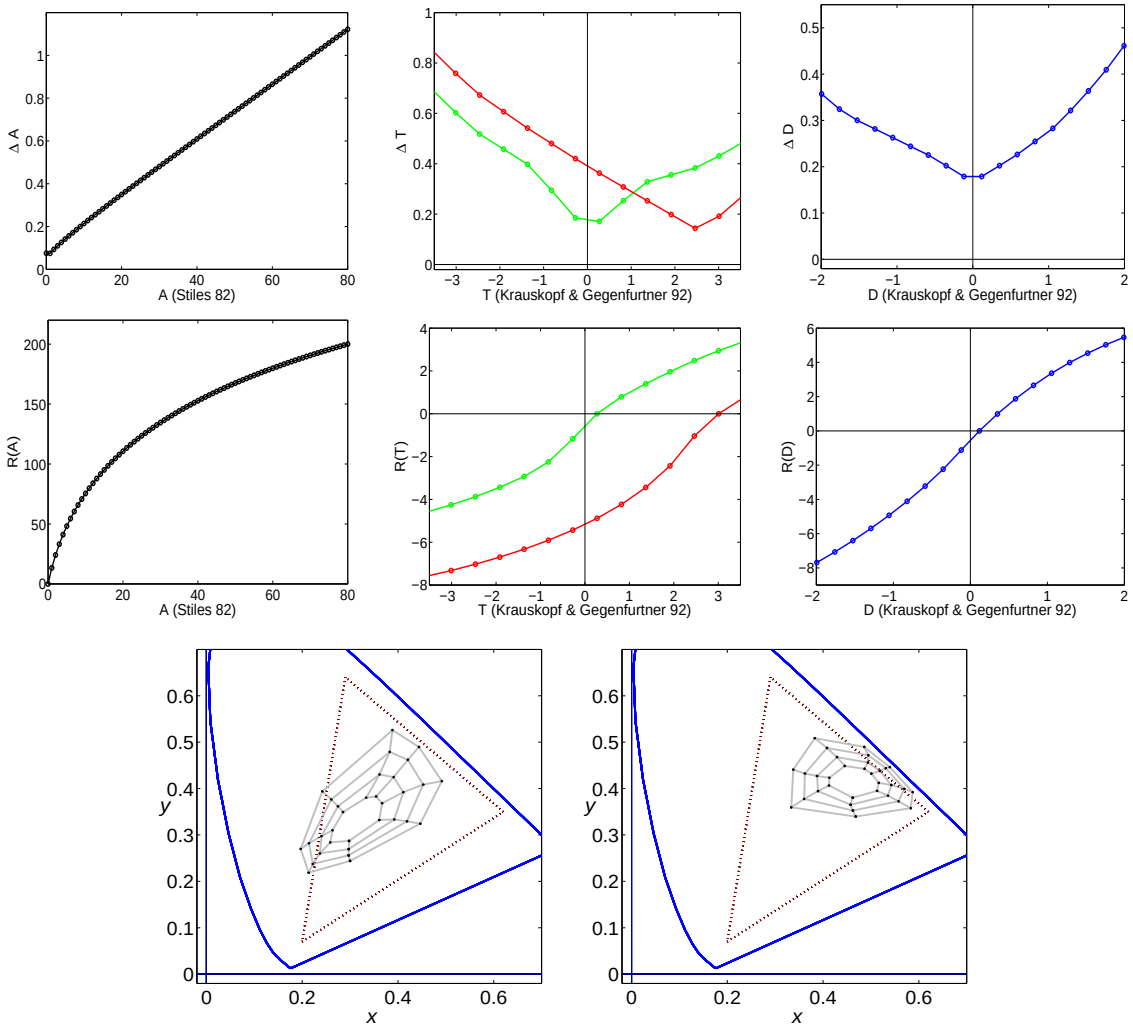


Figure 4.2: Summary of psychophysical results. Experimental color discrimination thresholds in ATD channels (top row) and the corresponding nonlinear responses (middle row). From left to right: channel A (Achromatic), channel T (Green-Red), and channel D (Yellow-Blue). In the case of the T channel, the green and the red curves represent the thresholds in two adaptation conditions: white adaptation point and reddish adaptation point respectively. Bottom row shows the experimental corresponding colors (CIE xy chromaticities) under CIE D65 illuminant (left) and under CIE A illuminant (right).

1987; M. Luo et al., 1991; M. Luo & Rhodes, 1999]. Corresponding colors reveal the color constancy ability of human observers under change of illumination conditions: despite the change in the linear measurements, the corresponding pairs are perceived as equal. A chromatic adaptation transform should be able to predict corresponding colors. The chromatic diagrams in Fig. 4.2 show the data compiled by Luo et al. [M. Luo et al., 1991; M. Luo & Rhodes, 1999] regarding corresponding colors under CIE D65 and CIE A illu-

minants[1]. Note that the corresponding stimuli under CIE A illuminant have reddish-yellowish chromaticities with regard to those under CIE D65 indicating the illuminant compensation ability of human viewers.

4.1.2 Sensor design by learning nonlinear data representations

The responses of sensory systems devoted to describe some phenomenon of interest have to convey as much information as possible about the phenomenon while minimizing the representation error for every possible input. The first of these related, but not exactly equivalent, requirements is usually known as *Information Maximization* [Bell & Sejnowski, 1995; Laughlin, 1983; Lee et al., 2000; Linsker, 1988]. Here we will refer to the second requirement [Lloyd, 1982; D. MacLeod & Twer, 2003; D. A. MacLeod, 2003; Twer & MacLeod, 2001] as *Error Minimization*.

Additionally, the sensors should discount variations in linear measurements coming from non-interesting sources: even though eventual modifications of the measurement conditions may give rise to changes in the PDF of the linear measurements, such as the ones illustrated in Fig. 4.1, the internal representation (the perception) has to be invariant to these changes. In the psychophysics literature, this ability is usually known as *adaptation* [M. Fairchild, 2005], in the signal processing literature as *adaptive modeling and filtering* [Haykin, 2002], while in the machine learning literature this has been recently referred to as *domain adaptation* [Pan et al., 2009; Storkey, 2009].

The manifold learning method proposed in this section is motivated by the *infomax* and the *error minimization* principles, and by the need of learning systems capable of dealing with the *adaptation* or *dataset shift problem* in the specific context of the color statistics.

Nonlinear sensory systems design: infomax and error minimization principles

Processing input observations $\mathbf{x} \in \mathbb{R}^n$ requires the design of an *appropriate* set of n, sensors that respond according to the mapping R, which transforms points $\mathbf{x}$ to $\mathbf{r} \in \mathbb{R}^n$. Physical sensors may have limited resolution or may be subject to internal noise in such a way that the responses are corrupted according to a sort of quantization Q,

$$\mathbf{x} \underset{R^{-1}}{\overset{R}{\rightleftarrows}} \mathbf{r} \overset{Q}{\rightarrow} \mathbf{r}^{\star} \,. \tag{4.3}$$

The *infomax* sensory organization principle states that the mapping R has to be selected to maximize the transferred information from $\mathbf{x}$ to $\mathbf{r}^{\star}$. This requirement induces different constraints on the Jacobian of the response transform $\nabla R(\mathbf{x})$ in the noise-free and noisy scenarios depending on the PDF of the input measurements, $p(\mathbf{x})$, as will be reviewed

[1] Data available on-line at `http://colour.derby.ac.uk`

below and extensively reported elsewhere [Bell & Sejnowski, 1995; Gersho & Gray, 1992; Laughlin, 1983; Lee et al., 2000; Lloyd, 1982; D. A. MacLeod, 2003]. In general, the sensitivity of the system (the slope of the response) in each region of the input space has to be related to the population in that region. Additionally, assuming that the internal representation of the sensory system Euclidean, as done in psychophysics [Hillis & Brainard, 2005; Laming, 1997], the system induces a *perceptual* metric in the input domain, $M(\mathbf{x})$, related to the Jacobian of the response transform, see [Dubrovin et al., 1982; Epifanio et al., 2003; Laparra, Marí, & Malo, 2010; Malo et al., 2006]:

$$M(\mathbf{x}) = \nabla R(\mathbf{x})^\top \cdot \nabla R(\mathbf{x}), \tag{4.4}$$

which follows from considering an Euclidean metric in the response domain, i.e. the sensory system considers all distortions in the same way[2]. Accordingly, relations between the sensitivity of the system and the population of the input space will give rise to relations between the induced metric in the input space and $p(\mathbf{x})$. The information maximization and error minimization criteria impose different restrictions on both the Jacobian and the metric.

On the one hand, in the noise-free case, the *infomax* principle to set R reduces to looking for transforms that lead to responses with maximal entropy or equivalently, to independent responses [Bell & Sejnowski, 1995; Lee et al., 2000]. This scenario implies a restriction on the Jacobian of the transform:

$$|\nabla R(\mathbf{x})| \propto p(\mathbf{x}), \tag{4.5}$$

which, according to Eq. (4.4), leads to the following determinant of the induced metric:

$$|M(\mathbf{x})| \propto p(\mathbf{x})^2. \tag{4.6}$$

On the other hand, the minimization of the representation error in sensory systems subject to internal noise or limited resolution leads to a different constraint on the Jacobian. In particular, in [D. A. MacLeod, 2003], MacLeod and Twer show that, in that situation, the optimal sensitivity in mean-square-error terms has to be,

$$|\nabla R(\mathbf{x})| \propto p(\mathbf{x})^{1/3}, \tag{4.7}$$

which is consistent with the classical optimal MSE distribution of discrete perceptions in Vector Quantization [Gersho & Gray, 1992; Lloyd, 1982]. According to Eq. (4.4), the determinant of the induced metric should be:

$$|M(\mathbf{x})| \propto p(\mathbf{x})^{2/3}. \tag{4.8}$$

The exponent accompanying the PDF in the Jacobian will be hereafter referred to as γ.

[2] In the situation described in Eq. (4.3), distances induced by small distortions in the input and the response domains, $\Delta\mathbf{x}$ and $\Delta\mathbf{r}$, may be described by local metrics $M(\mathbf{x})$ and $M(\mathbf{r})$: $d(\mathbf{x}, \mathbf{x}+\Delta\mathbf{x})^2 = \Delta\mathbf{x}^\top \cdot M(\mathbf{x}) \cdot \Delta\mathbf{x} = \Delta\mathbf{r}^\top \cdot M(\mathbf{r}) \cdot \Delta\mathbf{r} = d(\mathbf{r}, \mathbf{r}+\Delta\mathbf{r})^2$. Assuming that the response transform R is differentiable, the distortion in the response may be approximated by $\Delta\mathbf{r} \approx \nabla R(\mathbf{x}) \cdot \Delta\mathbf{x}$, which, assuming $M(\mathbf{r}) = I$, leads to Eq. (4.4).

Particular solutions for the response transform

The above constraints on the Jacobian do not lead to a unique solution for the transform. It is well-known that independent responses from input signals following a certain PDF (the *infomax goal*) may be obtained in many different ways [Hyvärinen & Pajunen, 1999]. A straightforward solution such as the equalization of the slices of the joint PDF at the input representation is not possible in practice from a finite set of samples due to the curse of dimensionality.

Iterative approaches related to Projection Pursuit [Huber, 1985] may circumvent this problem. In fact, as pointed out in [Laparra, Camps-Valls, & Malo, 2011], obtaining independent responses with deep neural networks is possible even using random rotations in the linear stages. However, in general, these iterative approaches lead to non-intuitive (or even meaningless) transform domains since R is not constrained to preserve the local geometry of the input space. According to this non-uniqueness, the *infomax principle* and the nonlinear ICA goals are not enough to determine the sensors that reveal the intrinsic coordinates of data. Nevertheless, a wide range of unsupervised manifold learning techniques has been proposed to extract the latent coordinates from raw measurements, although not exactly in the context of nonlinear ICA.

Self-Organizing Maps [Kohonen, 1982] and variants [Bishop et al., 1998] are based on tuning a predefined topology in such a way that the nonlinearities of the sensors and the *complete* lattice of discrete responses are obtained simultaneously. These approaches are not feasible in highly dimensional situations since the number of nodes in the lattice explodes with dimensionality. Another group of techniques is based on the eigen-analysis of graphs and kernels related to the local structure of the data in the manifold [Schölkopf et al., 2000; Tenenbaum et al., 2000; Weinberger & Saul, 2004], or on sparse matrices describing the local topology of the data [Belkin & Niyogi, 2002; Roweis & Saul, 2000]. Though efficient in many tasks, these *spectral* methods do not generally yield intuitive mappings between the original and the intrinsic curvilinear coordinates of the low dimensional manifold. In addition, even though a metric can be derived from particular kernel functions [Burges, 1999], the interpretation of the transformation is hidden behind implicit mappings and out-of-sample extensions are typically difficult, if not impossible. An alternative family of manifold learning methods consider complicated manifolds as a mixture of local models [Kambhatla & Leen, 1997] that are identified and conveniently merged into a single global representation [Brand, 2003; Roweis et al., 2002; Teh & Roweis, 2003; Verbeek et al., 2002]. The explicit direct and inverse transforms to the intrinsic representation can be derived from the obtained mixture model.

Enforcing coordination between neighboring local models may be seen as reducing multi-information between variables in the coordination (or unfolding) operation. This relates NL-ICA with techniques based on coordination of local models. However, in [Brand, 2003; Roweis et al., 2002; Teh & Roweis, 2003; Verbeek et al., 2002], the effect of these local

operations in the (eventually point-dependent) metric or line element was not explicitly analyzed. In the context of NL-ICA, an alternative way of merging locally disconnected representations was proposed in [Malo & Gutiérrez, 2006]. In that case, the global representation was based on the fact that the global NL-ICA at a certain point, $R(\mathbf{x})$, may be differentially approximated by the local linear ICA separating matrix, $W(\mathbf{x})$, [Chang & Lin, 2001a]. The issue was posed as an initial value problem and the global representation was obtained by integrating the local separating matrices in *arbitrary paths*. Note that, in the particular case of a mixture of Gaussians, the factorization of local models is consistent with (1) the Mahalanobis distance, and (2) the relation between the probability, the response and the metric under the noise-free infomax assumption[3]. However, the coordination by integrating the differential behavior in arbitrary paths as proposed by Malo & Gutiérrez [2006] only works for manifolds where the set of local basis functions fulfills the Stokes' theorem in the sense used in conservative vector fields. Moreover, the invertibility of the transform was not addressed therein [Malo & Gutiérrez, 2006].

In conclusion none of the above learning techniques is readily applicable to the simultaneous explanation of the non-linearities and adaptation of color vision mechanisms.

Our proposal for the response transform

The method proposed in Section 4.1.3 is based on the assumption of mixture of local models as classical methods based on vector quantization [Kambhatla & Leen, 1997] and the variants that enforce model coordination [Brand, 2003; Roweis et al., 2002; Teh & Roweis, 2003; Verbeek et al., 2002]. However, no explicit mixture of models is computed in our approach. On the contrary, as in [Malo & Gutiérrez, 2006], we propose to merge the local models by integrating some differential behavior, ∇R. However, unlike [Malo & Gutiérrez, 2006], the integration is done along a particular sequence of successive Principal Curves, similar to propossed by Delicado [2001], instead of using arbitrary paths. In this way, fulfillment of the Stokes' theorem in the manifold is no longer required, and a meaningful transformed domain is obtained since the differential behavior is integrated along meaningful trajectories in the manifold thus preserving the local topology of the input space. Easy interpretation of the features defined by the Principal Curves solves the interpretability problem of nonlinear ICA techniques related to Projection Pursuit where the independent representation may be even random [Laparra, Camps-Valls, & Malo, 2011]. Moreover, here we propose an explicitly tunable local metric according to the local PDF to achieve different goals such as *infomax*, as in Eq. (4.6), or *error minimization*, as in Eq. (4.8).

[3]If the local models are assumed to be Gaussian, local factorization is achieved by local PCA and whitening. Specifically, if the local covariance can be decomposed as $\Sigma(\mathbf{x}) = B(\mathbf{x})\Lambda(\mathbf{x})B(\mathbf{x})^\top$, the local separating matrix is just $W(\mathbf{x}) = \Lambda(\mathbf{x})^{-1/2}B(\mathbf{x})^\top$. In that case, the metric is $M(\mathbf{x}) = W(\mathbf{x})^\top \cdot W(\mathbf{x}) = B(\mathbf{x})\Lambda(\mathbf{x})^{-1}B(\mathbf{x})^\top = \Sigma(\mathbf{x})^{-1}$, i.e. the local Mahalanobis metric. Note also that $|\Sigma(\mathbf{x})|^{-1/2}$ is inversely proportional to the volume of the local Gaussian support, thus, in this case, $|\nabla R(\mathbf{x})| = |W(\mathbf{x})| \propto p(\mathbf{x})$ as in Eq. (4.5).

Finally, the proposed transform is readily invertible which is a key issue to reproduce chromatic adaptation (see Section 4.1.5). Accordingly, the proposed response is suitable to reproduce the experimental facts reviewed in Section 4.1.1 using the optimality criteria reviewed in Section 4.1.2.

4.1.3 Sequential Principal Curves Analysis (SPCA) with local metric

This section presents a manifold learning method, SPCA, that gives rise to an invertible transform, R. The technique can be seen as a method to design a set of eventually nonlinear sensors optimized according to the different goals reviewed in Section 4.1.2. The method is first motivated by the particular characteristics of smooth manifolds. Then, we present the direct and inversion transforms, and finally study the impact of the metric on the solution.

Motivation

SPCA is based on the following characteristics of curved manifolds:

1. Representing the data in curvilinear coordinates defined by Principal Curves (PCs) yields a representation where the data are unfolded. Intuitively, the dimensions become more meaningful in the sense that each one isolates a distinct feature of the signal (i.e. they are more independent). In [Laparra, Jiménez, et al., 2011b] it is shown that formulating PCs as a set of local rotations and alignments reduces multi-information in curved manifolds. Specifically, unfolding along a PC is an adequate step towards independence since it makes equal the first moment of all the conditional PDFs along the curve.

2. Additional local processing after unfolding is required to achieve, either independence or minimum representation error, by using local expansions or compressions (i.e. locally changing the metric) of the unfolded domain.

The diagram in Fig. 4.3 illustrates the intuitive ideas behind the proposed SPCA, that will be confirmed in the example of Fig. 4.4. We assume the existence of a transformation defined by a curvilinear lattice made of recursively defined PCs along all dimensions. This lattice is similar to the topology assumed in Self-Organizing Maps (SOMs) and variants. In our context, each dimension of the lattice can be seen as the optimal feature for an eventually non-linear sensor in the original domain, defining a canonical direction in the transformed domain. However, unlike SOM, the whole lattice has not to be explicitly computed in order to find the transformation (response) of a particular point (stimulus). To do so, we propose to integrate a local differential behavior, $\nabla R(\mathbf{x})$, along a particular path.

The proposed integration path first follows the *standard PC of the set* [Hastie & Stuetzle, 1989] up to the geodesic projection of the point $\mathbf{x}$ on this *first* PC, $\mathbf{x}^1_\perp$. The *first* PC pro-

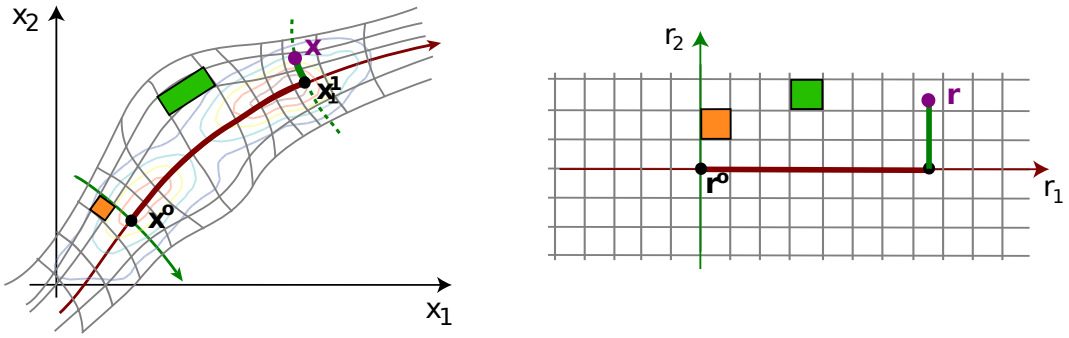


Figure 4.3: Illustration of the SPCA with Local Metric. Left plot represents the input domain $\mathbf{x}$ and right plot represents the response domain $\mathbf{r}$. Colored contours represent the underlying PDF, $p(\mathbf{x})$. The assumed curvilinear lattice (gray lines) is not explicitly computed. The proposed differential behavior (Eqs. (4.9) and (4.11)) implies that highly populated regions (such as the orange area) are expanded while lower density regions (such as the green area) are shrunk in the response domain (right figure). Given an origin of coordinates, $\mathbf{x}^o$, in the first PC (red line) and some point of interest, $\mathbf{x}$, the response for the point of interest is given by the lengths (the integrals in Eq. (4.18)) along the path consisting of successive PCs: the first (or standard) PC in red, and the second PC (in green) in the orthogonal subspace at $\mathbf{x}^1_\perp$, which is the (geodesic) orthogonal projection of $\mathbf{x}$ on the first PC.

vides a global summary of the whole dataset but the residual structure in the hyperplanes orthogonal to this first PC may also be worth to be described. In principle, any set of $(d-1)$ linearly independent vectors living in the corresponding hyperplanes would be equally suited to form a linear basis to describe this structure. However, as noted by Delicado [Delicado, 2001], the structure at those hyperplanes may also be nonlinear so it makes sense to draw secondary PCs to capture this residual structure. These ideas imply that geodesic projections according to the local structure of the manifold can be obtained by secondary PCs [Laparra, Jiménez, et al., 2011b]. After following the first PC, the path follows the *second PC* [Delicado, 2001] at $\mathbf{x}^1_\perp$, i.e. the PC of the orthogonal subspace with regard to the first PC at the point $\mathbf{x}^1_\perp$. In this second segment, the path goes up to the geodesic projection of the point $\mathbf{x}$ on the second PC, $\mathbf{x}^2_\perp$. This *sequence* is continued until the last dimension. The lengths of the curved segments represent the projections in each dimension of the new representation and may be seen as the response of d sensors tuned to curved features.

The proposed differential behavior, ∇R, is based in the above mentioned characteristics of smooth manifolds, and can be expressed as:

$$\nabla R(\mathbf{x}) = D(\mathbf{x}) \cdot \nabla U(\mathbf{x}), \tag{4.9}$$

where $\mathbf{u} = U(\mathbf{x})$ is the unfolding transform that consists of concatenated local rotations along the proposed path made of a sequence of PCs, and the diagonal matrix $D(\mathbf{x})$ rep-

resents the local length of the line element along this path (change of metric). Note that $\nabla U(\mathbf{x})$ is orthonormal for all $\mathbf{x}$ since the unfolding U can be formulated as a set of concatenated local rotations. In fact, in the selected method to draw one PC [Laparra, Jiménez, et al., 2011b], the curve consists of aligned rotations estimated by using local PCA. This is consistent with the fact that other PC algorithms use local PCA to estimate the tangent to the curves [Delicado, 2001; Einbeck et al., 2005].

In order to adapt the metric to the density, we set the elements of D using the marginal PDF on the unfolded coordinates and an appropriate exponent $\gamma \geq 0$:

$$D(\mathbf{u})_{ii} \propto p_{u_i}(u_i)^{\gamma}, \tag{4.10}$$

where the marginal on each direction is estimated following k-neighborhood rule. The metric induced in the input space is:

$$M(\mathbf{x}) = \nabla U(\mathbf{x})^{\top} \cdot D(\mathbf{x})^2 \cdot \nabla U(\mathbf{x}). \tag{4.11}$$

Assuming that local clusters can be factorized by the local rotations, and taking into account that $|\nabla U(\mathbf{x})| = 1$, we have

$$|M(\mathbf{x})| = |D(\mathbf{x})|^2 \propto \prod_{i=1}^{n} p_{u_i}(u_i)^{2\gamma} = p(\mathbf{u})^{2\gamma} = p(\mathbf{x})^{2\gamma} |\nabla U(\mathbf{x})|^{-2\gamma} = p(\mathbf{x})^{2\gamma}, \tag{4.12}$$

which, with the appropriate choice of γ, is the behavior required in Eqs. (4.6) or (4.8).

Unfolding along Principal Curves: the cumulants perspective

Unfolding along a principal curve and cluster alignment imply a step in the right (independence) direction but it is not enough since a metric change is needed. This is easy to see by looking at the cumulant expansion of PDFs. Unfolding along a principal curve with parameter, u_1, implies independence with regard to orthogonal subspaces *iff* it gives rise to:

$$p(u_2, \ldots, u_d | u_1) = p(u_2, \ldots, u_d) \tag{4.13}$$

Therefore, the cumulant generating functions of both sides of the above equation should be equal:

$$1 - j\boldsymbol{\omega}^{\top}\mathbf{m}_1 + \frac{1}{2}\boldsymbol{\omega}^{\top}\boldsymbol{m}_2\boldsymbol{\omega} - \ldots = 1 - j\boldsymbol{\omega}^{\top}\mathbf{m}'_1 + \frac{1}{2}\boldsymbol{\omega}^{\top}\boldsymbol{m}'_2\boldsymbol{\omega} - \ldots \tag{4.14}$$

where $\mathbf{m}_i$ and $\mathbf{m}'_i$ are the ith-order moments of each PDF, and $\boldsymbol{\omega}$ is the parameter of the characteristic functions. Independence holds if $\mathbf{m}_i = \mathbf{m}'_i, \forall i$.

One principal curve should satisfy the parametric equation [Hastie & Stuetzle, 1989]:

$$f(u_1) = \mathbb{E}[\mathbf{x} | \lambda(\mathbf{x}) = u_1], \tag{4.15}$$

where $\lambda(\mathbf{x})$ is the orthogonal projection of $\mathbf{x}$ on the curve, so $\{\mathbf{x}|\lambda(\mathbf{x}) = u_1\}$ is the orthogonal subspace at the curve point u_1. According to this, the curve passes through the average of the orthogonal subspace (the origin of the subspace in the unfolded representation):

$$\mathbb{E}[u_2, \ldots, u_d|u_1] = \mathbf{0}, \forall u_1 \Rightarrow \mathbb{E}[u_2, \ldots, u_d] = \mathbf{0}, \tag{4.16}$$

which means that *unfolding along a principal curve* makes $\mathbf{m}_1 = \mathbf{m}'_1 = \mathbf{0}$. However higher order moments may not be equal along the curve.

In the infomax context, the effect of the suggested local change of the metric along the curve (setting the line element, or local equalization) should be achieving a constant PDF along the curve thus ensuring the equality of all higher order moments.

Direct transform

Given an arbitrary origin of coordinates on the first PC, $\mathbf{x}^o$, assumed to give zero response, $\mathbf{r}^o = \mathbf{0}$, and some point of interest, $\mathbf{x}$, the corresponding response is given by the following integration along the path on PCs described above (cf. Fig. 4.3):

$$\mathbf{r} = R(\mathbf{x}) = C \cdot \int_{\mathbf{x}^o}^{\mathbf{x}} \nabla R(\mathbf{x}') \cdot d\mathbf{x}' = C \cdot \int_{\mathbf{x}^o}^{\mathbf{x}} D(\mathbf{x}') \cdot \nabla U(\mathbf{x}') \cdot d\mathbf{x}', \tag{4.17}$$

where C is just a constant diagonal matrix that independently scales each component of the response. The selected path implies displacements in one PC at a time. According to this, in each segment of the path, the vectors $d\mathbf{u}' = \nabla U(\mathbf{x}')d\mathbf{x}'$ have only one non-zero component: the one corresponding to the considered PC at the considered segment. Therefore, the response of each sensor to the point $\mathbf{x}$ is just the length on each Principal Curve in the path from $\mathbf{x}^o$ to $\mathbf{x}$, measured according to the metric related to the local density with the selected exponent,

$$r_i = C_{ii} \cdot \int_{\mathbf{x}_\perp^{i-1}}^{\mathbf{x}_\perp^i} D(\mathbf{x}') \cdot \nabla U(\mathbf{x}') \cdot d\mathbf{x}' = C_{ii} \int_0^{u_{i\perp}^i} p_{u_i}(u_i')^\gamma \, du_i', \tag{4.18}$$

SPCA is initialized by setting (i) the origin of the coordinate system, and (ii) the scale of the different dimensions, C_{ii}, and the order in which they will be visited by the sequential algorithm. Sensible choices for the origin are those suggested in other bottom-up Principal Curve algorithms [Delicado, 2001; Einbeck et al., 2005]: the most dense point of the distribution (if known) or the mean of the data. Then a set of d locally orthogonal principal curves is drawn at the selected origin, which will be used to set the order and the relative scale of the dimensions. In our case, we set the scaling constants C_{ii} according to an information distribution criterion: we use the number of quantization bins *per* dimension given by classical bit allocation results in transform coding [Gersho & Gray, 1992]. This is consistent with sorting the curvilinear dimensions according to the marginal entropy

(higher entropy first). Note that, other criteria could be used, as for instance the total standard deviation of the projected data (as in global PCA) or the total Euclidean length of the curvilinear axes.

Once the dimensions have been sorted and the global scaling is set, SPCA obtains the transform of an arbitrary point $\mathbf{x}$ by *sequentially* applying the next two steps. Step 1 traces a principal curve in the i-th direction from the previous starting point ${\mathbf{x}_\perp}^{i-1}$ (${\mathbf{x}_\perp}^0$ is the origin of the coordinate system). Step 2 defines the line element in the drawn principal curve using the marginal PDF along the curve, $p_{u_i}(u_i)^\gamma$. The response r_i will be the integral of the line element from ${\mathbf{x}_\perp}^{i-1}$ to the geodesic projection of $\mathbf{x}$ into the principal curve, which is ${\mathbf{x}_\perp}^i$ (cf. Eq. 4.18). Details on the iterative refinement procedure to obtain the geodesic projections from orthogonal projections are given in [Laparra, Jiménez, et al., 2011b]. Since SPCA requires that the individual principal curves are drawn in particular directions from particular points, appropriate algorithms to draw individual curves should operate in a bottom-up manner, as those in [Delicado, 2001; Einbeck et al., 2005] or the particular one used here [Laparra, Jiménez, et al., 2011b]. The local-to-global behavior in the selected algorithm to draw each PC is necessary to identify the structure around ${\mathbf{x}_\perp}^{i-1}$ in the subspace locally orthogonal to the previous PC.

A Matlab implementation of SPCA with worked examples is available on-line[4]. Figure 4.4 illustrates the performance of SPCA in a practical situation.

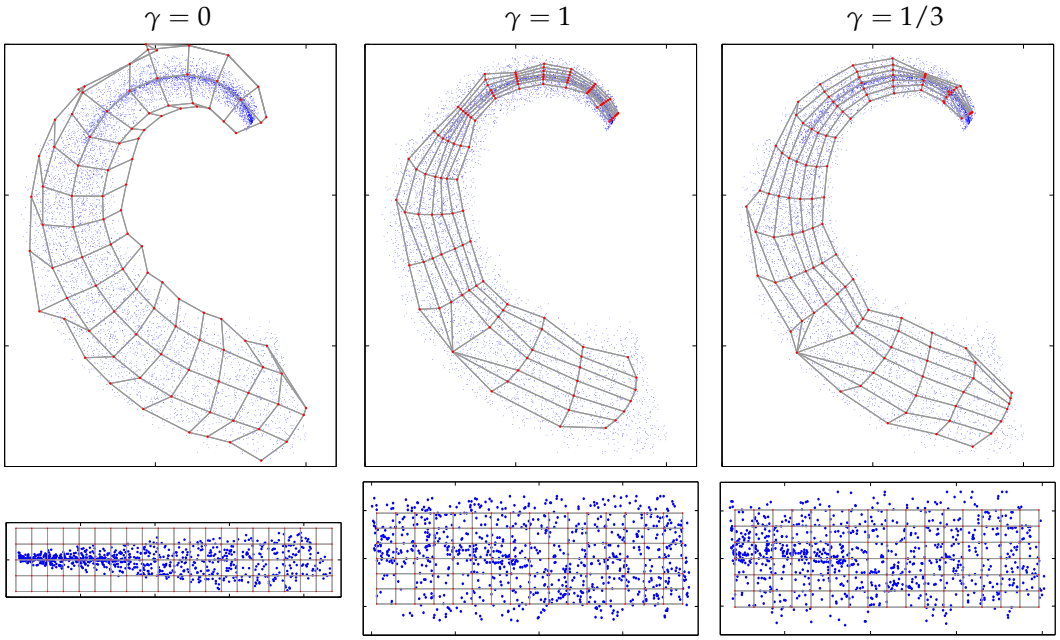
Inverse transform

A distinctive property of the method is the possibility of computing the inverse of the transform. Given a set of samples from the same source, the origin in the input space, $\mathbf{x}^o$, and the scale and order of the dimensions, the computation of the inverse, $\mathbf{x} = R^{-1}(\mathbf{r})$, is very simple. It involves drawing the first PC through the origin and taking the length r_1 on this curve, measured according to $p_{u_1}(u_1)^\gamma$. Displacement on the first curve by the length r_1 leads to the first projection $\mathbf{x}_\perp^1$. Then, the second locally orthogonal curve is drawn from $\mathbf{x}_\perp^1$, and one takes a second displacement r_2 on this second PC leading to the second projection, $\mathbf{x}_\perp^2$. This process is repeated sequentially in every dimension until the desired point $\mathbf{x}$ is found by taking the displacement r_d from $\mathbf{x}_\perp^{d-1}$ on the d-th principal curve.

Infomax and error minimization through SPCA

Here we present a synthetic experiment that stresses the usefulness of SPCA in sensor design, and study the effect of using different metrics ($\gamma = 0, 1$ and $\frac{1}{3}$). We generated 10000 samples from an illustrative curved manifold with changing PDF: half of the manifold has an increasing variance Laplacian distribution while the other half follows an increasing variance uniform distribution (Fig. 4.4).

[4] `http://isp.uv.es/spca.html`



RMSE	0.55	0.56	0.53
MI	0.27	0.05	0.06

Figure 4.4: Infomax and error minimization through SPCA. Samples of the sets in the first row were transformed using SPCA (second row) with different γ value. Additionally, Cartesian lattices in the response domain were inverted back into the input domain giving rise to the curved lattices in the top row. Results are analyzed in terms of independence (Mutual Information in bits), and reconstruction error (root-mean-square-error, RMSE). In each case, MI was computed in the corresponding transform domain, while RMSE values refer to the quantization error in the original domain using the corresponding lattices as codebook. For the sake of reference, in the original domain results were $MI = 0.75$ bits and RMSE=0.63 (using uniform scalar quantization). Note how $\gamma = 1$ obtains better results in independence while $\gamma = \frac{1}{3}$ is better for RMSE minimization.

The advantage of using Principal Curves to design a set of sensors is that their flexibility makes them suitable to describe curved manifolds, as pointed out in the example. No matter the metric used, an unwrapped representation of the data is obtained. When using $\gamma = 0$ the data is unfolded and the original local metric is preserved (e.g. the different distributions inside the manifold remain the same). When using $\gamma = 1$, we obtain a representation where the different distributions are almost uniformized leading to a representation where the different dimensions are almost independent. Finally, when using $\gamma = \frac{1}{3}$, the reconstruction error is minimized. In this latter case, redundancy is certainly reduced with regard to the input domain, however, the kurtotic structure of the Laplacian is more visible than in the second case. Note also the differences in the distribution of the

inverted lattices: while in the $\gamma = 0$ case, lattice cells are approximately uniform no matter the local population (local metric independent of the PDF), in the other cases, the size is related to the population, e.g. the $\gamma = 1$ case results in tighter slices around the peak of the Laplacian. As anticipated in this section, unfolding alone ($\gamma = 0$) in general is not enough to remove redundancies, but additional processing, i.e. local changes in the metric related to the local PDF, are required to achieve independent components.

4.1.4 Simulation of color psychophysics using SPCA

This section describes the procedure to simulate the experimental nonlinearities and the adaptation results described in Section 4.1.1 using SPCA on suitable ensembles of natural colors. Since the available experimental data involve adaptation under specific white and reddish illuminations, a new database was required.

