Efficient and Robust Signal Approximations

Doru Cristian Balcan

CMU-CS-09-129

May 2009

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Michael S. Lewicki, Chair
Manuel Blum
Jelena Kovačević
Gary Miller
Markus Püschel

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

This research was sponsored by the National Science Foundation under contract nos. IIS-0413152, IIS-0238351, and IIS-0705677, the National Geospatial Intelligence Agency under contract no. HM1582-04-C-0053, the Office of Naval Research (Stanford University) under contract no. 1968767038469A, and MPC (CNBC Lab) under contract no. 93-03. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Efficient and Robust Signal Approximations	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Carnegie Mellon University,School of Computer Science,Pittsburgh,PA,15213**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**84**	

Keywords: signal processing, image compression, independent component analysis, sparse representations, adaptive signal coding, shiftable-kernel dictionaries, robust coding

To my family, as a modest reward for their patience.

Abstract

Representation of natural signals such as sounds and images is critically important in a broad range of fields such as multimedia, data communication and storage, biomedical imaging, robotics, and computational neuroscience. Often it is crucial that the representation be efficient, *i.e.*, the signals of interest are encoded economically. It is also desirable that the representation be robust to various types of noise. In this thesis, we advocate several ways to expand current signal encoding approaches via the framework of adaptive representations.

In recent decades, the multiresolution paradigm has provided powerful mathematical and algorithmic tools to signal encoding. In spite of widely proven effectiveness, such methods ignore statistical structure of the class of signals they should represent. On the other hand, high computational costs artificially confine standard linear adaptive statistical models to relatively small block-based encoding scenarios. We show that a good tradeoff between computational complexity and coding efficiency can be achieved via a hybrid encoding scheme: Multiresolution ICA. When applied to natural images the new method significantly outperforms JPEG2000, the current compression standard, which indicates adaptivity as a source of practical improvement for modern coders.

Sparsely encoding large signals via a set of adaptive variable-size shiftable kernels has been studied in several contexts, like efficient auditory coding. One important merit of this paradigm is that, besides efficient adaptive coding, it also provides a direct approach towards an (approximately) shift-invariant representation. This is especially desirable in modeling encoding systems robust to signal shifts, such as biological sensory systems. We study this problem in the case of images and provide contributions leading to fast and superfast algorithms, significantly improving the complexity of the kernel learning process.

The third part of this thesis is a mathematical study of Robust Coding - the problem of optimal linear coding with limited precision units. We characterize optimal solutions in the case of Gaussian channel noise and arbitrarily many encoding units, and derive efficient and stable algorithms for their computation. By expressing the limit of optimization as a closed-form bound, we provide a formal justification of the intuition that noisy encoding units can preserve signal information if sufficiently many are used - a case very relevant to modeling neural encoding systems.

Acknowledgments

Going down the Ph.D. path is an extraordinary adventure. Truly valuable research gems are never easy to find, and many monsters (frustration, impatience, procrastination, despair) await for the weak to slip into their traps. I am deeply grateful to my advisor Mike Lewicki for inspiring me to discover nothing less than the most precious stones, for subtly helping me become more and more independent, as well as for trying to make sure that I don't get scared along the way. Not a typical C.S. advisor[1], he did more than his fair share as fearless lab leader. He exposed me to fascinating problems, suggested ideas I wouldn't have come across in a million years, and was there when I needed him. (Also, every now and then, he'd help me up without my even beginning to realize I need to start asking.)

Along the way, I discovered several other people which were so kind to lend me their valuable insight, their spectacular vision, and huge chunks of their precious time. Jelena "with the magic touch" Kovačević helped me discover entire research topics, pointed me to breathtaking papers, invited me to her own lab meetings, and injected me with many shots of confidence. Manuel Blum overwhelmed me with his wisdom, humor, kindness, respect for students, and appetite for elegance in thought, word, and deed. Among multi-faceted computer scientists, one of the best is Gary Miller; he offered competent advice to me on several dozen different topics, always being a step ahead and always suggesting yet another really cool idea to try. Probably the best faculty role model for me has been Markus Püschel; I have *never* enjoyed any other course in my entire "career" more than I did his *Algebraic Signal Processing Theory*, and few times did I have so much fun (and apparent ease!) when working on a research problem as I had during our collaboration. Danke schön!

When considering getting a Ph.D., it is extremely useful to have someone lay out a solid plan for what you should expect ahead (and how to handle it); in my case, that someone is George Necula. Without his generous and wise advice, there's no telling where I'd gotten. A great contribution to my "A Ph.D. is not enough!" wake-up call was that of Justinian Roşca, my internship supervisor, who recommended the book with such self-explanatory title and offered his valuable assistance every time, without hesitation. I would probably be doing something completely different today without the inspiration and tremendous support of Prof. Luminiţa State, my advisor at the University of Bucharest; her enthusiasm and charisma left me helplessly in love with Artificial Intelligence and so much more.

But to get to know all of the people above, something else had to have happened first: meeting the *greatest* couple of Math teachers on the face of the Earth, Dana and Eugen Radu (aka *naşii*); they knew exactly how to push my buttons[2] so that I'll always love and appreciate the geometrical beauty of thinking clearly. Each of them explained to me more than just about my ABC's, or even ABCD's; by their humor, warm decency, mind-blowing

[1] Right... they never are!

[2] An operation that I now call "to program".

short proofs, and (amazingly!) being right all the time, I learned that there's nothing I'd like better than to follow in their footsteps. I would like to thank my high-school Rom.-lit. prof Mrs. Emanuela Crişan for the delightful and sparky in-class arguments, always on the same theme (in many disguises): what's more important - human reason, or divine revelation? I won every time: the prize I took away were her always beautiful and uplifting thoughts.

All members of CPLab[3] at CMU (aka *MikeLab*) contributed in unique ways to maintaining a friendly research environment. Our lab meetings lead us to most interesting discussions, as well as to new and creative ways to explain our ideas. I owe a lot to the super-amazing Yan Karklin, the elegant Evan Smith, the sunny Sofia Cavaco, the playful Jing Chen, the conscientious Daniel Leeds, and the awesome Woo-Young Lee. Determinant for me was the company of Eizaburo Doi; I am grateful to him for introducing me to Robust Coding and for having the patience of smoothing the heck out of my roughest rough edges (research-wise). Arigato! Another extraordinary office mate was Dr.[4] Andrew Gilpin, who semi-supervisedly learned Romanian to become an even better office mate (also, to understand what was going on in the office behind his back). I'll never forget his first *Ce faci?*[5] uttered at me, or *Mă dau jos la tine!*[6] playfully addressed to our other (more restless) Romanian office mate. He even got to use this knowledge outside of the office, in his visit to Romania, amazing and flattering everybody (particularly my folks). Mulţumesc!

I know now that life in academia is more than "publish or perish". Actually, it's more like "appreciate all your collaborators and co-authors for all they do to make it easier and fun for you to sit around and look smart... or perish". At CMU, I was blessed to have met an amazing group of fellow students who influenced much of what I know, and how I (should) think. I will only pick two out of this remarkable crowd. It's hard to describe how much fun it is to think about the most impossible of problems when Gowri Srinivasa is in the house. Compared to that, even her generous and cheerful help giving, or contagiously enthusiastic energy seem to fall short (although not by much). Her new students might not know yet under what a lucky star they have been gathering (but for sure they'll find out!). An extraordinary guy is my fellow Eastern European co-author Aliaksei Sandryhaila. Dynamic and determined, friendly and brilliant, practical and funny, he often seems like nothing can stop him. So far nothing has, and I'm willing to bet that nothing will. A great deal of my gratitude and appreciation goes to Justin Romberg and Nick O'Donoughue for promptly and graciously agreeing to help out when I was recently unable to physically present two of my posters (any grad student's nightmare).

The work in this thesis would have not been possible without the care and dedication of competent and nurturing staff. Thank you, Sharon Burks and Deb Cavlovich, for the many last-minute support letters and for everything else in between! As for Ms. Anna Hegedus, the "know-it-all supreme" title in computer... everything, is not so far from the truth.

In Pittsburgh, you are fortunate if you have around many Romanians like the ones I had. From them I learned that a hearty community can be stronger than a country. Here is to Alina Oprea, Radu Niculescu, and Cristina Cânepă. Special thanks go to Florin Oprea for always being Mr. Helpful, Mr. Available, Mr. Reliable, and generally the best soccer player I've ever aspired to be. Speaking of which, I'd like to thank all my soccer team mates whose

[3] *Laboratory for Computational Perception and Statistical Learning*
[4] At the moment of this writing, the newest Doctor around!
[5] "How are you doing?" (Rom).
[6] "I'm coming down to get you!"; as in, "If you're not good..." (Rom).

efforts over the years inspired me to get serious about winning the 2008 CMU Intermediate Intramural Indoor Tournament: Leonid, Aaron, Vince, David, Daniel, Lucian, Rob, Will, and the other thoroughbred "Real Mellons".

In the past few years, my old envy of friends with siblings got cured; the long wait finally paid off big time due to my ever-so-studious brother-in-law Marius and to my smart, beautiful, and witty sister-in-law Aliona. (Now others can envy me!) They - in fact all my family and friends back home - deserve my sincere apologies for my not writing or calling as often as I had liked, as well as my gratitude for their going ahead and asking what's up, how much longer is the PhD supposed to take, or when is that (ever elusive!) next visit home likely to happen - so that they can properly plan and be completely available when it comes to satisfying my every whim.

Last, but not least, I would like to thank my parents, Marieta and Nelu, for all the unconditional love and support they offered me throughout my life. My mother encouraged me to explore more of this world so that I could know more about who I am and who I should want to become. She is also the first person in my family with a Ph.D. From my father I learned how important it is to be strong, stubborn, patient, pragmatic, but relaxed and (at least from time to time) extremely funny. Our fishing trips are excellent lessons about how (mostly his) preparation can meet (sometimes my) opportunity; if only I had realized earlier that this is what "fisherman's luck" is all about.

Dr. Nina Balcan is beyond any acknowledgment. If it were only for her moral contribution to this thesis, her name should be engraved in golden capital letters on every page[7]. Ever since I met her, she is constantly showing me by the power of personal example how to be better; how to constantly want more; how to get more; how willpower beats all adverse odds and prejudice eventually; how respecting and loving others does not exclude always demanding nothing less than the best from them; how to not forget anything good (or bad!). She's my twenty-four-hours-a-day advisor, best friend, hero, favorite researcher, and loving wife. And I feel that's only the beginning...

[7]I am not sure, but I suspect this goes against the department's policy, or dissertation guidelines, or both.

Contents

List of Figures

Chapter 1

Introduction

In everyday life, a change of perspective about a particular problem that we confront is likely to reveal entirely new aspects which both enrich our understanding of the problem and improve the way we solve it. Sometimes, a different perspective might be even critical in discovering the most efficient solution. This high-level principle lies at the foundation of science and discovery in general, but in signal processing it has a concrete, low-level analog: the choice of signal representation.

1.1 Motivation

Deriving efficient representations of natural signals is an important and challenging research topic. There are multiple ways to quantify progress, many of them employed as the "working standard" by an entire beneficiary community. For instance, to multimedia users, having better signal representations means more music and video on their portable devices, and consequently more entertainment. To robot manufacturers, it translates into a better chance for a robot to navigate and operate within a new environment. Diverse applications in communications, earth sciences, and medicine can greatly benefit from representations with good descriptive and computational properties.

Besides the attraction towards practical applications, often associated with financial gratification, there exists the human drive to understand nature. One of the greatest challenges posed to science has been to explain the function of the brain: what are the principles that govern the phenomena taking place here? There is still much to be learned about representing and processing information in the brain, even in relatively specialized subsystems. For instance, by observation and experimentation we learned that the visual system has the role of analyzing and combining the various information content of the perceived images, as it is transmitted from the retina all the way to the cortex. This is much more than to merely format the visual stimulus into spikes and distribute it to the processing areas; it also involves "splitting" complex scenes into features that later combine into higher-level concepts.

Although the exact mechanisms and computational principles driving these processes are not fully understood (or maybe *because* of it), the brain is rightfully considered the ultimate high-

end signal processor. It outperforms by far any known artificial system at tasks that involve abstract concept manipulation – analyzing complex scenes, navigating unknown environments, extracting specialized types of features like text, etc. It is then perhaps not surprising that several security protocols are heavily relying on this (see, for instance [118])!

In this thesis, we investigate several aspects related to natural signal representation while focusing on one main class of applications: visual signal encoding. We address several problems generated by existing signal representation frameworks and propose novel extensions by embracing the adaptive encoding point of view. For each of the instances hereby studied, we follow two main goals: to clearly identify the theoretical principles which govern the representation's optimality, and to identify the most efficient algorithm to compute it.

1.2 Thesis outline

The thesis is organized as follows.