Database of calibrated natural color images

Calibrated measurements (tristimulus values instead of digital counts) and controlled white and reddish illumination on the same objects are needed to ensure the appropriate statistical adaptation conditions in the simulation of the psychophysics. Unfortunately, the current available color image databases do not fulfill such requirements because of different reasons:

1. Spectro-radiometric natural image databases, such as those used in [Brown, 1994; D. Ruderman & Chiao, 1998; J. et al., 1994; Nascimento et al., 2002; Parraga et al., 1998; M. Webster & Mollon, 1997], may be used to estimate the reflectance of natural surfaces under the flat Lambertian assumption. Then, these reflectance values can be used to obtain new tristimulus values under different illuminants. However, such procedure neglects the nonlinearities induced by geometric factors and mutual illumination, which are relevant factors to induce non-uniformities within the PDF support (as illustrated in Fig. 4.1).

2. Databases where the illumination is modified on the same objects, include spectro-radiometric examples [Brainard et al., 2000] and uncalibrated examples [Barnard et al., 2002; Geusebroek et al., 2005]. The problem in these cases is that either the database consists of a very restricted set of artificial objects (unnatural clusters in the color space) [Barnard et al., 2002; Brainard et al., 2000], or that the database is not calibrated [Barnard et al., 2002; Geusebroek et al., 2005].

3. Calibrated natural image color databases, such as [Doi et al., 2003; Olmos & Kingdom, 2004; Parraga et al., 2009], are not suitable for the simulation of color adaptation because either they do not include the same surfaces under the required controlled

illuminants or they are not wide enough to find samples with the appropriate illumination.

4. A large database, such as that in [Ciurea & Funt, 2003], does include a wide range of scenes, a subset of which could match the desired white and reddish adaptation conditions but, unfortunately, it has been acquired with an uncalibrated video camera.

These shortcomings led us to compile a new color image database of natural objects in controlled illumination conditions. We used a Macbeth light chamber equipped with standard CIE D65 and CIE A illuminants and we took the CIE XYZ pictures using a calibrated image colorimeter Lumicam1300. The accuracy of the illuminants and the measurements was checked by taking pictures of 10 hue pages of the Munsell's Book of Color and comparing the results with theoretical tristimulus values computed from the reflectance of the samples and the radiances of the illuminants. The database consists of 75 scenes of size 1000×1280. For each scene, two pictures were taken under CIE A and CIE D65 illuminants. The scenes include plants and flowers, natural terrain and materials, samples of colored fabric, office material, and Munsell chips.

The database is publicly available on-line[5] and it is suitable for other accurate experiments on color constancy and chromatic adaptation. Details on the experimental procedure to gather the database are given in the dedicated web site. In our specific experiments, we used 50 images excluding the Munsell chips and the pictures of the (too flat) artificial objects. This amounts to $64 \cdot 10^6$ color samples for each illumination. Figure 4.5 shows the pictures considered in our experiments. Transformation from CIE XYZ values to VGA digital counts for visualization purposes in Fig. 4.5 was done using standard display calibration data [Malo & Luque, 2000]. This may introduce some color reproduction errors in Fig. 4.5. However, note that these eventual errors do not affect the simulations, which were done from the raw CIE XYZ measurements.

Procedure for the simulation of color mechanisms behavior using SPCA

Simulation of nonlinearities Nonlinearities along the A, T and D dimensions of the color space and their variations under adaptation changes can be reproduced by computing the response of the SPCA mechanisms on the corresponding axes and the appropriate adaptation environment (CIE D65 set or CIE A set). Figure 4.6 (top row) shows the points considered in the simulation in the Ingling and Tsou ATD space [Ingling & Tsou, 1977]. This space is selected as the input linear representation instead of the MacLeod and Boynton ATD space [D. MacLeod & Boynton, 1979] used in [Krauskopf & Gegenfurtner, 1992] because it better reproduces basic psychophysical data such as color matching functions

[5] `http://isp.uv.es/databasecolor.html`

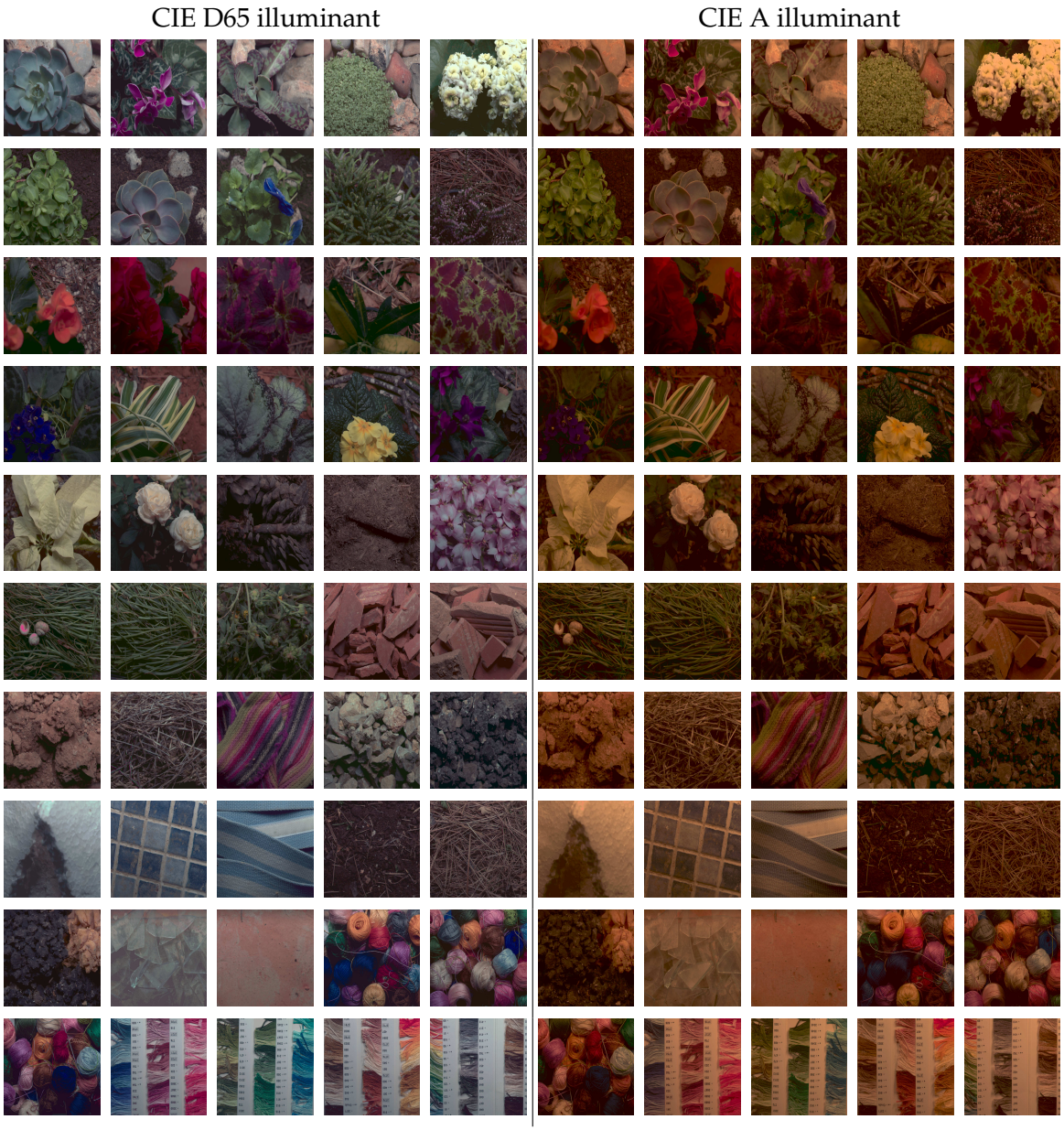


Figure 4.5: Scenes used in the statistical simulations under different illumination.

similar to Jameson and Hurvich hue cancellation curves, and appropriate orientation of the McAdam's ellipse at the white point [Capilla et al., 1998].

Experimental results on nonlinearities can be simulated in two different ways that we will refer to as the *psychophysical paradigm* and the *physiological paradigm*. In the *physiological paradigm*, we assume we have access to the response of each mechanism as in an ideal neuron recording. In this case, we can register the response of the corresponding sensor (the first sensor in the A case, the second in the T case, and the third in the D case), and we can simulate the incremental thresholds of these mechanisms from the derivative (slope)

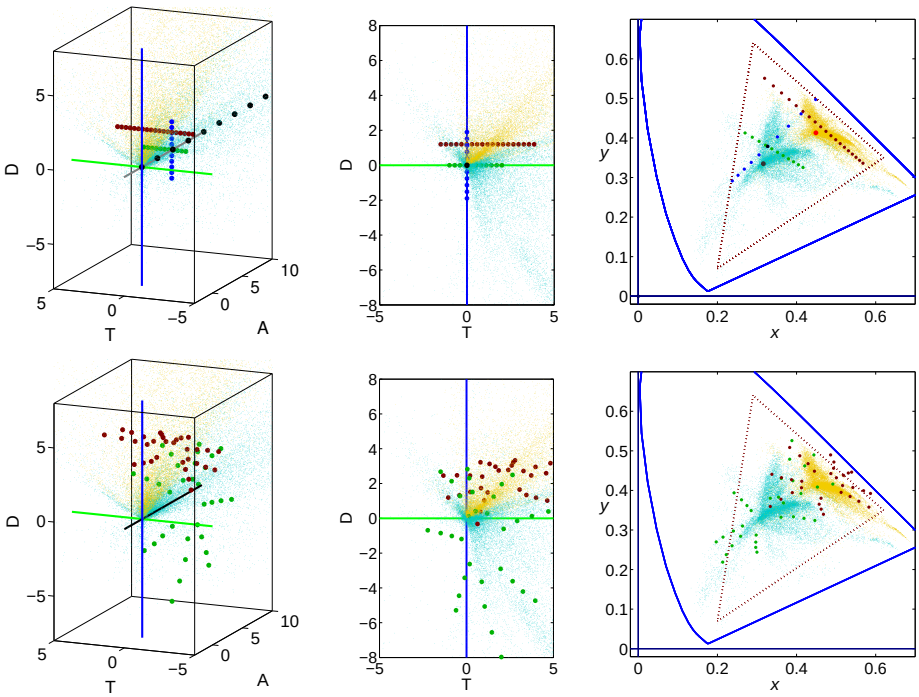


Figure 4.6: Training and test points for the simulation of the psychophysics. Small points represent the training samples of the database (cyan for CIE D65 and yellow for CIE A) and large points represent the different sets of test points. In these plots only a subset of randomly selected training points is shown for better visualization. **Top:** Black, green and blue dots in the top row are the considered points to simulate the nonlinearity of the A, T, and D mechanisms respectively. In these cases the response is computed using the CIE D65 training set (gray dots). Dark red dots are the considered points to simulate the nonlinearity in a reddish environment using the CIE A training set (light red dots). Top left plot shows the data in the tristimulus ATD space. A zoom around the origin is shown here for better visualization. However note that test points along the achromatic axis spread up to A = 80. Top center plot show the data in the (T, D) plane. Top right plot shows the data in the CIE xy chromaticity diagram. In this case, the chromatic coordinates of the CIE D65 and CIE A illuminants are also shown for reference (larger gray and red dots respectively). **Bottom:** Green and red points in the bottom row are the considered points to compute the corresponding pairs using the CIE D65 training set (cyan dots) and the CIE A training set (yellow dots).

of the responses at the considered points. In the *psychophysical paradigm*, the isolated responses are assumed to be inaccessible. On the contrary, we assume a certain summation of the variations in the responses of the sensors (e.g. the Euclidean norm for simplicity). The incremental threshold is reached when this norm achieves some prefixed value. In this way, we can simulate the thresholds and, by integrating their inverse, the underlying response can be derived as in psychophysics, cf. Eq. (4.2).

As in any finite color database, a certain bias is expected [Koenderink, 2010]. Fig-

ure 4.6 displays the existing bias in the collected database. Note that the maximum of the PDFs (the statistical adaptation points) in each case do not match the CIE D65 and the CIE A chromaticities due to the particular objects in the database. Moreover, the most dense points are also shifted from the origin in the considered linear representation (intersection point between the T and D axes in the top right plot in Fig. 4.6). This will introduce the corresponding bias in the results but it does not reduce the generality of the results, as recognized in previous statistical studies also dealing with biased databases [D. MacLeod & Twer, 2003; D. A. MacLeod, 2003].

Simulation of adaptation Our proposal for domain adaptation using SPCA as response transform, R, is inspired by the *corresponding pair procedure* framework used in chromatic adaptation models to predict corresponding stimuli [Capilla et al., 2004]. In this framework, linear measurements (e.g. CIE XYZ, LMS or ATD tristimulus values, $\mathbf{x}$), obtained in different conditions, C, are transformed according to an invertible color appearance model described by a transform, R, to a canonical space, e.g. the space of perceptual descriptors, $\mathbf{r}$, related to brightness, hue and colorfulness. The direct and inverse transforms depend on the measurement conditions, C. Measurement conditions may include information about the environment (e.g. spectral illumination, geometry), or information about the properties of the measurement system (e.g. normal or defective observers):

$$\mathbf{r} = R_C(\mathbf{x}). \tag{4.19}$$

Once a given point acquired in situation B is transformed to the canonical representation of perceptual descriptors, it can be transformed back into the input domain of situation A by using the inverse of the transform for situation A (see Eq. (1) in [Capilla et al., 2004]):

$$\widehat{\mathbf{x}}_{\mathbf{A}} = R_A^{-1}(\mathbf{r}_{\mathbf{B}}) = R_A^{-1}(R_B(\mathbf{x}_{\mathbf{B}})). \tag{4.20}$$

In problems where changes in the PDF due to non-interesting sources are smooth such as the ones found in color vision, we conjecture that transforms to canonical domains defined by the meaningful latent variables of the manifold can be used to solve the dataset shift problem. Changes in the PDF may give rise to nonlinear deformations of the curvilinear coordinates of the manifold and to changes on the length measures on them. By using the technique proposed in Section 4.1.3, one should be able to arrive to the same response thus achieving a canonical invariant representation. The results in Fig. 4.7 illustrate this concept.

In the simulations of the empirical chromatic adaptation data, we transform one of the sets of the corresponding color data (e.g. colors at the bottom row in Fig. 4.6) using the learned transform with SPCA with the appropriate adaptation environment (e.g. the CIE D65 training set). Then, the obtained responses are inverted back into the ATD space by using the inverse SPCA with the other environment set (e.g. the CIE A training set). In

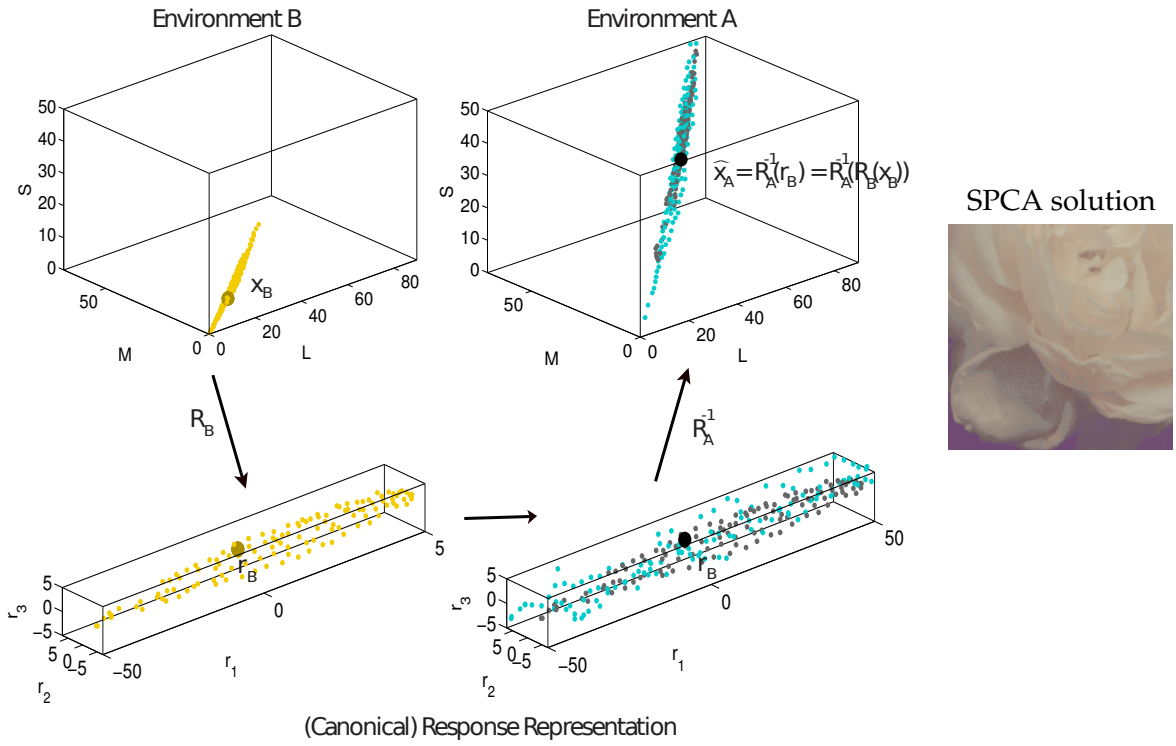


Figure 4.7: Dataset shift compensation by using the corresponding pair concept and SPCA. Transforms leading to the corresponding latent coordinates of the manifold in the environments A and B may be used to estimate the position in environment A of new points measured in environment B. Unlike in the linear adaptation cases in Fig. 4.1, the proposed nonlinear transform not only removes the yellowish appearance, but additionally the shadows are reduced as expected in a better PDF matching. In particular, note how the highlighted point $\mathbf{x}_B$ (the same one as in Fig. 4.1) results in a white, *higher luminance* corresponding point $\hat{\mathbf{x}}_A$.

our simulations, the procedure was applied in both directions: from CIE D65 to CIE A, and viceversa. In each case, the computed colors have to be compared with the experimental data in Fig. 4.2 (bottom row). Figure 4.6 (bottom row) shows the training and test data for the corresponding pairs experiment.

4.1.5 Numerical results and discussion

This section shows how both nonlinearity and adaptation phenomena emerge from tristimulus samples using the proposed SPCA. In particular, we show the results for the nonlinear behavior along the ATD directions and the corresponding data reproduction using SPCA with the *error minimization* and the *infomax* strategies (exponents $\gamma = 1/3$ and $\gamma = 1$ respectively).

Parameters for drawing a principal curve

The parameters associated to the particular algorithm used to draw individual Principal Curves refer to the rigidity of the assumed underlying grid, or equivalently, to the freedom to find curved axes far from the global linear PCA solution. In our implementation, rigidity is controlled with the locality k, the step size τ, and the stiffness q (details in [Laparra, Jiménez, et al., 2011b]). In the color statistics problem, the manifold is not globally curved and changes in spectral illumination induce almost linear rotations. In this situation, the relevant nonlinearities come from the non-uniform data distribution inside the PDF support basically due to the statistics of reflectance and geometric issues such as oblique illumination (cf. Fig. 4.1). These non uniformities are not taken into account by the rigidity parameters of the particular PC algorithm (k, τ and q) but by the non-Euclidean metric used in the SPCA framework (i.e. by the *infomax* or the *error minimization* strategies). According to this, in the problem at hand, the relevant comparison is between these strategies, which incidentally is the biologically interesting issue.

In our case, the rigidity constraints of the principal curves algorithm have been optimized for best performance and applied in the same way in both *infomax* and *error minimization* cases. Optimization of rigidity parameters has been done by exhaustive search in a discrete grid in the parameter space. The best values found were: $k = 0.2$ (20 % of the samples in the neighborhoods), $\tau = 15$ in Euclidean units in the considered ATD space, and $q = 16$ for the stiffness parameter. These parameters imply assuming a relatively rigid underlying grid, which makes sense in the color statistics problem.

Results

Figures 4.8 and 4.9 show the SPCA results for the *error minimization* and the *infomax* strategies respectively. In the reproduction of the thresholds and nonlinearities, we used both the *physiological* and the *psychophysical* paradigms. Since results are very similar for both, we just show the *physiological-like* result in each case. Black lines in the plots indicate the axes in the input ATD space. The deviation of the responses from the origin comes from the bias of the database. This just stresses the fact that the algorithm is adapted to the environment represented by the PDFs, which are biased with regard to the particular adaptation conditions used in the experiments. As stated above, this kind of bias does not represent a failure of the model, but the (necessarily) restricted nature of the database [D. MacLeod & Twer, 2003; D. A. MacLeod, 2003].

As convenient reference to assess the quality of the statistical results, which use no perceptual information, we also show the performance of several psychophysically-based Color Models with chromatic adaptation transform and nonlinearities in opponent channels: CIELab [Robertson, 1977], SVF [Seim & Valbereg, 1986], RLAB [M. D. Fairchild, 1996], LLab [M. R. Luo et al., 1996], and CIECAM [Moroney et al., 2002]. See [M. Fairchild, 2005]

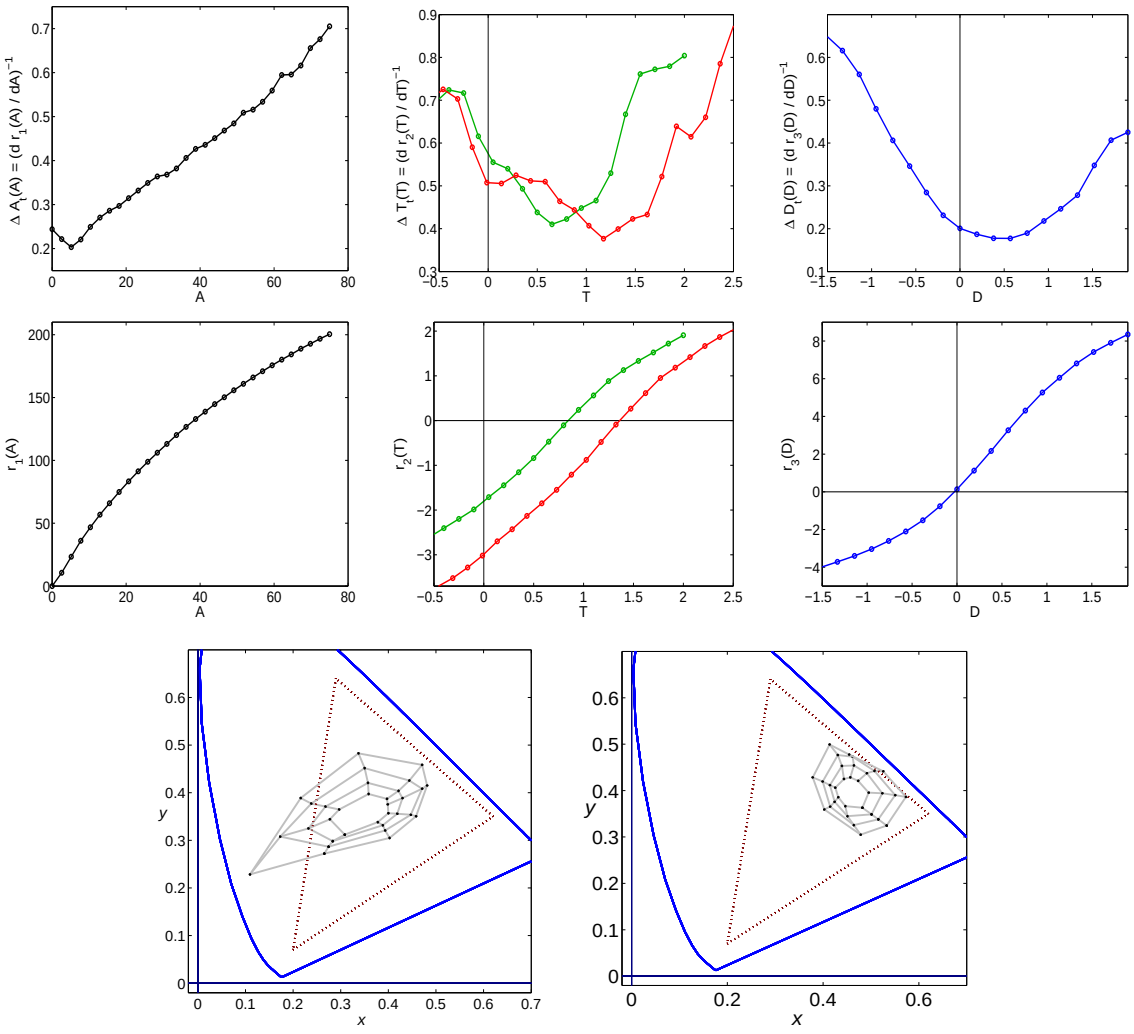


Figure 4.8: Simulation of psychophysics with SPCA using the *error minimization* strategy.

for recent collective comparison of these models. A Matlab implementation of the considered color appearance models is available on-line [Malo & Luque, 2000]. Figures 4.10, 4.11 and 4.12 show the results for CIELab, LLab and CIECAM, respectively. These results illustrate the general trend when using empirical models and stress the challenge represented by the simultaneous reproduction of nonlinearities and color adaptation data: widely used traditional models such as CIELab, LLab, RLab and SVF fail to simultaneously reproduce both aspects of the phenomenology. They reproduce either the color adaptation (as in the CIELab case) or the nonlinear behavior (as in the LLab case). Only the more recent CIECAM model is able to approximately account for both psychophysical aspects.

Interestingly, the results show that both SPCA strategies (*error minimization* and *infomax*) qualitatively reproduce the trends in both aspects of the phenomenology. SPCA

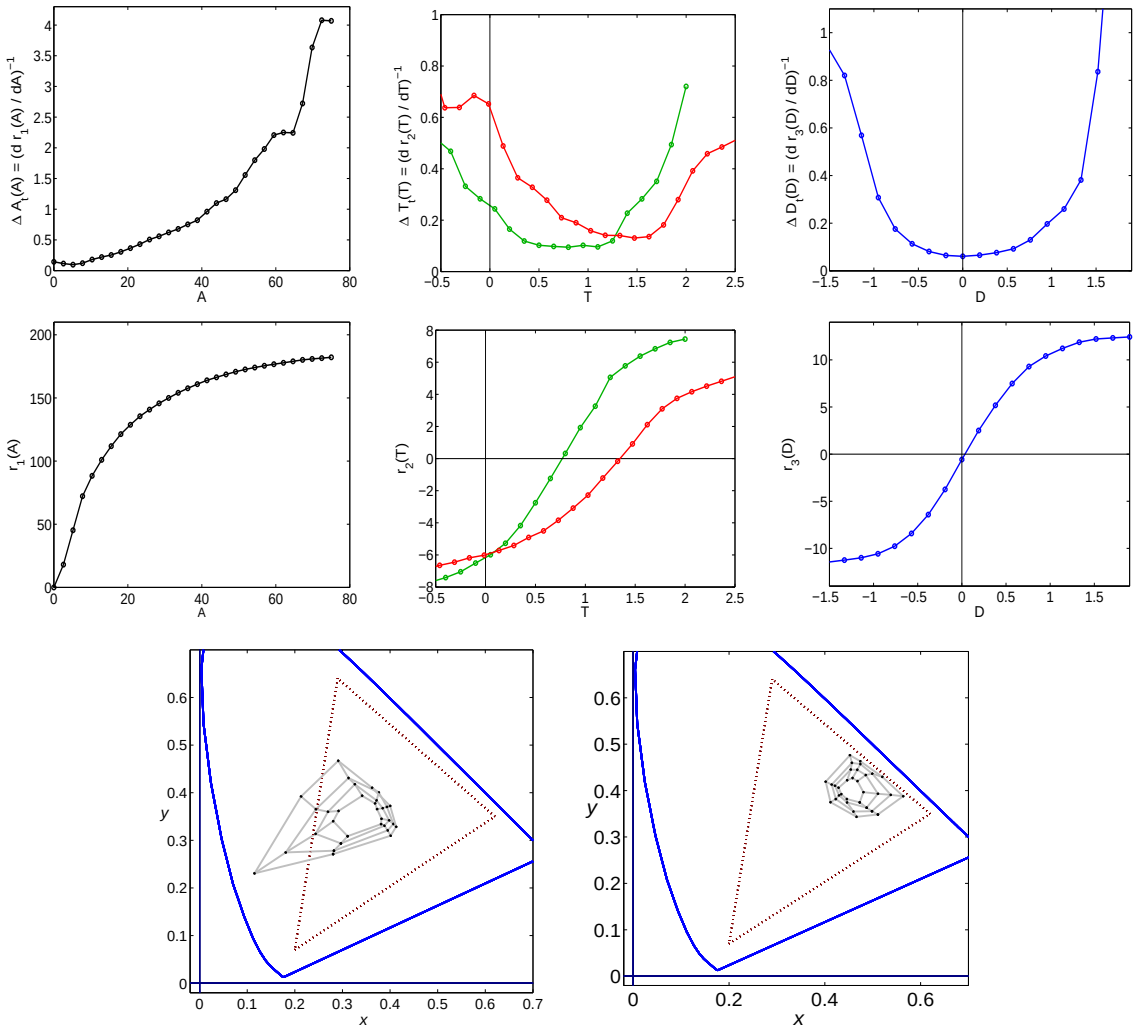


Figure 4.9: Simulation of psychophysics with SPCA using the *infomax* strategy.

gives rise to nonlinear responses in the ATD directions that shift in the appropriate way when changing the adaptation environment from white to reddish illumination. Moreover, SPCA qualitatively reproduces the shift in the corresponding colors and the orientation of chroma circles, both in the CIE D65 from CIE A data and viceversa. This general behavior comes from the fact that the proposed algorithm follows the changes in the PDFs, and has increased resolution, higher sensitivity, in the more populated regions (as illustrated in the example of Fig. 4.4). Note that the minima in the thresholds in the T and D axes (Figs. 4.8 and 4.9, top row) coincide with the corresponding maximum in each PDF (Fig. 4.6).

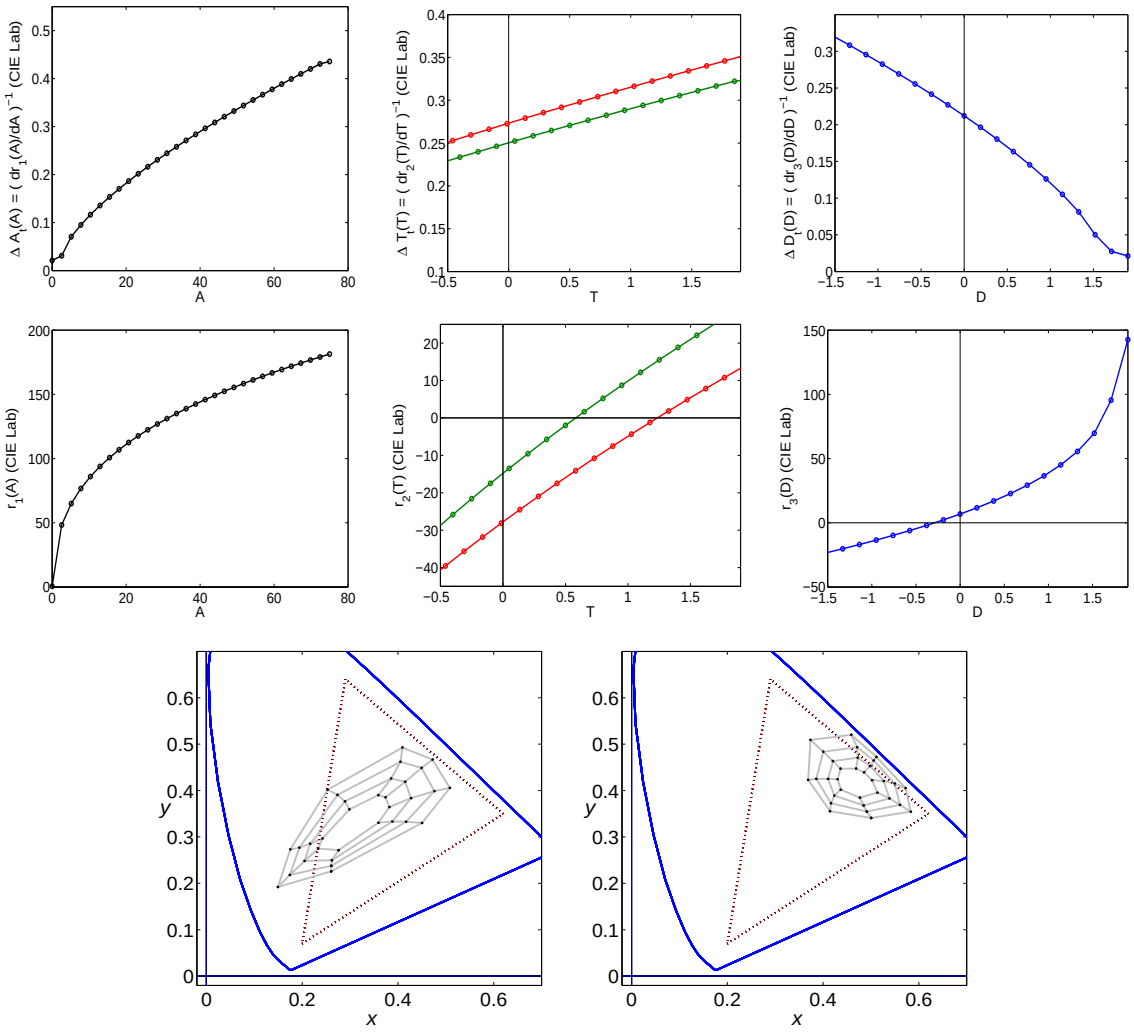


Figure 4.10: Simulation of psychophysics with CIELab color appearance model.

Discussion

The proposed statistical explanation is more general than previous statistical approaches only focused on one of the two phenomena. On the one hand, PCA-based approaches such as [Atick et al., 1993; M. Webster & Mollon, 1997] do reproduce the shift and scaling of Luo et al. corresponding colors (results not shown), but their linear nature implies that they cannot reproduce the nonlinearities in ATD. We did not check the performance of more recent linear-ICA-based approaches such as [Doi et al., 2003; Wachtler et al., 2001] in reproducing corresponding colors, but in any case, they inherently suffer from the same limitation with regard to the Weber's Law and the chromatic nonlinearities. On the other hand, Laughlin and MacLeod et al. certainly introduced strategies to account for the nonlinearities [Laughlin, 1983; D. MacLeod & Twer, 2003; D. A. MacLeod, 2003; Twer & MacLeod, 2001] but they did not explicitly propose a multidimensional transform to perform the

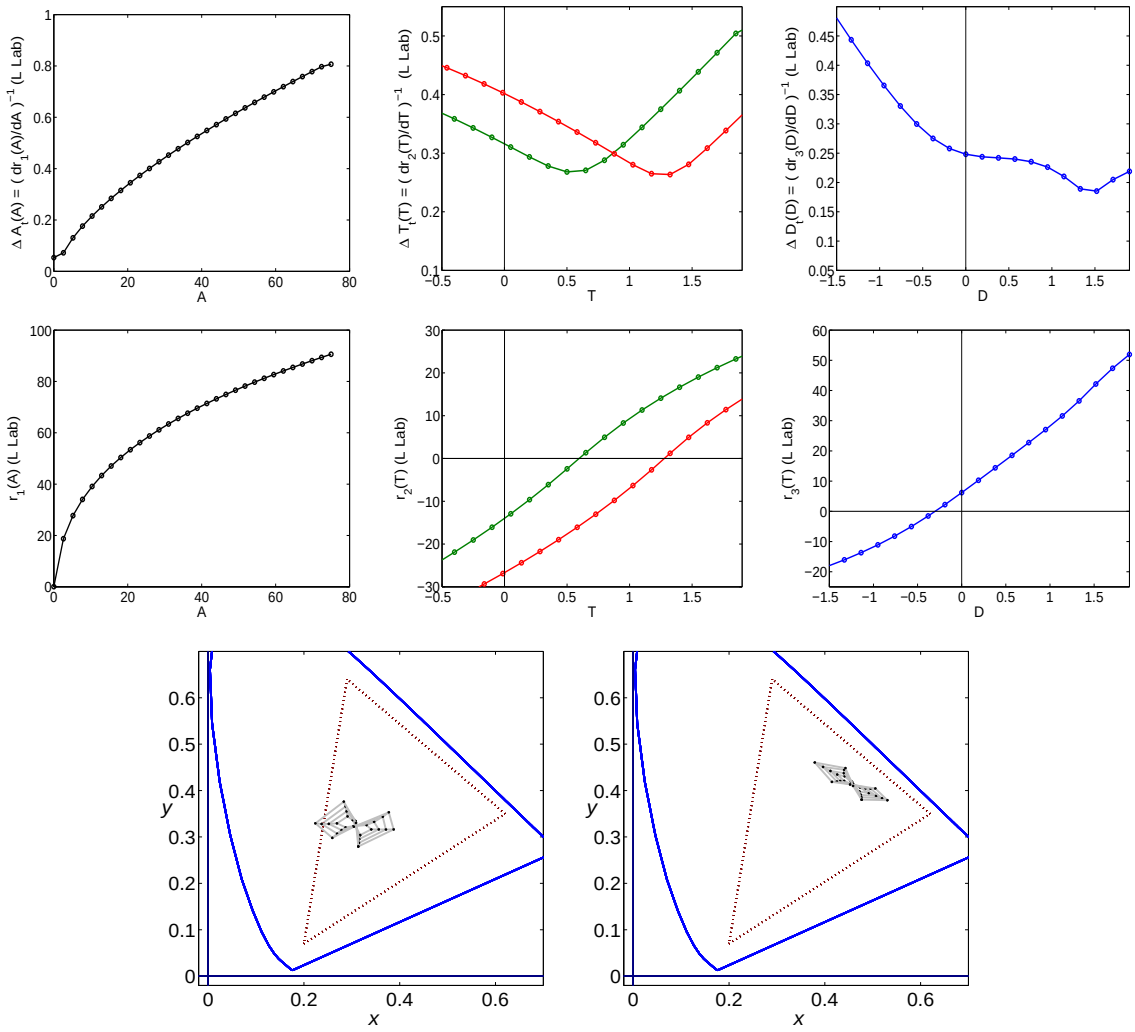


Figure 4.11: Simulation of psychophysics with LLAb color appearance model.

analysis, so their ideas cannot be used to straightforwardly derive the corresponding colors dataset.