- **Background.** In Chapter 2, we review most of the fundamental notions and theoretical concepts used in the remainder of the thesis. We start by introducing basic signal processing notions such as bases, frames, and dictionaries, then continue by presenting the concept of multiresolution and several signal representations of that family. Next, we address the issue of adaptivity and illustrate it with two finite-dimensional linear models: ICA and Robust Coding. After a short introduction to sparse signal approximations with a particular emphasis on greedy encoding methods, we explain the concept of shiftable kernel dictionary representations, and illustrate a way to obtain adaptive dictionaries of this type.

- **Multiresolution ICA.**[1] In Chapter 3, we study the problem of efficient and adaptive representation of large-scale images. In recent decades, the multiresolution paradigm has provided powerful mathematical and algorithmic tools to signal encoding. In spite of widely proven effectiveness, such methods ignore the statistical structure of the class of signals they represent. On the other hand, because of usually high computational costs, standard linear adaptive statistical models have been confined artificially to relatively small block-based encoding scenarios. We show that a good tradeoff between computational cost and coding efficiency can be achieved via a hybrid encoding scheme: Multiresolution ICA. When applied to natural images the new method significantly outperforms JPEG2000, the current compression standard, which indicates adaptivity as a source of practical improvement for modern coders.

- **Point Coding.**[2] In Chapter 4, we review the problem of sparsely encoding large signals via a set of adaptive variable-size shiftable kernels. An important merit of this approach is that it produces a very efficient adaptive code, which is explained by the flexibility of such a

[1] Parts of this chapter have been published in [11].

[2] The work in this chapter has been published in [12].

dictionary, compared with conventional representations, such as wavelets. In addition, this provides a more direct way to obtain an (approximately) shift-invariant signal representation. This is especially desirable in modeling encoding systems robust to signal shifts, such as biological sensory systems. We study this problem in the case of images and provide contributions leading to efficient algorithms, significantly improving the complexity of the kernel learning process. Specifically, we show that we can employ fast and superfast algorithms for the learning step, by formulating the problem as a least-squares problem with a highly structured (Toeplitz-mosaic) matrix. Encoding is implemented via a fast version of Matching Pursuit based on exploiting the relatively small sizes of the kernels compared to the signal and by using appropriately designed data structures to speed up computations.

- **Robust Coding.**[3] In Chapter 5 we provide a detailed mathematical study of Robust Coding - the problem of optimal linear coding with limited precision units. We characterize optimal solutions in the case of Gaussian channel noise and arbitrarily many encoding units, and derive efficient and stable algorithms for their computation. By conveniently expressing the limit of optimization as the closed-form bound, we formally explain the intuition that noisy encoding units can preserve signal information if sufficiently many are used - a case very relevant to modeling neural encoding systems.

- **Conclusions.** Chapter 6 contains a summary of this thesis, together with a list of directions and ideas we intend to pursue in the future.

In completing the research projects reported here, we employed the principles of learning and optimization to the design of signal representations. In each case, the first step was to identify some clear computational objective and express it in the most convenient mathematical form. The next step was to choose the algorithm most suitable to exploit the intrinsic structure of the problem. Finally, by implementation and simulation, hypotheses were validated and often new insights appeared.

For the sake of coherence, we decided not to include in the body of this thesis several published results, which are nevertheless useful in exploring and understanding various aspects of signal representation design. One significant direction is artificial bandwidth extension of speech signals (see [9, 103]). There we consider the practical problem of enhancing speech whose TF content is missing, either as a result of bandpass filtering (like in telephony) or as a by-product of certain source separation algorithms. We address the problem of "filling the spectral holes" a case of statistical estimation with missing data. A second research direction we mention is the problem of designing polynomial signal transforms asymptotically approximating the Discrete-Time Fourier Transform, but which do not require the periodicity assumption usually associated with DFT (see [10]). To answer this question, arising from the general algebraic signal processing theory [99], we identify a fairly large class of such finite polynomial transforms by defining polynomial families whose set of roots approximately converges to the complex unit circle (with perhaps finitely many exceptions).

[3] Parts of this chapter have been published in [43, 44].

Chapter 2

Background

The focus of this thesis is the study of signal representations. We investigate several ways in which existing approaches can be extended and improved, but first we need to lay out the mathematical groundwork of our construction.

To start, let us specify that we only consider discrete-time/space signals. Although several of the problems we treat in the following do not necessarily require it, for simplicity we restrict to finite signals. Accordingly, a signal shall usually be regarded as a vector $\mathbf{x}$ in a finite-dimensional space $\mathcal{V}$ (*e.g.*, $\mathbb{R}^N$ or $\mathbb{C}^N$). Its representation shall be defined with respect to *dictionary* $\mathbf{\Phi}$, a finite subset of $\mathcal{V}$, as a vector $\mathbf{s}$ such that

$$\mathbf{x} = \mathbf{\Phi}\mathbf{s}. \tag{2.1}$$

Assuming the dictionary has M elements, the representation is thus a M-dimensional vector of *coefficients*. Computing the signal from the representation (when the dictionary is fixed) is called *decoding* or *synthesis*, and by definition is a simple linear operation. The reverse operation (*encoding*, or *analysis*) may be a more complicated process is; for example, there may exist infinitely many vectors $\mathbf{s}$ (or even none at all!) that satisfy eq. 2.1.

A central issue of this thesis is *adaptivity*. When the dictionary is optimized in some respect to represent signals of a given class, *i.e.*, it reflects either deterministic or statistical properties of the class, we say it is adapted to the class and the induced representation shall be called *adaptive*; otherwise, we label it as *fixed*.

2.1 Bases, Frames, and Dictionaries

The canonical, sample-based representation is often not appropriate to describe compactly the complicated structure of many classes of signals. The most common source of redundancy is the strong local dependency between samples; in the case of images, this would correspond to neighboring pixels having similar values, due to similar color or intensity. Other types of regularities, such as texture patterns, and even "frequent" *irregularities* or other discontinuities, like edges, could be handled more efficiently by alternative representations.

The idea of choosing the appropriate signal representation for the task at hand is generating

much of the effort (and the progress) in signal processing and many results from linear algebra have contributed to great advancements in the field. Essentially, searching for linear representations of finite signals is searching for dictionary matrices $\mathbf{\Phi}$ whose columns span the entire vector space $\mathcal{V}$. If the columns of such a matrix form a spanning set for the whole space and they are linearly independent, we say that the representation is complete[1]. In case the vectors lack linear independence but still span the space, the representation is called overcomplete[2]. Otherwise, we shall denote the representation as undercomplete[3].

The ideas about changing coordinates to analyze structure go a long way back, originating in Physics, but spectral methods and their applications to signal processing have flourished after the (re)discovery of the fast algorithms for their implementation. Unfortunately, the Discrete Fourier Transform, the many kinds of Discrete Trigonometric Transforms, or any linear basis of $\mathbb{C}^N$ with only global structure for that matter, do not offer a compact description of signals with local spatial structure, which is a consequence of the uncertainty principle. The need for more specialized descriptors was behind the multiscale revolution, and the wavelet frenzy. The representational advantage and the low computational cost of applying the Discrete Wavelet Transform have lead to the design of current image coding standards such as JPEG2000 [109] (also see [86, 117]).

The various applications often require representations having specific analytical and computational properties. Generally, the process of computing such bases requires an objective function and a set of constraints. In come cases, the solutions of these optimization problems are unique, while in others they are not; moreover, it is also possible that the optimization problem be overconstrained and that no solution exists. For example, in frame design the objective is to find an overcomplete basis with specified properties, such as prescribed length columns and minimum inner product between different column vectors. If we throw in additional constraints, for example those regarding the set of singular values of the matrix, we may get a rather difficult problem to solve. Luckily, some of these problems can be handled numerically by the use of efficient optimization algorithms on structured spaces, as opposed to having closed-form, analytic solutions. Depending on the context, we will be satisfied with such a particular numerical solutions, or we shall search for specifications of the whole space of solutions. In this thesis, we will pursue both possibilities, and specify the advantages for adopting each point of view.

In certain cases, it is possible to implement the process of signal analysis or that of synthesis of a without the *explicit* use of a matrix-vector product. This happens frequently when faster algorithms exist which do not require all the entries of the basis matrix. Just to give an example, computing the Discrete Fourier Transform of a signal (either the direct or the inverse one) is nowadays almost synonymous with using the Fast Fourier Transform (FFT) algorithm (see *e.g.*, [56]). Other widely-known signal processing operations (*e.g.*, convolution) can benefit from this aspect. We will point out the distinction between explicit and implicit linear operations when we

[1]We call the set of vectors, and by extension the matrix itself - a basis. For finite dimensional vector spaces, basis matrices are always square.

[2]To describe such a system of vectors, we will use the term "frame". The formal definition of a frame (see *e.g.*, [34]) is equivalent to this one in the case of finite sets of vectors spanning a finite dimensional space.

[3]We will slightly abuse the term "basis" in this case, using it even if the vector set $\mathbf{\Phi}$ does not span the space $\mathcal{V}$.

discuss implementation aspects, otherwise we will keep this issue transparent.

In the following, we gradually introduce several strategies for obtaining desirable representations as well as for computing the corresponding signal coefficients. Our generic goal will be efficient coding, which also can be formulated as compact signal description. Usually, this involves computing a minimal set of nonzero coefficients; the advantage of these so-called *sparse* representations is that we only need to store or compress a small set of numbers, as opposed to the whole set of samples, for (approximately) reconstructing the signal. Then, we review adaptive linear methods, whose goal is to improve the descriptive properties of the dictionary with respect to the observed data. Next, we describe the problem of deriving adaptive linear representations that are robust to coefficient perturbations. Finally, we briefly review the so-called Spike Coding model – a method for sparse one-dimensional signal representation using an adaptive shiftable-kernels dictionary.

2.2 Adaptive Models: ICA and Robust Coding.

Independent component analysis (ICA) has appeared in the signal processing community as a general method to separate a number of sources, assumed mutually independent, when several (linear) combinations of these are available [68]. The particular case when the sources are Gaussian distributed had already been known as principal component analysis (PCA), but this could not handle and explain the many examples of signal distributions that are not proper Gaussian. In a relatively short time, the field also was extended theoretically and equivalence between the apparently different settings has been revealed (see [23, 31]; extensive treatment of ICA also can be found in [30, 64, 79]).

Let $x_1, x_2, \ldots, x_m$ be samples drawn from a distribution with pdf p over $\mathbb{R}^N$. The goal of ICA is to compute the $N \times N$ matrices $\mathbf{W}$ such that vectors $\mathbf{s}_j = \mathbf{W}\mathbf{x}_j$ are the realizations of a random vector whose components are as statistically independent as possible, according to formal criteria described below. In other words, we search for the linear mapping allowing us to best approximate the data distribution by a product of marginals. Alternatively, ICA can be viewed as a method to describe the data by a linear combination of vectors (or "basis functions"):

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{2.2}$$

such that the components of random vector $\mathbf{s}$ are maximally independent. Matrix $\mathbf{A}$ is called the *mixing* matrix, while $\mathbf{W}$ is referred as the *demixing* matrix. A standard assumption is that the data has been processed to have zero mean, and unit covariance (possibly by dimensionality reduction), which implies that matrices $\mathbf{A}$ and $\mathbf{W}$ are nonsingular, and inverse to each other. When the underlying data distribution is Gaussian, coefficient independence is equivalent to decorrelation, and thus, when the coefficients are indeed independent, the ICA and PCA bases will coincide. In general, PCA will search for the best *orthogonal* basis to represent the data, while ICA does not have this constraint.

There are several mathematical objective functions associated with independence, and each leads to one formulation of the ICA problem. For example, one such objective is to minimize

the mutual information of the representation components. Specifically, for any invertible linear transformation $\mathbf{s} = \mathbf{W}\mathbf{x}$, the mutual information among the components of $\mathbf{s}$ is defined as:

$$I(s_1, \ldots, s_n) = \sum_{i=1}^{n} H(s_i) - H(\mathbf{x}) - \log|\det \mathbf{W}|. \tag{2.3}$$

By an appropriate scaling of $\mathbf{W}$, mutual information can be viewed as the difference between the sum of the marginal entropies and the joint entropy. This can be interpreted further in terms of the Kullback-Leibler divergence between the joint probability and the product of marginals, or in terms of the negentropy of the projections, or in terms of data likelihood [31, 64]. For instance, if we denote the pdf's of the projections by $p_i(\cdot)$, the expectation of the log-likelihood can be written as[4]

$$\frac{1}{m} E\{\log L(\mathbf{W})\} = \sum_{i=1}^{n} E\{\log p_i(\mathbf{w}_i \cdot \mathbf{x})\} + \log|\det \mathbf{W}| \tag{2.4}$$

which is the negative of mutual information, except for a constant term (entropy of the data). Equivalently, this can be viewed as maximizing various measures of non-Gaussianity (*e.g.*, the kurtosis) of the entire ensemble and many practical ICA algorithms are based on methods that attempt to maximize higher order moments of the coefficients. Finally, another direction to approach independence of the coefficients is by diagonalization of certain matrix functionals. Methods of this family thus translate independence into simultaneously solving a series of eigen-problems.

Generally, existing ICA optimization algorithms can be grouped into several categories: gradient-based [2, 13, 23, 66, 81], fixed-point[5] [63, 65, 119], joint-diagonalization [24, 25]. A very efficient procedure based on Relative Trust-Region Optimization has been presented recently in [29]; due to its excellent behavior, we decided to employ this algorithm for all the ICA-related experiments presented in this thesis.

To represent images in the ICA model, we can regard them as samples drawn from an unknown distribution, over a linear space (the so-called *image space*). The fact that independent components of natural images have a very sparse (or "thorny") distribution makes ICA highly suitable for image representation and coding. The immediate consequence of a sparse marginal distribution is a low entropy, and thus the possibility of achieving a short average code length, which ultimately yields better compression. In fact, among all linear models the ICA representation is optimal in the entropy minimization sense, which thus recommends it for compression tasks.

A well known limitation of ICA is its poor scaling behavior. Due to the relatively high computational cost, it cannot be applied directly to large dimensional data. For example, in case of $d \times d$ images computing a complete basis to span the space implies estimating d^4 parameters. Even for a moderate value of d (say 100) the memory requirements are tremendous. Moreover, the computational cost of typical gradient optimization is $\Omega(d^{2\log_2 7})$, which is prohibitive.

[4]Here $\mathbf{w}_i$ is the ith row vector of matrix $\mathbf{W}$.

[5]Although it has been proven that some of these algorithms were wrongly classified as "fixed-point" [119], for convention purposes we chose to leave them in this category.

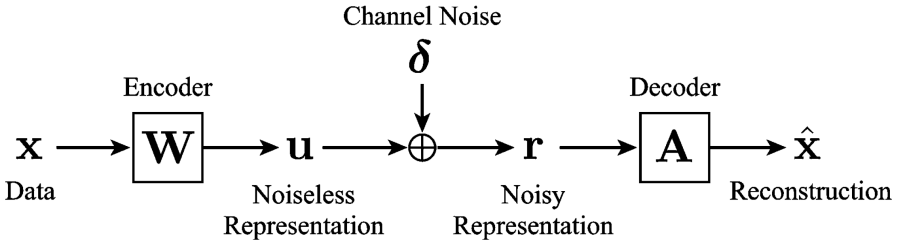


Figure 2.1: Diagram of the Robust Coding Model.

Therefore, this approach is feasible only if the problem is reduced to a smaller space (as in PCA), which in turn limits the signal structure we are able to represent. From the image coding point of view, the quantization or sparsification of the coefficients in this block transform approach leads to blocking artifacts in the reconstruction. Block processing also leads to artifacts for more general signal processing algorithms, such as image denoising. In chapter 3, we shall present a method to compute an quasi-ICA basis for large images by solving several (typically much smaller) ICA problems.

Optimizing for coding efficiency is a very desirable goal in applications such as storage and communication. An equally important aspect is the resilience of the signal representation to various types of noise. Reliable communication over noisy channels is the most fundamental problem of information theory. Out of the many variations on this theme, we shall focus on the problem of finding finite-dimensional linear representations that optimally preserve information in the transmitted signals when the representation has limited precision. Proposed by E. Doi et.al. in [43] and further analyzed in [44], the Robust Coding scheme uses arbitrarily many coding units to minimize reconstruction error, by explicitly introducing redundancy in the code to compensate for channel noise.

The above problem was pointed out to be of particular relevance to the mathematical modeling of neural representations. This is not at all surprising; cells can be regarded as communication channels for the traveling neural spikes, and their coding precision is limited by intrinsic biological constraints to as low as 1-2 bits per spike (see [16], [43] and references therein). By identifying the short time activity of a neuron with a real value, the limited information capacity of the encoding unit can be modeled effectively by additive Gaussian noise.

To describe the problem formally, let us consider our signals as samples drawn from an N-dimensional zero-mean data distribution, with known full-rank covariance matrix $\mathbf{\Sigma_x}$. We shall search for *analysis matrix* $\mathbf{W} \in \mathbb{R}^{M\times N}$ and *synthesis matrix* $\mathbf{A} \in \mathbb{R}^{N\times M}$ that maximally reduce the effect of additive Gaussian noise, independent of the signal and having the same power σ_δ^2 on each channel. If we denote $\epsilon = \mathbf{x} - \hat{\mathbf{x}} = (\mathbf{I}_N - \mathbf{AW})\,\mathbf{x} - \mathbf{A}\delta$, our objective is to minimize the reconstruction MSE

$$\begin{aligned}
\langle ||\epsilon||_2^2 \rangle_{\mathbf{x},\delta} &= \langle \epsilon^T \epsilon \rangle = \mathbf{tr}(\langle \epsilon\epsilon^T \rangle) = \mathbf{tr}\left\{ \langle ((\mathbf{I}_N - \mathbf{AW})\,\mathbf{x} - \mathbf{A}\delta)((\mathbf{I}_N - \mathbf{AW})\,\mathbf{x} - \mathbf{A}\delta)^T \rangle \right\} \\
&= \mathbf{tr}\left\{ \langle (\mathbf{I}_N - \mathbf{AW})\,\mathbf{x}\mathbf{x}^T\,(\mathbf{I}_N - \mathbf{AW})^T \rangle_{\mathbf{x}} + \langle \mathbf{A}\delta\delta^T\mathbf{A}^T \rangle_\delta \right\} \\
&= \mathbf{tr}\left\{ (\mathbf{I}_N - \mathbf{AW})\mathbf{\Sigma_x}(\mathbf{I}_N - \mathbf{AW})^T \right\} + \sigma_\delta^2 \mathbf{tr}\left\{ \mathbf{A}\mathbf{A}^T \right\}
\end{aligned}$$

In chapter 5 we shall provide a more thorough description of the problem, as well as a complete characterization of the optimal encoder/decoder pair.

2.3 Sparse Approximations

Data compression is built around the principle of employing compact descriptions of a seemingly complex signal. Computing sparse linear representations of signals is useful for many applications, ranging from data communication to statistical data analysis and machine learning.

In general, it may not be possible to choose a small set of nonzero coefficients to represent any signal exactly in a given basis. Besides an intrinsic measure theoretic difficulty, the limitation can persist even if we relax the exactness and settle for a sparse *approximate* coefficient set. The properties of the dictionary can help significantly if they match the statistical properties of the signal. We will address this issue in the following subsection, but now it is important to focus on two questions. The first is: how can we compute the sparsest representation in a given, fixed dictionary? The second question is concerned with the validation of our answer to the first one: how close are we to the sparsest representation? Unfortunately, in the most general case both problems are NP-hard (see [90], as well as [37, 38, 39]), which means that it is unlikely to compute optimal solutions in polynomial time. In spite of this shortcoming, it is possible to obtain approximate solutions in (pseudo-)polynomial time. Next, we describe the most frequently used approaches to obtain such sparse linear approximations.

The most straightforward idea is *thresholding*. More precisely, out of a complete set of coefficients corresponding to a linear combination of atoms, we only keep the ones with the largest absolute values, while the rest are assumed to be zero. Various strategies for determining the "appropriate" number of coefficients were studied, and the success of this approach to applications like compression and denoising was investigated in the context of wavelet dictionaries.

The proper *greedy* approach to the sparsest set selection problem has become known in signal processing by the name of Matching Pursuit (MP) [87]. Unlike thresholding, which assumed that all the coefficients of a transform would have to be available in order to choose the largest ones, MP is a sequential algorithm, allowing us to individually pick dictionary atoms that are most correlated with the signal. Mathematically, if we consider a signal $\mathbf{x}$ and a dictionary $\mathbf{\Phi}$ made of atoms $(\phi_i)_{i\in I}$, with I a given set of indices, and define $R^0\mathbf{x} = \mathbf{x}$, then at iteration step $k \geq 1$ the new residual $R^k\mathbf{x}$ is computed by the rule

$$R^k\mathbf{x} = R^{k-1}\mathbf{x} - \mathbf{s}_k\phi_{i_k} \tag{2.5}$$

where $\phi_{i_k} = \arg\max_{i\in I} \left| < R^{k-1}\mathbf{x}, \phi_i > \right|$, and $s_k =< R^{k-1}\mathbf{x}, \phi_{i_k} >$. The atom selected at each step is orthogonal to the newly computed residual and thus, by Pythagora's theorem, the energy of the residual error decreases strictly which in turn guarantees asymptotical convergence. The computational cost of this procedure is in general $O(MN)$ per iteration, where N the dimension of $\mathbf{x}$ and M is the number of atoms in the dictionary (the bottleneck lies in updating the correlation coefficients at each step). If the inner products of all pairs of atoms are available, then the cost can be reduced to $O(M)$. This could be useful in a setting where MP is called for many

signals $\mathbf{x}$, using the same dictionary; for this scenario to be practical the number of calls should be at least M. If the dictionary is very big however, this approach would fail because of memory constraints.

One fundamental criticism of the Matching Pursuit algorithm is the suboptimal criterion for choosing each atom. Although at each iteration step the atom most correlated with the signal is chosen, the distance from the current residual to the linear span of the atoms chosen so far is not (as we would expect) given by the current coefficients, unless the dictionary is an orthogonal basis. To address this problem, Orthogonal Matching Pursuit (OMP) [95] performs an additional step (orthogonalization) to insure that the computed residual is orthogonal to *all* of the already selected atoms. One effect is that the OMP algorithm is forced to stop after a number of steps smaller or equal to the dimension of the space, unlike MP which could potentially run infinitely. Nevertheless, OMP uses the same (rather local) criterion for atom selection. It was the so-called Optimized Orthogonal Matching Pursuit (OOMP) procedure introduced in [102] which addressed this issue. Namely, if we denote by $\mathcal{V}_k$ the vector space spanned by the first k selected atoms, and for all remaining atoms α_j we define[6] $\gamma_j = \alpha_j - P_{\mathcal{V}_k}\alpha_j$, then the OOMP approach selects the index $i_k = \arg\max_j |<\gamma_j, R^{k-1}\mathbf{x}>|/||\gamma_j||,\ \ ||\gamma_j|| \neq 0$, unlike MP and OMP, which pick $\arg\max_j |<\gamma_j, R^{k-1}\mathbf{x}>|$. Further extensions of this method emerged, *e.g.*, Backward-Optimized OMP (BOOMP [6, 101]), Swapping-Based OMP (Swap-OOMP [5]), which are heuristics meant to improve the chances of the greedy algorithm to approach the optimally sparse solution (regarding implementation issues, see also the OOMP tutorial [4]). The computational price paid for such refinements is sometimes hard to accept though, and for many situations when a faster, reasonably sparse approximation is sufficient, MP is the standard choice.

Finally we mention another approach which is frequently used in approximation algorithms to tackle hard combinatorial problems, by reducing them to the continuous domain: *relaxation*. In the case of the sparse selection of dictionary atoms, this was known as Basis Pursuit (BP) [28]. Specifically, instead of minimizing the number of nonzero coefficients of the decomposition of signal $\mathbf{x}$ into dictionary $\mathbf{\Phi}$, equivalent to solving

$$\min_{\mathbf{s}} \quad ||\mathbf{s}||_0 \quad s.t. \quad \mathbf{x} = \mathbf{\Phi}\mathbf{s} \tag{P_0}$$

we attempt to solve the relaxed variant in ℓ_1 sense:

$$\min_{\mathbf{s}} \quad ||\mathbf{s}||_1 \quad s.t. \quad \mathbf{x} = \mathbf{\Phi}\mathbf{s} \tag{P_1}$$

The obvious advantage of this change of objective is that now we are dealing with a convex problem, which can be conveniently formulated as a linear program. Less obviously, in certain conditions the optimal solutions of these two seemingly different problems coincide (see the multitude of results on this topic contained in *e.g.*, [28, 48, 49, 112]). Moreover, numerical experiments have also confirmed the merits of this approach, displaying its superiority to greedy algorithms in applications such as signal denoising. In spite of this fact, BP fails to offer

[6] By $P_{\mathcal{V}}\mathbf{x}$ we denote the orthogonal projection of $\mathbf{x}$ onto the linear space $\mathcal{V}$.

satisfactory computational speed when compared against Matching Pursuit, even for relatively structured dictionaries.

To conclude, we remark that computing sparse representations is desirable, but in general remains computationally challenging. In the following, we will describe an instance of a successful compromise between sparsity and speed.