The performance of our non-analytic technique is consistent with the general conclusions found by Abrams et al. [Abrams et al., 2007] where discrimination and color constancy are simultaneously considered. They found that analytic models based on Von-Kries adaptation, color opponent transforms and dimension-wise nonlinearities can be simultaneously optimal in discrimination and adaptation under spectral illumination changes, but not when the reflectance ensemble is substantially changed. Here we did not try to address the optimality of the proposed technique in terms of ROC as in [Abrams et al., 2007], but it is obvious by the construction of SPCA that compensation of observation conditions is not going to be possible for our technique if the objects giving rise to the different adaptation ensembles are very different from each other. Our technique needs wide enough

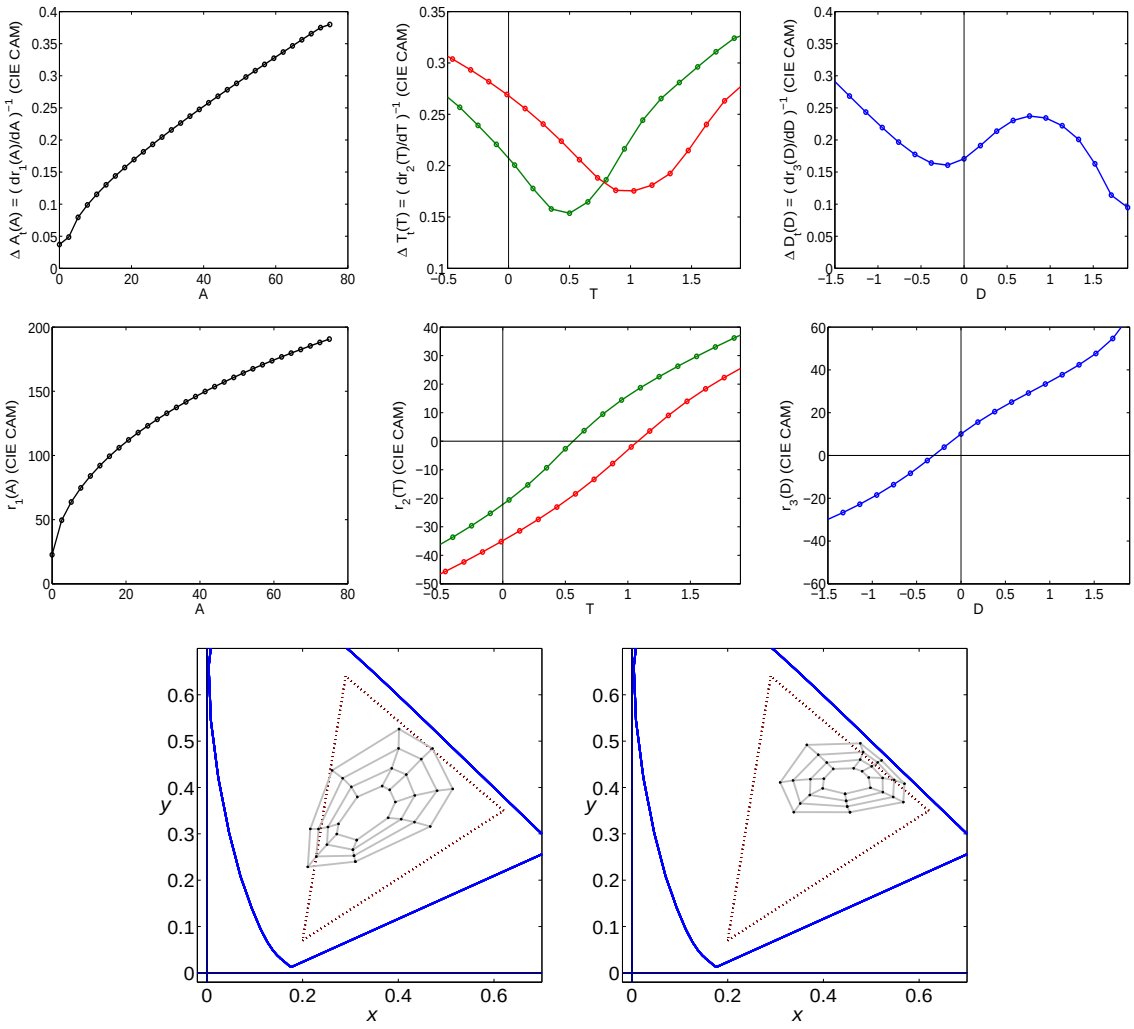


Figure 4.12: Simulation of psychophysics with CIECAM color appearance model.

reflectance ensembles for a proper adaptation: if one tristimulus manifold comes from a wide set of objects while the other comes from a (restricted) set of, say, mainly reddish objects, the manifolds do not qualitatively match, so color compensation results are not going to be accurate. In Abrams et al. terms, a different set of parameters (a different mechanism) is needed in this situation. Our results represent a data-driven alternative to Abrams et al. approach since, in our case, no analytic model is assumed in advance. Here, the nonlinear and adaptive behavior (and its limitations) strictly emerge from data, and not from a statistically fitted model with a particularly convenient functional form.

An additional issue is answering to the question of what strategy is using the brain in encoding color information at this abstraction level. In this respect, *with the considered color image database*, better agreement with the experimental data is obtained using the *error minimization* strategy. As expected from its design, the *infomax* principle gives rise to steeper

nonlinear responses, while the experimental nonlinearities (and the *error minimization* solutions) are smoother. Better reproduction of corresponding pairs is also obtained with the *error minimization* strategy. These results seem to favor the MacLeod suggestions on *error minimization* in front of the *infomax* principle.

A possible objection to such preliminary conclusion would be related to its dependence on the particular dataset. Note that the steepness of the statistically derived chromatic responses depends on the relative concentration around the achromatic axis in the considered database: if the database would be strongly biased towards achromatic objects, the higher concentration around the achromatic axis would favor the *error minimization* strategy. And the other way around for a database of highly saturated objects. Even though we subscribe the dependence of the results on the database, it does not seem that our database is particularly biased towards achromatic objects (cf. Fig. 4.5). In fact, neglecting the cluster towards saturated green (due to over representation of plants), the curved cluster visible in the CIExy diagram under D65 is quite consistent with the theoretical predictions in [Koenderink, 2010]. Therefore, the database does not seem to be specifically favoring the *error minimization* strategy. Nevertheless, given the practical impossibility of achieving a truly unbiased database [Koenderink, 2010], the definitive way to confirm these suggestions on the optimality strategy is extending the Webster and Mollon's measures [M. Webster & Mollon, 1997] performing both color discrimination and corresponding-pairs experiments in which observers are adapted to the same (controlled) statistics as the ones used in the numerical simulations. Recent experiments in color discrimination seem to follow this direction [Giesel et al., 2009; Hansen et al., 2008]. In such experiments, the analysis we proposed here could be used to obtain some insight into the question of the particular optimality criterion applied by the brain in these tasks.

4.2 Complex Independent Component Analysis of Images

In recent years, the advances in natural image statistics have been mainly in describing the signals statistics after the linear stage [Eichhorn et al., 2009; Hyvärinen & Hoyer, 2000; Hyvärinen & Köster, 2007; Lyu & Simoncelli, 2009; Malo & Laparra, 2010a; Portilla et al., 2003]. A common point of these models is that they focus on the total magnitude of the sensor (simple cell) outputs. Often, a combination of the squared outputs of simple cells is learned, leading to something like complex cells. However, there is evidence that relative magnitude, or phase, of simple cells plays an important role. A simple example about the relative importance of the magnitude and phase can be found in [Oppenheim & Lim, 1981]. In this example the magnitude and the phase in the Fourier domain of two images were exchanged, and the images which were perceptually more similar to the originals were the ones that carried the phase information. Moreover there is experimental evidence of phase coupled Gabor-like filters in V1 [Pollen & Ronner, 1981; Touryan et al., 2005]. For this reason, J. G. Daugman [1993] suggested that the receptive field in the first stage could be seen as Gabor sensors defined in the complex domain: the real and the imaginary part are essentially the same Gabor filter but with phases in quadrature.

Despite the evidences of the importance of the phase, not too much progress has been made in modeling the phase of the signals after the simple cell step statistically. The contributions in this field are restricted to models with a fixed linear stage, the wavelet transform [Cadieu, 2009; Portilla & Simoncelli, 2000]. Although this led to interesting results about the distribution of natural images, the statistics used in this kind of modeling could depend on the particular choice of using the wavelet transform as first linear stage.

Here, we aim at both modeling the phase distribution and learning the first linear stage from the data. For that purpose, we are proposing a extension of complex independent component analysis (cICA) [Bingham & Hyvärinen, 2000]. The proposed extension deals with explicit modeling the phase of non-circularly symmetric sources as an alternative to [J. Eriksson & Koivunen, 2005], which does consider non-symmetric sources but it does not model the lack of symmetry.

The present work falls naturally in two parts. In Section 4.2.1, we review cICA and point out its limitations in modeling the phase distribution. Section 4.2.2 shows how cICA can be extended to better capture the distribution of the phase variable. The extension includes the version of [Bingham & Hyvärinen, 2000] as special case. Although we focus here on natural images, the extension can be applied to all kinds of data.

4.2.1 Complex Independent Component Analysis and its limitations

As in Independent Component Analysis (ICA) for real variables, the goal in complex ICA (cICA) [Bingham & Hyvärinen, 2000] is to find a linear transformation W such that, when applied to some vector of signals x, the elements of the output vector $s = W^H x$ are statis-

tically as independent as possible. The difference to real ICA is that W, x, and hence also s are complex valued. Furthermore, instead of the transpose W^T, the transposed, complex conjugate W^H is used.

In ICA, one approach to find such a W is to first whiten the data and then to maximize the kurtosis, or a statistically more robust contrast function. In cICA, the same approach can be taken by appropriately defining whitening and choosing an appropriate contrast function.

For complex variables, the random vector x is white if both the real and imaginary parts can be defined to be white and if the real and imaginary parts are uncorrelated. An equivalent condition is that $\mathbb{E}\{xx^H\} = I$ and $\mathbb{E}\{xx^T\} = 0$. Denoting a column of W by w_i, in [Bingham & Hyvärinen, 2000], cICA can be performed by optimization of J_G,

$$J_G(W) = \sum_{i=1}^{n} \mathbb{E}\{G(|w_i^H x|^2)\}, \tag{4.21}$$

under the constraint $W^H W = I$. Depending on the nature of the sources, J_G needs to be maximized or minimized. The contrast function G must be a smooth even function and x is assumed to be white. Possible candidates include $G(y) = -\sqrt{a + y^2}$ for a small constant a. In the simulations in the next section, we will use this contrast function with $a = 0.1$. Note that the objective function depends only on the moduli $r_i = |w_i^H x|$ of the complex variable $s_i = w_i^H x$, no matter the choice of G. For sparse sources, maximization of this G leads to consistent estimators [Bingham & Hyvärinen, 2000].

An alternative viewpoint of cICA is based on maximum likelihood estimation of the statistical model $x = Ws$ where x and s are white and $W^H W = I$. Assuming independence of the sources in $s = (s_1, \ldots, s_n)$, the log-likelihood is

$$\ell(W) = \sum_t \sum_{i=1}^{n} \log p_{s_i}(w_i^H x_t), \tag{4.22}$$

where x_t is the t-th observation of x and p_{s_i} is the density of the sources s_i. Since the variables are complex valued, $p_{s_i}(s_i)$ is a bidimensional distribution that can be written as $p_{r\phi}(r_i, \phi_i)/r_i$, where r_i is the modulus and ϕ_i is the phase of s_i. Assuming further that the modulus and the phase are independent and that, importantly for the next sections, the distribution of the phase is a uniform distribution, maximization of ℓ becomes maximization of

$$J_2(W) = \sum_t \sum_{i=1}^{n} \left(\log p_r(r_{it}) - \log r_{it}\right). \tag{4.23}$$

The term p_r denotes the distribution for the moduli r_i, where we assume that all of them follow the same distribution. Replacing sample average by expectation, we obtain the objective function in Eq.(4.21) with $G(r^2) = \log p_r(r) - \log r$. Note further that the distribution p_q of the squared modulus $q = r^2$ is $p_q(q) = p_r(r)/(2r)$. This means that $G(q) = \log p_q(q) + \log 2$. Hence, the contrast function G used in cICA can be directly

related to the distribution of the squared moduli of the complex sources. In particular, we can relate the contrast function $G(q) = -\sqrt{a+q^2}$, to the choice of p_q being a Gamma distribution,

$$p_q(q) = q^{k-1}\frac{\exp\left(\frac{-q}{\theta}\right)}{\Gamma(k)\theta^k}, \tag{4.24}$$

with $k = 1$. Then, $\log p_q(q) = -q + \text{const}$, which is, up to additive constants, the same as the above contrast function when a is small.

Simulations with natural images

Here we apply cICA on natural images in the Fourier domain. The natural images are 16×16 patches extracted from the data base in [Olmos & Kingdom, 2004]. The data x on which we apply cICA are the complex Fourier coefficients. For the visualization, we show the learned W combined with the whitening matrix and the Fourier transform.

Figure 4.13 shows the results. The real and the imaginary part of the complex filters obtained are shown in pairs from left to the right. Real and imaginary parts in Figure 4.13 display a quadrature-phase relationship. This statistical result is consistent with measurements in V1 [Pollen & Ronner, 1981; Touryan et al., 2005] and related empirical models [J. G. Daugman, 1993]. Complex ICA results essentially replicate those obtained by independent subspace analysis [Hyvärinen & Hoyer, 2000], but the complex-valued formalism automatically creates two-dimensional subspaces in ordinary linear ICA.

Checking model assumptions

Here we check whether, for natural images, the obtained complex sources s_i follow the assumption in cICA that the (squared) moduli follow a Gamma distribution and the phases are uniformly distributed.

Fitting gamma distributions to the empirical distributions of the modulus of the sources leads to good fits, see Figure 4.14. In contrast, the empirical distributions of the phases do not follow the model assumptions, as shown in Figure 4.15. The clearly visible oscillations in the phases violate the assumption of uniformity in cICA. These roughly bimodal histograms may be modeled by a modified Von Mises distribution to account for the two peaks,

$$p_\phi(\phi|k,\mu) = \frac{1}{2\pi I_0(k)} e^{k\cos(2(\phi-\mu))}, \tag{4.25}$$

where $I_0(k)$ is the Bessel function of order 0. In contrast to the ordinary von Mises distribution, we have here introduced the factor 2 inside the cosine to model the two-peaked distributions seen in Fig. 4.15. Note that this distribution correspond to a uniform distribution when the parameter $k = 0$. In figure 4.15 we can see how fitting this distribution to

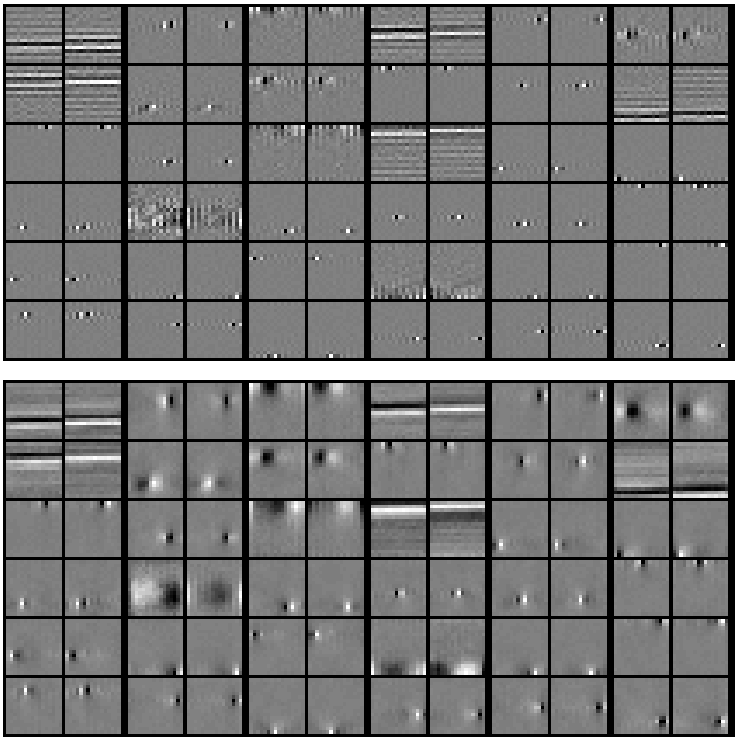


Figure 4.13: Filters and features (defined by the pseudoinverse of the filter matrix) obtained with cICA using the algorithm in [Bingham & Hyvärinen, 2000], ordered according to contrast function value (first 36 of 126). Filters and features are shown in pairs, with the real part at the left and the imaginary part at the right. Top: complex filters. Bottom: complex features.

the empirical distribution of the phase is much more precise than fitting a uniform distribution.

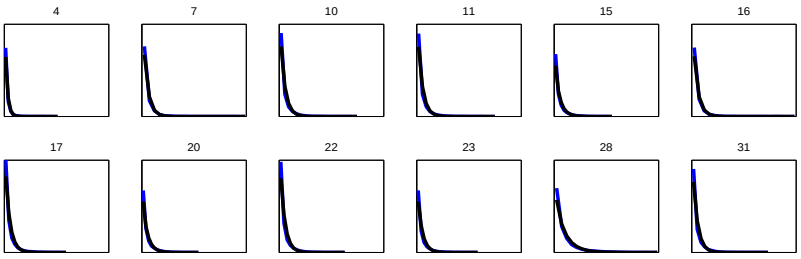


Figure 4.14: Selection of distributions of the modulus of the cICA sources (blue) and a fitted gamma distribution (black). The curves are strongly overlapping and thus not clearly visible. Numbers refer to the corresponding sensor in the figure 4.13 (left to right, top to bottom).

4.2.2 Extension of complex ICA

In this section we propose an extension of cICA. The extension builds on the maximum likelihood approach to cICA in Eq. 4.22. It will take into account that the distribution of the phase variables can be non-uniform, as found in natural images (Eq. 4.25 and Fig. 4.15).

As in the previous section, we write in Eq. 4.22 p_{s_i} as $p_{r\phi}(r_i, \phi_i)/r_i$, where r_i is the modulus and ϕ_i is the phase of s_i. Also as previously, we assume that the modulus and the phase are independent. However, instead of assuming a uniform distribution for the phases, we assume the distribution in Eq. 4.25. Since this distribution includes the uniform distribution, our extension includes the conventional cICA as a special case. With these assumptions, the maximum likelihood principle leads us to maximize the following objective function

$$J_{GQ}(W) \quad = \quad \sum_i \mathbb{E}\{G(r_i) + Q(\phi_i, k_i)\}. \tag{4.26}$$

Here, r_i is the modulus of the complex number $w_i^H x$, and ϕ_i is its phase. As before, w_i denotes a column of the matrix W and we have the constraint $W^H W = I$. The function G is, as before, related to the distribution of the squared modulus. A possible choice is $G(y) = -\sqrt{a + y^2}$. The function Q is related to the distribution of the phase and is given by $Q(\phi_i) = k_i \cos(2\phi_i)$, and depends on the w_i and the shape parameters k_i. Here, we can set $\mu = 0$ because this phase localization parameter is redundant: the phase of the oscillations will be determined by the estimated features anyway.

This modification of cICA can also be considered from an information theoretical point of view. The main goal of all ICA-based algorithms is to obtain independent sources, which is equivalent to reduce the mutual information (MI) between them. Therefore, as $MI(s_1, s_2, ..., s_n) = \sum_i \{h(r_i) + h(\phi_i)\} - h(s_1, s_2, ..., s_n)$, where $h(\cdot)$ is the entropy. This result can be derived by using the same assumptions as in section 4.2.1. Accordingly, we have to reduce the entropy of r_i and ϕ_i, (the joint entropy is invariant under unitary transforms). Note that the uniform distribution is the one with maximum entropy when the domain is bounded. Therefore, anything different to a uniform phase distribution will have less

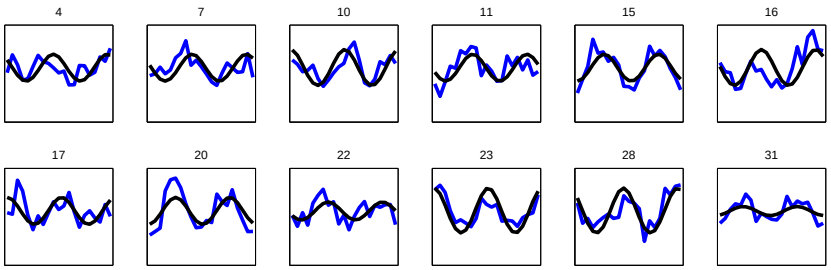


Figure 4.15: Selection of distributions of the phases of the cICA sources (blue) and a fitted modified Von Mises distribution (black). Numbers refer to the corresponding sensor in the figure 4.13 (left to right, top to bottom).

entropy, which means less MI between the variables, and hence more independent sources.

Figure 4.16 shows the results when the above extended cICA is applied to natural images (same setup as before). Note how the shape of the filters is more elongated (especially the highest-ranked ones) and spatially more extended than for the classical cICA. In figure 4.17 we can see the distribution of phases of the sources obtained with the proposed algorithm. The distributions are similar to the proposed modified Von Mises distribution.

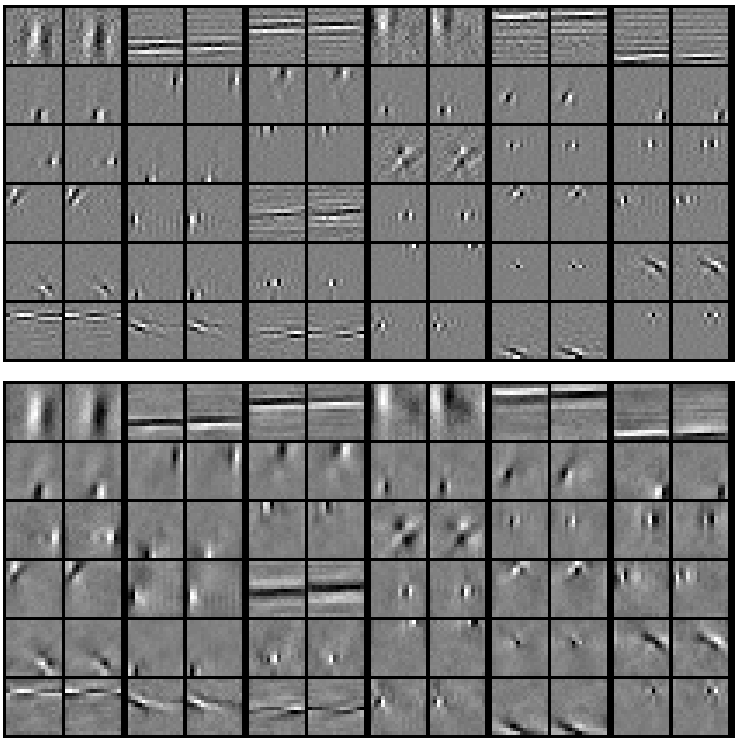


Figure 4.16: Filters and features obtained with the extended cICA, ordered according to contrast function value (first 36 of 126). Filters and features are shown in pairs with the real part at the left and the imaginary part at the right. Top: complex filters. Bottom: complex features.

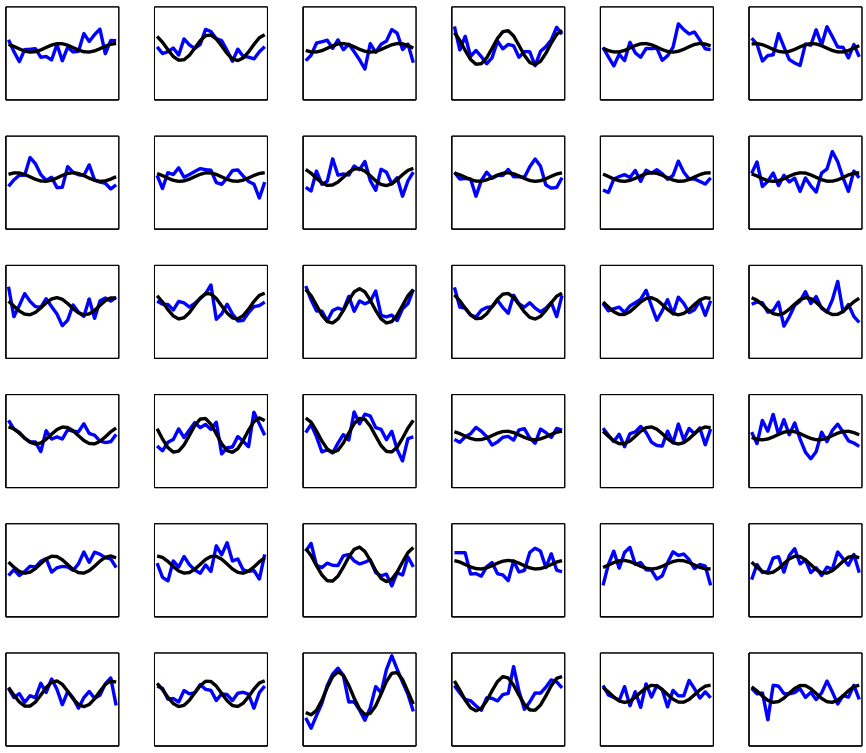


Figure 4.17: Phase distributions of the sources obtained using the modified cICA algorithm corresponding to the filters and features of the figure 4.16. Empirical distributions in blue and fitted modified Von Mises distribution in black

4.3 Ability of Linear Transforms in Removing Dependencies

Lately, a lot of works have been developed under the assumption that Independent Components Analysis (ICA) is a good tool to obtain independent components over natural images. Moreover, a number of suggestions about how the brain works are based on this ability of ICA. In [Bethge, 2006] the ability of Principal Components Analysis (PCA) and ICA in removing dependencies in natural images was analyzed. Surprisingly the ability of ICA of obtaining independent components was set only around 5% better than PCA in mutual information terms. Here we analyze this fact in detail. Specifically, since *natural images* display very different features, here we assess the effect of the linear transforms more widely used in natural image statistics, Discrete Cosine Transform (DCT), PCA and ICA, when dealing with textures of different nature. Moreover, the effect of adapting these transformations to the specific kind of image is also analyzed.

Section 4.3.1 reviews how to measure the mutual information difference after performing a linear transformation. Moreover the accuracy of the estimator is shown. The two main experiments are presented in section 4.3.2. The first experiment consists of measuring the amount of redundancy reduction achieved when the DCT, PCA and ICA are adapted to different natural textures. The second experiment consists of measuring the amount of redundancy reduction achieved when the DCT and a generic ICA basis are applied over natural textures. Generic ICA basis are trained for a wide set of natural images. This experiment remarks the importance of adapting the basis to the particular situation.

4.3.1 Measuring dependencies

Measuring multi-information is a difficult task, sometimes impossible, due to the fact that it is implicitly based on the estimation of multidimensional distributions. However, measuring the difference of multi-information before and after a transformation can be expressed in terms of unidimensional entropy measures and the Jacobian of the transformation:

$$\begin{aligned} \nabla I &= I[\vec{x}] - I[\vec{y}] \\ &= \textstyle\sum_i h[x_i] - \sum_i h[y_i] + \log|\nabla F(x)| \end{aligned} \tag{4.27}$$

where $I[\cdot]$ is the multi-information, $h[\cdot]$ is the differential entropy, $\vec{y} = F(\vec{x})$, and $\nabla F(x)$ is the Jacobian of the transformation.

If F is a linear transformation, $\nabla F(x)$ reduces to $|F|$, which can be easily computed. In our case we will restrict to $|F| = 1$. Therefore, ∇I requires only to estimate the differential entropies in each dimension before and after the transformation.

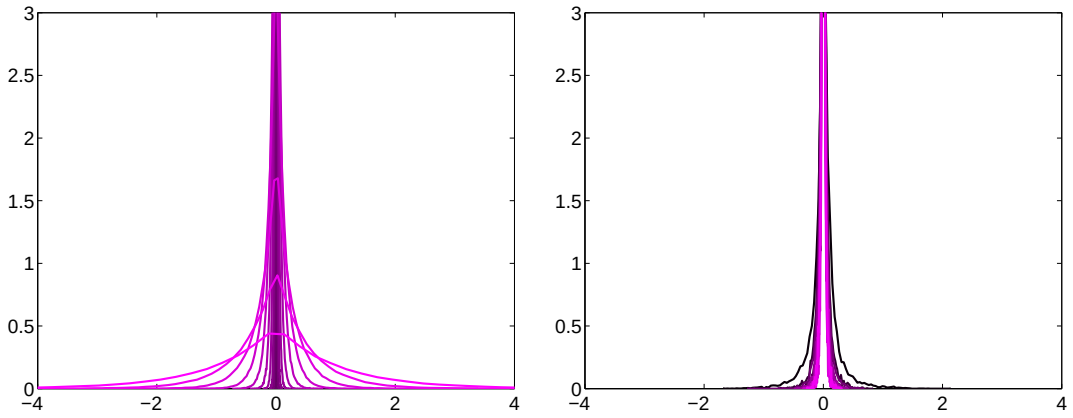


Figure 4.18: Left: example of Laplacian distribution with different λ values. Right: distribution of natural images DCT coefficients for different dimensions.

Testing the entropy estimator

In order to compute the marginal entropies we use the MLE entropy estimator with the Miller-Maddow correction [Miller, 1955]. In this section the accuracy of this entropy estimator is tested. We assume that the marginal distribution of a natural image dataset transformed with DCT, PCA or ICA closely follows a Laplacian distribution. Figure 4.18 shows theoretic Laplacian distributions with different λ values and empirical distributions for DCT coefficients. PCA and ICA coefficients have similar distributions (results not shown).

We have tested the committed error when using the entropy estimator in a random data drawn from a Laplacian distribution with different values of the λ parameter. We used 20.000 samples, which is the same number used in all the experiments of this work, and the same λ values as in the theoretic distributions of Fig. 4.18. Results shown in Fig. 4.19 suggest that the error committed using this estimator is negligible respect to the obtained multi-information differences.

4.3.2 Measuring dependencies on natural textures

In this section, we evalue the differences in multi-information when applying DCT, PCA and ICA over a set of natural textures in order to: 1) quantify the difference in redundancy reduction when the data (image) has different features, and 2) know what is the effect of applying generic linear transforms on different images. For these tasks we have selected 32 different natural textures from McGill Calibrated Color Image Database[6] (Figure 4.20). FastICA algorithm [Hyvärinen, 1999a] has been used in order to obtain the ICA basis.

[6] http://tabby.vision.McGill.ca

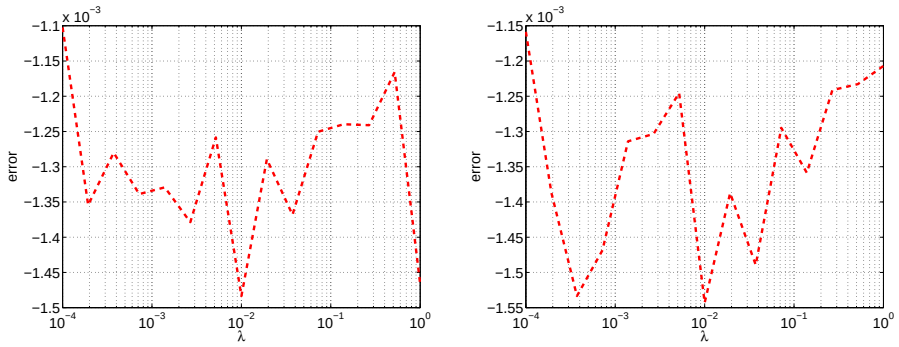


Figure 4.19: Errors of using the proposed estimator for different λ values and patch size of 7×7 (left) and 14×14 (right).

Experiment 1: Adaptive linear transforms

Here we analyze the effect on redundancy reduction when widely used linear transformations are applied to different textures. In this section PCA and ICA transforms are trained for each texture. The experiment consists of applying the following sequence to each texture:

1. 20.000 patches are randomly extracted from the current texture.
2. DCT is performed over each patch.
3. The difference in multi-information is measured using the formula (4.27).
4. The DC component (mean patch luminance) is removed, and PCA transform is computed and applied over the data.
5. The difference in multi-information is measured using the formula (4.27).
6. ICA transform is trained and applied over the data.
7. The difference in multi-information is measured using the formula (4.27).

Results for 7×7 and 14×14 patch size are shown in figure 4.21.

Experiment 2: Fixed linear transforms

Similarly to the previous section, here we analyze the effect on redundancy reduction when the most used linear transformations are applied over different textures. However, in this section we do not adapt the transforms to each texture but we use the same transformations for the whole set. We use DCT and a generic ICA transform. Note that generic PCA would obtain similar basis to the DCT. The idea is to analyze how much redundancy

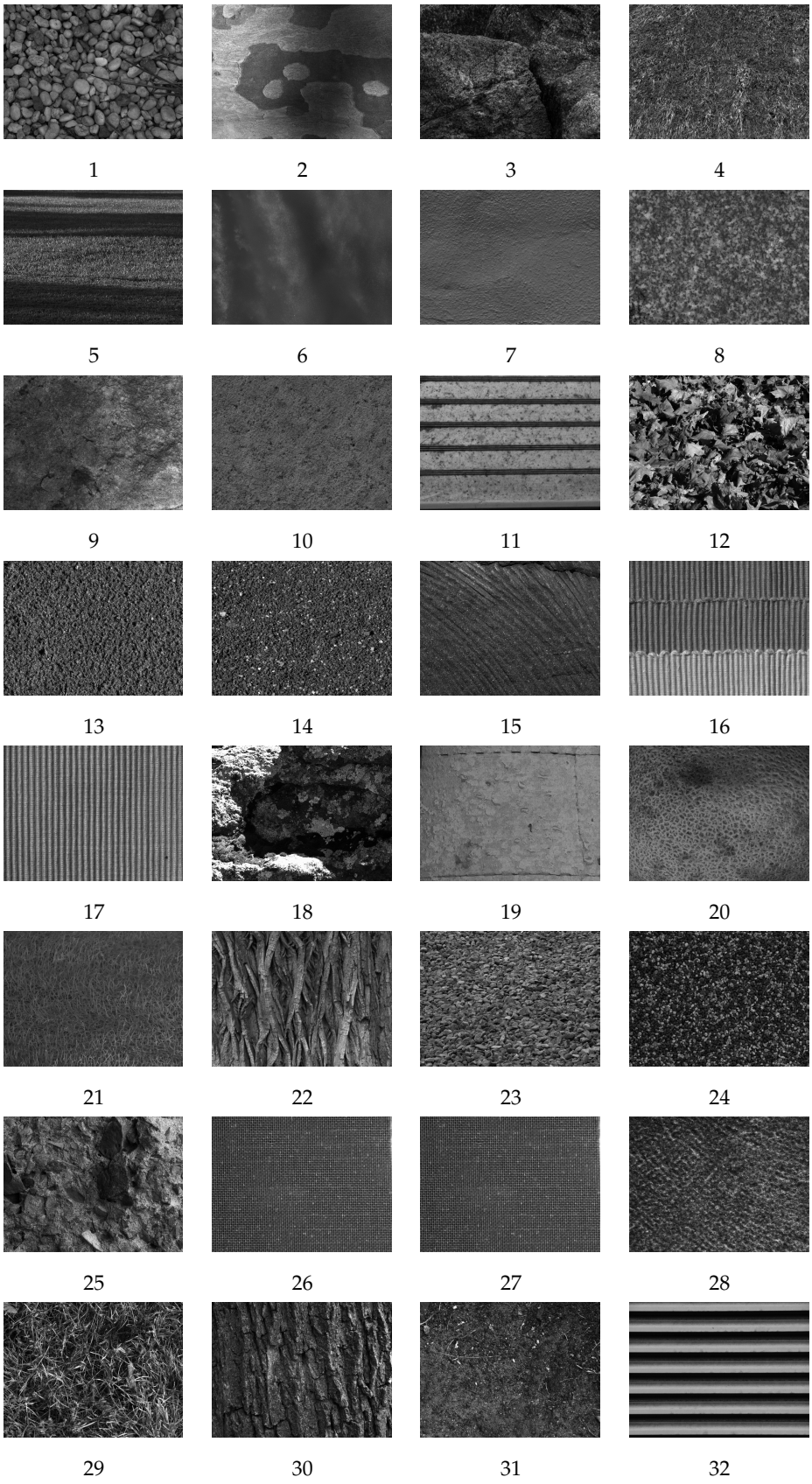


Figure 4.20: Textures used in the experiments.

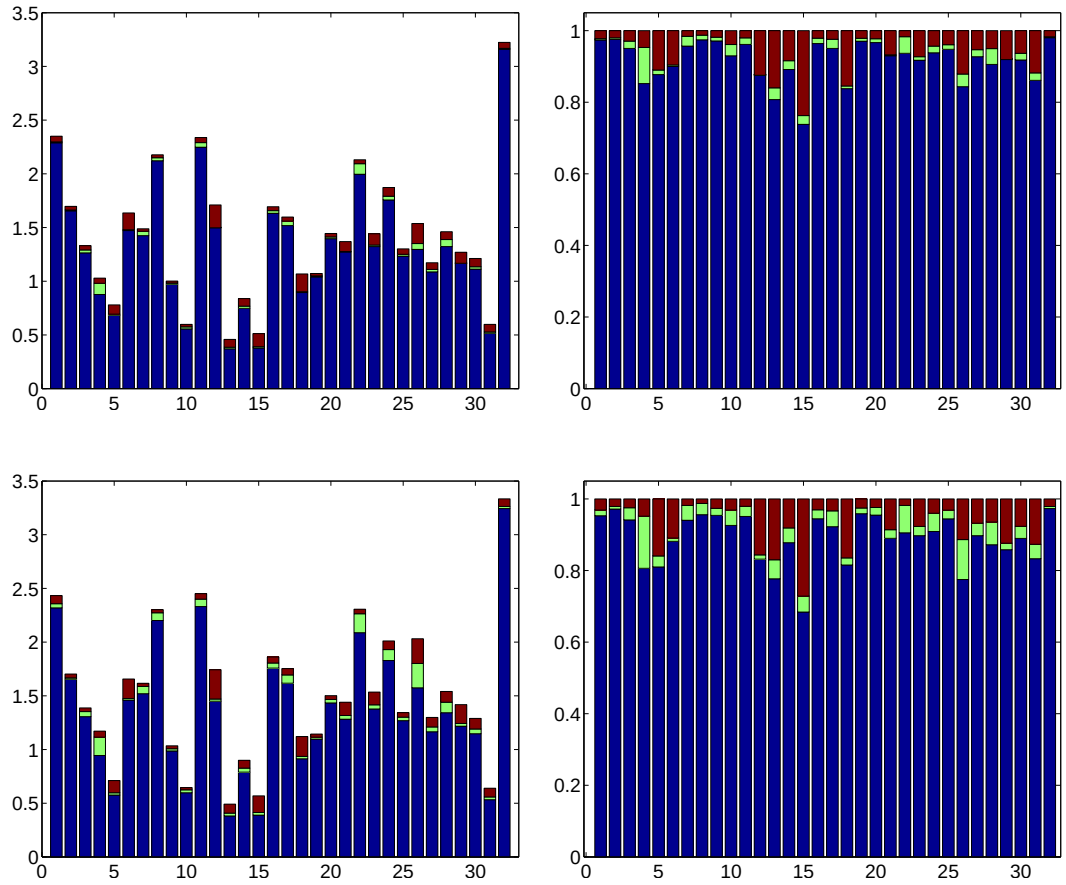


Figure 4.21: Results of multi-information reduction when PCA and ICA are adapted to each texture. Reduction of DCT is shown in blue, for PCA in green and, for ICA in red. In the left panel results are shown in bits per pixel (bpp). In the right panel results are shown relative to the reduction achieved by ICA. These results are obtained for 7×7 (top) and 14×14 (bottom) patches.

reduction can be achieved when fixed basis, optimized for processing natural images, are applied on particular images. Note that, unlike in the first experiment, here the ICA basis are not trained/adapted for each texture. However a generic ICA basis is computed from a set of 20.000 patches randomly selected from a wide set of natural images extracted from [Olmos & Kingdom, 2004]). The experiment consists of applying the following sequence to each texture:

1. 20.000 patches are randomly selected from a texture.
2. DCT is performed over each patch.
3. The difference in multi-information is measured using formula 4.27.
4. The DC component is removed, and the generic ICA transform is applied over the

data

5. The difference in multi-information is measured using formula 4.27.

Results for 7×7 and 14×14 patch size are shown in figure 4.22.

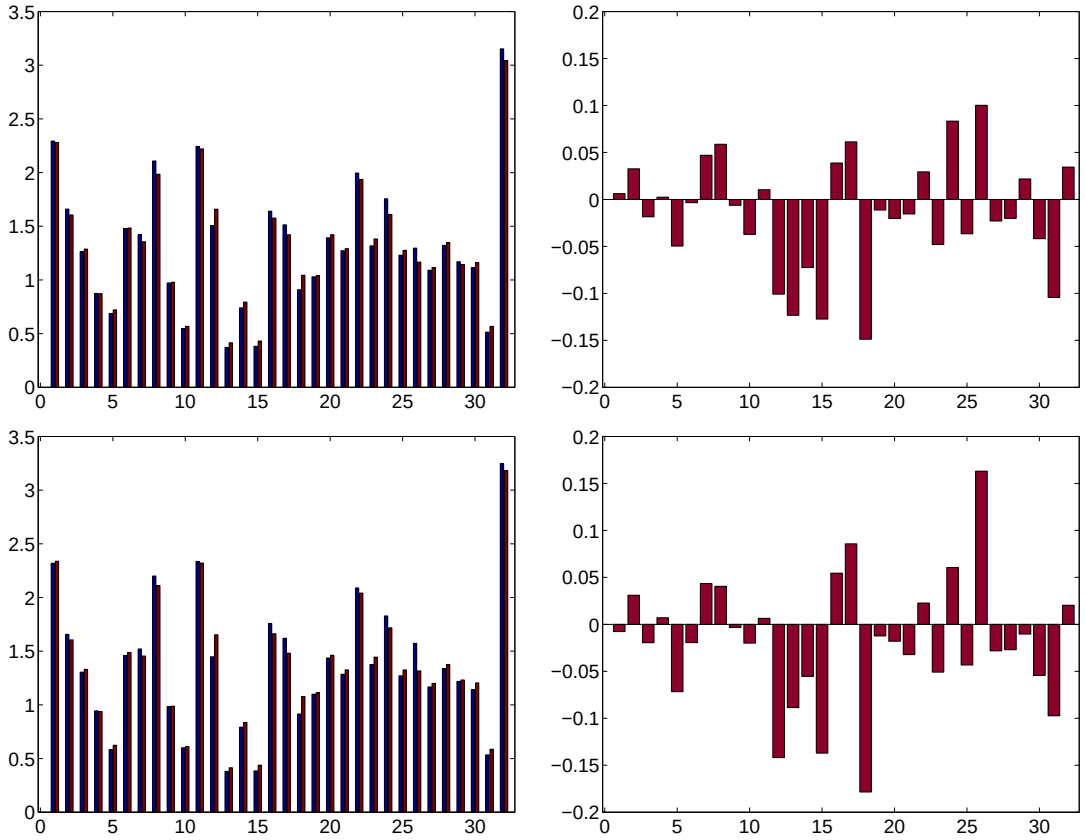


Figure 4.22: Results of multi-information reduction when using DCT (blue) and generic ICA (red). Right: Reduction achieved in bits/pixel. Note that are not shown in stacked way as in previous figures because for some textures reduction achieved by generic ICA is less than with DCT. Left: Reduction achieved by generic ICA, shown relative to the reduction achieved by DCT. These results are obtained for 7×7 (top) and 14×14 (bottom) patches.

4.4 Chapter conclusions

In section 4.1 we have shown that the basic features of color vision sensors, namely their nonlinearities, the variation of their response under change of adaptation conditions, and their ability to compensate for the changes in spectral illumination emerge from the description of the manifolds of tristimulus values of natural objects under different illuminations. To this end, we have proposed a new nonlinear manifold description technique, the

Sequential Principal Curves Analysis with local metric. SPCA is better suited to the color statistics problem than previous manifold description techniques since it is readily invertible and can be easily tuned for the *infomax* or the *error minimization* principles by simply selecting the appropriate population-dependent metric. In addition, a new accurate color image database has been collected that could be eventually used for further accurate experiments on color constancy and chromatic adaptation.

The proposed technique generalizes previous statistical explanations of color perception that just account for a subset of the data [Atick et al., 1993; Doi et al., 2003; Laughlin, 1983; D. MacLeod & Twer, 2003; D. A. MacLeod, 2003; Twer & MacLeod, 2001; Wachtler et al., 2001; M. Webster & Mollon, 1997]. Moreover, it also generalizes the results in [Abrams et al., 2007], which simultaneously analyze discrimination and adaptation, because we do not assume an explicit functional form of the model. Consistently with the results of Abrams et al. on the performance of statistically fitted analytic models [Abrams et al., 2007], the proposed non-parametric technique also requires similarity between the reflectance ensembles for an accurate color adaptation.

The simulation of perceptual results with the considered image database suggests that color vision mechanisms may be guided by an *error minimization* strategy. However, in order to confirm this conjecture, new psychophysical data are required in which the observers adaptation is determined by particular statistics. In that case, SPCA could be applied to obtain some insight about the goal used by the brain in encoding color information.

We have started with modeling natural images with complex Independent Component Analysis (cICA). This led to the emergence of complex filter where the real and the imaginary part have the same Gabor-like shape (same orientation and same frequency) but a difference in the phases of $\frac{\pi}{2}$, which are similar to the complex cells in the V1 visual area.

Checking the model assumptions in cICA, we have noticed that the assumption of uniformity of the phases is often violated for natural image data. This led us to formulate an extension of cICA which models also the phase distributions. Simulations with natural images showed that the empirical distribution of the phases provide a good match to the assumptions of the extended model.

This research has the potential for more extensions. For instance, the assumption of the independence between modulus and phase should be investigated more carefully.

In section 4.3 the potential for redundancy reduction of the most popular local linear transforms used in image processing (DCT, PCA and ICA) has been evaluated for different textures.

In the first experiment Sec. 4.3.2, where ICA basis are adapted to each texture, we found only a small average improvement of ICA over PCA or even the fixed DCT representation,

which further corroborates the results in [Bethge, 2006]. Furthermore, the absolute ΔI between DCT and ICA is always smaller than 0.5 bits/pixel for all textures, and smaller than 0.3 bits/pixel between PCA and ICA. In contrast, the redundancy reduction achieved with the fixed DCT basis is highly dependent on the type of texture, ranging from 0.4 to 3.2 bits/pixel. However, the redundancy reduction achieved by PCA and ICA with regard to DCT remains more or less constant. This effect makes that the relative redundancy reduction achieved by ICA regard DCT is very different depending on the texture. In the presented experiments for a particular texture the reduction is 30% and for other particular texture 2%. Results show that for a patch size of 14×14 the reduction achieved by PCA and ICA is bigger than for patches of 7×7.

The second experiment analyzes the redundancy reduction achieved when using generic ICA basis (trained for a set of natural images) on particular textures. The generic ICA basis yield an average improvement of about only 0.5% of ΔI for 7×7 patches, and even no reduces but increases the mutual information (3.5%) for 14×14 patch size. These results raises the question about the ability of fixed ICA basis in obtaining an independent representation. This is even clearer when analyzing particular textures: for some textures, generic ICA is even 20% worst in independence terms than DCT. It has been argued that the shape of the filters in V1 is due to its capacity to obtain a representation were the components are independents, experiments were ICA algorithms are trained using natural images data show that the ICA filters have similar shape to the V1 filters [Bell & Sejnowski, 1997; Olshausen & Field, 1996]. However, the results reported in the second experiment suggest that this idea should be revisited. On the other hand, these results together with those reported in the experiment 1 of Sec.4.3.2, stress the importance of adaptation or overcompleteness in the V1 representation. Therefore, these last results suggest that (1) the framework of linear transformations is too limited to achieve a large reduction of redundancy, even if the basis functions are adapted to visually homogeneous regions such as textures, (2) the idea that the shape of the V1 filters is due to its optimization to obtain an independent representation, and (3) adaptation or overcompleteness is necessary in order to obtain an optimal linear representation for natural images.

Chapter 5

From Statistics to Applications

5.1 Denoising with Kernels Based on Image Relations

Denoising requires representing the distorted signal in a domain where signal and noise display different enough behavior. In such a representation, noise is removed by imposing the known properties of the signal to the distorted samples. In image denoising, classical regularization techniques are used to impose smoothness in the spatial domain since noise is typically white [Banham & Katsaggelos, 1997]. Smoothness in the spatial domain means predictability of the signal from the neighborhood, and thus classical approaches exploit the low-pass behavior of the power spectrum to rely on band-limitation or autoregressive models of the signal [Andrews & Hunt, 1977; Banham & Katsaggelos, 1997; Bertero et al., 1988]. Several image denoising methods working in the spatial domain have been presented in the literature, either based on splines [Takeda et al., 2007], patch-based approximations [Kervrann & Boulanger, 2007], local auto-regressive models [Gutiérrez et al., 2006], or support vector regression [Ginneken & Mendrik, 2006; Kai Tick Chow & Lee, 2001] to perform smooth (regularized) approximations of the noisy signal. Recently, successful methods use adaptive local basis representations [K. Dabov & Egiazarian, 2007]. Approaches to the problem using local basis is qualitatively related to wavelet descriptions. In fact, wavelet representations have been recognized as quite appropriate domains for image denoising.[1]

Wavelet representations are convenient in image denoising because natural image samples have a very specific statistical behavior in this domain. On the one hand, smoothness is represented by a strong energy compaction in coarse scales. On the other hand, the combination of smooth regions with local, high contrast features, such as edges, gives rise to sparse activation of wavelet sensors. This leads to very particular, heavy-tailed,

[1] In the 2007 IEEE International Symposium on Information Theory (ISIT2007), the tutorial "Recent Trends in Denoising" (`http://www.stanford.edu/~slansel/tutorial/summary.htm`) pointed out that state-of-the-art methodologies are usually defined in the wavelet domain.

marginal probability density functions (PDFs) of the wavelet coefficients [Burt & Adelson, 1983; Field, 1987; Hyvärinen, 1999a; E. Simoncelli, 1997]. These basic features were incorporated in the classical wavelet-based image denoising techniques [Donoho & Johnstone, 1995; Figueiredo & Nowak, 2001; E. P. Simoncelli, 1999]. Classical techniques such as hard and soft thresholding [Donoho & Johnstone, 1995] have been derived by using Bayesian approaches in non-redundant wavelets, looking for either *Maximum a Posteriori* (MAP) or *Bayesian Least Squares* (BLS) estimators, in combination with simple marginal models and assuming statistical independence among coefficients [Figueiredo & Nowak, 2001; E. P. Simoncelli, 1999].

It is well-known, however, that marginal models in the wavelet domain are not enough for a proper signal characterization: relevant relations among coefficients still remain after wavelet transforms [E. P. Simoncelli, 1999]. For instance, edges lead to strong coupling between the energy of neighboring wavelet coefficients of natural images. These relations among wavelet coefficients have proven to be a key issue in applications such as image coding [Camps-Valls et al., 2008; Malo et al., 2006], texture analysis and synthesis [Portilla & Simoncelli, 2000] or image quality metrics [Laparra, Marí, & Malo, 2010]. The use of these relations is in the roots of the most recent and successful image denoising approaches as well [Goossens et al., 2009; Portilla et al., 2003; Simei & Simoncelli, 2007]. In this case, more complex image models explicitly including the relations among coefficients have to be plugged and fitted into the Bayesian framework to obtain the image estimates.

Unfortunately, all these model-based Bayesian techniques have three common problems:

1. They critically depend on the accuracy of the image model, whose definition is not trivial;

2. MAP or LS estimations can only be derived analytically for particular, typically Gaussian, noise sources. For different noise sources, the whole technique has to be reformulated which may not be analytically tractable;

3. The estimation of the parameters of the image model from the noisy observation is difficult in general.

Conversely, non-parametric approaches can include the above qualitative properties in an indirect way without the restriction of being analytically attached to particular image or noise models. These approaches are based on *learning* the underlying model directly from the image samples.

In this work we apply support vector regression (SVR) in a redundant (overcomplete) wavelet domain to the noisy image. The proposed method has the following advantages in front of the Bayesian framework:

1. It does not use a particular parametric image model to be fitted, for example, no analytical PDF is required.

2. Its solution may be found for arbitrary noise sources even without knowing the functional form of the noise PDF since it can work with just noise histograms. Therefore, the procedure does not need to be reformulated for different noise sources.

3. It is capable to take into account the relations among wavelet coefficients of natural images through the use of a suitable kernel. In this way, the method preserves the relevant relations among the coefficients of the true signal and better removes the degradation.

The proposed method does not assume independence among the signal coefficients in the wavelet domain, as opposed to [E. P. Simoncelli, 1999] and [Figueiredo & Nowak, 2001], nor an explicit model of signal relations, as done in [Portilla et al., 2003]. Therefore, the proposed machine learning approach can be seen as a more flexible (model-free) alternative to the explicit description of wavelet coefficient relations for image denoising. Even though the selection of a particular SVR may be seen as a signal parametrization, the model is still non-parametric in the sense that no functional form of the signal (or noise) characteristics (e.g., the PDF) is assumed.

Non-explicit use of dependencies in local frequency domains for denoising was also introduced in [Gutiérrez et al., 2006]. In that case, relations were embedded into a perceptual model used for non-parametric spectrum estimation, and offered better results than local parametric autoregressive models not including these relations. Here we pursue the same goal (a model-free technique including local frequency relations), but with a completely different framework (SVR instead of perceptual information). The idea of using SVR regularization in the wavelet domain for image denoising has been recently introduced in [Kai Tick Chow & Lee, 2001], [Cheng et al., 2004] and [Ginneken & Mendrik, 2006]. However, in these works, (1) the qualitative effect of the different parameters of the SVR was not analyzed, (2) these parameters were set without plausible justification of their values, and more importantly, (3) the relevance of the relations among the wavelet coefficients of the signal was not an issue, so the ability of SVR to take these relations into account in the kernel was neither assessed nor compared to other methods that do consider them. In fact, a trivial isotropic Gaussian kernel was used in all cases. On the contrary, in this work we address the key following issues:

- **Natural images features in redundant wavelet domains.** Interesting insight about the problem can be obtained by analyzing the mutual information between the coefficients of wavelet representations [Buccigrossi & Simoncelli, 1999; Liu & Moulin, 2001]. However, in redundant domains, it is strictly necessary to discern what are the relations specific to the signal and those due to the transform.

- **General constraints of the SVR parameters in image denoising.** Generic recommendations about the SVR parameters have been adapted to propose specific subband-dependent profiles for the insensitivity and the penalization parameters, and to propose a mutual information based kernel.
- **Effect of the SVR parameters.** We show the qualitative effect of varying the values of the parameters under the constrained parameter space.
- **Procedure to optimize the SVR parameters.** We propose an automatic procedure to select the SVR parameters based on the Kullback-Leibler divergence, under certain assumptions on signal and noise.

Even though this methodological framework is proposed in the context of achromatic image denoising, it can be readily extended to other denoising problems in which wavelet coefficients exhibit particular relations, such as in color or multispectral images, speech signals, etc.

The remainder is outlined as follows. In Section 5.1.1, we point out relevant signal features in redundant wavelet domains through mutual information measurements. These key properties will be used by the proposed algorithm presented in Section 5.1.2. In Section 5.1.3, the effect of SVR parameters and the validity of the proposed criterion for its selection is addressed experimentally. Section 5.1.6 shows the performance of the proposed method compared to standard denoising methods in the wavelet domain. Several experiments dealing with different amount and nature of noise illustrate the capabilities of our proposal.

5.1.1 Features of natural images in the Steerable Wavelet Domain

The starting hypothesis for image denoising is that signal and noise display different characteristics and thus it is possible to separate them in a certain domain. Natural images show non-trivial relationships among wavelet transform coefficients. In the following, we review the reported statistical properties of natural images in orthogonal wavelet domains, and then analyze them in the redundant steerable wavelet domain selected in our implementation. Specifically we will use mutual information (MI) to assess the statistical relations among wavelet coefficients of natural images as in [Buccigrossi & Simoncelli, 1999] and [Liu & Moulin, 2001].

Intraband versus interband signal relations in Orthogonal Wavelets

Dependencies among *orthogonal wavelet* coefficients were measured using mutual information in [Liu & Moulin, 2001]. The dependencies were studied at interband and intraband levels, and the results suggested that the mutual information between intraband neighbors is typically *larger* than the interband relations for several models and types of interaction.

In [Buccigrossi & Simoncelli, 1999], the authors analyzed the linear predictability of a coefficient's magnitude from a conditioning coefficient set, either its parent, neighbors (left and upper), cousins (coefficients at the same location but in different orientation subbands), or aunts (cousins of the parent). After an exhaustive mutual information analysis, the parent provided less information content than the neighbors. These evidences suggest that the dependencies among spatial neighboring coefficients (intraband) in orthogonal wavelet descriptions are stronger than the interband dependencies.

Natural images relations in Steerable Wavelets

Redundant wavelet representations may be more suited to image denoising applications since is easier to describe the invariant statistical image features. Specifically, some representations are designed to be translation or rotation invariant [Coifman & Donoho, 1995; Freeman & Adelson, 1991; Kingsbury, 2006]. This behavior is convenient to ensure that a particular feature in different spatial regions (or with different orientations) gives rise to the same neighboring relations. Some translation invariant wavelets have also a smoother rotation behavior than non-redundant transforms [E. Simoncelli & Freeman, 1995]. This justifies applying the same processing all over a particular subband and dealing with the different orientations in similar ways. Besides, this prevents aliasing artifacts appearing in critically-sampled wavelets. In this work we choose a redundant steerable pyramid representation [E. Simoncelli & Freeman, 1995] to take advantage of these properties.

Despite the reported results on the relations of signal coefficients in orthogonal transforms, a number of questions have to be answered in the case of redundant representations, and in particular, in the steerable wavelet domain:

1. How relevant are the relations among coefficients of natural images in this domain?

2. How relatively important are interorientation, interscale and intraband signal relations?

3. How is the spatial arrangement of these signal relations?

The first question is particularly important since, even though the steerable transform may intensify the relations among signal coefficients, its redundant nature may also introduce relations which could be due to the transform but not to the signal. The second question allows us to focus on the most significant relations. Answering the third question is crucial to design suitable kernels for image denoising.

In the following, we get some insight on these concerns by performing two experiments on a representative database of 920 achromatic images of size 256×256 extracted from the McGill Calibrated Colour Image Database [Olmos & Kingdom, 2004].

Signal relations are specific to the signal

In our first test, following [Liu & Moulin, 2001], we computed the mutual information among steerable wavelet coefficients of the data set for different spatial, orientation, and scale distances. We used a steerable pyramid with 8 orientations and 4 scales. The mutual information was estimated from the uniformly binned empirical data (256 bins) by computing the histogram of all available sample pairs (721280 samples) for the three considered neighborhoods. In addition, as stated above, in redundant domains it is necessary to know whether these relations come from the images or they are due to the transform. Note that, considering i.i.d. signals, any relation among the coefficients after a linear transform will be due to the transform no matter their PDF in the original domain. Therefore, in order to assess the amount of relations due to the transform, we compared the MI among natural images coefficients, and the MI among the coefficients of an i.i.d. signal (Fig. 5.1). The relations displayed by i.i.d. signals in the transformed domain may be seen as a lower bound for the mutual information of signal coefficients. From Fig. 5.1, it can be noticed that, in every case, relations found in natural images are bigger than those introduced by the transform.

Intraband signal relations dominate over interscale or orientation

Besides, the results show that intraband relations in the signal are also more important than interorientation or interscale relations. Note that mutual information measures are defined to depend on logarithms of probability so that comparisons have to be done by subtraction, not by division. Beyond consistency with previously reported results for orthogonal wavelet transforms [Buccigrossi & Simoncelli, 1999; Liu & Moulin, 2001], it has been observed that the relations are specific to the signal and not just induced by the transform.

Intraband relations are strongly oriented

In our second test, we studied the spatial arrangement of the relations among intraband coefficients since they display the most relevant relations. To this end, we computed the mutual information in a 2D 5×5 neighborhood for the different orientations and scales. Figure 5.2[top] shows the above mentioned results for the set of natural images (finest scale). We also provide the relations introduced by the transform (i.i.d. signal, Fig. 5.2 [bottom]). Similar results were obtained for the other (coarser) scales. Again, the relations among the signal coefficients are higher than those introduced by the transform. Another key issue observed in Fig. 5.2 [top] is the specific spatial arrangement of these relations: the presence of oriented structures in natural images gives rise to strong anisotropic intraband relations in the different subbands. Coefficients following these relations are expected to

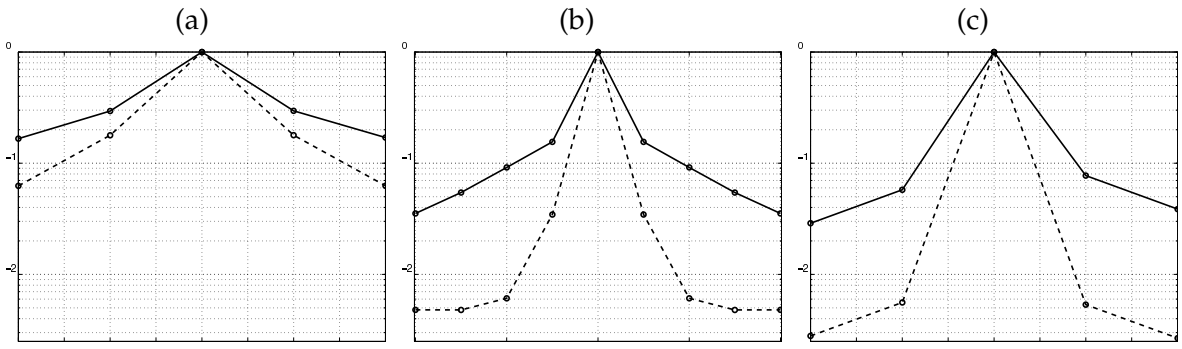


Figure 5.1: Comparison between redundancy of natural image coefficients in the steerable wavelet representation (solid), and the redundancy due to this representation (dashed). Redundancy is measured in terms of relative mutual information in logarithmic scale among (a) spatial (b) orientation and (c) scale neighbors.

be representative of natural features. These mutual information results match recently reported results on autocorrelation of intraband wavelet coefficients [Goossens et al., 2009]. The results obtained in these experiments will be further used in Section 4 to design specific kernels that take into account the *observed* natural image relations.

Summarizing, natural images have singular features in the steerable wavelet domain (Figs. 5.1 and 5.2): given a distorted image, enforcing these singular oriented relations among coefficients in every subband (with the appropriate kernels) will eventually preserve the natural signal relations and remove the noise. Of course, the bigger the difference between the shape of the intraband relations in signal and noise the better the results are expected to be.

5.1.2 Restoring Wavelet relations with SVR

The effect of noise in the wavelet domain is introducing artificial deviations from the original signal and hiding the natural relations among the coefficients (see an illustrative example in Fig. 3). In the more general case, the degraded observation, $\mathbf{i_d}$, can be written as the result of the addition of a certain realization of noise, $\mathbf{n}$, to the original signal, $\mathbf{i}$:

$$\mathbf{i_d} = \mathbf{i} + \mathbf{n} \tag{5.1}$$

Note that this (convenient) way to state the problem does not necessarily mean that the physical degradation has to be additive. In fact, the nature of the degradation should ideally be expressed through a probabilistic noise model that may depend on the original signal, $p(\mathbf{n}|\mathbf{i})$. The other desirable piece of information is a probabilistic model of the

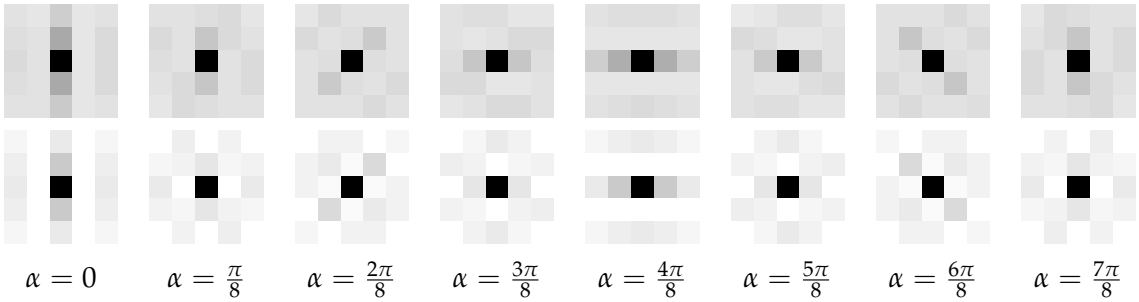


Figure 5.2: Mutual information among the central coefficient and its spatial neighbors in the same subband (intraband) in the steerable wavelet domain. Darker gray values indicate higher mutual information. Top row shows the results for the different orientations of the finest scale of the natural image database, and bottom row shows the equivalent results for Gaussian noise.

signal, $p(\mathbf{i})$. However, in most practical situations, the complete probabilistic description of the problem, that is, having $p(\mathbf{i})$ and $p(\mathbf{n}|\mathbf{i})$, is not available in analytical form.

In order to avoid this lack of information, we propose to use the regularization ability of SVRs. In this section, first we review the capabilities of the SVR for signal approximation. Afterwards, general constraints to the SVR parameter space are given for the particular problem of natural image denoising. Finally, we present an automatic procedure to choose the appropriate SVR parameters (from the above restricted space) to be used for any combination of image and noise.

Capabilities of SVR for signal estimation

Throughout this work, a wavelet transform, matrix T, is applied to the observed image, leading to a set of (noisy) coefficients, $\mathbf{y} = T \cdot \mathbf{i_d}$. The original set of wavelet coefficients, $\mathbf{x} = T \cdot \mathbf{i}$, has to be estimated from the distorted observation, $\mathbf{y}$. Due to the observed strong intraband relations, we will use the SVR in the wavelet domain in patches inside each subband. Subbands are decomposed into non-overlapping 16×16 patches, leading to sets of $N = 256$ samples. Now, given input-output pairs $\{p_i, y_i\}_{i=1}^{N}$, where p_i are the wavelet indices and y_i are the corresponding noisy wavelet coefficients in a patch, we train the *adaptive* SVR [Camps-Valls et al., 2001; Gómez et al., 2005; Navia-Vázquez et al., 2001] to approximate the signal.

Let $\boldsymbol{\phi}$ be a non-linear mapping to a higher dimensional feature space, then the adaptive SVR computes the weights $\mathbf{w}$ to obtain the estimation, $\hat{x}_i = \boldsymbol{\phi}^\top(p_i)\mathbf{w}$, by minimizing the following regularized functional:

$$\|\mathbf{w}\|^2 + \sum_i C_i \, \xi_i \tag{5.2}$$

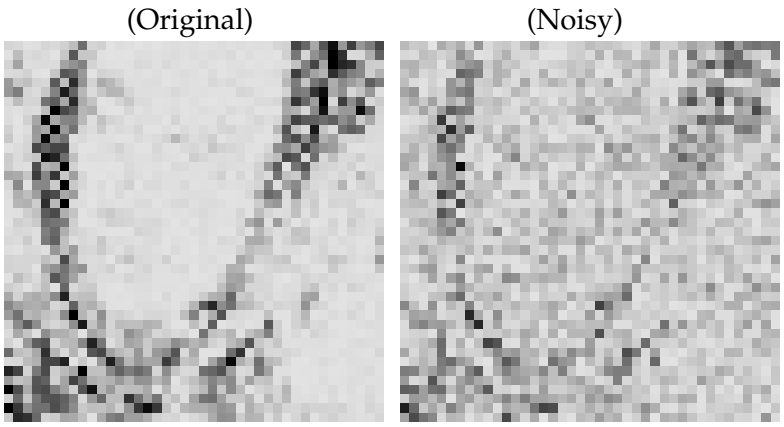


Figure 5.3: Effect of noise on the wavelet coefficients. Patch of a subband of a wavelet representation of the original image Barbara (left) and its noisy version (right). Darker values indicate higher amplitudes.

subject to $|y_i - \boldsymbol{\phi}^\top(p_i)\mathbf{w}| \leq \varepsilon_i + \xi_i, \forall i = 1, \ldots, N$, where ξ_i are the magnitude of the deviations of the estimated signal from the observed noisy data outside the (sample-dependent) insensitivity zones ε_i. Sample-dependent penalization parameters, C_i, tune the trade-off between fitting the model to the observed noisy data (minimizing the deviations) and keeping model weights $\|\mathbf{w}\|$ small (enforcing flatness in the feature space).

This adaptive SVR differs from the standard formulation [Smola & Schölkopf, 2004], in two aspects: (1) the loss function given by (ε_i, C_i) is sample-dependent, which is convenient in wavelet domains where signal and noise variances strongly depend on the subband, and (2) the usual bias term in SVM formulations has been intentionally dropped to account for the fact that the expected value of wavelet coefficients is zero. The appropriate design of C_i and ε_i profiles is analyzed in Section 5.1.3.

Explicitly working with the non-linearity $\boldsymbol{\phi}$ is no longer necessary since the whole formulation can be expressed in the form of dot products of the mapping functions called *kernels*, $K(p_i, p_j) = \boldsymbol{\phi}(p_i)^\top \boldsymbol{\phi}(p_j)$. In this case, the estimation is given by $\hat{\mathbf{x}} = K \cdot \boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is the dual representation of weights $\mathbf{w}$ [Smola & Schölkopf, 2004]. The kernel matrix can be seen as a similarity matrix between samples (or coefficients), and should reflect the relations between them. Many kernel functions have been proposed in the literature [Smola & Schölkopf, 2004]. In the image denoising case in wavelet domains, we focus on the basic structure of the generalized Radial Basis Functions (RBF) kernel since the relationship among the wavelet coefficients corresponding to spatial neighbors within a subband is local. However, as it will be analyzed in Section 5.1.3, the kernel will be adapted to incorporate the anisotropic signal relations studied in Section 5.1.1, see Fig. 5.2.

5.1.3 General constraints on SVR parameter space in image denoising

As stated above, SVR signal approximation will depend on the penalization parameters, C_i, the insensitivities, ε_i, and the kernel K. In the following, we restrict the range of possible values of these parameters, $\theta = (C_i, \varepsilon_i, K)$, in the particular case of image denoising in wavelet domains:

Penalization factor. In general, the penalization factor of SVRs should be related to the standard deviation of the signal [Cherkassky, 2004]. In the denoising problem considered here, the signal variance substantially differs in each wavelet scale. According to this, it is strictly necessary to set a different penalization factor *per* scale, $C_i = C\,k_i$, where k_i is a scale-dependent profile. This profile k_i was obtained by averaging the standard deviation of wavelet coefficients over 100 images from the database used in Section 5.1.1. This profile was multiplied by a factor, C, varied in the range $[10, 10^4]$, which did not show a strong impact on the results provided a sufficiently large value. This is consistent with the suggestions reported in [Chalimourda et al., 2004] in a more general context. Note that, for instance, in the examples of the next section (Fig. 5.3), indistinguishable results are obtained for a large enough C. In our experiments, we found that a reasonable prescription for the global factor on the penalization profile is $C \approx 10^3$.