2.4 Shiftable Kernel Representations.

As we argued previously, the need for linear adaptive signal representations is motivated by many families of computational procedures involving "intelligent" ways to represent data. On the other hand efficiency has been intuitively associated with sparse representations. For natural sounds and images, as argued by Simoncelli and coll. [51], an additional desirable property is shift-invariance. Combining these different requirements has recently lead to new sparse signal representations based on adaptive shiftable-kernel dictionaries.

Introduced and further studied by Smith and Lewicki [107, 108], Spike Coding is such a method, which proved particularly successful in providing a biologically plausible, nonparametric acoustic model for the mammal auditory system (see also [106]). The computational principles are independent of the modality and they apply for sounds, as well as for images or video. (For example, spike-based models were proposed by Perrinet et.al. [96], and Rozell et.al. [104].

In the following, we will give a basic mathematical description of the model (as derived for 1D signals), and address several aspects regarding its implementation. The objective of Spike Coding is linear approximation of a signal $\mathbf{x} \in \mathbb{R}^L$ with a set of K shiftable kernels $\mathbf{\Phi} = \{\phi_k\}_{1 \leq k \leq K}$, such that the representation is as sparse as possible. Thus, the optimization problem can be expressed as

$$\begin{aligned} &\min_{\mathbf{s}} ||\mathbf{s}||_0 \\ &s.t.\ \ ||\mathbf{x}(\cdot) - \sum_{k,t} s_{k,t}\phi_{k,t}(\cdot)||_2 < \epsilon. \end{aligned}$$

As this problem is NP-hard (see section 4.2), we can only hope to approximate its solution. One approach which is computationally very attractive, is Matching Pursuit [87]. Atoms in the signal representation are obtained by selecting the shifted version of a kernel most correlated with the signal, thus picking out the single most informative feature available at each step. More precisely, if $R^0\mathbf{x} = \mathbf{x}$, the *residual* signal corresponding to the n-th iteration is computed as:

$$R^n\mathbf{x} = R^{n-1}\mathbf{x} - s_n\phi_{k^n,p^n} \tag{2.6}$$

where $\phi_{k^n,p^n} = \arg\max_{k,p} |< R^{n-1}\mathbf{x}, \phi_{k,p} >|$, and $s_n =< R^{n-1}\mathbf{x}, \phi_{k^n,p^n} >$.

It is worth pointing out that this encoding method is close to being shift invariant. By translating the whole signal by a sample in any direction, most of its correlation with the kernels will suffer an appropriate change in position, but not in value. In fact the only places where this does

not hold true is at the boundaries, and their number is less and less significant as the size of the signal gets larger. Due to the particular form of the dictionary we are effectively combining a convolution-based method with the greedy procedure of picking a very sparse coefficient set. Efficient implementations of MP in the one-dimensional case have been proposed independently by Sallee [105], and Gribonval et al. [74, 75], which reduce the computational complexity of an MP iteration to $O(K\ \log L)$, instead of $O(KL)$.

It is possible to reduce the size of the representation (in ℓ_0 sense) by adapting the kernels to the class of signals they should best represent, and thus by increasing their descriptive power. In the following we will regard the optimal dictionary for a given set of points, in terms of searching for the mode of the posterior distribution in a similar fashion to [91, 107]:

$$p(\mathbf{x}|\mathbf{\Phi}) = \int p(\mathbf{x}|\mathbf{\Phi}, \mathbf{s})p(\mathbf{s})d\mathbf{s} \tag{2.7}$$

where the integration is made by marginalizing over all possible point sets. If the integral above is approximated by $p(\mathbf{x}|\Phi, \mathbf{s}')p(\mathbf{s}')$, where $\mathbf{s}'$ is the set of coefficients produced by Matching Pursuit, and assume that ϵ, the representation noise, is distributed according to $\mathcal{N}(0, \sigma_\epsilon \mathbf{I})$, then for every kernel ϕ_k we can find:

$$\begin{aligned}
\frac{\partial}{\partial \phi_k} \log(p(\mathbf{x}|\mathbf{\Phi})) &= \frac{\partial}{\partial \phi_k}\{\log(p(\mathbf{x}|\mathbf{\Phi}, \mathbf{s}')) + \log(p(\mathbf{s}'))\} && (2.8)\\
&= \frac{-1}{2\sigma_\epsilon}\frac{\partial}{\partial \phi_k}||\mathbf{x} - \sum_{j=1}^{K}\sum_{p=1}^{n_j} s'_{j,p} \cdot \phi_{j,p}||_2^2 && (2.9)\\
&= \frac{1}{\sigma_\epsilon}\sum_{p=1}^{n_k} s'_{k,p} \cdot [\mathbf{x} - \widehat{\mathbf{x}}(\mathbf{s}', \mathbf{\Phi})]_{k,p} && (2.10)
\end{aligned}$$

where the expression $[\mathbf{x} - \widehat{\mathbf{x}}(\mathbf{s}, \mathbf{\Phi})]_{k,p}$ denotes the restriction of the error signal to the support of $\phi_{k,p}$. Thus, we have a learning rule for the MAP dictionary. On the other hand, maximizing the posterior with respect to $\mathbf{\Phi}$ is equivalent to minimizing the (squared) reconstruction error, which is simply a quadratic form of the ensemble $\mathbf{f}$, the concatenation of all the kernels in the dictionary. We shall further develop this issue in chapter 4.

Chapter 3

Multiresolution ICA

3.1 Introduction

The problem of efficiently describing visual structure has been of great importance to many research fields spanning from biology to engineering (see *e.g.*, [92, 117]). Best existing coders (notably JPEG2000 [109]) rely on the flexibility of multiresolution (MR) transforms to capture structure in natural images by exploiting their intrinsic multiscale character [54, 85]. The most important and practical analysis tool employed to access this structure was the Discrete Wavelet Transform (DWT) [34, 86, 117]

In spite of their success, wavelets do have well-known limitations in terms of modeling or detecting two-dimensional, sharp, arbitrarily-oriented (ridge-like) discontinuities. Various types of MR representations emerged in the past decade in computational harmonic analysis, which provably outperform wavelets in approximating particular classes of signals [22, 42]). Because of their great diversity, it is not clear what makes an optimally efficient code for images. Common intuition that optimal image features are smooth surfaces and short straight edges may be accurate for some classes (*e.g.*, natural scenes [54]), but not for others (faces, textures, cartoons, fingerprints, and medical images of all sorts).

Separating signal content into different subbands, and concentrating the relevant information into a small set of non-zero coefficients, seems a natural recipe for achieving efficiency. However, a representation is inherently suboptimal unless it can capture the probability density of the data, according to Shannon's source coding theorem. As such, optimal efficiency can only be achieved by adapting the representation to the statistical structure of the target image class. When searching for the "most compact" code, one method to employ is independent component analysis (ICA) [64]. Generally speaking, the goal of ICA is to derive a data dependent linear mapping such that the coefficients in this new representation are maximally independent. Therefore, a suitable mathematical cost to minimize is the mutual information among coefficients. Due to its poor computational scalability with respect to data dimensionality, ICA has been traditionally applied to images (either for analysis, encoding, or denoising) by extracting relatively small image patches to be used as training samples, followed by block-transforming the image. Unfortunately, the arbitrary alignment of the blocks with the image and the insufficient capacity

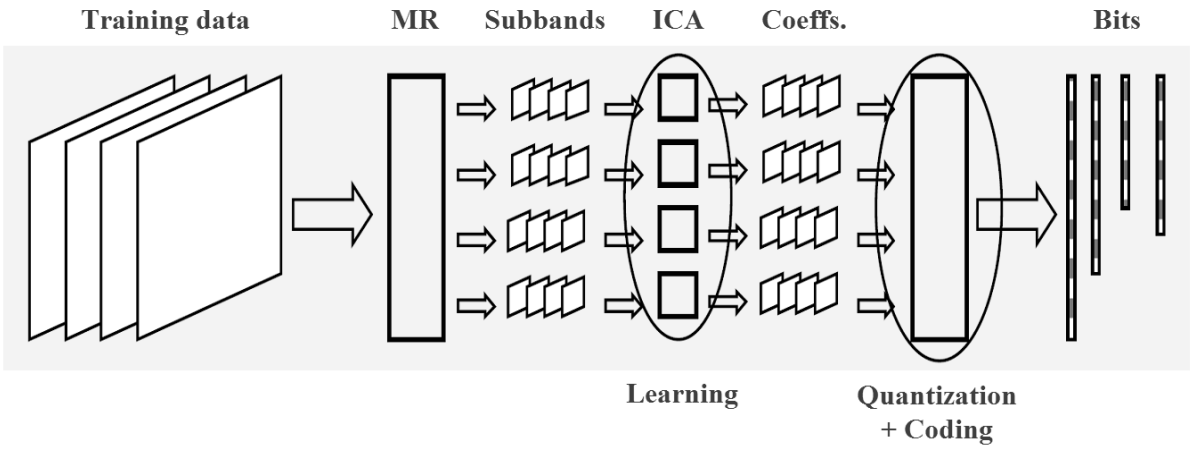


Figure 3.1: Multiresolution ICA flowchart.

to represent image structure spread across blocks produce artifacts at reconstruction.

In this chapter, we propose an ICA-like image representation, which overcomes the artificial block confinement and computational obstacles. Our method consists of a preliminary MR (*e.g.* wavelet) decomposition step, followed by learning an ICA basis for each of the resulting subbands. The purpose of the MR step is to allow easier access to structure at each scale, while reducing the bulk of image information to the coarsest scale; indirectly, this helps in eliminating blocking artifacts. Since the learned ICA bases provide the most compact linear code for each subband, we can conclude that this hybrid Multiresolution-ICA procedure (henceforth referred as MrICA) gives an improvement over both types of representations. For a flowchart of the MrICA procedure described above, see fig. 3.1.

Efficient coding has been a very suitable paradigm in attempting to explain how biological systems cope with processing complex information. The resemblance of the optimally derived linear features learned from natural scenes to the receptive fields of simple cells in primary visual cortex (V1) has led to very interesting hypotheses about the role and function of the brain's sensory systems [14, 92, 115]. A probabilistic modeling approach aimed directly at optimal efficient coding of natural images [80] has revealed that the average entropy improvements of adaptive linear representations over fixed ones (Fourier, DCT, wavelets, Gabor functions [35, 36]) are too important to neglect. However, due to the computational constraints, their representation was derived for relatively small image patches and thus a comparison to multiscale bases was limited. We can mention here two other block-based ICA approaches to image compression [53, 89]. Both compare favorably to JPEG (for faces and natural images), and the first one even outperforms the FBI Wavelet Scalar Quantizer (WSQ) coder [17, 18] for fingerprint images at low rates; however, they do not fully exploit the potential of multiresolution. Modeling subband information statistically for image coding has been performed in [19]; the resulting coder (EPWIC) explicitly exploits statistical relationships between coefficients in different wavelet subbands via a parameterized model. Another parametric adaptive multiscale method has been presented in

[93]; there the objective is to adapt the parameters of a certain wavelet-based transform to better fit natural images. In contrast to their approach, MrICA derives an adaptive *non-parametric* multiscale image representation by letting the subband ICA basis functions be learned from scratch, therefore keeping them unconstrained. Finally, a different multiscale framework for blind separation is presented in [71], and [120]. The essential difference between their work and the MrICA lies in the nature of the mixtures: we regard the images as being *sample points/vectors* drawn independently from a certain distribution, while in their case the images are the *mixtures* and the samples correspond to sets of pixels drawn from all images, at identical spatial locations.

Structure of the chapter. In Section 3.2, we describe the main components of our image encoder. We address the issues of computational complexity of MrICA in section 3.3. Section 3.4 describes in detail the experimental results illustrating the encoding performance of the proposed adaptive method, while section 3.5 concludes the chapter.

3.2 Adaptive Multiresolution Coding

In this section, we shall describe the proposed method for MR adaptive image encoding. We shall start by assuming we have a set of sample images $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ assumed to be drawn *iid* from a common distribution over $\mathbb{R}^N$ (here N is their common size). We shall decompose these images by a fixed MR transform and then for each of the resulting subband spaces we shall learn an ICA basis, by using the subband coefficient sets of all the images in the sample as training data. Finally, we shall use the subband ICA coefficients to design a quantizer. Every new image will be transformed and quantized, and finally output into a bitstream via an arithmetic coder. In the following, we provide details on each of the modules of our system.

Multiresolution Transforms. The first step of our hybrid method aims at separating image content by projecting images on scale-orientation subbands. The most widely used MR transforms are based on wavelets and this due their theoretical and computational properties. For instance, JPEG2000 (Part 1) uses the Cohen-Daubechies-Feauveau 9/7 biorthogonal wavelet filters [34], as its only supported "irreversible" wavelet transform [109]. We also chose to employ this wavelet because we wanted to test the coding efficiency of our method against that of the most common fixed MR transform. To share further similarities with existing image coders, we applied this separable decomposition method, using whole-point symmetric edge handling. The implementation we used in our experiments was that of Matlab Wavelet Toolbox 4.2. (For the sake of completeness, let us mention that the CDF 9/7 wavelets are referred there as `'bior4.4'`.)