Adaptive insensitivity zone. In general, the insensitivity has to be related to the standard deviation of the noise [Kwok & Tsang, 2003]. In transformed domains, the effect of the transform has to be taken into account in order to estimate the corresponding standard deviation. In redundant wavelet representations, this standard deviation is coefficient dependent. Thus it is strictly necessary to introduce a subband-dependent ε_i profile [Camps-Valls et al., 2001; Gómez et al., 2005]. The transformed standard deviations can be estimated either (1) empirically from noise samples, or (2) computed from the noise covariance matrix if it is known. In the empirical case, noise samples can be experimentally obtained by applying the noise source to a large enough set of images, and writing the noise as in Eq. 5.1. In our experiments, we used the natural image database used in Section 5.1.1, and we obtained fairly stable results for the profile by considering 100 images. In the case that the noise covariance is known, the corresponding matrix in the selected wavelet domain can be obtained from the noise covariance matrix in the spatial domain, Σ_n, and the transform T [Stark & Woods, 1994]. Therefore, the insensitivity profile can be computed as:

$$\varepsilon_i = \tau\; diag(T \cdot \Sigma_n \cdot T^\top)_i^{1/2} \tag{5.3}$$

In the case of white noise, $\Sigma_n = \sigma_n^2 \cdot I$, and thus Eq. (5.3) reduces to:

$$\varepsilon_i = \tau\; \sigma_n\; diag(T \cdot T^\top)_i^{1/2}, \tag{5.4}$$

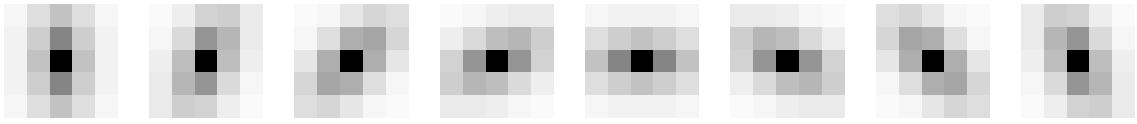


Figure 5.4: Anisotropic kernel functions used in the support vector regression method for the eight considered orientation subbands.

where σ_n^2 is the noise variance in the spatial domain, and τ is a scaling factor to be adapted for each particular image and noise combination. The scaling factor, τ, should be in the range [0.5, 3] according to the known relationship between the ε-insensitivity zone and the noise standard deviation [Kwok & Tsang, 2003]. Note that (5.3) may cope with colored noise. Considering the off-diagonal elements of the covariance matrix (neglected in (5.3) and (5.4)) would give rise to coupling ε-insensitivities among samples. This issue has been already considered and solved in the context of image coding by using an additional non-linear transform and a constant ε in the transformed domain [Camps-Valls et al., 2008]. However, here we restrict ourselves to the approximated diagonal case.

Including signal relations in the kernel. In the kernel methods literature, the use of *prior* knowledge about the problem can be encoded through bagged, cluster, or probabilistic kernels [Jebara et al., 2004; Weston et al., 2004]. In our case, we propose to take into account image coefficient relations by analyzing a large (representative) database and taking the (oriented) mutual information among samples as core distance measure. However, using these empirical measures to set the kernels is not straightforward since the kernels have to fulfill Mercer's Theorem [Mercer, 1905]. According to this, we propose to use generalized Gaussian kernels. In particular, we fitted anisotropic Laplacian kernels to the MI measures to consider the intraband oriented relations within each subband:

$$K_\alpha(p_i, p_j) \quad = \quad \exp\big(-((p_i - p_j)^\top G(\alpha)^\top \Sigma^{-1} G(\alpha)(p_i - p_j))^{1/2}\big), \tag{5.5}$$

where $\Sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$, σ_1 and σ_2 are the widths of the kernels, $p_i \in \mathbb{R}^2$ denotes the spatial position of coefficient y_i within a subband, and $G(\alpha)$ is the 2D rotation matrix with rotation angle, α, corresponding to the orientation of each subband (see Fig. 5.4). Note that these set of oriented kernels describe the signal relationships that emerge from experiments in Section 5.1.1 (cf. Fig. 5.2[top]).

We obtained proper values for the widths σ_1 and σ_2 by fitting the above kernel to the MI measures among coefficients described in Section 5.1.1 ($\sigma_1 = 2\sigma_2$, and $\sigma_1 = 4.8$

in spatial coefficient units). The kernel was further normalized in the standard way. Since this width comes from direct measures from images, it describes a fundamental property of natural images so it can be kept fairly constant.

The conclusion of this section is that in the case of image denoising in wavelet domains, an appropriate analysis of the signal variance, the noise variance, and the relations among the wavelet coefficients of the signal can be used to strongly reduce the dimensionality of the SVR parameter space. After this analysis, the only SVR parameter that remains fixed is the global scaling, τ, to be applied to the insensitivity profile.

5.1.4 Procedure for automatic SVR selection

In the more general case, applying SVRs with a given set of parameters, $\grave{}$, to a noisy image leads to a certain image estimate, $\hat{\mathbf{i}}_\theta = T^{-1} \cdot \hat{\mathbf{x}}_\theta$. From this image estimate, and the convenient additive notation for the noise (Eq. (5.1)), a noise estimate can be obtained: $\hat{\mathbf{n}}_\theta = \mathbf{i_d} - \hat{\mathbf{i}}_\theta$. In this section we propose a procedure to select the SVR parameters, θ, that better approximates the noise free image, using the available information.

In the more general situation the only available information is the noisy image. However, as stated above, denoising methods usually assume that additional probabilistic information on the signal and noise is available: $p(\mathbf{i})$ and $p(\mathbf{n}|\mathbf{i})$. Note that this knowledge is equivalent to the knowledge of the joint signal and noise distribution since $p(\mathbf{i},\mathbf{n}) = p(\mathbf{n}|\mathbf{i})\, p(\mathbf{i})$.

Let us momentarily assume that this information is available to propose the general procedure to set the SVR parameters. Afterwards, we will relax the requirements by considering an approximation that can be easily applied in practical situations.

In order to enforce solutions that closely follow the (assumed to be known) statistics of signal and noise, we propose to select the SVR that minimizes the k-th order Kullback-Leibler divergence (KLD) [Cover & Tomas, 1991] between the joint PDF of signal and noise, and the joint PDF of the estimated signal and the estimated noise:

$$\theta^* = \arg \min_\theta \left\{ D_{KL}\left[\, p(\hat{\mathbf{i}}_\theta, \hat{\mathbf{n}}_\theta) \;\|\; p(\mathbf{i},\mathbf{n}) \,\right] \right\} \tag{5.6}$$

The underlying idea is that the SVR that minimizes the divergence between the above PDFs is the one that better captures the features of the true signal and better removes the degradation.

Although in ideal situations the application of this procedure would obtain the best results in statistical terms, in practical situations the full probabilistic description of the problem is not available. A number of approximations are done in practical situations. For instance, thermal noise in CCD cameras is not independent from the input signal since diffusion increase with the irradiance. However, thermal noise is usually assumed to be

independent of the input signal. Additional assumptions as additivity or certain analytical marginal PDF of the noise are also widely used.

In our case, we assume independence between signal and noise:

$$p(\mathbf{i}, \mathbf{n}) = p(\mathbf{i})\, p(\mathbf{n}) \tag{5.7}$$

However, no analytical model for these PDFs is assumed. Under this independence assumption, it is easy to see that eq. 5.6 reduces to:

$$\theta^* = \arg \min_{\theta} \left\{ D_{KL}\left[\, p(\hat{\mathbf{i}}_\theta) \parallel p(\mathbf{i}) \,\right] + D_{KL}\left[\, p(\hat{\mathbf{n}}_\theta) \parallel p(\mathbf{n}) \,\right] \right\} \tag{5.8}$$

This means that the selected SVR parameters are those that give rise to both signal and noise estimates probabilistically similar to the true signal and noise respectively. Note that this similarity does not require analytical models of the PDFs since it can be computed from histograms (or signal and noise samples).

Of course, the independence assumption does not hold in general, however, as it will be shown in, this is not a critical fact for a good behavior of the method even in non-additive cases in which the noise is clearly signal-dependent. Moreover, the independence assumption simplifies the practical application of the criterion for SVR selection since, for a limited number of samples, histogram estimations of $p(\mathbf{i})$ and $p(\mathbf{n})$ are far more reliable than histogram estimations of $p(\mathbf{i}, \mathbf{n})$, which implies the duplication of the dimensionality (in an already high dimensional situation).

In the examples we restricted ourselves to second order KLD measures due to the lack of samples, yet capturing the second order structure of signal and noise. The optimization in Eq. (5.8) was carried out by exhaustive search.

Summary of the proposed denoising method

The proposed denoising method can be summarized as follows. First the noisy image is transformed by a steerable wavelet filter bank. Then, a set of SVRs is applied to the patches of the subbands of the transform. These SVRs use the profiles for the penalization factor and the insensitivity computed from signal and noise samples as described in Section 5.1.3. The SVRs use the kernel based on MI that captures signal relations in the wavelet domain as described in Section 5.1.3. While the scaling of the penalization profile and the kernels are kept fixed as indicated in Section 5.1.3, the scaling of the insensitivity profile is automatically selected following the procedure described in section 3.3.

5.1.5 Behavior of the proposed method

In this section, we show an illustrative example of how the SVR parameters affect the estimated solution. Moreover we validate the proposed automatic procedure for SVR selec-

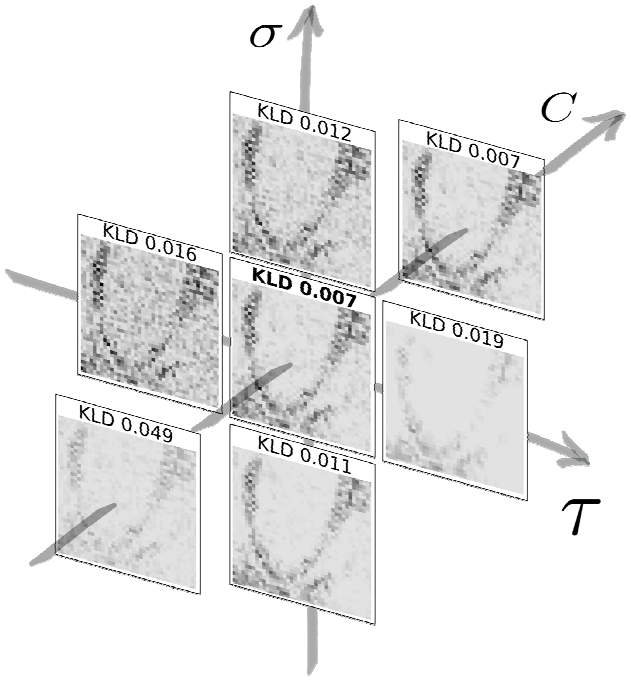


Figure 5.5: Effect of SVR parameters on the noisy wavelet patch of Fig. 5.3. The values of the KL-divergence criterion between the estimated and the actual PDFs of noise and signal are given in each case (see text in Section 3.3).

tion considering examples with different noise sources including non-additive and signal dependent cases.

Impact of SVR parameters in image denoising

As stated above, the regularization behavior of the SVR depends on $\theta = (C_i, \varepsilon_i, K)$. Here we show the qualitative effect of the global penalization scaling C, the global insensitivity scaling τ, and the kernel width σ assuming a generalized RBF kernel. Figure 5.5 shows the qualitative effect of SVR estimation as a function of these parameters. Compare the results with the original and noisy subbands shown in Fig. 5.3.

Increasing the kernel width, σ (vertical direction), introduces too strong relations among coefficients in such a way that spurious energy appears in the reconstruction. Increasing the insensitivity, τ (horizontal direction), a sparser solution is obtained, leading to information loss and thus relevant features of the signal are discarded. On the contrary, a too small insensitivity value gives rise to overfitting, and hence noise is not removed. Small values of the C parameter gives rise to over-regularized estimations. Large enough values of C give rise to similar behavior (see comments in Section 5.1.3).

Of course, interactions among these parameters occur, and have been studied in other contexts elsewhere [Chalimourda et al., 2004; Cherkassky, 2004; Cherkassky & Ma, 2003]. In the image denoising case, the deviation from an *appropriate* solution in combined direc-

tions of the parameters gives rise to different solutions that combine the negative effect of the departure in each direction.

The above example suggests that *appropriate* SVRs can certainly recover the underlying structure of the original signal from the noisy observation, which is the rationale of the proposed method.

Validation of the automatic procedure for SVR selection

In this section, we validate the previous SVR selection procedure in two different ways. Firstly, note that KLD values in the example of Fig. 5.3 *qualitatively* illustrate the usefulness of the proposed procedure: the minimum divergence solution (central subband patch) gives also a reasonable trade-off between smoothness and detail preservation of the original subband patch.

Secondly, we *quantitatively* show that the SVR that enforces the similarity between the estimated and actual signal and noise joint PDFs (in KLD terms) is not far from the SVR that maximizes the structural similarity between the estimated and the original image. In order to do so, we compare the KLD measures for different SVRs, with the corresponding distortion measured with the Structural SIMilarity (SSIM) index [Wang et al., 2004a]. The SSIM index is a widely used similarity measure, which is better related to human quality assessment than Euclidean measures, such as MSE or PSNR. Note that while KLD values are available in real situations (provided the noise histogram and a generic natural images histogram are known), distortion measures are not available since the original image is unknown. Consequently, the SSIM results next presented are for mere comparison purposes.

In this experiment, the SVM parameter space is reduced to the scaling factor on the insensitivity profile as recommended in Section 5.1.3. Accordingly, Fig. 5.6 shows the KLD and distortion (1-SSIM) results as a function of τ (see Eq. (5.4)). Curves are shown for different kinds of (Gaussian and non-Gaussian) noise sources corrupting a particular image (details on the noise sources are given in Section 5.1.6).
For the Gaussian noise case, two different variances are shown. It is worth noting that (1) the KLD criterion (solid) closely follows the actual distortion curve (dashed), and (2) the minima for low and high noise regime curves are very similar. These facts suggest that, in the Gaussian noise case, the proposed criterion is quite robust and provides consistent results: the higher the noise (red curves) the higher the ε zone minima. Besides, the linear relation between ε and the noise standard deviation, reported in [Kwok & Tsang, 2003], is confirmed here: as expected, the scaling factor keeps fairly constant, $\tau \approx 2.5$, for both $\sigma_n^2 = 200$ and $\sigma_n^2 = 400$. Obviously, higher noise levels imply more distorted estimations. For other (non-Gaussian) noise sources, similar results are obtained. For the JPEG and JPEG2000 quantization noise sources, the KLD criterion also matches SSIM performance. For the case of more complex noise sources, such as vertical striping (VS) and Infra Red

Imaging System (IRIS) noise, the criterion gives close-to-optimal solutions in SSIM terms. Note that, remarkably, the KLD criterion is better suited to the error minimization when the signal and noise independence assumption holds (Gaussian case). Therefore there is room to further improve the SVR selection criterion. The above results suggest that the proposed SVR selection procedure can be considered as a convenient approximation to distortion minimization (which is not possible in real situations).

5.1.6 Denoising experiments and discussion

In this section, we evaluate the performance of the proposed method in different scenarios for image denoising. Our algorithm is compared to several wavelet-based denoising methods using standard 256×256 images ('Barbara', 'Boats', 'Lena') with different levels and sources of degradation. In the following, we first give details on implementation issues of the considered algorithms. Then, we analyze their performance for several kinds of noise sources:

- Experiment 1. Additive Gaussian noise of different variances ($\sigma_n^2 = \{200, 400\}$).
- Experiment 2. Coding noise: JPEG and JPEG2000 at different quantization coarseness.
- Experiment 3. Acquisition noise: vertical striping and Infra Red Imaging System (IRIS) noise.

Note that the noise PDF is in general unknown, except for the academic case of Gaussian noise, but the histogram can be computed from samples in all cases.
All results are compared numerically by using the standard (yet not perceptually meaningful) RMSE, and the perceptually meaningful SSIM index [Wang et al., 2004a]. Moreover, representative examples are shown in every case for visual inspection. For proper visualization, all the results are equalized in the same way by truncating the values outside the $[0, 255]$ range.

Implementation details

The denoising algorithms used for comparison that do not use information about the inter-coefficient relations are straightforward to implement and have few parameters to tune [Donoho & Johnstone, 1995; Figueiredo & Nowak, 2001; E. P. Simoncelli, 1999]. All these methods use orthogonal wavelet representations. In our particular implementation, we used 4-scale QMF wavelets from MatlabPyrTools.[2] In every case, we followed authors' prescriptions to choose these parameters for the best performance:

[2] See `http://www.cns.nyu.edu/~eero/software.php`.

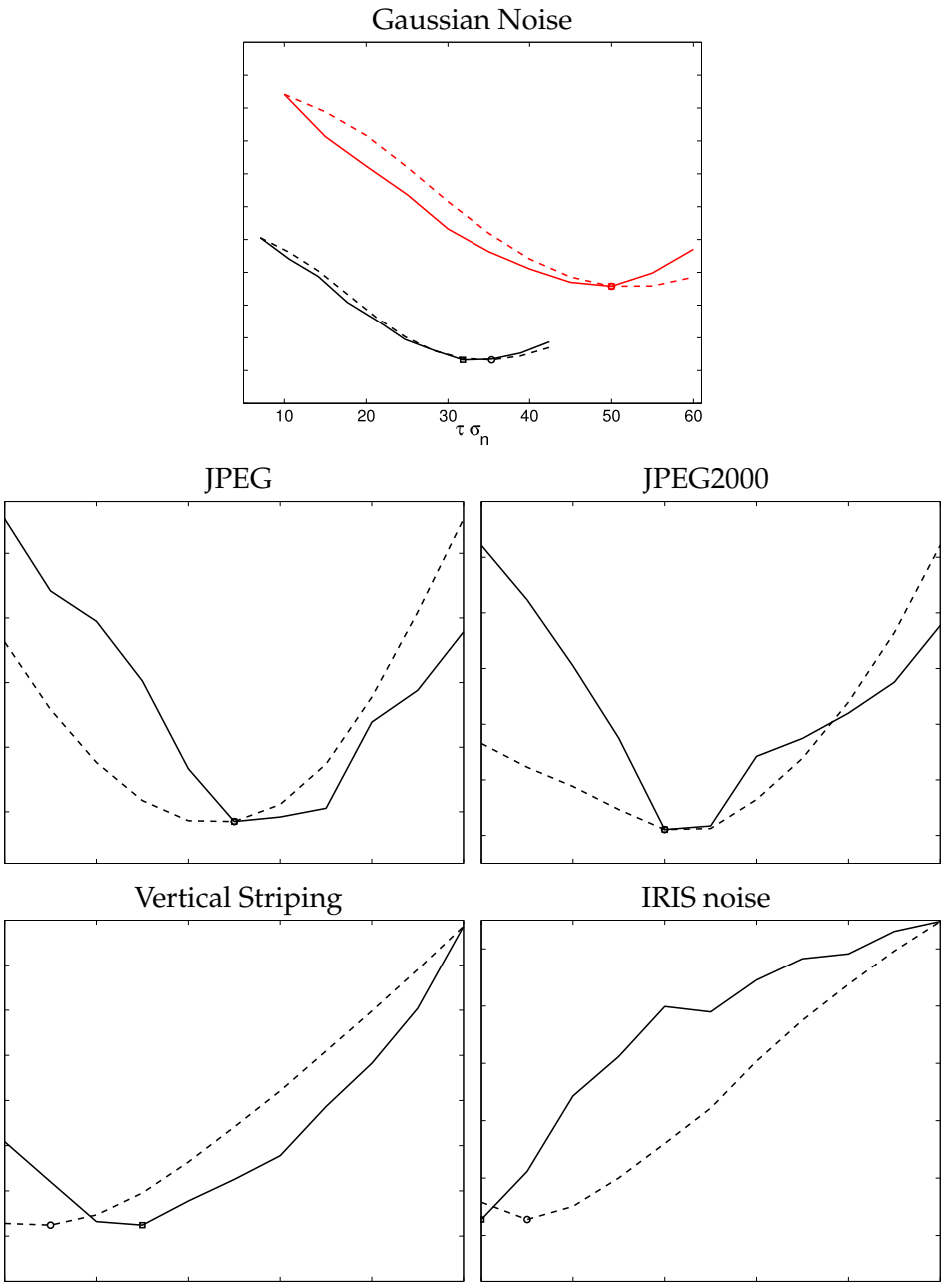


Figure 5.6: Validation of the proposed KLD criterion to adjust SVR parameter ε (or equivalently τ, see text). In every distortion case, solid lines represent the KLD criterion and dashed lines represent the distortion (1-SSIM). For proper visualization, KLD curves were normalized to fall in the same range as the distortion. In the Gaussian noise case, two different noise variances are considered: $\sigma_n^2 = 200$ (black lines) and $\sigma_n^2 = 400$ (red lines). As can be seen, the minima of the KL distance (squares) are always in the same region as the minima of the distortion (circles), thus giving rise to similar SSIM performance.

- *Hard Thresholding (HT)*. The key parameter is the threshold value λ. We use the noise variance to set the threshold, $\lambda = 3\sigma_n$, as suggested in [Donoho & Johnstone, 1995].

- *Soft Thresholding (ST)*. In our implementation, the threshold in each subband is derived from the standard deviation of the noise, σ_n, using optimized values to minimize the mean square error (MSE) in a set of 100 natural images. Threshold values were optimized for the σ_n^2 in the range [0,1600].

- *Bayesian Laplacian (BL)*. In this case, the parameters of the Laplacian distribution (s and p in [E. P. Simoncelli, 1999]) for the marginal PDFs in each subband are estimated by maximum likelihood (ML), as suggested by the author.

- *Bayesian Gaussian (BG)*. The threshold value was set according to the function of noise variance provided in [Figueiredo & Nowak, 2001].

On the other hand, in the case of the Gaussian Scale Mixture (GSM) [Portilla et al., 2003], which does consider inter-coefficient relations, we used the implementation provided by the authors.[3] We have used (1) the same representation as in the proposed method (4-scale, 8-orientation steerable pyramid), and (2) we also provided the average noise power spectral density (PSD) to achieve the best possible performance of the GSM method.

Details of the proposed SVR method are included in previous Section 5.1.3. A Matlab implementation of the algorithm is available online.[4] Since the C_i and ε_i profiles are computed off-line, the computational cost of the proposed method is mainly constrained by the SVR training. In our current implementation, we used the IRWLS algorithm in Matlab [Pérez-Cruz et al., 2000] in order to drop the bias term and incorporate the insensitivity and penalization profiles easily. These modifications are not trivial in faster implementations [Huang & Kecman, 2004; Kecman et al., 2004]. As a result, our Matlab implementation takes about 30 seconds[5] for each image/noise estimation for a set of SVR parameters. Three strategies can be carried out for speeding up the optimization: (1) using faster SVR implementations [Chang & Lin, 2001b; Platt, 1999; Tsang et al., 2005], (2) alternative procedures to exhaustive search on convex error surfaces [Lewis & Torczon, 2002; Torczon, 1997; Vishwanathan et al., 2006], and (3) restricting the dimension of the parameter space (as done in Section 5.1.3).

Experiment 1. Additive Gaussian noise

Table 5.1 shows the distortion results for the three considered images and the two noise variances, $\sigma_n^2 = 200$ and $\sigma_n^2 = 400$.. The best SSIM values in each case are highlighted. In every case, we also provide the SVR^{opt} result, which is the best result the proposed

[3] See `http://decsai.ugr.es/~javier/denoise/`.

[4] See `http://www.uv.es/vista/vistavalencia/denoising_SVR/`.

[5] Computations were carried out in a 2.8GHz processor with 4GB RAM.

	'Barbara'		'Boats'		'Lena'	
Method	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE
HT	0.77	16.48	0.76	13.62	0.73	18.97
ST	0.78	14.37	0.79	10.26	0.74	12.59
BG	0.80	14.14	0.79	11.70	0.76	12.75
BL	0.81	12.95	0.83	8.30	0.78	11.66
GSM	**0.90**	8.94	**0.87**	8.94	**0.83**	13.61
SVR	0.87	10.11	0.84	10.16	0.81	12.54
SVR^{opt}	0.87	10.11	0.85	10.36	0.82	12.30
	'Barbara'		'Boats'		'Lena'	
Method	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE
HT	0.67	24.52	0.68	20.15	0.67	20.22
ST	0.69	19.04	0.71	16.16	0.66	19.72
BG	0.70	20.40	0.70	19.17	0.67	19.26
BL	0.73	16.52	0.77	10.26	0.67	18.45
GSM	**0.86**	11.02	0.80	17.40	**0.79**	15.95
SVR	0.83	13.13	**0.81**	10.73	0.78	14.50
SVR^{opt}	0.83	13.13	0.81	10.73	0.78	14.06

Table 5.1: Results for the Gaussian noise: distortions for different images and methods are given at $\sigma_n^2 = 200$ (top) and $\sigma_n^2 = 400$ (bottom).

method can get in SSIM terms. This is useful to assess the performance of the proposed divergence-based criterion and to give an upper bound of method's performance. Results show that our algorithm performs better than the methods that neglect signal relations (HT, ST, BG and BL), and obtains similar (yet slightly lower) numerical results than the one which incorporates them (GSM). It is not surprising that the GSM method achieves the best performance in this case, since it is analytically suited to deal with Gaussian noise. The SVR performance is consistent through all images and noise variances, thus suggesting that the guiding criterion is robust. Finally, it must be noted that, in the most difficult case of $\sigma_n^2 = 400$, GSM and SVR offer more similar results, and clearly outperform the rest of the methods.

Figure 5.7 shows representative visual results in the challenging situation of $\sigma_n^2 = 400$. It can be noticed that thresholding methods (HT, ST) and Bayesian generalizations not including signal relations in the model (BG, BL) show poor performance, producing images either grained or corrupted by too salient wavelet functions. Even though SVR yields slightly lower numerical scores than GSM, global visual performance is comparable.

Experiment 2. Coding noise: JPEG and JPEG2000

In this section, we focus on restoring grayscale images after JPEG or JPEG2000 compression, which induces non-Gaussian noise: it produces heavy tailed marginal error PDFs

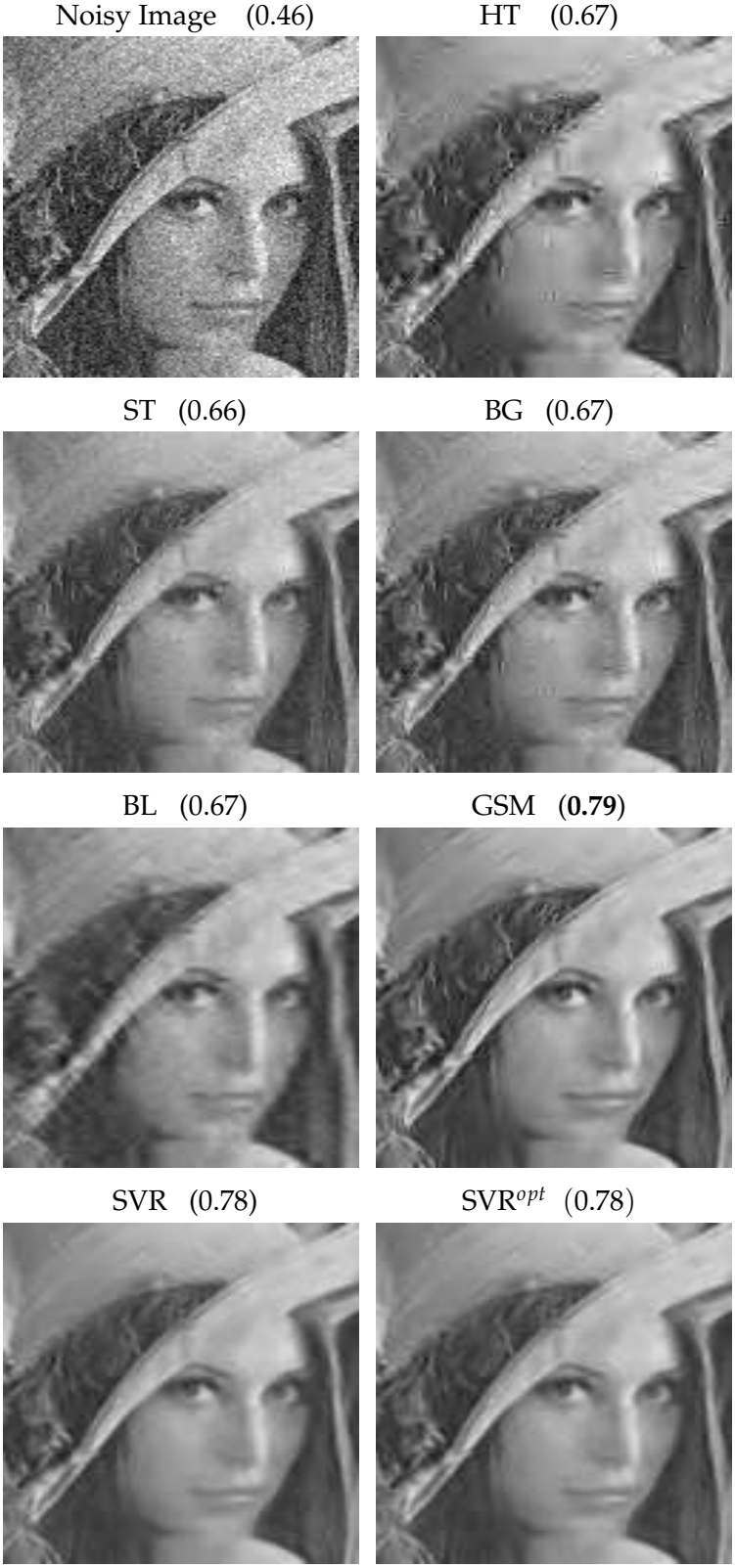


Figure 5.7: Visual results for the 'Lena' image corrupted with Gaussian noise, $\sigma_n^2 = 400$. SSIM values are given in parentheses.

in the spatial domain with non-negligible relations among the pixels (see comments in Section 5.1.7). Quantization noise is an illustrative example of how the proposed method can cope with non-Gaussian, colored and signal-dependent noise. In order to obtain the necessary samples to build the noise histograms, we used 100 images from the database described in Section 2 encoded by JPEG and JPEG2000. In the first case, the Matlab implementation of the JPEG algorithm with quality factors $Q = 9$ (small distortion) and $Q = 7$ (large distortion) was used. In the second case, scalar quantization of the QMF wavelet domain using standard JPEG2000 bit allocation tables [Taubman & Marcellin, 2001] was used. Different values of quantization coarseness, that will be referred to as Δ_1 (small distortion) and Δ_2 (large distortion) were applied.

Table 5.2 shows the quantitative results for all considered methods for the three images at different quantization levels. It can be noticed that again the SVR method outperforms the thresholding methods (HT, ST) and those not including signal relations in the model (BG, BL). SVR yields similar numerical scores than GSM in JPEG (Fig. 5.8). However, in JPEG2000 better numerical (Table 5.2 [bottom]) and visual (Fig. 5.9) results are obtained with SVR. In general, high frequency details are better preserved by our method, while GSM yields over smoothed solutions, particularly in JPEG2000.

Experiment 3. Acquisition noise: Vertical Striping and IRIS

Real imaging systems introduce complex forms of noise depending on the acquisition process, so assuming a particular PDF for all cases is far from being realistic. For instance, variation of the intensity between neighboring elements of the CCD typically leads to vertical striping noise in pushbroom sensors [Barducci & Pippi, 2001; Mouroulis et al., 2000]. Other typical acquisition noise source is observed in infrared imaging cameras, which is a complex mixture of different noise sources. In this section, we pay attention to these two particular non-Gaussian realistic acquisition noises through controlled experiments:

1. *Vertical striping noise.* We simulated this noise by modifying 4% of the image columns selected randomly. The luminance of the selected columns was modified by a random factor following a uniform distribution between 0.8 and 1. Spatial coherence was forced by attaching groups of contiguous 5 to 10 strips.

2. *InfraRed Imaging System (IRIS) noise.* Inspired in the observed characteristics of a representative number of acquired images by a commercial IR camera, the noise was modeled by a combination of four noise sources: low-variance Gaussian noise ($\sigma_n^2 \approx 50$), 'salt-and-pepper' noise (with a percentage of corrupted pixels about 0.05%), some spatially coherent missing pixels (black patches), and interlaced lines all over the image.

In both cases, we computed the contrast noise PDF, $p(\mathbf{n})$, from 100 noisy images. In the next Section 5.1.7, the non-Gaussian nature of these acquisition noise PDFs is shown.

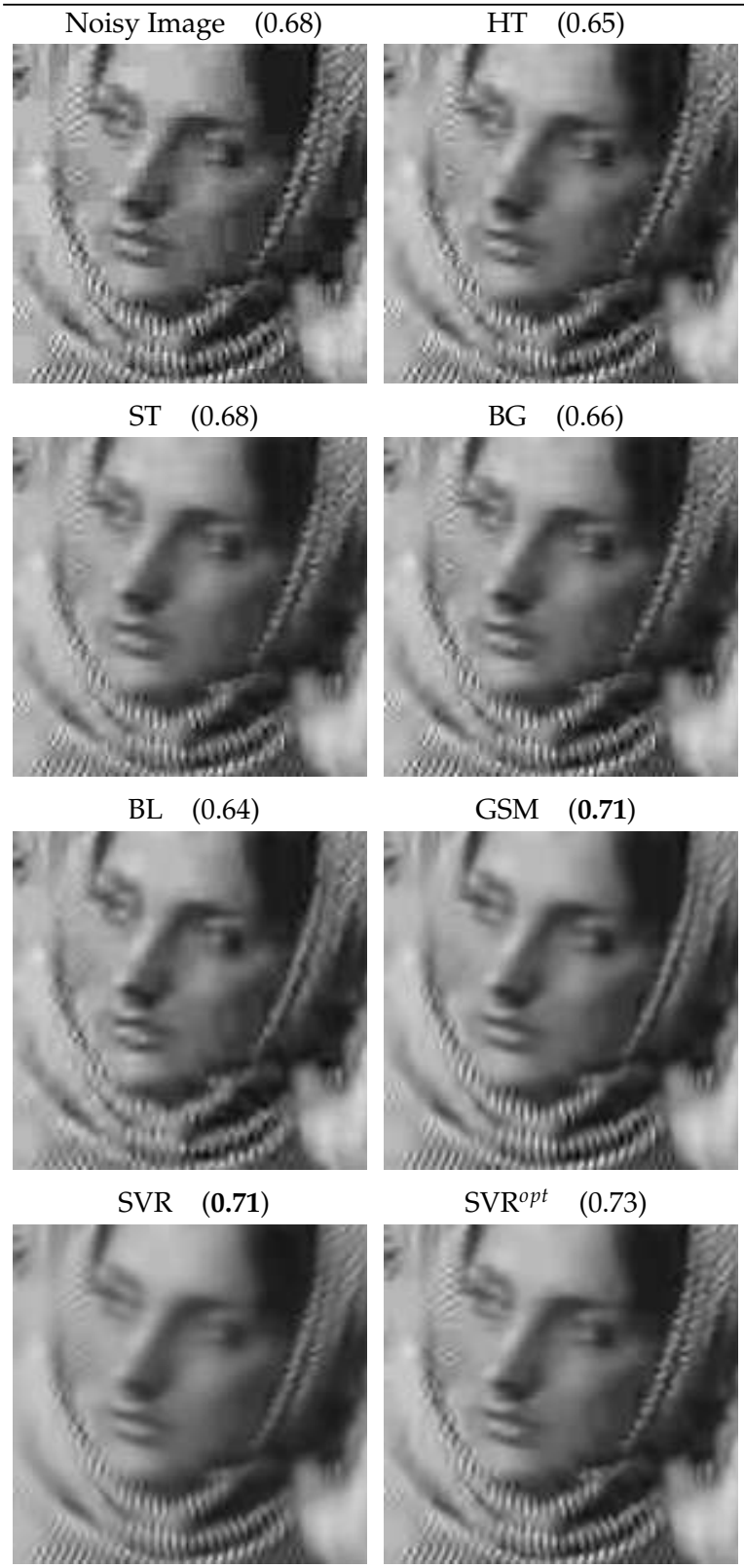


Figure 5.8: Visual results for the 'Barbara' image with JPEG quantization noise ($Q = 7$). SSIM values are given in parentheses.

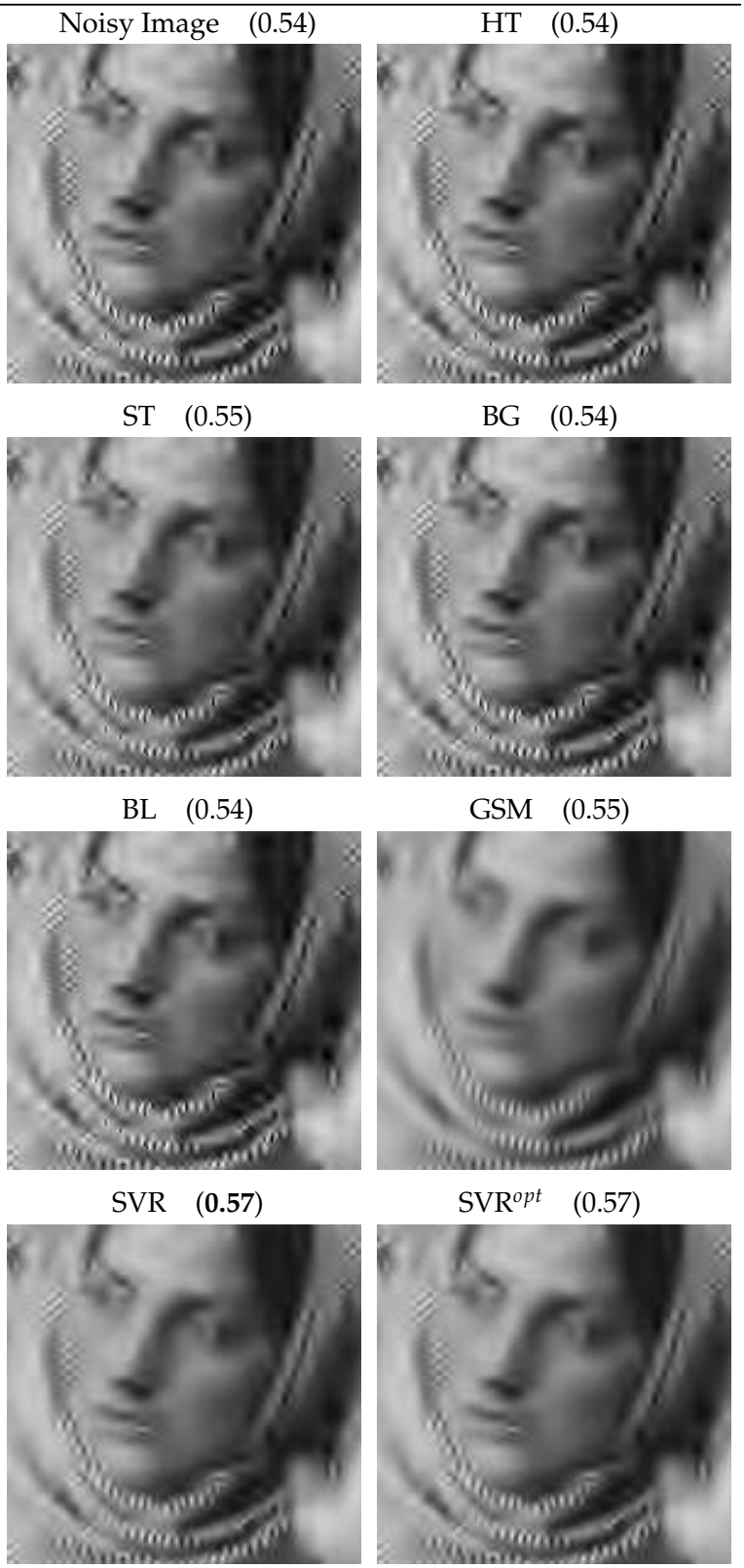


Figure 5.9: Visual results for the 'Barbara' image with coarse quantization JPEG2000 noise. SSIM values are given in parentheses.

JPEG	$Q = 9$						$Q = 7$					
	'Barbara'		'Boats'		'Lena'		'Barbara'		'Boats'		'Lena'	
Method	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE
HT	0.70	20.05	0.75	13.07	0.70	18.40	0.65	22.11	0.71	16.34	0.65	24.99
ST	0.73	17.51	0.78	11.59	0.73	15.13	0.68	19.71	0.75	12.72	0.68	18.77
BG	0.72	18.76	0.77	12.30	0.72	16.27	0.66	21.57	0.74	13.32	0.67	21.05
BL	0.71	20.37	0.77	13.43	0.73	16.52	0.64	21.67	0.74	14.70	0.69	17.65
GSM	0.77	15.50	**0.80**	11.15	**0.75**	13.66	**0.71**	18.56	**0.77**	12.18	**0.71**	17.45
SVR	**0.78**	14.89	0.78	12.13	0.74	13.22	**0.71**	18.42	0.76	12.84	**0.71**	15.68
SVR^{opt}	0.78	14.89	0.80	11.35	0.75	13.97	0.73	18.28	0.76	12.89	0.71	15.72
JPEG2000	Δ_2						Δ_1					
	'Barbara'		'Boats'		'Lena'		'Barbara'		'Boats'		'Lena'	
Method	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE
HT	0.54	30.81	0.55	26.23	0.51	32.66	0.67	24.82	0.59	25.18	0.56	28.25
ST	0.55	28.83	0.55	25.15	0.51	31.24	0.68	22.52	0.60	23.69	0.56	27.47
BG	0.54	30.37	0.55	26.08	0.51	32.45	0.67	24.16	0.59	24.92	0.56	28.10
BL	0.54	30.30	0.55	25.87	0.51	29.05	0.67	24.35	0.59	24.79	0.56	28.12
GSM	0.55	28.47	**0.57**	20.92	**0.52**	25.84	0.68	20.54	**0.64**	17.94	0.58	23.64
SVR	**0.57**	25.31	**0.57**	21.88	**0.52**	29.32	**0.71**	17.23	**0.64**	18.27	**0.59**	21.55
SVR^{opt}	0.57	25.31	0.57	21.74	0.52	25.35	0.72	17.04	0.64	18.27	0.59	21.55

Table 5.2: Results for the coding noise: distortions at different quality levels of JPEG ($Q = \{9,7\}$) and JPEG2000 (coarseness Δ_1 and Δ_2) are given for different images and methods.

Table 5.3 shows the obtained numerical results for all images and both acquisition noise sources. In both complex scenarios, the proposed SVR-based method outperforms GSM and the rest of methods numerically. A noticeable gain in SSIM is observed, which is confirmed when looking at the restored images in Figs. 5.10 and 5.11. It is worth noting that in the vertical striping noise (Fig. 5.10), SVR yields a sharper (and more realistic) reconstruction while GSM produces an over-blurred solution. In the case of the IRIS noise, only SVR removes the interlacing noise contribution, producing better visual results. Including the average PSD information in GSM, as we do in the experiments, improves its performance. However, it is not enough to remove the interlacing artifact due to the particular nature of IRIS noise. IRIS noise is difficult because the PSD and variance of each particular realization of the noise may substantially differ from the (estimated) averages. On the contrary, the proposed SVR method uses an adaptive cost function learned from the noisy image. Here, nevertheless, the upper bound of performance is not met, suggesting that there is still room for improving the selection criterion proposed, possibly considering the joint density.

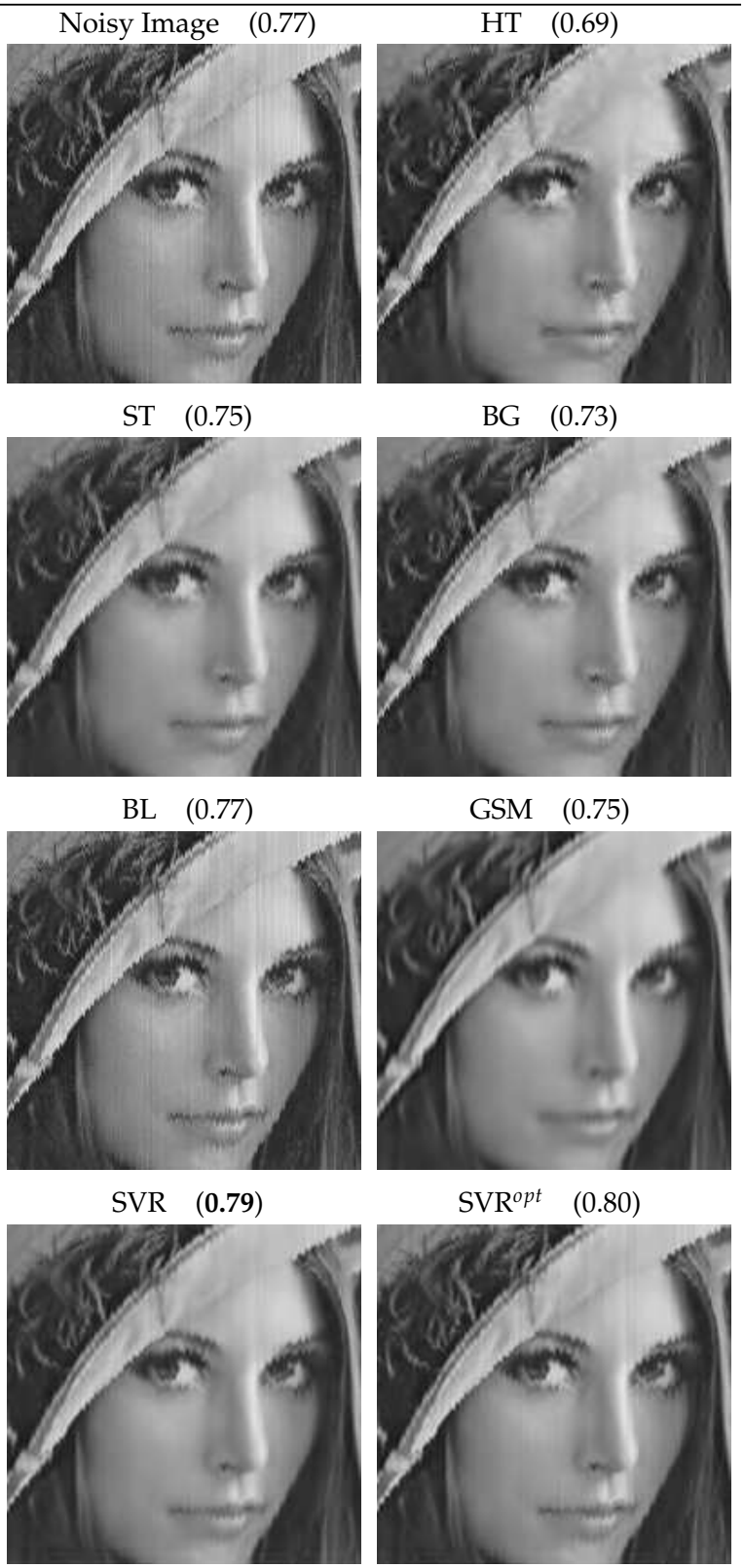


Figure 5.10: Visual results for the 'Lena' image with vertical striping noise. SSIM values are given in parentheses.

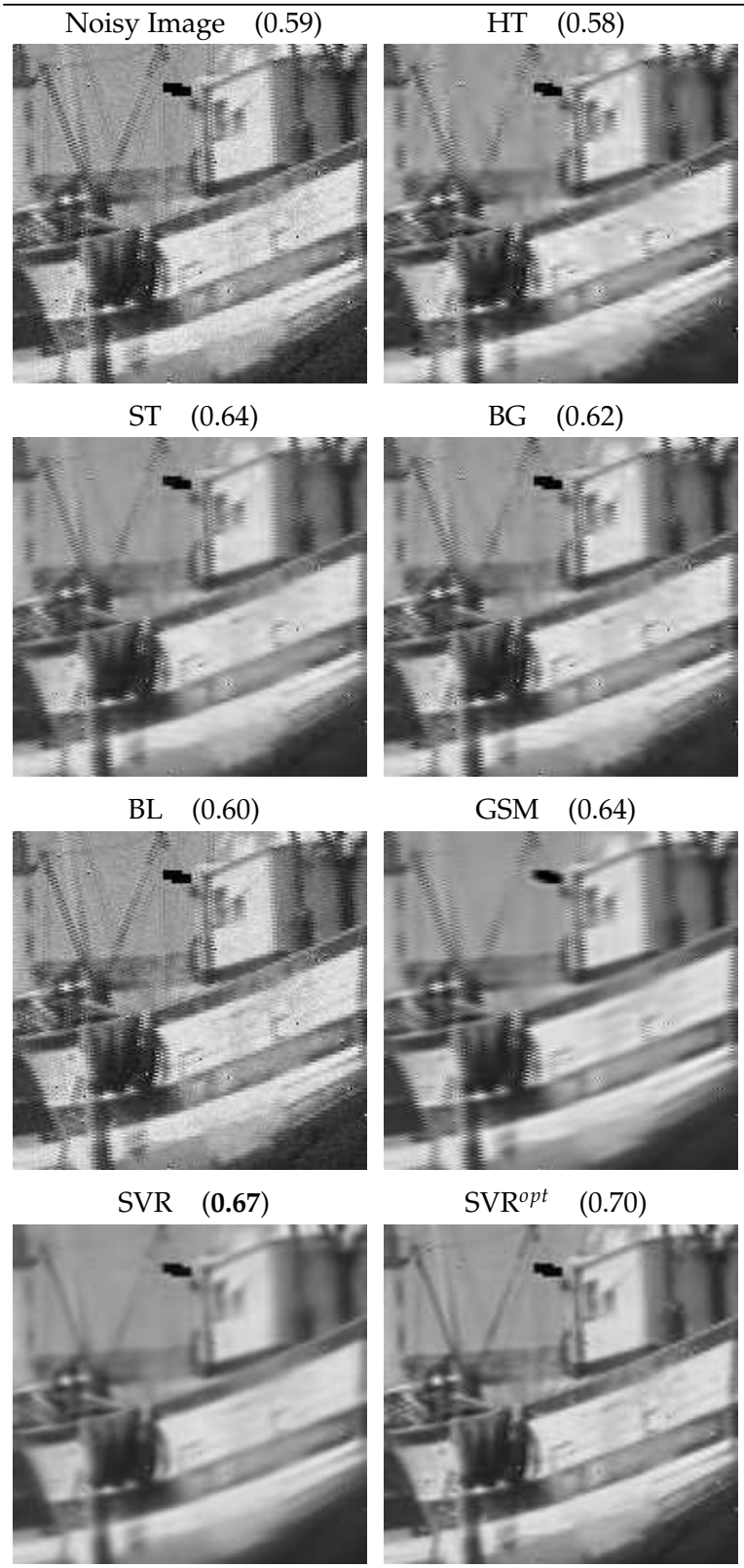


Figure 5.11: Visual results for the 'Boats' image with IRIS noise. SSIM values are given in parentheses.

Method	'Barbara'		'Boats'		'Lena'	
	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE
HT	0.73	17.43	0.73	15.99	0.69	18.07
ST	0.77	15.71	0.78	14.04	0.75	14.08
BG	0.76	16.01	0.76	14.75	0.73	15.14
BL	0.77	16.56	0.81	14.96	0.77	14.64
GSM	0.79	14.83	0.79	14.36	0.75	14.45
SVR	**0.80**	15.66	**0.80**	13.47	**0.79**	13.18
SVR^{opt}	0.80	15.45	0.82	14.25	0.80	13.31
HT	0.50	30.80	0.58	28.70	0.56	28.81
ST	0.55	27.02	0.64	23.48	0.60	24.40
BG	0.54	28.40	0.62	25.44	0.59	26.20
BL	0.50	28.74	0.60	21.77	0.55	24.08
GSM	0.53	30.51	0.64	25.92	0.61	30.99
SVR	**0.59**	31.07	**0.67**	21.44	**0.66**	31.44
SVR^{opt}	0.60	30.71	0.70	24.56	0.66	32.05

Table 5.3: Acquisition noise: vertical striping (top) and IRIS noise (bottom). Distortions for different images and methods.

5.1.7 Analysis of the residuals

Further qualitative insight in the obtained solutions can be achieved by comparing the estimated and actual PDFs of signal and noise with the different methods and noise sources. Since we are restricting ourselves to second order KLD criterion, this comparison reduces to assess the difference between 2D histograms (in the spatial domain).

It is widely known that the PDF of pairs of neighbor pixels in natural images is an oriented ellipsoid reflecting the strong correlation among luminance values in the spatial domain [Clarke, 1985]. The corresponding restored images (even for the worse performing algorithms) also display such strong local correlation. Therefore, no relevant conclusion is gained by direct inspection of these histograms (results not shown). On the contrary, the 2D histograms of the noise are more suitable for direct inspection because (1) actual noise histograms are quite different for the different noise sources, and (2) the estimated histograms strongly depend on the denoising method.

Figure 5.12 represents the distribution of the actual and estimated noise PDFs by all the considered methods in the spatial domain. It can be noticed that, for the Gaussian noise, all methods reproduce quite well the shape and extent of the PDF, as expected for the parametric models, which use a proper Gaussian noise model. Note that the SVR method also succeeds in approximating the energy of the noise even without using the Gaussian assumption explicitly.

For non-Gaussian noise sources, the behavior of the methods markedly differ. For instance, the quantization noise induced by JPEG/JPEG2000 follows a non-Gaussian, oriented joint distribution (the central dark area is actually an oriented ellipsoid), indicating correlation among noise samples. In the case of JPEG, this central ellipsoid is better reproduced by hard thresholding and the proposed SVR method. The other methods slightly underestimate the variance of the noise. For the case of JPEG2000, methods not considering signal relations dramatically underestimate the noise variance. In the case of more complex noise sources, such as vertical striping or IRIS, none of the methods reproduce the low probability structure (light gray regions). However, the central peak is poorly reproduced by marginal methods, either overestimating (HT, ST, BG) or underestimating (BL) the width. On the contrary, GSM and SVR give more reasonable width estimation. To conclude, methods assuming an (inadequate) Gaussian noise model do not match, in general, the noise distribution, so they should be reformulated for each particular noise source, which may be complicated or even impossible. GSM constitutes an exception to this statement, since results suggest that the quality of the signal model compensates the unsuitability of the noise model. On the contrary, this is not necessary for the proposed method, which only needs examples of noisy images to *learn* from.

Figure 5.12: 2D histograms of the residuals in the spatial domain for all methods and noise sources (darker regions indicate higher probability). In all cases, we considered a pixel and its right-hand neighbor (one pixel shift). All the histogram values have been exponentiated to 0.25 for better visualization.

5.2 Iterative Gaussianization Framework

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) [Banham & Katsaggelos, 1997; R. O. Duda et al., 2000; Gersho & Gray, 1992; Hastie et al., 2003; Portilla & Simoncelli, 2000]. 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 [Scott, 1992]. 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) [Jolliffe, 1986], that reduces to DCT in many natural signals such as speech, images and video, assumes a Gaussian source [R. O. Duda et al., 2000; Jolliffe, 1986]. More recently, linear ICA, that reduces to wavelets in natural signals, assumes that observations come from the linear combination of independent non-Gaussian sources [Hyvärinen, 1999b]. 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 [Portilla & Simoncelli, 2000], image coding [Buccigrossi & Simoncelli, 1999] or image denoising [Portilla et al., 2003]. Non-linear methods using non-explicit statistical models have been also proposed to this end in the denoising context [Gutiérrez et al., 2006; Laparra, Gutiérrez, et al., 2010] and in the coding context [Camps-Valls et al., 2008; Malo et al., 2006]. 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 [R. O. Duda et al., 2000; Hastie et al., 2003; Jolliffe, 1986].

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 is organized as follows. In Section 5.2.1 we present the underlying idea that motivates the proposed approach to Gaussianization. In Section 5.2.2, 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 5.2.3 discusses the similarities and differences of the proposed method and Projection Pursuit (PP) [Chen & Gopinath, 2000; Friedman & Tukey, 1974; Huber, 1985; Rodríguez-Martínez et al., 2010]. Links to other techniques (such as single-step Gaussianization transforms [Erdogmus et al., 2006; Lyu & Simoncelli, 2009], one-class support vector domain descriptions [Tax & Duin, 1999], and deep neural network architectures [Hinton & Salakhutdinov, 2006]) are also explored. Section 5.2.4 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.

5.2.1 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 [Stark & Woods, 1986]:

$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{y}}(\mathcal{G}(\mathbf{x})) \left| \frac{d\mathcal{G}(\mathbf{x})}{d\mathbf{x}} \right| = p_{\mathbf{y}}(\mathcal{G}(\mathbf{x})) |\nabla_{\mathbf{x}} \mathcal{G}(\mathbf{x})|, \tag{5.9}$$

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 [Erdogmus et al., 2006; Lyu & Simoncelli, 2009]. 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 [Cardoso, 2003; Comon, 1994], 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[6], $J(\mathbf{x}) = \mathrm{D_{KL}}\left(p(\mathbf{x})|\mathcal{N}(\mathbf{0}, \mathbf{I})\right)$. Negentropy can be decomposed as the sum of two non-negative quantities, the multi-information and the marginal negentropy:

$$J(\mathbf{x}) = I(\mathbf{x}) + J_m(\mathbf{x}). \tag{5.10}$$

This can be readily derived from Eq. (5) in [Cardoso, 2003], by considering as contrast PDF $\prod_i q_i(x_i) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The multi-information is [Studeny & Vejnarova, 1998]:

$$I(\mathbf{x}) = \mathrm{D_{KL}}(p(\mathbf{x})|\prod_i p_i(x_i)) \tag{5.11}$$

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:

$$J_m(\mathbf{x}) = \sum_{i=1}^{d} \mathrm{D_{KL}}\left(p_i(x_i)|\mathcal{N}(0,1)\right) \tag{5.12}$$

Given a data distribution from the unknown PDF, in general both I and J_m will be non-zero. The decomposition in (5.10) 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 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.

[6]This usage of the term negentropy slightly differs from the usual definition [Comon, 1994] 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.

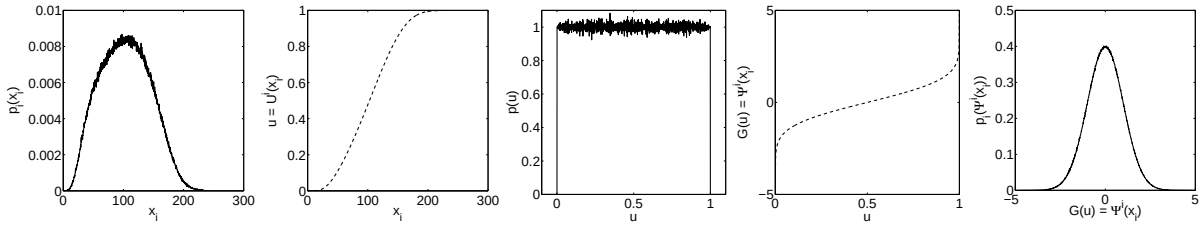


Figure 5.13: Example of marginal Gaussianization in some particular dimension i. From left to right: marginal PDF of x_i, uniformization transform $u = U^i(x_i)$, PDF of the uniformized variable $p(u)$, Gaussianization transform $G(u)$, and PDF of the Gaussianized variable $p_i(\Psi^i(x_i))$.

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 [Studeny & Vejnarova, 1998]. 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 [Chen & Gopinath, 2000; Huber, 1985]. 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 [Learned & Fisher, 2003] would lead to extremely slow Gaussianization algorithms whereas relatively more convenient alternatives such as FastICA [Hyvärinen, 1999a] 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 point-wise nonlinearities [Squartini et al., 2006; K. Zhang & Chan, 2005], or after *ad hoc* speech-oriented feature extraction steps [Xiang et al., 2002]. 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.

5.2.2 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.

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:

$$\mathcal{G} : \mathbf{x}^{(k+1)} = \mathbf{R}_{(k)} \cdot \mathbf{\Psi}_{(k)}(\mathbf{x}^{(k)}) \tag{5.13}$$

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. (5.9). 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.

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. (5.9). Invertibility guarantees that the transform is bijective which is a necessary condition to apply Eq. (5.9). 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 univari-

ate Gaussian: $\mathbf{\Psi}^i_{(k)} = G \odot U^i_{(k)}$, where:

$$u = U^i_{(k)}(x_i^{(k)}) = \int_{-\infty}^{x_i^{(k)}} p_i(x_i'^{(k)})\, dx_i'^{(k)} \tag{5.14}$$

$$G^{-1}(x_i) = \int_{-\infty}^{x_i} g(x_i')\, dx_i' \tag{5.15}$$

and $g(x_i)$ is just a univariate Gaussian. Figure 5.13 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 [Hastie et al., 2003; Silverman, 1986]. 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 (5.13), it can be shown that the inverse transform is given by:

$$\mathcal{G}^{-1} : \mathbf{x}^{(k)} = \mathbf{\Psi}^{-1}_{(k)}(\mathbf{R}^\top_{(k)} \cdot \mathbf{x}_{(k+1)}). \tag{5.16}$$

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, 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:

$$\nabla_{\mathbf{x}}\mathcal{G} = \prod_{k=1}^{K} \mathbf{R}_{(k)} \cdot \nabla_{\mathbf{x}^{(k)}} \mathbf{\Psi}_{(k)} \tag{5.17}$$

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

agonal matrix,

$$\nabla_{\mathbf{x}^{(k)}} \mathbf{\Psi}_{(k)} = \begin{pmatrix} \frac{\partial \mathbf{\Psi}^1_{(k)}}{\partial x^{(k)}_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\partial \mathbf{\Psi}^d_{(k)}}{\partial x^{(k)}_d} \end{pmatrix} \tag{5.18}$$

According to the two equalization steps in each marginal Gaussianization, Eq. (5.15), each element in $\nabla_{\mathbf{x}^{(k)}} \mathbf{\Psi}_{(k)}$ can be easily computed by applying the chain rule on u defined in Eq. (5.14):

$$\begin{aligned} \frac{\partial \mathbf{\Psi}^i_{(k)}}{\partial x^{(k)}_i} &= \frac{\partial \mathcal{G}}{\partial u} \frac{\partial u}{\partial x^{(k)}_i} = \left(\frac{\partial \mathcal{G}^{-1}}{\partial x_i} \right)^{-1} p_i(x^{(k)}_i) \\ &= g(\mathbf{\Psi}^i_{(k)}(x^{(k)}_i))^{-1} p_i(x^{(k)}_i) \end{aligned} \tag{5.19}$$

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

Convergence properties

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

Property 1 (Negentropy reduction). *Marginal Gaussianization reduces the negentropy and this is not modified by any posterior rotation:*

$$\Delta J = J(\mathbf{x}) - J(\mathbf{R}\Psi(\mathbf{x})) \geq 0, \forall \mathbf{R} \tag{5.20}$$

Proof. Using Eq. (5.10), 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 multi-information is invariant under dimension-wise transforms such in our case Ψ [Studeny & Vejnarova, 1998], and the marginal negentropy of a marginally Gaussianized variable is zero,

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

□

Property 2 (Redundancy reduction). *Given a marginally Gaussianized variable,* $\Psi(\mathbf{x})$*, any rotation reduces the redundancy among coefficients,*

$$\Delta I = I(\Psi(\mathbf{x})) - I(\mathbf{R}\Psi(\mathbf{x})) \geq 0, \forall \mathbf{R} \tag{5.21}$$

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 Eq. (5.10) 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,

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

□

The above properties suggest the convergence of the proposed Gaussianization method. Property 1 (Eq. (5.20)) 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 2 (Eq. (5.21)) 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.

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 1 and 2 provide some intuition on the suitable class of rotations. By using (5.20) and (5.21) in the sequence (5.13), one readily obtains the relations:

$$\Delta J_{(k)} = J_m(\mathbf{x}^{(k)}) = \Delta I_{(k-1)}, \tag{5.22}$$

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 [Cardoso, 2003], 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 [León et al., 2006]. Other possibilities include extension to complex variables [Novey & Adali, 2008].

As an illustration, Table 5.4 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 5.2.4 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. (5.20)) is always strictly positive, except for the case that the Gaussian PDF has been finally achieved. This is consistent with previous results [Chen & Gopinath, 2000]. Moreover, the convergence rate is optimal for ICA since it gives rise to the maximum $J_m(\mathbf{x})$ (indicated in Table 5.4 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 5.4 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, $\mathbf{R}_{(k+1)}$, 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

Table 5.4: Properties of the Gaussianization method for different rotations (see comments in the text).

Rotation	Closed -form	Theoretical convergence	Convergence rate	CPU cost† [7]
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.

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 [Feller, 1968]). 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 [Golub & Loan, 1996]. In this case, the computation time of the rotation is negligible compared to that of the marginal Gaussianization.

5.2.3 Relation to other methods

In this section we discuss the relation of RBIG to previously reported Gaussianization methods. Specifically, iterative Projection Pursuit techniques [Chen & Gopinath, 2000; Friedman & Tukey, 1974; Huber, 1985] and direct approaches suited for particular PDFs [Eichhorn et al., 2009; Erdogmus et al., 2006; Lyu & Simoncelli, 2009]. Additionally, relations to other machine learning tools are also considered, Support Vector Domain Description [Tax & Duin, 1999] and deep neural networks [Hinton & Salakhutdinov, 2006].

Iterative Projection Pursuit Gaussianization

As stated above, the aim of Projection Pursuit (PP) techniques [Friedman & Tukey, 1974; Huber, 1985] 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 [Chen & Gopinath, 2000]:

$$\mathcal{G} : \mathbf{x}^{(k+1)} = \mathbf{\Psi}_{(k)}(\mathbf{R}^{\text{ICA}} \cdot \mathbf{x}^{(k)}) \tag{5.23}$$

As stated in Section 5.2.1, this is *Approach 1* to the Gaussian goal. Unlike PP, RBIG aims at the Gaussian goal following *Approach 2*. The differences between (5.23) and (5.13) (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[8], RBIG follows a *disordered* transition to disorder.

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* [Erdogmus et al., 2006], or (2) elliptically symmetric PDFs so that the final Gaussian can be achieved by equalizing the length (norm) of the whitened samples [Eichhorn et al., 2009; Lyu & Simoncelli, 2009].

The method proposed in [Erdogmus et al., 2006] 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 '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 [Portilla et al., 2003] and related models [Malo & Laparra, 2010a] share this symmetry). Radial Gaussianization (RG) was specifically developed to deal with these particular kind of models [Lyu & Simoncelli, 2009]. 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 [Eichhorn et al., 2009]. 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 5.14 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 5.2.4. Both RG and RBIG strongly reduce

[8] 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.

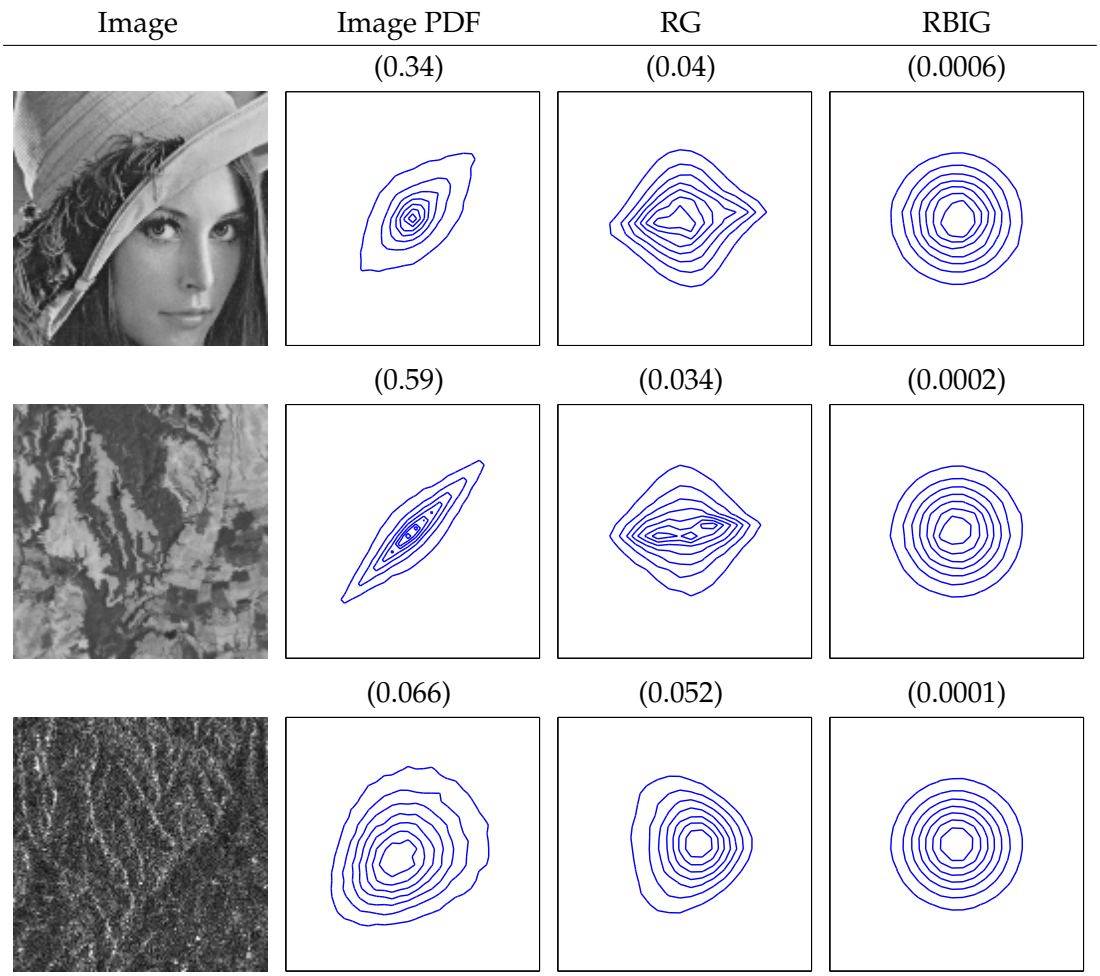


Figure 5.14: Gaussianization of pairs of neighbor pixels from different images with RG and RBIG: natural image (top row), remote sensing Landsat channel in the optical range (middle row) and intensity of a ERS2 synthetic aperture radar (SAR) image (bottom row). Contour plots show the PDFs in the corresponding domains. The estimated mutual information (in bits) is given in parenthesis.

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.

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 [Tax & Duin, 1999]. 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.

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 [Hinton & Salakhutdinov, 2006], 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 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.

5.2.4 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.

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. (5.21). 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 5.15 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. 5.16. 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. (5.20) 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. 5.16 give the result of the multivariate Gaussianity test presented by Székely & Rizzo [2005]: 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

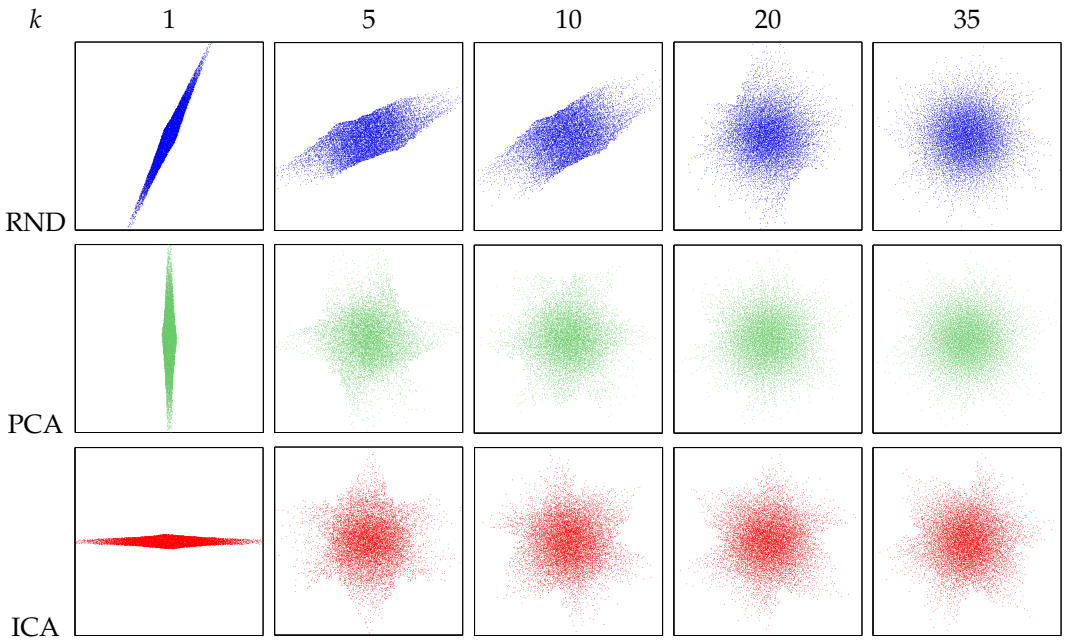


Figure 5.15: Scatter plots of a 2D data in different iterations for the considered rotation matrices: RND (top), PCA (middle) and ICA (bottom).