Formally, we consider that the wavelet transform is represented by a $N \times N$ invertible matrix $\mathbf{M}$, partitioned (row-wise) into $N_k \times N$ matrices $\mathbf{M}_k$, with $1 \leq k \leq K$. In our case, the submatrices correspond to different scale/orientation subbands. For each subband k, and image $\mathbf{x}_j$, let $\mathbf{x}_j^{[k]} = \mathbf{M}_k \mathbf{x}_j$ be the coefficients of the image over the subband. The goal MrICA is to derive for each k the $N_k \times N_k$ matrices $\mathbf{W}_k$ such that $\mathbf{s}_j^{[k]} = \mathbf{W}_k \mathbf{x}_j^{[k]}$ are realizations of a random vector with maximally independent components. In other words, within each subspace defined by the partition $(\mathbf{M}_k)_k$ we search for the linear mapping allowing us to best approximate the projected

data distribution by a product of marginals.

Adaptation. In unsupervised learning, the problem of separating signals into independent linear components can be formulated as follows: given a set of N-dimensional vectors $(\mathbf{y}^{(k)})_{1\leq k\leq K}$, search for a linear transform $\mathbf{A}$ such that the observed vectors are linear mixtures (induced by $\mathbf{A}$) of realizations of an M-dimensional random vector $Z = (z_1, \ldots, z_M)^T$ whose components are as independent as possible. We can express this model compactly as:

$$\mathbf{Y} = \mathbf{A}\mathbf{Z} \tag{3.1}$$

where $\mathbf{Y} \in \mathbb{R}^{N\times K}$, $\mathbf{A} \in \mathbb{R}^{N\times M}$, and $\mathbf{Z} \in \mathbb{R}^{M\times K}$. The ICA computational objective is then to find the linear transform $\mathbf{A}$, such that the mutual information among the coefficients z_i is minimized. To simplify the description, we will assume that $\mathbf{A}$ is square and invertible (that is, $M = N$) and if we denote its inverse by $\mathbf{W}$, the problem is reduced to minimizing:

$$I(z_1, \ldots, z_N) = \sum_{j=1}^{M} H(z_j) - H(\mathbf{Y}) - \log|\det \mathbf{W}| \tag{3.2}$$

Since the entropy of the observed mixture $\mathbf{Y}$ is constant, imposing $|\det \mathbf{W}| = 1$ causes the quantity we seek to minimize to be the sum of the coefficients' marginal entropies; that is, ICA searches for the transformation giving the (potentially) most compact linear code of the data.

When the size of the training images is very high let us observe that the size of the subbands at the first decomposition level (roughly one quarter of that of the original image) is still very large and we cannot learn a complete basis for the subband. However, by applying a variant of our method called *modified MrICA* we can overcome this obstacle. Namely, we impose that for all subbands up to some decomposition level L' we learn ICA bases in a block-based fashion, while for the coarsest subbands we perform MrICA as usual. As we will later point out, the computational savings will be tremendous since we only need to solve a linear number of conveniently small ICA problems. Moreover, since most of the wavelet coefficients are already very small, there will be virtually no blocking artifact in the reconstruction. (Let us point out that this is *not* the same as decomposing the image into moderately large blocks (*e.g.*, 64×64) and applying MrICA to the blocks.)

Quantization and Coding. Next, we shall describe the subband coding procedure employed to transform the coefficients into bitstreams, for both the wavelets and MrICA. For a group of images from the training set, we group the MR coefficients belonging to the same subband and from the whole group, we estimate a scalar quantizer. Note that scalar quantization is justified in the case of MrICA, since coefficients within each subband are as independent as possible. To design the subband quantizers, individual bit rates are allocated according to the relative energy within each subband. Since we are interested in the potential improvement of the adaptive representation, and less so in the great many practical issues of image coding, we will compute the "optimal" entropy-constrained scalar quantization [52] for each subband. This should provide a reliable *upper bound* for the performance of each representation. After quantizing the coefficients, we use Matlab Communication Toolbox's arithmetic coder to construct the bitstreams

and record the total bitstream length and the reconstruction SNR for each test image. Then, we take the average over the whole test set to estimate the coding efficiency of the distribution. We repeat this procedure for various target rates, and by interpolation we construct the rate-distortion curve.

3.3 Complexity Issues of MrICA

An important aspect that motivates our hybrid method is the lack of ability to compute an ICA basis for large-dimensional data. As we have mentioned in the background chapter (see section 2.2) the computational cost becomes prohibitive because of the large number of parameters we need to estimate. In the following we shall explain how this problem is handled by MrICA.

For simplicity, let us assume that we employ a wavelet basis for the MR step; for one decomposition level, this will produce three detailed subbands (horizontal, vertical, and diagonal) and one approximation subband whose dimension will be 1/4 of the original image size. Let us denote by $T(n)$ the computational cost of performing one ICA iteration when the size of the data is $n \times n$; suppose the original image size is $d \times d$ and say we use the wavelet decomposition with L resolution levels. Then, the total cost of one ICA step *across subbands* is:

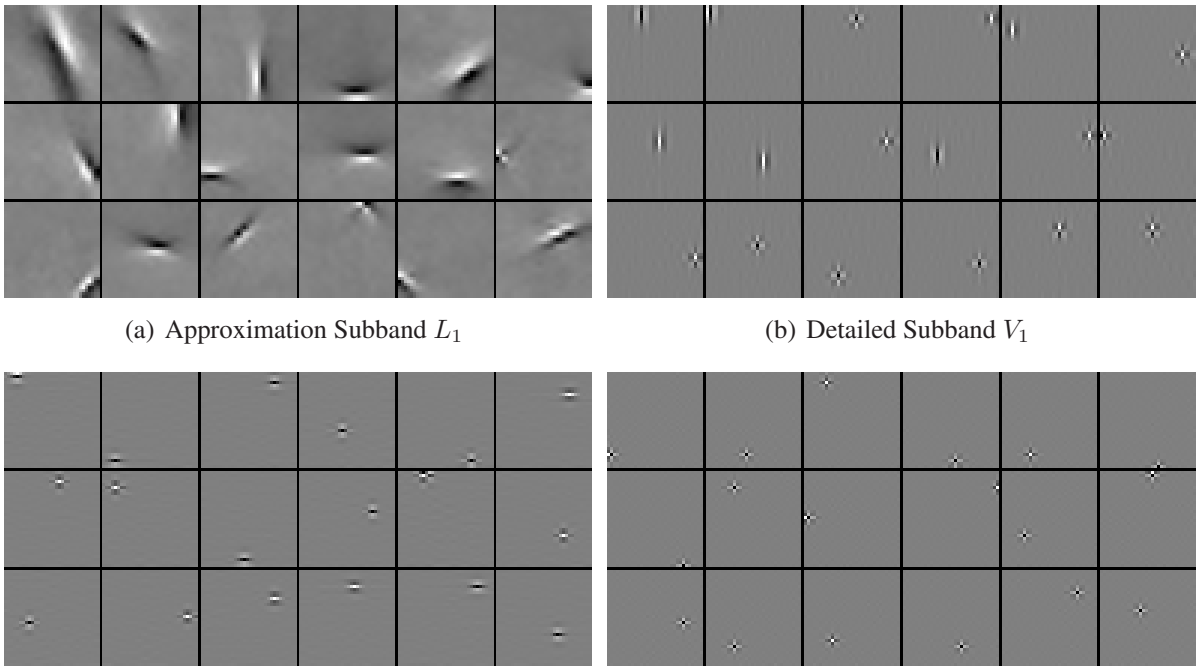
$$3\left[T(\frac{d}{2}) + T(\frac{d}{2^2}) + ... + T(\frac{d}{2^L})\right] + T(\frac{d}{2^L}) \tag{3.3}$$

The iteration cost of a typical off-line ICA algorithm which involves a matrix multiplication, (or an inversion, or a re-orthogonalization) is of the order $n^{2\log 7}$. Thus a rough estimate of the ratio between the cost of a MrICA iteration (again, for all subbands) and $T(d)$ is

$$3\sum_{t=1}^{L} \frac{1}{(2^{2\log 7})^t} = 3\sum_{t=1}^{L} \alpha^t \approx \frac{3\alpha}{1-\alpha} \tag{3.4}$$

where $\alpha = 2^{-2\log 7} \approx 0.0204$. Thus, the computational savings are significant, the entire process getting to be roughly 16 times faster. If we also take into account the fact that for a smaller problem we generally need fewer iterations to converge, we realize that even higher savings are possible. Let us point out that in the case of the modified MrICA, which treats the most detailed subbands up to level $L' < L$ in a block-based fashion, the approximate cost ratio computed like above becomes $(3L' + 4)\alpha^{L'}$). For a very large image, and even a moderate limit level L', the cost is reduced tremendously.

We should emphasize that for each image class that we are interested to represent, we need to pay the computational price of learning only once. That is, having learned the MrICA basis for our class we will be able to use it whenever is necessary. This is a common practice in (off-line) machine learning and thus applies to our adaptive signal coding setting.



(a) Approximation Subband L_1 (b) Detailed Subband V_1

(c) Detailed Subband H_1 (d) Detailed Subband D_1

Figure 3.2: Basis functions computed by MrICA with 1 MR decomposition level for 32x32 log-scale natural images (see text). For each subband, a random set of basis functions are displayed.

3.4 Experimental Results

We shall illustrate the encoding performance of MrICA by plotting the average rate-distortion curves generated by coefficient quantization at various levels of precision, for both the fixed and the adaptive MR transforms. First of all, however, let us comment on the features learned by our method when applied to natural images.

We applied MrICA to encoding natural scenes randomly cropped from van Hateren's database of natural stimuli [116]. We tested our method on images of two sizes, 32×32 and respectively, 64×64 pixels[1]. Instead of working with the pixel intensities, we took the logarithm of these intensities before any further processing; as explained in [116], the reasons for this operation are to incorporate contrast invariance of natural scenes, get better first-order statistics of the natural image data, and better mimic the operations performed by the first stages of visual systems. On each of these logarithmically transformed images, we applied the discrete wavelet decomposition and learned the subband ICA matrices, as described in the previous section.

Interpreting the ICA objective as maximizing the (log-)likelihood of the data under the linear model, or as minimizing the Kullback-Leibler divergence between the joint probability and the product of marginals has produced several families of ICA algorithms (see section 2.2). For the results reported in this section we employ the Relative Trust-Region algorithm [29]. Figure 3.2

[1]As they are similar to JPEG2000 standard code blocks sizes, we considered these image sizes relevant to use for comparison purposes.

displays a random set of such ICA basis functions learned from the 32×32 data set with one MR decomposition level. MrICA basis functions of the approximation subband retain the aspect of classic image ICA basis functions (relatively low spatial frequency, all orientations) [14, 116], which is not a surprise considering that the approximation subband contains a low-resolution version of the original image. On the other hand, the detailed subbands basis functions look like localized features, preserving the dominant orientation of the subband. Besides the quantitative (mutual information) difference between the MrICA detailed bases and corresponding wavelets, we note that the adaptive features are also more diverse in shape (*i.e.*, they are not shifted copy of a single wavelet kernel).

Next, we illustrate the improvement in coding efficiency afforded by adaptiveness. In addition to comparing the adaptive and non-adaptive MR methods described in the previous section, we also compare both of them against JPEG2000; for this purpose, we used Jasper [83], a software package implementing JPEG2000. The coding cost of the adaptive and non-adaptive representations does not include the basis functions (respectively, the wavelets), as these can rightfully be considered part of the coder, and not of the code; also, in case of Jasper, we report only the codestream length (*i.e.*, not including the metadata). In the case of MrICA, the images included for the evaluation are taken from the testing set, that is, they belong to the same signal class as those in the training set but have not been used during learning. Let us point out that the JPEG2000 performs quantization for each individual image, and not over a whole sample set, unlike our method. In this respect, our quantizers take advantage of more information. On the other hand, JPEG2000 performs surprisingly well considering that we used an optimal ECSQ, and not a uniform one. The rate-distortion trade-off obtained for the 32×32 test images, with one MR level, for the three encoding methods, are presented in Figure 3.4. The top plot shows the relative coding gain, while the bottom plot shows the relative bit-rate difference of the three methods, taking the non-adaptive wavelet representation as reference. The better coding efficiency of MrICA (more apparent at low bit rates) has two important consequences: the same distortion (or SNR) can be achieved by the adaptive method for a significantly lower rate (*i.e.*, bit cost) and, reciprocally, for a fixed bit rate we can get a significantly better improvement in fidelity.

As it is well known, SNR is not a relevant measure of perceptual distortion; instead, evaluating the representational power of MrICA should also involve assessing the presence of reconstruction artifacts. For this purpose, we chose to display several test images from the two datasets, their encoded version via the adaptive and non-adaptive MR transforms, and the residual errors. Figure 3.5 illustrates the encoding results of six examples from the 32×32 dataset. Each image has been encoded to a quality of 25dB; the coding gains of MrICA over the non-adaptive wavelet method for these images are: 2.43 bpp, 0.62 bpp, 2.91 bpp, 2.78 bpp, 3.39 bpp, and 2.69 bpp. Figure 3.6 shows six examples of 64×64 images encoded at 20dB. The coding gain values of the adaptive method are in this case 1.5 bpp, 1.48 bpp, 0.33 bpp, 0.23 bpp, 0.45 bpp, and 1.23 bpp. (For both figures, the colormaps are maximally stretched to enhance visibility.) As a general conclusion, MrICA obtains a better coding rate than the fixed wavelet representation, with fewer reconstruction artifacts.

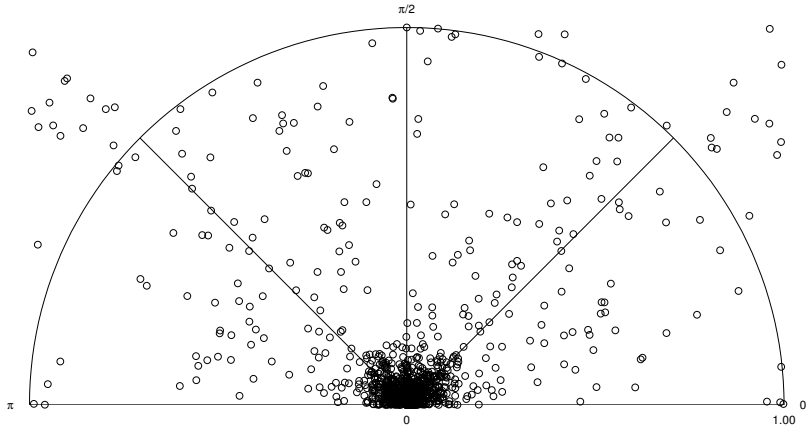
3.5 Concluding Remarks

We proposed MrICA, a hybrid method that combines the advantages of both adaptive and multiresolution representations. We illustrated the significant coding efficiency gain of MrICA over the wavelet transform when applied to natural images, which is explained by the ability of the new method to adaptively describe image structure at all scales. This suggests that an image coder devoted to a given class of signals should use not only multiresolution, but also adaptivity to optimize encoding performance.

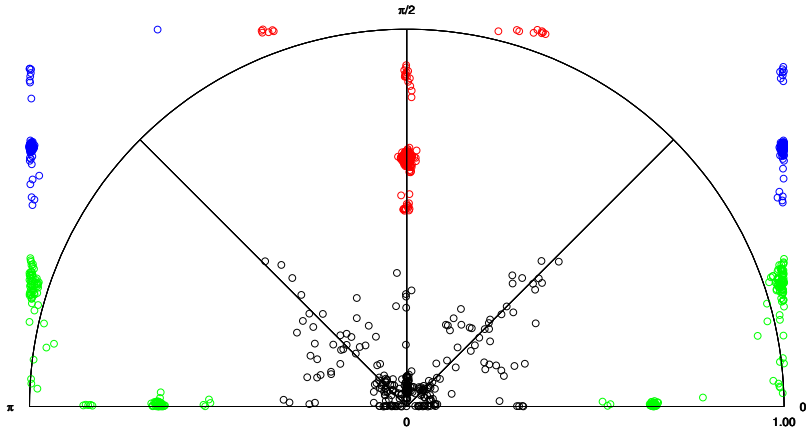
One particularly important issue that remains unsolved is the optimality guarantee: it would be very useful to find an explicit relationship between the lower bound of the "classic" ICA objective function, and the one achievable by MrICA, when applied to the same data. Since our method reduces only intra-band redundancies, we expect that in general this difference will not be negligible. This is definitely a very interesting direction that we intend to pursue in the future work, as it will help us measure the trade-off between coding efficiency and computational complexity more accurately.

Finally, let us mention that applying the method described here is by no means restricted to the class of natural images. Indeed, MrICA provides a general framework for data-dependent signal coding which could potentially be useful in representing more restricted image classes, such as faces[2] and medical images, and even to other modalities (*e.g.*, speech and video).

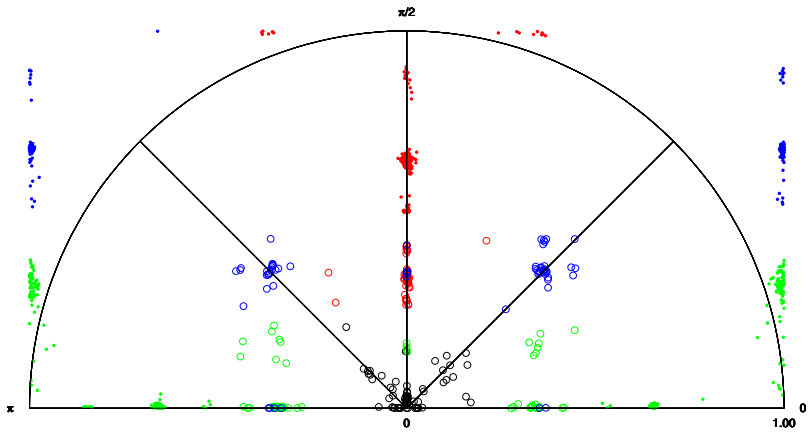
[2]One particularly interesting application we recently heard of is the Automatic Cameraman project at UCSD [55]



(a) 32×32 blocks without MR decomposition.



(b) 32×32 blocks, 1 MR decomposition level.



(c) 32×32 blocks, 2 MR decomposition levels.

Figure 3.3: Layout of the parameters corresponding to MrICA basis functions, in the Spatial frequency (radial) vs. Orientation (angular) domain. The black circles represent the parameters of MrICA basis functions computed for the approximation subband. Colored circles represent basis functions from the intermediate detailed subbands (red=horizontal, green=vertical, blue=diagonal). Colored dots represent basis functions from the highest resolution detailed subbands.

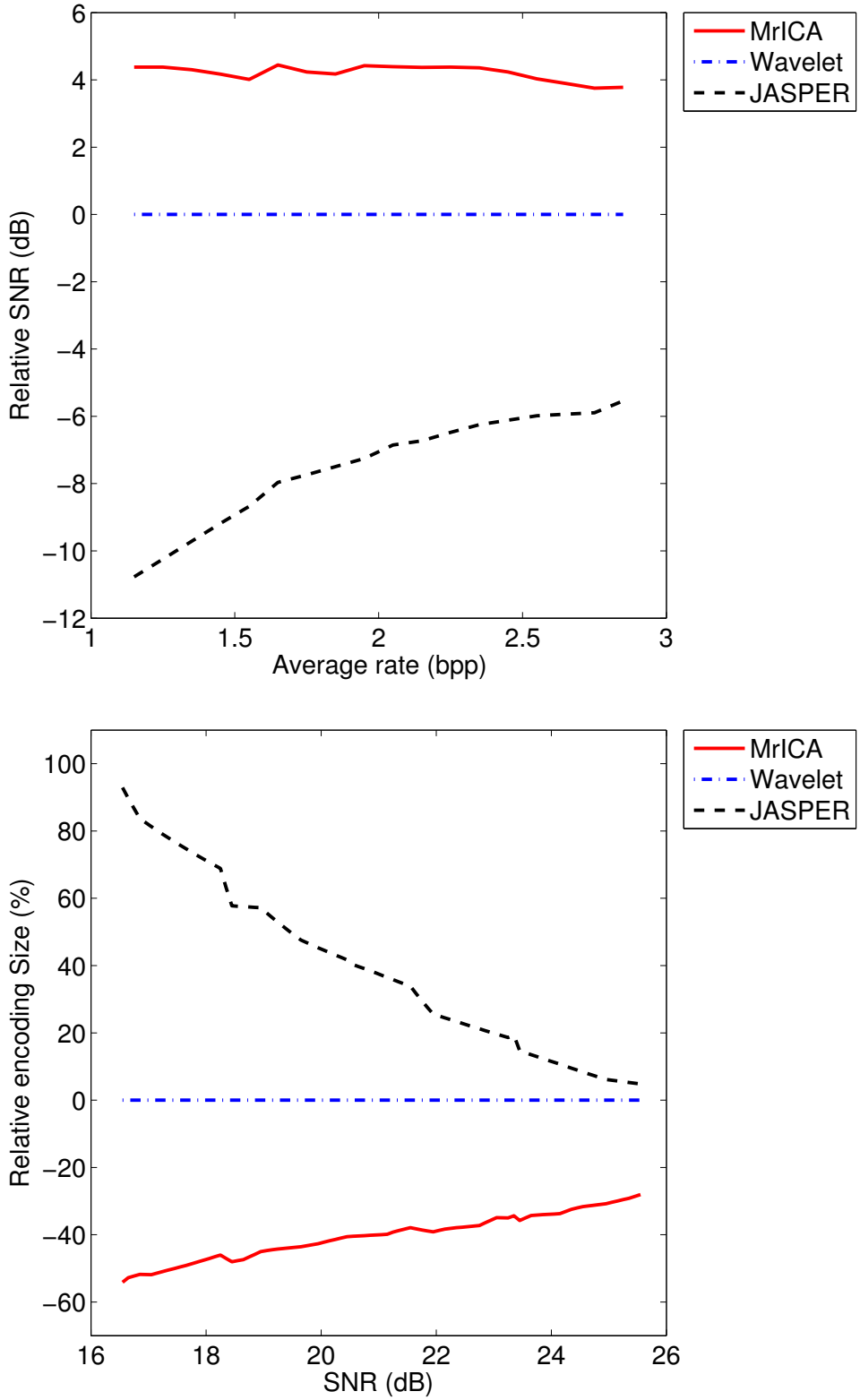


Figure 3.4: Relative rate-distortion performance of three methods (MrICA, wavelet, JASPER) computed for the 32×32 test images.

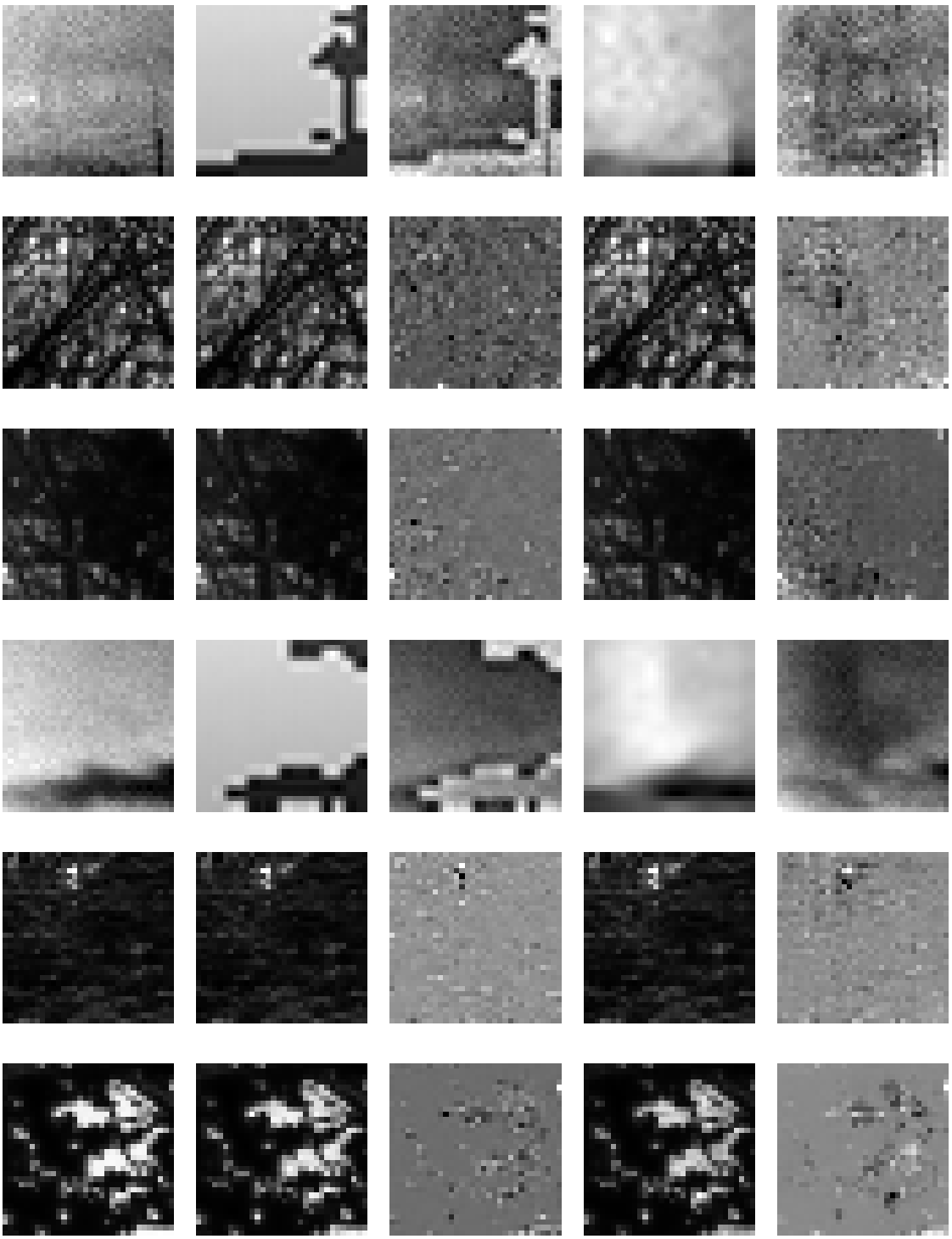


Figure 3.5: Examples of 32×32 images reconstructed at 25dB. Column 1.Original images; 2.Image encoded by non-adaptive method; 3.Error; 4.Image encoded by MrICA ; 5.Error