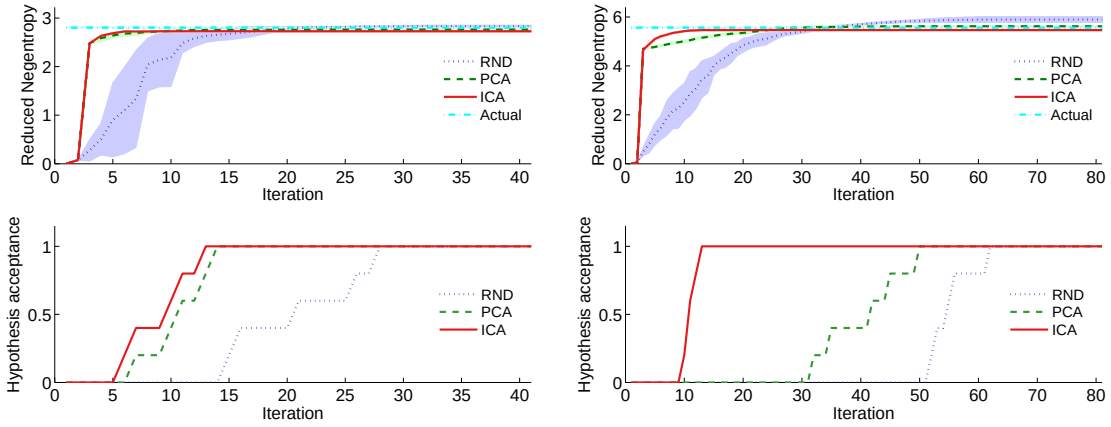


Figure 5.16: Cumulative negentropy reduction (top) and multivariate Gaussian significance test (bottom) for each iteration in 2D (left) and 4D (right) dimensional synthetic problem. Average and standard deviation results from 5 realizations is shown.

the parametric estimator in [Székely & Rizzo, 2005] which, in turn, confirms the analysis in Table 5.4.

Despite the previous conclusions, and as pointed out before, in practical applications, it

Table 5.5: 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

is not the length of the path to the Gaussian goal what matters, but the time required to complete this path. Table 5.5 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 5.5 are fairly consistent with the computational burden per iteration shown in Table 5.4 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 [Székely & Rizzo, 2005]. 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. However, we checked that such singular situation never happened by jointly using both criteria in each iteration.

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. (5.10)). 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

Table 5.6: 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

(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) [Moddemeijer, 1989] (see Table 5.6).
- A set of d-dimensional experiments, where RBIG results are compared to actual values (see Table 5.7).

Table 5.6 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.

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.

- Toy data

  Figure 5.17 shows examples of 2D non-Gaussian distributions (left column) transformed into a Gaussian (center column). The right column was obtained sampling

Table 5.7: 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

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.

- Face synthesis

  In this experiment, 2,500 face images were extracted from [Georghiades et al., 2001], 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 5.18 shows illustrative examples of original and synthesized faces with RG and RBIG.

  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.

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[9] [Gómez-Chova et al., 2006]. The considered test sites were the cities of Rome and Naples,

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

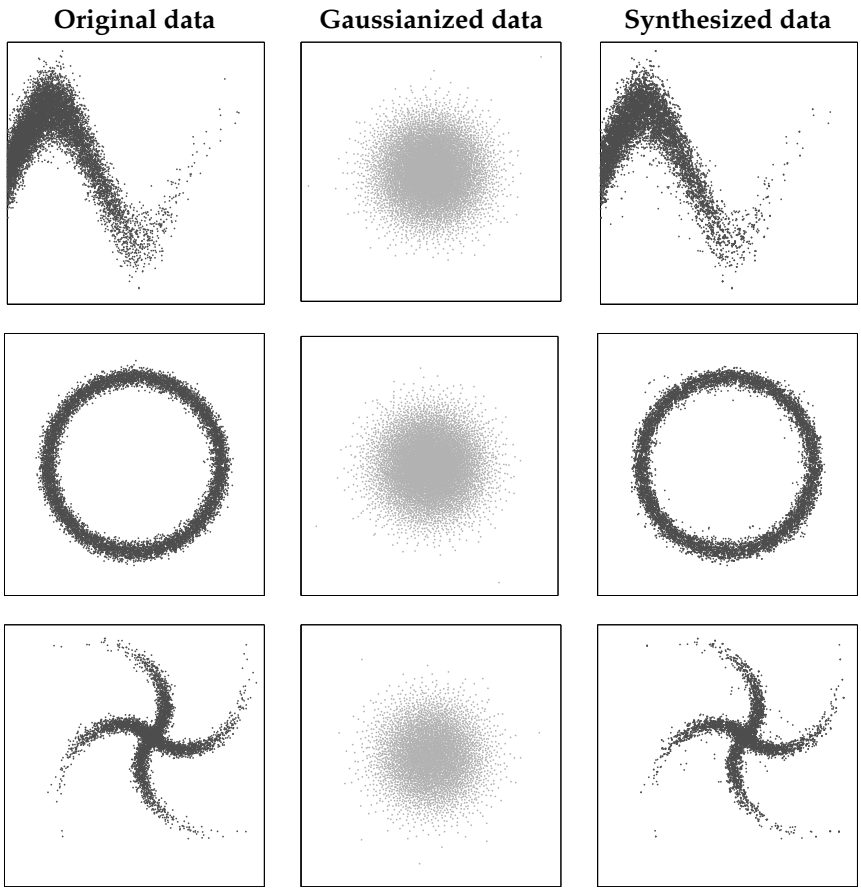


Figure 5.17: Toy data examples synthesized using RBIG.

Figure 5.18: Example of real (top) and synthesized faces with RG (middle) and RBIG (bottom).

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

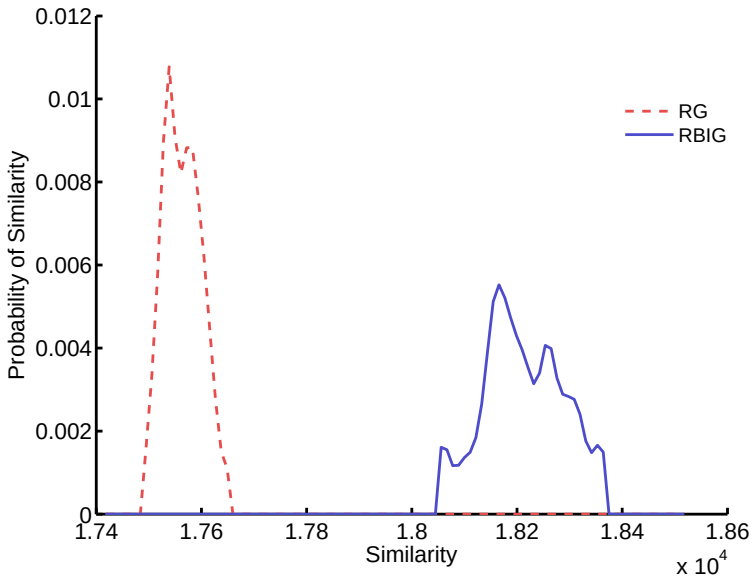


Figure 5.19: Histogram of the similarity (inner product) between the distribution of original and synthesized face images for 300 realizations. For reference, the average image energy (average inner product) in the original set is $1.81 \cdot 10^4$.

increase the urban areas discrimination [Gómez-Chova et al., 2006]. 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 [Tax & Duin, 1999]. 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 [Cohen, 1960]. 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 5.20 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

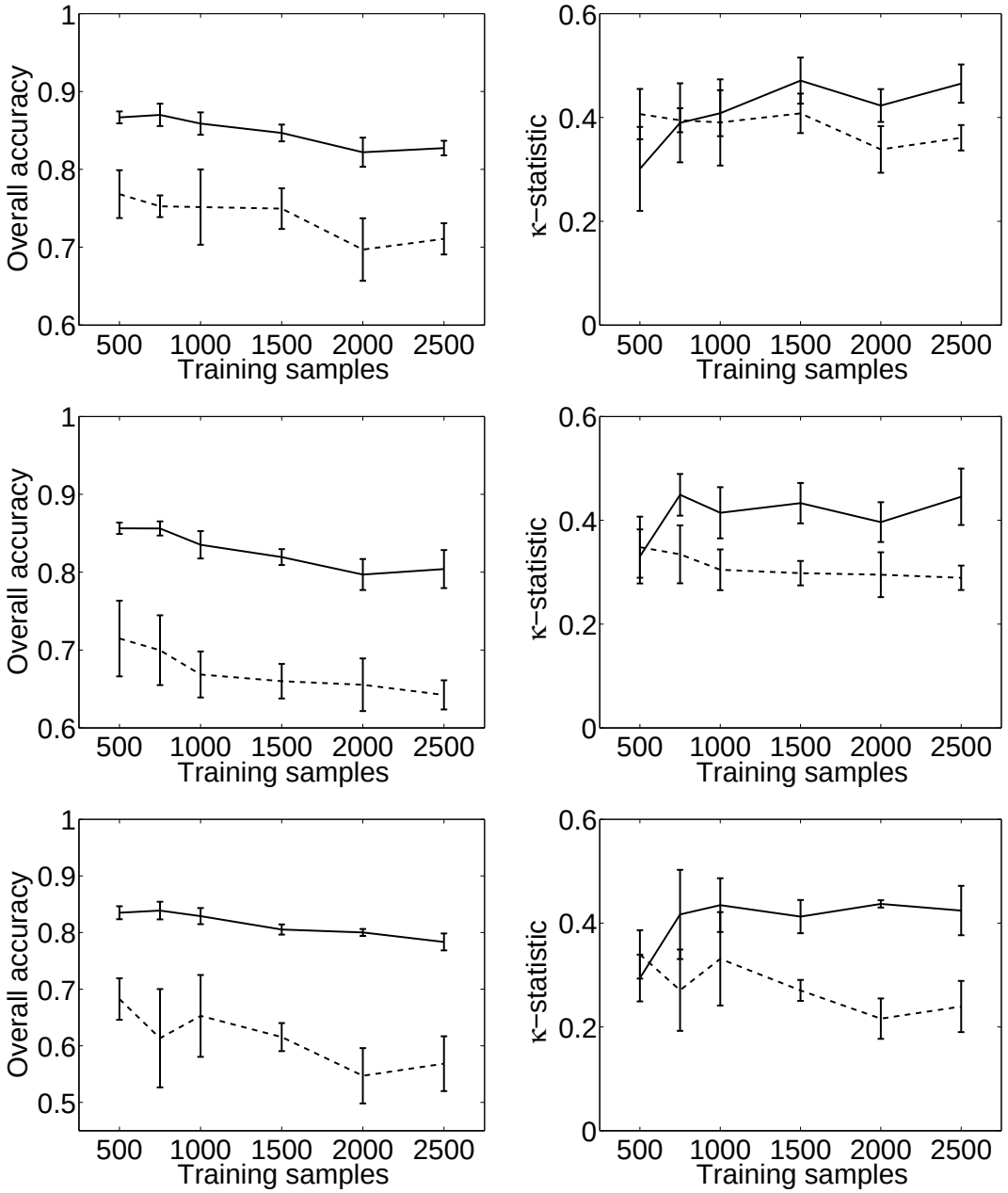


Figure 5.20: Overall accuracy (left) and kappa statistic, κ (right) for RBIG (solid line) and SVDD (dashed line) in different scenes: Naples 1995 (top), Naples 1999 (center) and Rome 1995 (bottom).

non-target class.

Figure 5.21 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

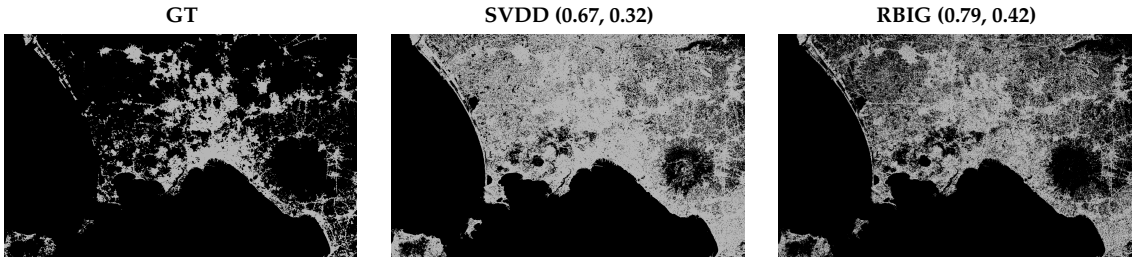


Figure 5.21: Ground-truth (GT) and classification maps obtained with SVDD and RBIG for the Naples 1995 scene. The white points represent urban area and the black points represent non-urban area. The corresponding overall accuracy and κ-statistic are given in parenthesis.

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.

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 image denoising methods have exploited the Bayesian framework to this end [Donoho & Johnstone, 1995; Figueiredo & Nowak, 2001; Portilla et al., 2003; E. P. Simoncelli, 1999]:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}^*}{\operatorname{argmin}} \left\{ \int \mathcal{L}(\mathbf{x}, \mathbf{x}^*) p(\mathbf{x}|\mathbf{x}_n) d\mathbf{x} \right\}, \tag{5.24}$$

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

$$p(\mathbf{x}|\mathbf{x}_n) = Z^{-1} p(\mathbf{x}_n|\mathbf{x}) p(\mathbf{x}), \tag{5.25}$$

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. (5.25) 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 [Olmos & Kingdom, 2004]. To do this, images were transformed using orthonormal QMF wavelet domain with four frequency scales [E. Simoncelli & Adelson, 1990], 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 [Laparra, Gutiérrez, et al., 2010; Liu & Moulin, 2001]. In order to evaluate Eq. (5.24), we sampled the posterior PDF at $8,000$ 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 [Bernardo & Smith, 1994]. Note that the classical hard-thresholding (HT) and soft-thresholding (ST) results [Donoho & Johnstone, 1995] 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 [E. P. Simoncelli, 1999].

Figure 5.22 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) [Wang et al., 2004a]. Note that RBIG method obtains better results (numerically and visually) than the classical methods due to the more accurate PDF estimation.

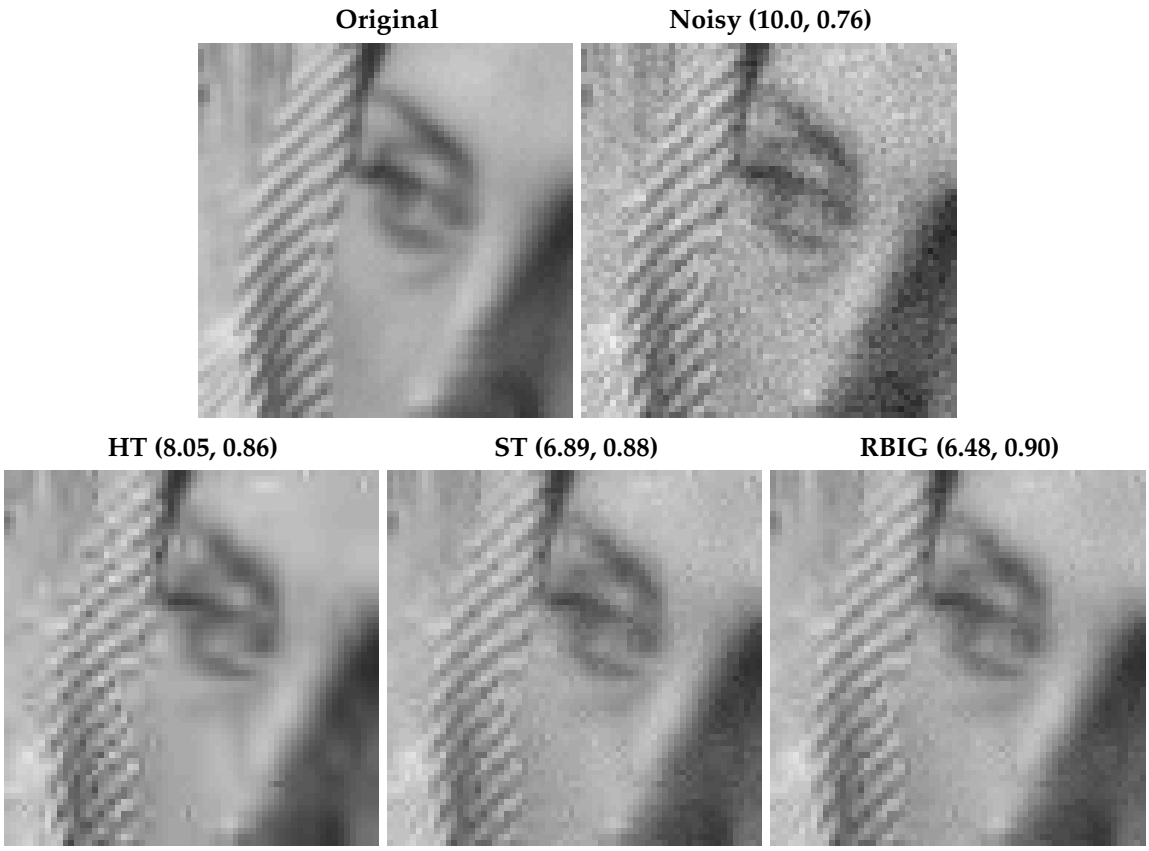


Figure 5.22: Original, noisy (noise variance $\sigma_n^2 = 100$) and restored 'Barbara' images. The root-mean-square-error (RMSE) and the perceptually meaningful Structural Similarity Measure (SSIM) [Wang et al., 2004a] are given in parentheses.

5.3 Chapter conclusions

In section 5.1 we proposed an alternative way to take into account the relations among natural image wavelet coefficients for denoising: we used SVR in the wavelet domain to enforce these relations in the estimated signal. The specific signal relations, which proved to be more relevant in intraband coefficients, are encoded in an anisotropic kernel based on mutual information computed from a representative image database. An adaptive SVR with different cost function *per* subband was developed: the subband-dependent ε_i and C_i are modeled by analyzing the particular signal and noise variances in a representative image database. By following general recommendations for the design of the kernel, ε_i and C_i, and adapting them to the particular image denoising problem, we restricted the class of appropriate SVRs. A KLD-based criterion was proposed to automatically select the SVR that best recovers the relevant wavelet coefficient relations of the true signal. The

criterion was quite consistent but there is still room for improvement, specially in the case of complex noise sources.

Results show that the performance of the proposed method is (1) better than conventional wavelet methods that assume coefficient independence, (2) similar to state-of-the-art methods that do explicitly include these relations when the noise source is Gaussian, and (3) numerically and visually better results are obtained when more complex realistic noise sources are considered. Therefore, the proposed SVR approach can be seen as a more flexible (model-free) alternative to the explicit description of coefficient relations. The important thing here is that no reformulation is needed for dealing with any other kinds of noise. Moreover, these results are an additional indication that relation between local frequency coefficients is a salient natural image feature that should not be neglected in denoising applications.

Future work is tied to the incorporation of new information in the kernels: here we focused on the consideration of signal relations in the kernel, but the particular structure of the noise could be eventually incorporated. Note that the denoising procedure is quite general and admits any kind of regression machine, see Gómez-Chova et al. [2011] for a review in regression kernel methods.

In section 5.2 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.

Chapter 6

Conclusions

THIS chapter summarizes the knowledge I acquired during the last years regarding the issues of this Thesis. For further details, I address the reader to the partial conclusions drawn in each section. Most of the conclusions presented here are focused on confirming the efficient coding hypothesis, which is already a well-established one. In this sense, this conclusions are *just another brick in the wall*.

One of the main conclusions is that the computational model of the HVS designed using physiological information, and fitted in a novel way using psychophysical data, has good statistical properties (chapter 2). Note that no statistical information has been used in the model definition. Which suggests that the HVS takes advantage of the statistical regularities of the visual environment. This is an *unconventional* approach to explain the relation between neuroscience and statistics, departing from neuroscience and studying the statistical properties of the HVS.

Regarding the *conventional* approach (from statistics to neuroscience) some of the HVS behavior has been derived by using only statistical information, i.e. optimizing models statistically through natural images data. The importance of these results falls in the fact that not too complex statistical models and data gathering are needed. The statistical models proposed in sections 4.1 and 4.2 obtain similar behavior to the HVS color and spatial mechanisms, respectively.

Section 4.1 presented a new non-linear method to design a sensory system in which the optimization criterion can be tuned. Results in natural images showed that a statistically optimal sensory system obtains similar behavior to the color mechanisms in the HVS. Moreover, results also suggested that color mechanisms in HVS may be guided by an *error minimization* strategy instead of a *redundancy reduction* strategy. Evidences against the independence goal have been also obtained for the spatial representation mechanisms (Sec. 4.3). Results suggested that the shape of the filters in V1 might not be due to an independence goal, and also is suggested that adaptation or overcompleteness properties are convenient at the linear stage in the V1 area.

Section 4.2 highlighted the need of analyzing the phase statistical relations between coefficients in the statistical models, which is still not well addressed nowadays. Moreover some insights on the phase treatment have been established. These results make also clear that it is necessary to introduce phase properties in the HVS model, which currently takes only the modulus information into account.

Another practical conclusion is that the HVS model of chapter 2 can be used to design image processing applications. Using this model as an image quality metric obtains similar results to the state-of-the-art image quality algorithms (chapter 3), and it could be easily used in other image processing tasks such as image compression or denoising.

Also regarding applications, one conclusion is that including statistical information (and even perceptual information) definitely improves image processing algorithms. In section 5.1, statistical regularities of natural images were included in a machine learning algorithm in order to impose these regularities to noisy images. This algorithm showed very good results (similar to well-established algorithms) in image denoising. The method presented in section 5.2 is able to obtain estimations of multidimensional PDFs circumventing the curse of dimensionality. The method was used in image problems by estimating PDFs of images and applying classical image processing formulas. Good results in a wide range of image processing problems have been obtained, comparable in some cases to those obtained with state-of-the-art algorithms. Results of both methods confirm that understanding higher order statistical relations is essential to improve existing image processing algorithms. Moreover, since these learning algorithms can be seen as inference machines, their behavior could be interpreted in order to explain the brain behavior, as done in section 4.1.

Although the relation between neuroscience and statistics was already well-established, this Thesis has been useful to clarify some ideas to me. All the works presented here highlighted this relation. It is clear to me now that the human brain (not only the HVS) has evolved to process information taking into account the statistical regularities of the world. Otherwise, the brain could not be able to process this amount of information with such high accuracy.

In the case of the HVS, I think these regularities are used for an error minimization goal. This strategy makes more sense to me, and moreover the results obtained in this Thesis make me feel that this is the goal that guide the early stages, instead of the redundancy reduction strategy. Figure 6.2 shows the codebook optimized for quantizing images (extracted from [Gersho & Gray, 1992]). These vectors resemble border filters similar to the filters obtained when optimizing for redundancy reduction, like in [Bell & Sejnowski, 1997; Olshausen & Field, 1996]. Moreover thinking about the brain as an inference machine, decision would be the main task of the HVS. Therefore, as suggested by the *statistical learning theory* [Vapnik, 1995], learning directly the decision function rather than modeling the PDF may be a good strategy for learning decision models. Therefore, this would be also a good strategy to follow in order to obtain some conclusions about the HVS

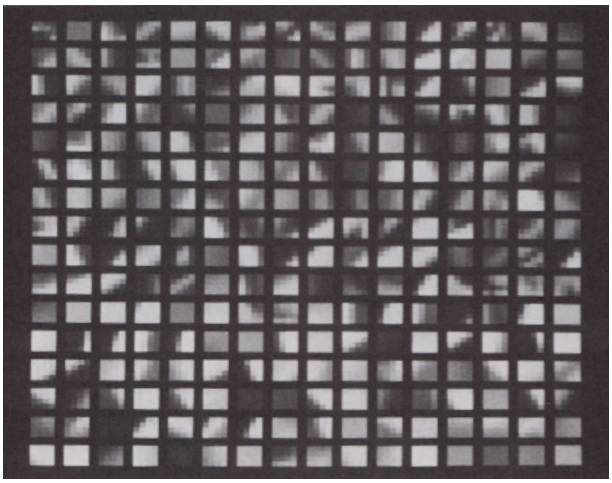


Figure 6.1: Images VQ codebook [Gersho & Gray, 1992].

behavior.

On the other hand, it is also clear to me that, in order to properly develop signal processing algorithms, it is necessary to understand what are the statistical regularities of the data. In the particular case of image processing, since the HVS is optimized statistically and it obtains near optimal performance for some tasks, it is a good reference in order to evaluate object recognition tasks or even more basic tasks as quantization or denoising. In this sense, understanding how the brain works is a field with lots of opportunities and applications.

Conclusiones

En este capítulo se resume el conocimiento relacionado con los temas de la Tesis que he adquirido durante estos años. Para más detalles sugiero al lector que lea las conclusiones parciales que hay en cada sección. La mayoría de las conclusiones presentadas aquí se centran en confirmar la hipótesis de la codificación eficiente, la cuál está actualmente muy bien establecida. En este sentido, estas conclusiones son *just another brick in the wall*.

Una de las conclusiones principales de la Tesis es que el modelo computacional del Sistema Visual Humano (SVH) descrito en el capítulo 2, el cual ha sido diseñado usando información fisiológica y ajustado de una forma novedosa usando datos psicofísicos, tiene unas buenas propiedades estadísticas. Nótese que no se ha usado ninguna información estadística en la definición modelo. Por lo tanto estas propiedades sugieren que el SVH utiliza las regularidades estadísticas del entorno visual. Esta es una aproximación *no convencional* para explicar la relación entre neurociencia y estadística, estudiar las propiedades estadísticas de un modelo de neurociencia.

Respecto a la aproximación *convencional*, desde la estadística a la neurociencia, en esta Tesis se han derivado comportamientos del SVH a partir de información estadística, es decir optimizando modelos estadísticos utilizando datos de imágenes naturales. La importancia de estos resultados recae en el hecho de que no son necesarios ni modelos estadísticos ni datos experimentales muy complejos. Los modelos estadísticos de las secciones 4.1 y 4.2 obtienen un comportamiento similar al de los mecanismos de color y espaciales del SVH, respectivamente. Concretamente, en la sección 4.1 se presenta un método nuevo para diseñar un sistema de sensores no-lineales en los cuales se puede seleccionar la métrica deseada. Los resultados sobre imágenes naturales muestran que los sistemas de sensores óptimos aprendidos obtienen un comportamiento similar al de los mecanismos de color del SVH. Además, estos resultados también sugieren que los mecanismos de color del SVH pueden estar guiados por una estrategia de *minimización de error* en lugar de una estrategia de *reducción de redundancia*. También se han encontrado evidencias contra la estrategia de independencia en la representación de los mecanismos espaciales (Sec. 4.3). Los resultados sugieren que, en términos de codificación eficiente, es necesaria algún tipo de adaptación o sobrecompletitud en la transformada lineal del área V1.

En la sección 4.2 se resalta la idea de la necesidad de analizar las relaciones entre las fases de los coeficientes de los modelos estadísticos de imagen, la cual no está desarrollada hoy en día. Además se introducen algunos nuevos conocimientos sobre aspectos de las fases. Los resultados dejan también clara la necesidad de introducir las propiedades de fase en los modelos del SVH, los cuales normalmente hacen uso únicamente de la información del módulo.

Otra conclusión de tipo práctico es que el modelo de SVH del capítulo 2 puede ser usado para diseñar aplicaciones de procesado de imágenes. Al usar este modelo para la medida de calidad de imágenes se obtienen resultados similares a los de los métodos actuales (capítulo 3). Además podría ser usado fácilmente en otras tareas de procesado de imágenes como compresión. Siguiendo con las aplicaciones, una conclusión importante es que incluir información estadística (e incluso perceptual) mejora los algoritmos de procesado de imagen. En la sección 5.1, se introducen regularidades estadísticas en un algoritmo de aprendizaje máquina para imponer dichas regularidades en imágenes con ruido. Este algoritmo muestra muy buenos resultados en limpieza de imágenes similar a los de algoritmos punteros. El método presentado en la sección 5.2 es capaz de paliar los problemas de la *maldición de la dimensión* y así obtener estimaciones de PDFs multidimensionales. Este método ha sido usado en aplicaciones de imagen haciendo una estimación de PDFs de imágenes y aplicando fórmulas clásicas de procesado de imagen. Los resultados obtenidos en diversos problemas son buenos en general y comparables en algunos casos a los obtenidos por algunos de los mejores algoritmos actuales. Los resultados de ambos métodos confirman que el conocimiento de las relaciones estadísticas de alto orden es esencial a la hora de mejorar los algoritmos de procesado de imágenes. Además, puesto que estos algoritmos pueden ser vistos como *máquinas de inferencia*, su funcionamiento podría ser interpretado para explicar el funcionamiento del cerebro, como se hace con el método propuesto en la sección 4.1.

Aunque la relación entre neurociencia y estadística ya estaba bien establecida, esta Tesis ha sido útil para clarificarme algunas ideas a mi mismo. Todos los trabajos presentados resaltan esta relación. Ahora mismo tengo claro que el cerebro humano (no sólo el SVH) se ha adaptado para procesar información teniendo en cuenta las regularidades estadísticas del mundo. De otro modo, no sería posible procesar la cantidad de información que procesa con la precisión con que lo hace.

En el caso del SVH, creo que dichas regularidades son usadas para minimizar el error de representación. Personalmente, creo que esta estrategia tiene mucho más sentido, y además los resultados obtenidos a lo largo de la Tesis me hacen pensar que este es el criterio que guía las primeras etapas del SVH, en lugar de la reducción de redundancia. La figura 6.2 muestra los representantes obtenidos cuando se optimiza un código de cuantización para imágenes (extraído de [Gersho & Gray, 1992]). Estos vectores son similares a los filtros de bordes obtenidos al optimizar el código de reducción de redundancia, como

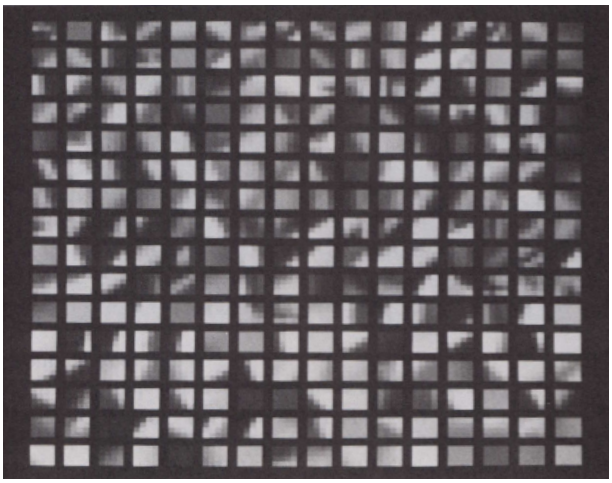


Figure 6.2: Images VQ codebook [Gersho & Gray, 1992].

en [Bell & Sejnowski, 1997; Olshausen & Field, 1996]. Además, pensando en el cerebro como una *maquina de inferencia*, tomar decisiones sería la tarea principal del SVH. Por tanto, como se sugiere en Vapnik [1995], aprender directamente la función de decisión, en lugar de modelar por separado la PDF y la función de coste, puede ser una buena estrategia para el aprendizaje de modelos de decisión. Por tanto, esta podría ser una buena estrategia a seguir para obtener conclusiones a cerca del funcionamiento del SVH.

Por otro lado, también tengo claro que para poder desarrollar buenos algoritmos de procesado de señal, es necesario entender cuales son las regularidades estadísticas de los datos a tratar. En el caso particular del procesado de imagen, puesto que el SVH está optimizado estadísticamente y obtiene un resultado óptimo en algunas tareas, es una buena referencia a a la hora testear tareas de reconocimiento de objetos o incluso tareas más básicas como cuantización o limpieza de ruido. En este sentido, entender cómo funciona el cerebro es un campo con una gran cantidad de posibilidades y aplicaciones.

References

Abrams, A., Hillis, J., & Brainard, D. (2007). The relation between color discrimination and color constancy: When is optimal adaptation task dependent? *Neural Computation*, *19*(10), 2610–2637.

Ahumada, A. (1993). Computational image quality metrics: A review. In J. Morreale (Ed.), *International symposium digest of technical papers* (Vol. 25, pp. 305–308).

Andrews, H., & Hunt, B. (1977). *Digital image restoration*. NY: Prentice Hall Technical Reference.

Armengot, M., Laparra, V., Goómez-Chova, L., Malo, J., & Camps-Valls, G. (2010). Adaptive kernel ridge regression for image denoising. In *Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on* (p. 432 -437).

Atick, J. (1992). Could Information Theory provide an ecological theory of sensory processing? *Network: Computational Neural Systems*, *3*(2), 213–251.

Atick, J., Li, Z., & Redlich, A. (1993). What does post-adaptation color appearance reveal about cortical color representation? *Vision Research*, *33*(1), 123–129.

Attneave, F. (1954). Informational aspects of visual perception. *Psychological Review*, *61*, 183-193.

Banham, M., & Katsaggelos, A. (1997). Digital image restoration. *IEEE Sig.Proc.Mag.*, *14*, 24–41.

Barducci, A., & Pippi, I. (2001). Analysis and rejection of systematic disturbances in hyperspectral remotely sensed images of the Earth. *Applied Optics*, *40*, 1464-1477.

Barlow, H. (1961). Possible principles underlying the transformation of sensory messages. In W. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Barlow, H. B. (2001). Redundancy reduction revisited. *Network: Computational Neural Systems*, *12*, 241–253.

Barnard, K., Martin, L., Funt, B., & Coath, A. (2002). A data set for colour research. *Color Research and Application*, *27*(3), 147–151.

Barten, P. (1990). Evaluation of subjective image quality with the square root integral method. *Journal of the Optical Society of America A*, *7*(10), 2024–2031.

Belkin, M., & Niyogi, P. (2002). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, *15*, 1373–1396.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159.

Bell, A. J., & Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. *Vision Research*, *37*(23), 3327–3338.

Bernardo, J. M., & Smith, A. F. M. (1994). Bayesian theory. *Measurement Science and Technology*, *12*(2), 221.

Bertero, M., Poggio, T., & Torre, V. (1988). Ill-posed problems in early vision. *Proceedings of the IEEE*, *76*(8), 869–889.

Bethge, M. (2006, June). Factorial coding of natural images: how effective are linear models in removing higher-order dependencies? *Journal of the Optical Society of America A*, *23*(6), 1253–1268.

Bingham, E., & Hyvärinen, A. (2000). A fast fixed-point algorithm for independent component analysis of complex valued signals. *International Journal of Neural Systems*, *10*, 1–8.

Bishop, C. M., Svensén, M., & Williams, C. K. I. (1998). GTM: The generative topographic mapping. *Neural Computation*, *10*, 215–234.

Brainard, D. H., Rutherford, M. D., & Kraft, J. M. (2000). *Hyperspectral image data base.* `http://color.psych.ucsb.edu/hyperspectral`.

Brand, M. (2003). Charting a manifold. In *Neural Information Processing Systems 15* (pp. 961–968). MIT Press.

Breneman, E. J. (1987, June). Corresponding chromaticities for different states of adaptation to complex visual fields. *Journal of the Optical Society of America A*, *4*, 1115-1129.

Brown, R. O. (1994). The world is not grey. *Investigative Ophthalmology & Visual Science*, *35*(4), 2165.

Buccigrossi, R. W., & Simoncelli, E. P. (1999). Image compression via joint statistical characterization in the wavelet domain. *IEEE Transactions on Image Processing*, *8*(12), 1688–1701.

Buchsbaum, G., & Gottschalk, A. (1983). Trichromacy, opponent colours coding and optimum colour information transmission in the retina. *Proceedings of the Royal Society B*, *220*(1218), 89-113.

Burges, C. J. C. (1999). Geometry and invariance in kernel based methods. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods— support vector learning* (p. 89-116). Cambridge, (MA): M.I.T. Press.

Burt, P., & Adelson, E. (1983). The Laplacian pyramid as a compact image code. *IEEE Transaction on Communications*, *31*, 532–540.

Cadieu, C. (2009). *Probabilistic models of phase variables for visual representation and neural dynamics*. Unpublished doctoral dissertation, UC Berkeley. Available from `http://redwood.berkeley.edu/cadieu/homepage/Home.html`

Campbell, F., & Robson, J. (1968). Application of Fourier analysis to the visibility of gratings. *Journal of Physiology*, *197*, 551–566.

Camps-Valls, G., Gutiérrez, J., Gómez, G., & Malo, J. (2008). On the suitable domain for SVM training in image coding. *Journal of Machine Learning Research*, *9*, 49–66.