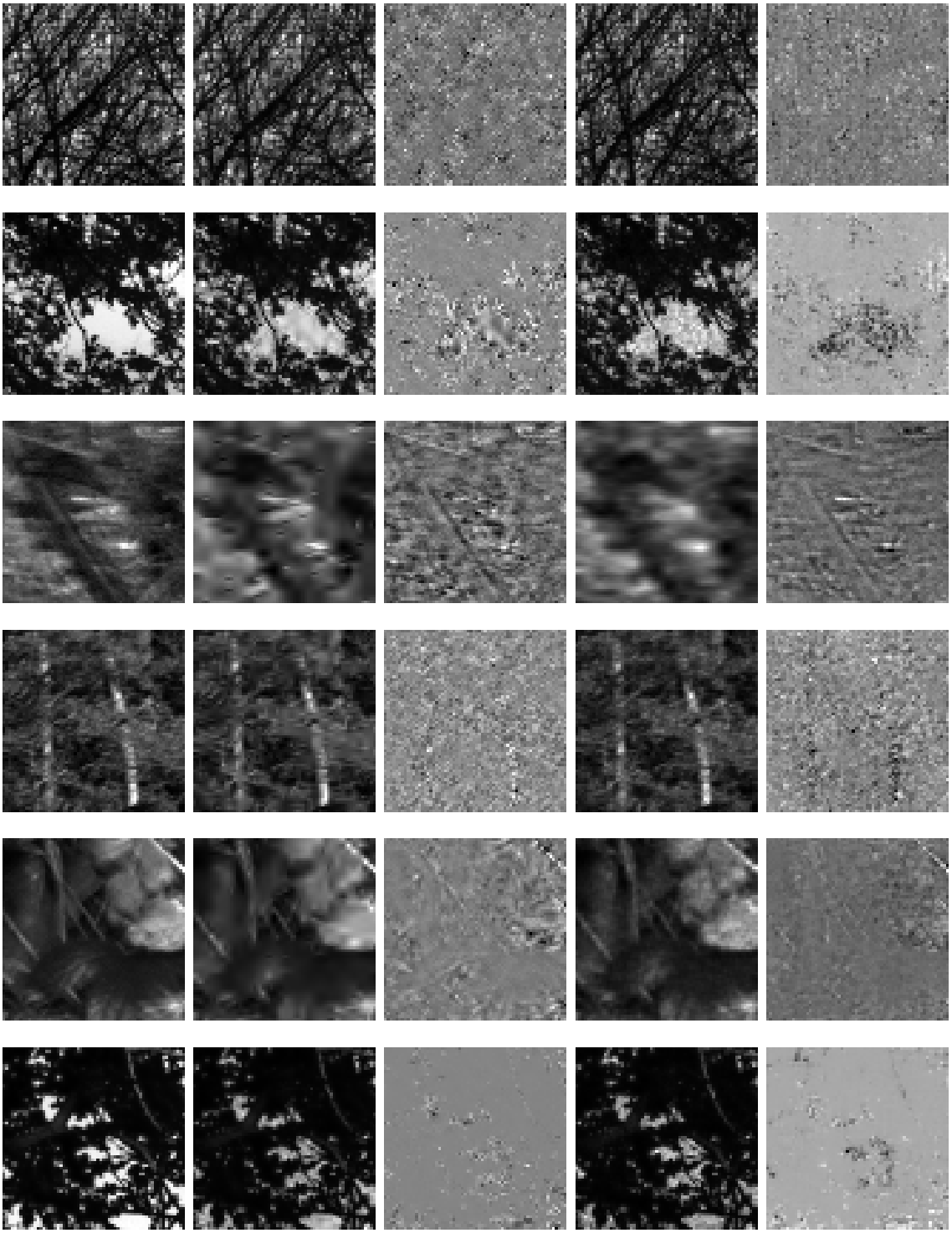


Figure 3.6: Examples of 64×64 images reconstructed at 20dB. Column 1.Original images; 2.Image encoded by non-adaptive method; 3.Error; 4.Image encoded by MrICA ; 5.Error

Chapter 4

Point Coding

4.1 Introduction

Efficient image representation is an important research problem, due to its practical applications [50] and to its potential as a principled approach to modeling natural vision [92]. Capturing the structure of visual signals and encoding it compactly is a challenging task. For example, relevant visual content can appear at any spatial position and scale, which explains the success of image coders based on multiscale representations such as wavelets [19, 109]. However, even wavelets prove suboptimal in modeling certain structure frequently occurring in images (among other things, sharp edges at arbitrary orientations) and recently several new representations have been designed to fill this gap (see for instance [20, 21, 42]).

In spite of spectacular progress, it still is not clear what constitutes an optimally efficient code for images in general: smooth surfaces and short straight edges may be the optimal features for some image classes (*e.g.*, natural scenes), but not for others (faces, cartoons, fingerprints, various types of texture, or medical images). Furthermore, according to Shannon's source coding theorem, a representation is inherently suboptimal for a given class of signals of interest, unless it captures the probability density of the data. This suggests that better representations can be obtained by learning more general and flexible dictionaries that reflect the statistical structure of particular image classes. In this chapter, we focus on adaptively deriving dictionaries defined by a set of relatively small image patterns (henceforth called "kernels"), shifted at arbitrary positions. The goal is to find such a set of kernels for which any signal in the class of interest has a sparse linear representation. Each coefficient in the sparse set represents a triple: one component is its scaling value (including sign), the second is the index k of a kernel, and the third one is a point p in signal space - the location of the shifted kernel k. Therefore, from now on we refer to the problem above as *Point Coding*. The usual approach to computing a solution is to minimize a two-term cost function: the first term measures the fidelity of reconstruction (usually, an *error* term), while the second term stands for some form of regularization (in our case, *sparsity*). By employing different choices for these two terms various algorithms can emerge, each with its own advantages and technical challenges.

This area of research has been particularly active in recent years, and several interesting di-

rections have been explored [3, 28, 67, 91]. However, this setting is by no means particular to images: approaches to sparse adaptive representation of general types of signals exist which focus on sounds [59, 84, 108], or combined audio-visual signals [88]. Our goal is to bring forward computationally efficient methods for designing very general, (approximately) shift-invariant adaptive image representations. The generality refers to the fact that the kernel sizes can be *distinct* and *arbitrary*; the user gives up the control over this issue to the kernel-learning procedure and the adaptive process may even produce a multiscale representation if optimality demands it (in a similar fashion to Spike Coding [106, 107]). Computational efficiency stems from exploiting the structure of the optimization problem, using tools imported from structured matrix algebra [61, 70, 94]. Namely, the coefficient extraction step uses a fast implementation of Matching Pursuit with essentially logarithmic cost per iteration, while dictionary update is performed by solving a highly structured least-squares problem, either by algebraic characterization of pseudoinverses of certain structured matrices [60], or by fast interpolation methods [76, 114].

The chapter is organized as follows. Section 4.2 contains the mathematical formulation of the problem. Next we describe the alternative steps of the numerical solution: section 4.3 reviews the sparse coefficient extraction and section 4.4 is concerned with the dictionary update. We include the experimental results in section 4.5 and present our conclusions in section 4.6.

4.2 The Point Coding Problem

In this section we shall formulate the Point Coding problem mathematically. Let us start by introducing the appropriate notation.

Let $\mathbf{\Phi} = \{\phi_1, ..., \phi_K\}$ be a set of two-dimensional (rectangular) kernels, of possibly different sizes $m_k \times n_k$, normalized to unit Frobenius norm, and let $\mathbf{f} = \left(\mathbf{vec}(\phi_1)^T, \ldots, \mathbf{vec}(\phi_K)^T\right)^T$ the ensemble obtained by concatenating their vectorized versions[1]; thus, if $D = \sum_{k=1}^{K} m_k n_k$, then $\mathbf{f} \in \mathbb{R}^D$. For each kernel $\phi_k \in \Phi$ and for $p \in I\!N^2$, we denote $\phi_{k,p}$ the translated version of the kernel such that its upper-left corner lies at position p. (Here, we shall work exclusively with finite-size images, which means that if ϕ_k is entirely contained within an $M \times N$ image, it can only be shifted into $(M - m_k + 1) \times (N - n_k + 1)$ positions.)

For all k and p, the coefficient of the shifted kernel $\phi_{k,p}$ will be denoted $s_{k,p}$. Under the linear additive noisy model assumption, for any image $\mathbf{x}$ we can write:

$$\mathbf{x} = \sum_{k=1}^{K} \sum_{p \in P_k} s_{k,p} \cdot \phi_{k,p} + \epsilon = \hat{\mathbf{x}}(s, \Phi) + \epsilon \tag{4.1}$$

where P_k is the set of all occurrences of kernel ϕ_k in the representation[2]. As a measure of

[1] For a matrix M, $\mathbf{vec}(M)$ is the set of all the entries in the matrix, stored column-wise.

[2] For brevity, we will further refer both to Φ and to $\mathbf{f}$ as the *encoding dictionary*. Also, we will refer to the coefficients $s_{k,p}$ as *points*. Note again that one point is determined not only by the value of the coefficient, but also by its corresponding kernel and by the position where it occurs.

representation accuracy, we hereby consider the (squared) reconstruction error:

$$F_{\mathbf{x}}(\Phi, \mathbf{s}) = ||\mathbf{x} - \hat{\mathbf{x}}(\mathbf{s}, \Phi)||_2^2 = ||\mathbf{x} - \sum_{k=1}^{K} \sum_{p=1}^{d_k} s_{k,p} \cdot \phi_{k,p}||_2^2 \tag{4.2}$$

where for all k, $d_k = |P_k|$ is the number of shifted versions of kernel ϕ_k. Optimizing for sparsity of the representation translates into minimizing the number of non-zeros in $\mathbf{s}$. Therefore, for a fixed signal $\mathbf{x}$, we should solve the following optimization problem

$$\begin{aligned} \min_{\Phi, \mathbf{s}} \quad & ||\mathbf{s}||_0 \\ \text{s.t.} \quad & F_{\mathbf{x}}(\Phi, \mathbf{s}) < \epsilon \end{aligned}$$

for $\epsilon \geq 0$ or equivalently:

$$\min_{\Phi, s} \; F_{\mathbf{x}}(\Phi, \mathbf{s}) + \lambda ||\mathbf{s}||_0 \tag{4.3}$$

for some $\lambda > 0$.

The optimization problems above are NP-hard (see for instance [37, 38, 39, 111]). Therefore, we attempt to approximately find a solution via an iterative, alternating procedure. First, we find a sparse set of points corresponding to a fixed dictionary and a preset level of precision; then, for a fixed set of coefficients, update the dictionary to better fit the data. Finding the sparsest linear approximation in a general dictionary is also NP-hard [90]; however, suboptimal approaches (like greedy) proved quite satisfactory in practice. Therefore, for the first step we choose to employ Matching Pursuit [87]. The second step, adapting the dictionary to the signal structure only requires solving a quadratic (*i.e.*, convex) optimization problem in $\mathbf{\Phi}$ (or $\mathbf{f}$). In the following, we shall describe each of the two steps separately.

4.3 Matching Pursuit

The Matching Pursuit (MP) algorithm [87] is a greedy iterative procedure whose goal is to identify a decomposition of a given vector as a linear combination of elements of a dictionary. If the dictionary is an orthogonal vector set and the signal is indeed a sparse combination of atoms, MP is guaranteed to find this sparse set. In general, this method only serves as an approximation to the sparsest set problem (see for instance [90]). For a detailed presentation of how MP can be used to compute sparse signal representations in a shiftable-kernel dictionary see section 2.4.