Camps-Valls, G., Laparra, V., Gómez-Chova, L., Muñoz-Marí, J., & Calbet, X. (2011). Kernel-based retrieval of atmospheric profiles from iasi data. In *Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International.*

Camps-Valls, G., Soria-Olivas, E., Pérez-Ruixo, J., Artés-Rodriguez, A., Pérez-Cruz, F., & Figueiras-Vidal, A. (2001, December). A profile-dependent kernel-based regression for cyclosporine concentration prediction. In *Neural information processing systems – Workshop on new directions in kernel-based learning methods.* Vancouver, Canada.

Camps-Valls, G., Tuia, D., Laparra, V., & Malo, J. (2010, july). Estimating biophysical variable dependences with kernels. In *Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International* (p. 828 -831).

Capilla, P., Díez, M., Luque, M. J., & Malo, J. (2004). Corresponding-pair procedure: a new approach to simulation of dichromatic color perception. *Journal of the Optical Society of America A*, *21*(2), 176–186. Available from `http://JournaloftheOpticalSocietyofAmericaa.osa.org/abstract.cfm?URI=JournaloftheOpticalSocietyofAmericaa-21-2-176`

Capilla, P., Malo, J., Luque, M., & Artigas, J. (1998). Colour representation spaces at different physiological levels: A comparative analysis. *Journal of Optics*, *29*, 324–338.

Carandini, M., & Heeger, D. (1994). Summation and division by neurons in visual cortex. *Science*, *264*(5163), 1333–6.

Carandini, M., Heeger, D. J., & Movshon, J. A. (1997, November 1). Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, *17*(21), 8621–8644.

Cardoso, J.-F. (2003). Dependence, correlation and gaussianity in independent component analysis. *Journal of Machine Learning Research*, *4*, 1177–1203.

Chalimourda, A., Schölkopf, B., & Smola, A. J. (2004). Experimentally optimal ν in support vector regression for different noise models and parameter settings. *Neural Networks*, *17*(1), 127–141.

Chandler, D., & Hemami, S. (2007). VSNR: A wavelet based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, *16*(9), 2284-2298.

Chang, C.-C., & Lin, C.-J. (2001a). LIBSVM: a library for support vector machines [Computer software manual]. (Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`)

Chang, C. C., & Lin, C. J. (2001b). LIBSVM: a library for support vector machines, *(http://www.csie.ntu.edu.tw/~cjlin/libsvm/)* [Computer software manual]. Available from `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

Chen, S., & Gopinath, R. (2000). Gaussianization. In *Neural information processing systems* (p. 423-429).

Cheng, H., Tian, J., Liu, J., & Yu, Q. (2004). Wavelet domain image denoising via SVR. *IEE Electronics Letters*, *40*.

Cherkassky, V. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks*, *17*(1), 113–126.

Cherkassky, V., & Ma, Y. (2003). Comparison of model selection for regression. *Neural Computation*, *15*(7), 1691–1714.

Ciurea, F., & Funt, B. (2003). A large image database for color constancy research. In *Proc. 11th color imaging conf.* (p. 160-164). IS&T and SID.

Clarke, R. J. (1985). *Transform coding of images*. New York: Academic Press.

Cohen, J. (1960, April). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Coifman, R. R., & Donoho, D. L. (1995). Translation-invariant de-noising. In A. Antoniadis & G. Oppenheim (Eds.), *Wavelets and statistics. lecture notes in statistics* (Vol. 103, pp. 125–150). Department of Statistics: Springer, Berlin.

Cole, G. R., Stromeyer, C. F., & Kronauer, R. E. (1990). Visual interactions with luminance and chromatic stimuli. *Journal of the Optical Society of America A*, *7*, 128–140.

Comon, P. (1994). Independent component analysis: A new concept? *Signal Processing*, *36*(3), 287–314.

Cover, T., & Tomas, J. (1991). *Elements of information theory*. New York: John Wiley & Sons.

Daly, S. (1990). Application of a noise-adaptive Contrast Sensitivity Function to image data compression. *Optical Engineering*, *29*(8), 977–987.

Darwin, C. (1859). *The origin of species* (J. Murray, Ed.).

Daugman, J. (1980). Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, *20*, 847–856.

Daugman, J. G. (1993, February 1). Quadrature-phase simple-cell pairs are appropriately described in complex analytic form. *Journal of the Optical Society of America A*, *10*(2), 375–377.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems* (1st ed.). MIT Press.

Delicado, P. (2001). Another look at principal curves and surfaces. *Journal of Multivariate Analysis*, *77*, 84–116.

Doi, E., Inui, T., Lee, T., Wachtler, T., & Sejnowski, T. (2003). Spatiochromatic receptive field properties derived from information-theoretic analysis of cone mosaic responses to natural scenes. *Neural Computation*, *15*(2), 397–417.

Donoho, D., & Johnstone, I. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, *90*, 432.

D. Ruderman, T. C., & Chiao, C. (1998). Statistics of cone responses to natural images: implications for visual coding. *Journal of the Optical Society of America A*, *15*(8), 2036–2045.

Dubrovin, B., Novikov, S., & Fomenko, A. (1982). Modern geometry: Methods and applications. In (chap. 3: *Algebraic Tensor Theory*). New York: Springer Verlag.

Duda, R., & Hart, P. (1973). *Pattern classification and scene analysis*. New York: John Wiley & Sons.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification (2nd edition)* (2nd ed.). Wiley-Interscience.

Eichhorn, J., Sinz, F., & Bethge, M. (2009, April). Natural image coding in V1: how much use is orientation selectivity? *PLoS computational biology*, *5*(4).

Einbeck, J., Tutz, G., & Evers, L. (2005). Local principal curves. *Statistics and Computing*, *15*, 301-313.

Epifanio, I., Gutiérrez, J., & J.Malo. (2003). Linear transform for simultaneous diagonalization of covariance and perceptual metric matrix in image coding. *Pattern Recognition*, *36*, 1799–1811.

Erdogmus, D., Jenssen, R., Rao, Y., & Principe, J. (2006). Gaussianization: An efficient multivariate density estimation technique for statistical signal processing. *J. VLSI Sig. Proc.*, *45(17)*, 67-83.

Fairchild, M. (1997). *Color appearance models*. New York: Addison-Wesley.

Fairchild, M. (2005). *Color appearance models, 2nd ed.* Chichester, UK: Wiley-IS&T.

Fairchild, M. D. (1996). Refinement of the RLAB color space. *Color Research and Applications*, *21*, 338–346.

Feller, W. (1968). *An introduction to probability theory and its applications, vol. 1* (3rd ed.). Wiley. Hardcover.

Field, D. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, *4*(12), 2379–2394.

Figueiredo, M., & Nowak, R. (2001). Wavelet-based image estimation: an empirical Bayes approach using Jeffrey's noninformative prior. *IEEE Transactions on Image Processing*, *10*(9), 1322-1331.

Foldiak, P. (1989, Jun). Adaptive network for optimal linear feature extraction. In *International joint conference on neural networks.* (Vol. 1, p. 401-405 vol.1).

Foley, J. (1994). Human luminance pattern mechanisms: Masking experiments require a new model. *Journal of the Optical Society of America A*, *11*(6), 1710-1719.

Freeman, W. T., & Adelson, E. H. (1991). The design and use of steerable filters. *IEEE Pattern Analysis and Machine Intelligence*, *13*(9), 891-906.

Friedman, J., & Tukey, J. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, *C-23*(9), 881-890.

Funt, B. V., & Drew, M. S. (1993). Color space analysis of mutual illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(12), 1319–1326.

Gegenfurtner, K., & Kiper, D. (1992, Nov.). Contrast detection in luminance and chromatic noise. *Journal of the Optical Society of America A*, *9*(11), 1880-1888.

Georghiades, A. S., Belhumeur, P. N., & Kriegman, D. J. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*, 643–660.

Gersho, A., & Gray, R. M. (1992). *Vector quantization and signal compression*. Boston: Kluwer Academic Press.

Geusebroek, J., Burghouts, G., & Smeulders, A. (2005). The Amsterdam library of object images. *Int. J. Comput. Vision*, *61*(1), 103–112.

Giesel, M., Hansen, T., & Gegenfurtner, K. (2009). The discrimination of chromatic textures. *J. Vision*, *9*(9), 1–28.

Ginneken, B. van, & Mendrik, A. (2006). Image denoising with *k*-nearest neighbor and support vector regression. In *International Conference on Pattern Recognition* (Vol. 3, p. 603-606). Hong Kong.

Goda, N., Koida, K., & Komatsu, H. (2009). Colour representation in lateral geniculate nucleus and natural colour distributions. *Lecture Notes in Computer Science*, *5646*, 23–30.

Golub, G. H., & Loan, C. F. van. (1996). *Matrix computations* (3rd ed.). The Johns Hopkins University Press.

Gómez, G., Camps-Valls, G., Gutiérrez, J., & Malo, J. (2005). Perceptual adaptive insensitivity for support vector machine image coding. *IEEE Transactions on Neural Networks*, *16*(6), 1574-1581.

Gómez-Chova, L., Fernández-Prieto, D., Calpe, J., Soria, E., Vila-Francés, J., & Camps-Valls, G. (2006, Mar). Urban monitoring using multitemporal SAR and multispectral data. *Pattern Recognition Letters*, *27*(4), 234-243. (3rd Pattern Recognition in Remote Sensing Workshop, Kingston Upon Thames, ENGLAND, AUG 27, 2004)

Gómez-Chova, L., Muñoz-Marí, J., Laparra, V., Malo-López, J., & Camps-Valls, G. (2011). *Optical remote sensing. advances in signal processing and exploitation techniques* (S. Prasad, L. M. Bruce, & J. Chanussot, Eds.). Heidelberg, Germany: Springer.

Goossens, B., Pizurica, A., & Philips, W. (2009). Removal of correlated noise by modeling the signal of interest in the wavelet domain. *IEEE Transactions on Image Processing*, *18*(6), 1153-1165.

Group, T. V. (2008). *Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, phase i* (Tech. Rep. No. 2.6). Video Quality Experts Group. Available from `http://www.its.bldrdoc.gov/vqeg/projects/multimedia/`

Gutiérrez, J., Ferri, F., & Malo, J. (2006). Regularization operators for natural images based on nonlinear perception models. *IEEE Transactions on Image Processing*, *15*(1), 189–200.

Hansen, T., Giesel, M., & Gegenfurtner, K. (2008). Chromatic discrimination of natural objects. *J. Vision*, *8*(1), 1–19.

Hastie, T., & Stuetzle, W. (1989). Principal curves. *J. Am. Stats. Assoc.*, *84*(406), 502-516.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2003). *The elements of statistical learning*. Springer. Hardcover.

Hateren, J. van, & Schaaf, A. van der. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings Royal Society of London*, *265*, 359–366.

Haykin, S. (2002). *Adaptive filter theory* (4th ed.). New Jersey, USA: Prentice-Hall.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*, 181–198.

Hillis, J. M., & Brainard, D. (2005). Do common mechanisms of adaptation mediate color discrimination and appearance? Uniform backgrounds. *Journal of the Optical Society of America A*, *22*(10), 2090–2106.

Hinton, G. E., & Salakhutdinov, R. R. (2006, July). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Huang, T.-M., & Kecman, V. (2004, April). Bias term *b* in SVMs again. In *12th european symposium on artificial neural network, ESANN 2004* (p. 441-448). Bruges, Belgium.

Huber, P. (1985). Projection pursuit. *Annals of Statistics*, *13*(2), 435-475.

Hyvärinen, A. (1999a). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neur. Nets.*, *10*(3), 626–634.

Hyvärinen, A. (1999b). Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation. *Neural Computation*, 1739–1768.

Hyvärinen, A., & Hoyer, P. (2000, July 1). Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural computation*, *12*(7), 1705–1720.

Hyvärinen, A., & Köster, U. (2007, June). Complex cell pooling and the statistics of natural images. *Network*, 1–20.

Hyvärinen, A., & Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, *12*(3), 429–439.

Ingling, C., & Tsou, B. (1977). Orthogonal combination of the three visual channels. *Vision Research*, *17*, 1075–1082.

J., V. M., R., G., & S., I. L. (1994). Measurement and analysis of object reflectance spectra. *Color Research and Application*, *19*, 4–9.

Jebara, T., Kondor, R., & Howard, A. (2004). Probability product kernels. *Journal of Machine Learning Research*, *5*, 819-844.

J. Eriksson, A. S., & Koivunen, V. (2005). Complex ICA for circular and non-circular sources. In *in proc. EUSIPCO.*

Jolliffe, I. T. (1986). *Principal component analysis*. Springer-Verlag.

Kai Tick Chow, D., & Lee, T. (2001). Image approximation and smoothing by support vector regression. In *International joint conference on neural networks, ijcnn'01.* (Vol. 4, p. 2427-2432). Washington, DC, USA.

Kambhatla, N., & Leen, T. (1997). Dimension reduction by local PCA. *Neural Computation*, *9*, 1493-1500.

Kayser, C., Kording, K. P., & Konig, P. (2003). Learning the nonlinearity of neurons from natural visual stimuli. *Neural Computation*, *15*, 1751–1759.

K. Dabov, V. K., A. Foi, & Egiazarian, K. (2007). Image denoising by sparse 3D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, *16*, 2080-2095.

Kecman, V., Huang, T. ming, & Vogt, M. (2004). Iterative single data algorithm for training kernel machines from huge data sets: Theory and. In *Performance, support vector machines: Theory and applications, springer-verlag,.studies in fuzziness and soft computing* (pp. 255–274).

Kervrann, C., & Boulanger, J. (2007, Feb). Local adaptivity to variable smoothness for exemplar-based image denoising and representation. *International Journal of Computer Vision*, *16*(2), 349-366.

Kingsbury, N. (2006). Rotation-invariant local feature matching with complex wavelets. In *Proceedings European Conference on Signal Processing*. Florence, Italy.

Koenderink, J. J. (2010). The prior statistics of object colors. *Journal of the Optical Society of America A*, *27*(2), 206–217.

Kohonen, T. (1982, January 1). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69.

Krauskopf, J., & Gegenfurtner, K. (1992). Color discrimination and adaptation. *Vision Research*, *32*(11), 2165-75.

Kwok, J. T., & Tsang, I. W. (2003, May). Linear dependency between ε and the input noise in ε-Support Vector Regression. *IEEE Transactions on Neural Networks*, *2130/2001*, 544-553.

Lagarias, J., Reeds, J., Wright, M., & Wright, P. (1998). Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM Journal of Optimization*, *9*(1), 112-147.

Laming, D. (1997). *The measurement of sensation (oxford psychology series)* (1st ed., Vol. 30). Oxford Medical Publications.

Laparra, V., & Bethge, M. (2011). *Redundancy reduction of linear transforms on image textures* (Tech. Rep.). In preparation.

Laparra, V., Camps, G., & Malo, J. (2009). PCA gaussianization for image processing. In *Proceedings of the IEEE International Conference on Image Processing.* IEEE.

Laparra, V., Camps-Valls, G., & Malo, J. (2011). Iterative Gaussianization: from ICA to Random Rotations. *IEEE Transactions on Neural Networks*, *22:4*, 537 - 549.

Laparra, V., Gutierrez, J., Camps-Valls, G., & Malo, J. (2008). Recovering wavelet relations using svm for image denoising. In *Proceedings of the IEEE International Conference on Image Processing* (p. 541 -544).

Laparra, V., Gutiérrez, J., Camps-Valls, G., & Malo, J. (2010, March). Image denoising with kernels based on natural image relations. *Journal of Machine Learning Research*, *11*, 873–903.

Laparra, V., Gutman, M., Malo, J., & Hyvärinen, A. (2011). Complex-valued independent component analysis of natural images. In *International Conference on Artificial Neural Networks.*

Laparra, V., Jiménez, S., Camps-Valls, G., & Malo, J. (2011a). Nonlinearities and adaptation of color vision from sequential principal curves analysis. *Neural Computation (submitted).*

Laparra, V., Jiménez, S., Camps-Valls, G., & Malo, J. (2011b). *Sequential principal curves analysis with local metric* (Tech. Rep.). Image Processing Lab, Universitat de Valencia. Tech. Report IPL, Univ. Valencia. `http://isp.uv.es/docs/spca_techrep_v1.pdf`.

Laparra, V., & Malo, J. (2008a). Color and luminance discrimination by non-linear PCA. In *Proceedings of the 1st Computational Vision Neuroscience Symposium.* Tubingen, Germany: MPI for Biological Cybernetics.

Laparra, V., & Malo, J. (2008b). Masking-like non-linearities from non-linear PCA. In *GRC Conference: Sensory Coding and The Natural Environment.* Lucca, Italy.

Laparra, V., Marí, J. Muñoz, & Malo, J. (2010). Divisive normalization image quality metric revisited. *Journal of the Optical Society of America A*, *27*(4), 852–864.

Laparra, V., Muñoz-Marí, J., Camps-Valls, G., & Malo, J. (2009, September). PCA Gaussianization for one-class remote sensing image classification. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* (Vol. 7477).

Laparra, V., Tuia, D., Jiménez, S., Camps-Valls, G., & Malo, J. (2011). Principal polynomial analysis for remote sensing data processing. In *Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International.*

Laplace, P. S. M. de. (1814). *A philosophical essay on probabilities*. New York : J. Wiley ; London : Chapman & Hall.

Laughlin, S. B. (1983). Matching coding to scenes to enhance efficiency. In *In Braddick, o.j. & sleigh, a.c. (eds) Physical and Biological Processing of Images* (pp. 42–52). Springer.

Learned, E., & Fisher, J. (2003). ICA using spacings estimates of entropy. *Journal of Machine Learning Research*, *4*, 1271-1295.

Le Callet, P., & Autrusseau, F. (2005). *Subjective quality assessment IRCCyN/IVC database.* (http://www.irccyn.ec-nantes.fr/ivcdb/)

Lee, T., Girolami, M., Bell, A., & Sejnowski, T. (2000). A unifying information-theoretic framework for ICA. *Comp. Math. Appl.*, *39*(11), 1–21.

Lehrer, J. (2010). The truth wears off: Is there something wrong with the scientific method? *The New Yorker, Annals of Science.*, *December 13*.

León, C. A., Massé, J.-C., & Rivest, L.-P. (2006). A statistical model for random rotations. *J. Multivar. Anal.*, *97*(2), 412–430.

Lewis, R. M., & Torczon, V. (2002). A globally convergent augmented Lagrangian pattern search algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Optimization*, *12*(4), 1075–1089.

Linsker, R. (1986). From basic network principles to neural architecture: emergence of orientation-selective cells. *Proceedings of the National Academy of Sciences of the United States of America*, *83*(21), 8390-8394.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, *21*(3), 105–117.

Liu, J., & Moulin, P. (2001). Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients. *IEEE Transactions on Image Processing*, *10*, 1647-1658.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137.

Lubin, J. (1993). *The Use of Psychophysical Data and Models in the Analysis of Display System Performance*. In A. Watson (Ed.), *Digital images and human vision* (pp. 163–178). Massachusetts: MIT Press.

Luo, M., Clarke, A., Rhodes, P., Scrivener, S., Schappo, A., & Tait, C. (1991). Quantifying colour appearance. Part I. LUTCHI colour appearance data. *Color Research and Application*, *16*, 166-180.

Luo, M., & Rhodes, P. (1999). Corresponding-colour datasets. *Color Research and Application*, *24*(4), 295-296.

Luo, M. R., Lo, M., & Kuo, W. (1996). The LLAB colour model. *Color Research and Application*, *21*, 412–429.

Lyu, S., & Simoncelli, E. P. (2009, May). Nonlinear extraction of 'independent components' of natural images using radial Gaussianization. *Neural Computation*, *21*(6), 1485-1519.

MacLeod, D., & Boynton, R. (1979). Chromaticity diagram showing cone excitation by stimuli of equal luminance. *Journal of the Optical Society of America*, *69*, 1183–1186.

MacLeod, D., & Twer, T. von der. (2003). The pleistochrome: optimal opponent codes for natural colours. In D. Heyer & R. Mausfeld (Eds.), *Colour perception: From light to object.* London, UK: Oxford University Press.

MacLeod, D. A. (2003, May). Colour discrimination, colour constancy, and natural scene statistics. In J. Mollon, J. Pokorny, & K. Knoblauch (Eds.), *Normal and defective colour vision* (p. 189-218). London, UK: Oxford University Press.

Malo, J., Epifanio, I., Navarro, R., & Simoncelli, E. (2006). Non-linear image representation for efficient perceptual coding. *IEEE Transactions on Image Processing*, *15*(1), 68–80.

Malo, J., & Gutiérrez, J. (2006). V1 non-linear properties emerge from local-to-global non-linear ICA. *Network: Comp. Neur. Syst.*, *17*(1), 85–102.

Malo, J., & Laparra, V. (2010a). Psychophysically tuned divisive normalization approximately factorizes the pdf of natural images. *Neural Computation*, *22*, 3179–3206.

Malo, J., & Laparra, V. (2010b). Visual cortex performs a sort of non-linear ica. In J. Sole-Casals & V. Zaiats (Eds.), *Advances in nonlinear speech processing* (Vol. 5933, p. 17-25). Springer Berlin / Heidelberg.

Malo, J., & Luque, M. J. (2000). *COLORLAB: A Matlab toolbox for color science and color image processing.* Available in `http://isp.uv.es`.

Malo, J., Pons, A., & Artigas, J. (1997). Subjective image fidelity metric based on bit allocation of the human visual system in the DCT domain. *Image & Vision Computing*, *15*(7), 535–548.

Malo, J., Pons, A., Felipe, A., & Artigas, A. (1997). Characterization of human visual system threshold performance by a weighting function in the Gabor domain. *Journal of Modern Optics*, *44*(1), 127–148.

Maloney, L. (1986). Evaluation of linear models of surface spectral reflectance with small numbers of parameters. *Journal of the Optical Society of America A*, *3*(10), 1673–1683.

Martínez-Uriegas, E. (1997). Color detection and color contrast discrimination thresholds. In *Proceedings of the OSA Annual Meeting ILS–XIII* (p. 81). Los Angeles.

Mercer, J. (1905, May). Functions of positive and negative type and their connection with the theory of integral equations. *Phylosophical Transactions of the Royal Society of London*, *CCIX*(A456), 215-228.

Miller, G. (1955, June). Note on the bias of information estimates. *Information Theory in Psychology*, *II-b*, 95-100.

Moddemeijer, R. (1989). On estimation of entropy and mutual information of continuous distributions. *Signal Processing*, *16*(3), 233–246.

Moroney, N., Fairchild, M. D., Hunt, R. W. G., Li, C., Luo, M. R., & Newman, T. (2002). The CIECAM02 color appearance model. In *Color Imaging Conference* (pp. 23–27).

Motoyoshi, I., Nishida, S., Sharan, L., & Adelson, E. (2007, April). Image statistics and the perception of surface qualities. *Nature*, *447*, 206–209.

Mouroulis, P., Green, R. O., & Chrien, T. G. (2000). Design of pushbroom imaging spectrometers for optimum recovery of spectroscopic and spatial information. *Applied Optics*, *39*, 2210-2220.

Mullen, K. T. (1985). The CSF of human colour vision to red-green and yellow-blue chromatic gratings. *J. Physiol.*, *359*, 381–400.

Nascimento, S., Ferreira, F., & Foster, D. (2002). Statistics of spatial cone-excitation ratios in natural scenes. *Jornal of the Optical Society of America A*, *19*, 1484–1490.

Navia-Vázquez, A., Pérez-Cruz, F., Artés-Rodríguez, A., & Figueiras-Vidal., A. (2001). Weighted least squares training of support vector classifiers leading to compact and adaptive schemes. *IEEE Transactions on Neural Networks*, *12*, 1047-1059.

Nill, N. (1985). A visual model weighted cosine transform for image compression and quality assessment. *IEEE Transactions on Communications*, *33*, 551–557.

Novey, M., & Adali, T. (2008). Complex ICA by negentropy maximization. *Neural Networks, IEEE Transactions on*, *19*(4), 596 -609.

Olmos, A., & Kingdom, F. (2004). *The McGill calibrated colour image database.* `http://tabby.vision.mcgill.ca`.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607–609.

Oppenheim, A. V., & Lim, J. S. (1981). The importance of phase in signals. *Proceedings of the IEEE*, *69*(5), 529–541.

Pan, S., Tsang, I., Kwok, J., & Yang, Q. (2009). Domain adaptation via transfer component analysis. In *Proceedings 21st IJCAI.* Paris, France.

Parraga, C., Brelstaff, G., Troscianko, T., & Moorhead, I. (1998). Color and luminance information in natural scenes. *Journal of the Optical Society of America A*, *15*(3), 563–569.

Parraga, C., Vazquez, J., & Vanrell, M. (2009). A new cone activation-based natural image dataset. *Perception (Suppl.)*, *36*, 180.

Pérez-Cruz, F., Navia-Vázquez, A., Alarcón-Diana, P. L., & Artés-Rodríguez, A. (2000, September). An IRWLS procedure for SVR. In *Proceedings of the EUSIPCO.* Tampere, Finland.

Platt, J. C. (1999). Sequential minimal optimization: A fast algorithm for training suppor vector machines. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods— support vector learning* (p. 185-208). Cambridge, (MA): M.I.T. Press.

Pollen, D. A., & Ronner, S. F. (1981, June 19). Phase relationships between adjacent simple cells in the visual cortex. *Science*, *212*(4501), 1409–1411.

Ponomarenko, N., Carli, M., Lukin, V., Egiazarian, K., Astola, J., & Battisti, F. (2008, Oct.). Color image database for evaluation of image quality metrics. *Proc. Int. Workshop on Multimedia Signal Processing*, 403-408.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal on Computer Vision*, *40*(1), 49–71.

Portilla, J., Strela, V., Wainwright, M., & Simoncelli, E. (2003). Image denoising using a scale mixture of Gaussians in the wavelet domain. *IEEE Transactions on Image Processing*, *12*(11), 1338–1351.

Pratt, W. (1991). Digital image processing. In (chap. 3: *Photometry and Colorimetry*). New York: John Wiley & Sons.

Rao, R. P., & Ballard, D. H. (1999, January). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*(1), 79–87.

Robertson, A. R. (1977). The CIE 1976 color-difference formulae. *Color Res. Appl.*, *2*, 7–11.

Rodríguez-Martínez, E., Goulermas, J., Mu, T., & Ralph, J. (2010). Automatic induction of projection pursuit indices. *IEEE Transactions on Neural Networks*, *21*(8), 1281 -1295.

Romero, J., García, J., Jiménez, L., & Hita, E. (1993). Evaluation of color discrimination ellipsoids in two color spaces. *Journal of the Optical Society of America A*, *10*(5), 827–837.

Roweis, S. T., & Saul, L. K. (2000, December). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323–2326.

Roweis, S. T., Saul, L. K., & Hinton, G. E. (2002). Global coordination of local linear models. In *Advances in neural information processing systems 14* (pp. 889–896). MIT Press.

Saghri, L., Cheatheam, P., & Habibi, A. (1989). Image quality measure based on a human visual system model. *Optical Engineering*, *28*(7), 813–819.

Sanger, T. (1989). Optimal unsupervised learning in a single-layer network. *Neural Networks*, *2*, 459–473.

Sanger, T. D. (1990). Analysis of the two-dimensional receptive fields learned by the generalized hebbian algorithm in response to random input. *Biological cybernetics*, *63*, 221-228.

Schölkopf, B., Smola, A. J., Williamson, R. C., & Bartlett, P. L. (2000, May). *New support vector algorithms* (Vol. 12; Tech. Rep.). Cambridge, MA, USA.

Schwartz, O., & Simoncelli, E. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience*, *4*(8), 819–825.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization (wiley series in probability and statistics)*. Wiley-Interscience.

Seim, T., & Valbereg, A. (1986). Towards a uniform color space: a better formula to describe the Munsell and OSA color scales. *Color Res. Appl.*, *11*(1), 11–24.

Seshadrinathan, K., & Bovik, A. (2008). Unifying analysis of full reference image quality assessment. In *IEEE International Conference Image Processing* (p. 1200-1203).

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 373–423.

Sharma, A., & Paliwal, K. (2007, July). Fast principal component analysis using fixed-point algorithm. *Pattern Recognition Letters*, *28*(10), 1151–1155.

Sheikh, H., & Bovik, A. (2006, Feb). Image information and visual quality. *IEEE Transactions on Image Processing*, *15*(2), 430-444.

Sheikh, H., Bovik, A., & Veciana, G. de. (2005, Dec). An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, *14*(12), 2117-2128.

Sheikh, H., Sabir, M., & Bovik, A. (2006, November). A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, *15*(11), 3440-3451.

Sheikh, H., Z.Wang, Cormack, L., & Bovik, A. (2006). LIVE image quality assessment database. Available from `http://live.ece.utexas.edu/research/quality`

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

Simei, L., & Simoncelli, E. (2007). Statistical modeling of images with fields of GSMs. In (Vol. 19).

Simoncelli, E. (1997). Statistical models for images: Compression, restoration and synthesis. In *31st Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA.*

Simoncelli, E. (2003). Vision and the statistics of the visual environment. *Current Opinion in Neurobiology*, *13*, 144–149.

Simoncelli, E., & Adelson, E. (1990). Subband image coding. In J. Woods (Ed.), (pp. 143–192). Norwell, MA: Kluwer Academic Publishers.

Simoncelli, E., & Freeman, W. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proc 2nd ieee int'l conference on image processing.*

Simoncelli, E., Freeman, W., Adelson, E., & Heeger, D. (1992). Shiftable multi-scale transforms. *IEEE Transactions on Information Theory*, *38*(2), 587-607.

Simoncelli, E., & Olshausen, B. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193–1216.

Simoncelli, E. P. (1999, Spring). Bayesian denoising of visual images in the wavelet domain. In P. Müller & B. Vidakovic (Eds.), *Bayesian inference in wavelet based models* (pp. 291–308). New York: Springer-Verlag.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*, 199-222.

Squartini, S., Bastari, A., & Piazza, F. (2006). A practical approach based on Gaussianization for post-nonlinear underdetermined BSS. In *ICCCSP.* Orlando, Florida.

Stark, H., & Woods, J. (1986). *Probability, random processes and estimation theory for engineers*. Englewood Cliffs, New Jersey: Prentice-Hall.

Stark, H., & Woods, J. (1994). *Probability, random processes, and estimation theory for engineers*. NJ: Prentice Hall.

Storkey, A. (2009). When training and test sets are different: Characterizing learning transfer. In *Dataset shift in machine learning.* Cambridge, MA: MIT Press.

Studeny, M., & Vejnarova, J. (1998, January). The multi-information function as a tool for measuring stochastic dependence. In M. I. Jordan (Ed.), *Learning in graphical models* (p. 261-298). Kluwer.

Székely, G. J., & Rizzo, M. L. (2005). A new test for multivariate normality. *J. Multivar. Anal.*, *93*(1), 58–80.

Takeda, H., Farsiu, S., & Milanfar, P. (2007, Feb). Kernel regression for image processing and reconstruction. *IEEE Transactions on Image Processing*, *16*(2), 349-366.

Taubman, D. S., & Marcellin, M. W. (2001). *JPEG2000: Image compression fundamentals, standards and practice*. Boston: Kluwer Academic Publishers.

Tax, D., & Duin, R. (1999). Support vector domain description. *Pattern Recognition Letters*, *20*, 1191-1199.

Teh, Y. W., & Roweis, S. (2003). Automatic alignment of local representations. In *Nips 15* (pp. 841–848). MIT Press.

Tenenbaum, J. B., Silva, V., & Langford, J. C. (2000, December 22). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323.

Teo, P., & Heeger, D. (1994). Perceptual image distortion. *Proceedings of the SPIE*, *2179*, 127–141.

Torczon, V. (1997). On the convergence of pattern search algorithms. *SIAM J. on Optimization*, *7*(1), 1–25.

Touryan, J., Felsen, G., & Dan, Y. (2005, March 3). Spatial structure of complex cell receptive fields measured with natural images. *Neuron*, *45*(5), 781–791.

Tsang, I. W., Kwok, J. T., & Cheung, P.-M. (2005). Core vector machines: Fast svm training on very large data sets. *J. Mach. Learn. Res.*, *6*, 363–392.

Twer, T., & MacLeod, D. A. (2001). Optimal nonlinear codes for the perception of natural colours. *Network: Comp. Neur. Syst.*, *12*(3), 395–407.

Van Hateren, J. H. (1992). A theory of maximizing sensory information. *Biological Cybernetics*, *68*(1), 23-29.

Van Hateren, J. H. (1993). Spatiotemporal contrast sensitivity of early vision. *Vision Research*, *33*(2), 257 - 267.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer–Verlag.

Verbeek, J. J., Vlassis, N., & Krose, B. (2002). Coordinating principal component analyzers. In *In Proceedings International Conference on Artificial Neural Networks* (pp. 914–919). Springer.

Vishwanathan, S. V. N., Schraudolph, N. N., & Smola, A. J. (2006). Step size adaptation in reproducing kernel Hilbert space. *Journal of Machine Learning Research*, *7*, 1107-1133.

Wachtler, T., Lee, T., & Sejnowski, T. J. (2001). Chromatic structure of natural scenes. *Journal of the Optical Society of America A*, *18*(1), 65–77.

Wang, Z., & Bovik, A. (2009, Jan.). Mean squared error: Love it or leave it? *IEEE Signal Processing Magazine*, 98–117.

Wang, Z., Bovik, A., Sheikh, H., & Simoncelli, E. (2004a). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004b). Perceptual image quality assessment: From error visibility to structural similarity. *IEEE Trans Image Processing*, *13*(4), 600–612.

Wang, Z., & Simoncelli, E. (2005b). Translation insensitive image similarity in complex wavelet domain. In *Proceedings IEEE International Conference on Acoustics, Speech & Signal Processing* (pp. 573–576).

Wang, Z., & Simoncelli, E. P. (2005a, 11-14 Sep). An adaptive linear system framework for image distortion analysis. In *Proc 12th ieee international conference on image processing* (Vol. III, pp. 1160–1163). Genoa, Italy: IEEE Computer Society.

Watson, A. (1983). Detection and recognition of simple spatial forms. In O. Braddick & A. Sleigh (Eds.), *Physical and biological processing of images* (Vol. 11, pp. 100–114). Berlin: Springer Verlag.

Watson, A. (1987). Efficiency of a model human image code. *Journal of Optical Society of America A*, *4*(12), 2401–2417.

Watson, A., & J.Malo. (2002). Video quality measures based on the standard spatial observer. *Proc. IEEE International Conference on Image Processing*, *3*, 41-44.

Watson, A., & Kreslake, L. (2001). Measurement of visual impairment scales for digital video. In *Proc. SPIE human vision, visual processing, and digital display* (Vol. 4299).

Watson, A., & Ramirez, C. (2000). A Standard Observer for Spatial Vision. *Investig. Opht. and Vis. Sci.*, *41*(4), S713.

Watson, A., & Solomon, J. (1997). A model of visual contrast gain control and pattern masking. *Journal of the Optical Society of America A*, *14*, 2379–2391.

Watson, A. B. (Ed.). (1993). *Digital images and human vision*. Cambridge, MA, USA: MIT Press.

Webster, M., & Mollon, J. (1997). Adaptation and the color statistics of natural images. *Vision Res*, *37*(23), 3283–3298.

Webster, M. A., & Mollon, J. D. (1991). Changes in colour appearance following post-receptoral adaptation. *Nature*, *349*, 235–238.

Weinberger, K. Q., & Saul, L. K. (2004). Unsupervised learning of image manifolds by semidefinite programming. In *Proc. IEEE CVPR* (pp. 988–995).

Weston, J., Leslie, C., Ie, E., Zhou, D., Elisseeff, A., & Stafford Noble, W. (2004). Semi-supervised protein classification using cluster kernels. *Bioinformatics*, *21*(15), 3241-3247.

Wyszecki, G., & Stiles, W. (1982). Color science: Concepts and methods, quantitative data and formulae. In (chap. 6: *Uniform Spaces*). New York: John Wiley & Sons.

Xiang, B., Chaudhari, U., Ramaswamy, G., & Gopinath, R. (2002). Short-time gaussianization for robust speaker verification. In *IEEE ICASSP.* Orlando, Florida.

Zarzoso, V., Comon, P., & Kallel, M. (2006). How fast is FastICA? In *European signal processing conference.*

Zhang, K., & Chan, L. (2005). Extended gaussianization method for blind separation of post-nonlinear mixtures. *Neural Computation*, *17*(2), 425-452.

Zhang, X., & Wandell, B. (1996). *A spatial extension of cielab for digital color image reproduction.*

Zhaoping, L. (2006). Theoretical understanding of the early visual processes by data compression and data selection. *Network: Computation in neural systems*, *17*(4), 301-334.