The main practical challenge in using Matching Pursuit with a high-dimensional, highly overcomplete dictionary is the large cost of the update and of identifying the next atom, maximally correlated with the residual. In the case of a 1-D shiftable-kernel dictionary with small kernels, approaches presented in [74, 75] and [105] shrink this cost to be $O(KL)$, *i.e.*, essentially logarithmic in the size of the signal (we assume that the number of kernels and their sizes are small constants compared to the signal size). The difference between the two approaches is that in MPTK [74, 75] the logarithmic cost is obtained by searching for the maximum coefficient within a balanced binary tree, while in Sallee's work [105] it is achieved by maintaining a

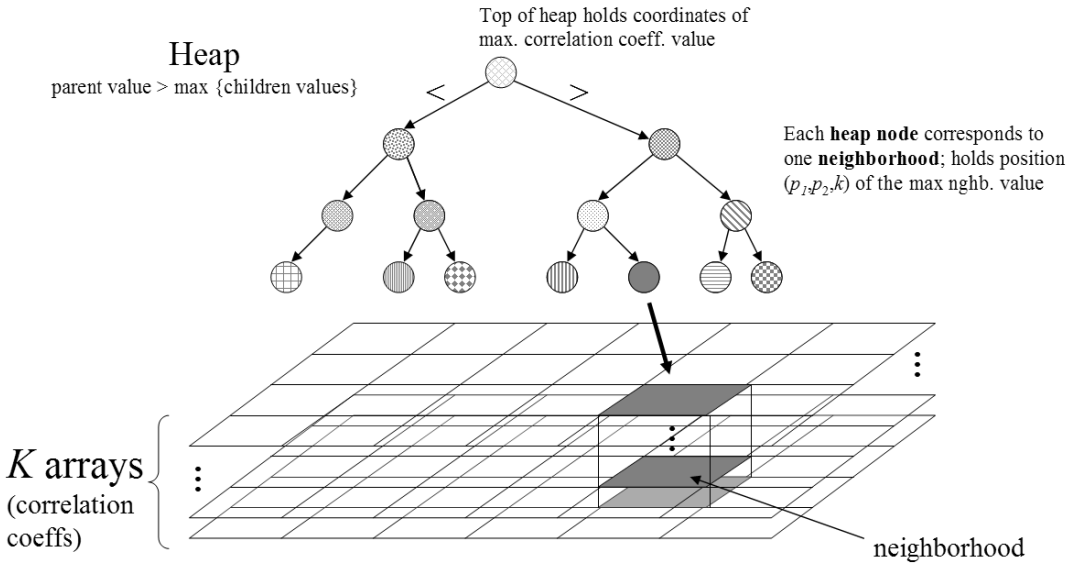


Figure 4.1: A graphical depiction of the data organization for the efficient Matching Pursuit algorithm.

heap [32] which holds the maximum correlation coefficients of the kernels with the signal over equal-length adjacent blocks. Thus, every pair of *delete-max* and *insert* heap operations, helps achieving the desired cost.

We employ the second approach, slightly adapting it to the particularities of the 2D case. Namely, we compute the correlation coefficients of the image with all the kernels in our dictionary and then divide these correlation maps into blocks of size $m_0 \times n_0$, where $m_0 = \max_k m_k$ and $n_0 = \max_k n_k$. At each step, we shall only update a small number of blocks (at most 4) and therefore we only need to search for a small number of *new maxima*. (Figure 4.1 contains a graphical description of this data structure.) The heap structure insures that we avoid most of the work required by searching the new maxima; also, careful storage of the correlation matrices further can help by enhancing data locality and thus avoid costly memory operations. As a typical result, decomposing a 256×256 image to 30dB reconstruction quality using a dictionary of 25 Gabor-looking 8×8 kernels can be executed in as little as 6.4 seconds on a G5 Mac computer (with a MEX C implementation of Matching Pursuit).

Once MP computes a sparse set of coefficients (now considered fixed), we proceed to optimizing the kernels to better fit the signals. Let us note that it is not trivial to adapt the "orthogonal" variants of MP (see 2.3) to the setting described above. Such methods attempt to remove the intrinsically suboptimal choice of MP by modifying the criterion (maximally correlated atom) as well as adding other (also greedy) heuristics as a subsequent refinement step. We are still investigating the possibility of adapting these ideas to the shiftable-kernel MP setting.

4.4 Dictionary Update

In the following we describe the computation of the optimal dictionary for a given set of points in terms of searching for the mode of the posterior distribution, in a similar fashion to [93, 107]:

$$p(\mathbf{x}|\mathbf{\Phi}) = \int p(\mathbf{x}|\mathbf{\Phi}, \mathbf{s}) p(\mathbf{s}) d\mathbf{s} \tag{4.4}$$

where for integration we marginalize over all possible point sets. We can approximate the integral above with $p(\mathbf{x}|\mathbf{\Phi}, \mathbf{s}')p(\mathbf{s}')$, where $\mathbf{s}'$ is the set of coefficients produced by Matching Pursuit. Then, assuming an additive Gaussian noise model $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon \mathbf{I})$, we can compute the gradient for every kernel ϕ_k as follows:

$$\begin{aligned}
\frac{\partial}{\partial \phi_k} \log(p(\mathbf{x}|\Phi)) &= \frac{\partial}{\partial \phi_k} \{\log(p(\mathbf{x}|\mathbf{\Phi}, \mathbf{s}')) + \log(p(\mathbf{s}'))\} && (4.5)\\
&= \frac{-1}{2\sigma_\epsilon} \frac{\partial}{\partial \phi_k} ||\mathbf{x} - \sum_{k'=1}^{K} \sum_{p=1}^{n_{k'}} s_{k',p} \cdot \phi_{k',p}||_2^2 && (4.6)\\
&= \frac{1}{\sigma_\epsilon} \sum_{p=1}^{n_k} s_{k,p} \cdot [\mathbf{x} - \widehat{\mathbf{x}}(\mathbf{s}, \mathbf{\Phi})]_{k,p} && (4.7)
\end{aligned}$$

where $[\mathbf{x} - \widehat{\mathbf{x}}(\mathbf{s}, \mathbf{\Phi})]_{k,p}$ denotes the restriction of the error image to the support of $\phi_{k,p}$. This immediately gives us a learning rule for the MAP dictionary and we could employ any (stochastic) gradient based method to perform the optimization.

Let us observe that maximizing the posterior with respect to Φ means minimizing the (squared) reconstruction error, which is simply a quadratic form of the ensemble $\mathbf{f}$. By a slight abuse of notation, we shall designate $\mathbf{x}_i$ as being both the training image $\mathbf{x}_i$ itself and its associate vector in $\mathbb{R}^{L_i}$ (L_i is the size of image $\mathbf{x}_i$). Now take to be $S^{(i)} = [S^{(i,1)}, \ldots, S^{(i,K)}] \in \mathbb{R}^{L_i \times D}$ the matrix corresponding to the linear mapping

$$\hat{\mathbf{x}}_i(\mathbf{s}^{(i)}, \mathbf{\Phi}) = S^{(i)} \cdot \mathbf{f} \tag{4.8}$$

and so the optimization problem we need to solve reduces to minimizing the following cost function:

$$\begin{aligned}
Q(\mathbf{f}) &= \sum_{i=1}^{I} ||\mathbf{x}_i - S^{(i)}\mathbf{f}||_2^2 && (4.9)\\
&= \sum_{i=1}^{I} \left(||\mathbf{x}_i||_2^2 - 2\mathbf{x}_i^T S^{(i)} \mathbf{f} + \mathbf{f}^T S^{(i)T} S^{(i)} \mathbf{f} \right) && (4.10)\\
&= ct. + \sum_{i=1}^{I} \left(-2\mathbf{x}_i^T S^{(i)} \mathbf{f} + \mathbf{f}^T S^{(i)T} S^{(i)} \mathbf{f} \right) && (4.11)\\
&= ct. - 2 \left(\sum_{i=1}^{I} \mathbf{x}_i^T S^{(i)} \right) \mathbf{f} + \mathbf{f}^T \left(\sum_{i=1}^{I} S^{(i)T} S^{(i)} \right) \mathbf{f} && (4.12)\\
&=: c - 2\mathbf{b}^T \mathbf{f} + \mathbf{f}^T \mathbf{A} \mathbf{f} && (4.13)
\end{aligned}$$

This quadratic form has a special structure: since matrix $S^{(i)}$ is a block-row whose blocks are each Toeplitz-Block-Toeplitz (TBT) matrices[3], it follows that matrices $S^{(i)T}S^{(i)}$ will be mosaic TBT matrices of identical block sizes. Consequently, matrix $\mathbf{A}$ will be a symmetric, positive semidefinite, $D\times D$ mosaic TBT matrix and therefore we can reduce the original problem to a structured least-squares problem.

Structured Least Squares. The advantage of working with structured matrices comes mainly from the fact that the number of parameters is much smaller than the actual dimension of the matrix and from the existence of fast and superfast algorithms that exploit the displacement rank of many such matrices [61, 70].

An algebraic characterization of pseudoinverses of Toeplitz and Hankel mosaic matrices has been presented in [60], which generalizes the well-known Gohberg-Semencul inversion formula for Toeplitz matrices, by using a general notion of Bezoutian. The effect is that fast and superfast algorithms can then be employed to compute the pseudoinverses, and consequently to solve the structured least-square problems.

Definition 1. *A (q,p)-mosaic matrix B is said to be a generalized Toeplitz (q,p,r)-Bezoutian if its generating function admits the representation*

$$\widehat{B}(\lambda,\mu)=\frac{1}{1-\lambda\mu}\widehat{U}(\lambda)\widehat{V}(\mu)^T. \tag{4.14}$$

where $\widehat{U}(\lambda)$ is a $q\times(p+q+r)$ and $\widehat{V}(\mu)$ is a $p\times(p+q+r)$ matrix polynomial.

Then, the following theorem provides a characterization of the pseudoinverse of a Toeplitz-mosaic matrix, which means that actually computing the pseudoinverse can be performed efficiently via several convolutions (in our case $p=q=K$, the number of kernels).

Theorem 4.4.1. *[60] The Moore-Penrose inverse of a $(p,q)-$Toeplitz mosaic matrix is a Toeplitz $(q,p,q+p)-$Bezoutian.*

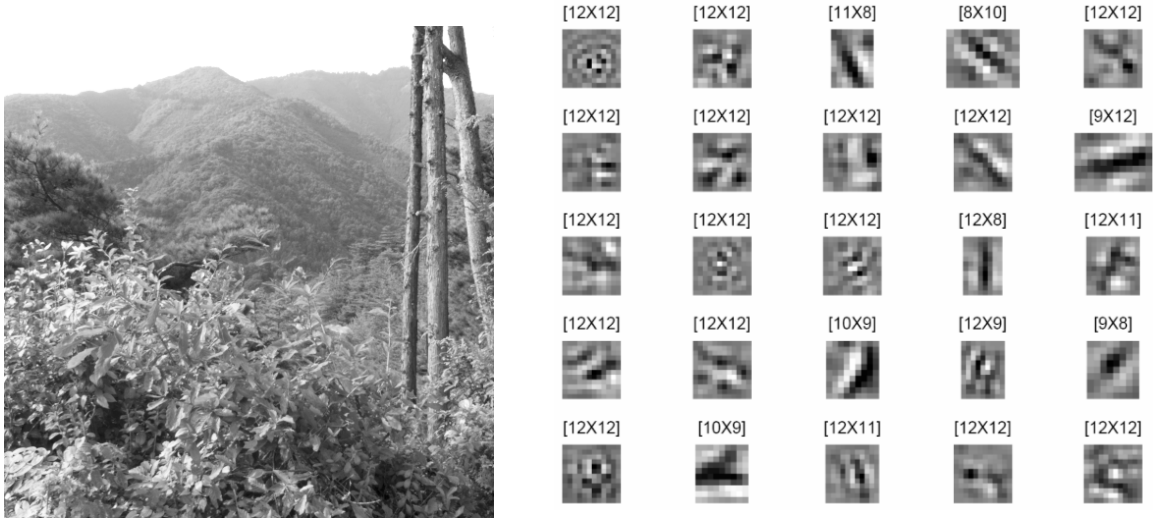
By adapting the result above to the case of mosaic TBT matrices, we obtain an analog characterization of matrix $\mathbf{A}$.

A different, but equally attractive solution to the structured problem above is suggested by the approaches in [113, 114]. Namely, they translate the Toeplitz least squares problem into an interpolation problem, which can then be solved by using a superfast method. Finding generators for the displacement of our particular mosaic TBT matrix is rather straightforward by using the algorithm in the appendix of [69]; this helps us reduce our own problem to a similar interpolation problem, which therefore can be solved efficiently in $\tilde{O}(D\log^2 D)$.

4.5 Experimental Results

In this section, we shall present the results of applying the above method to fairly different categories of images, which illustrates the importance of the signal class on the adapted dictionary.

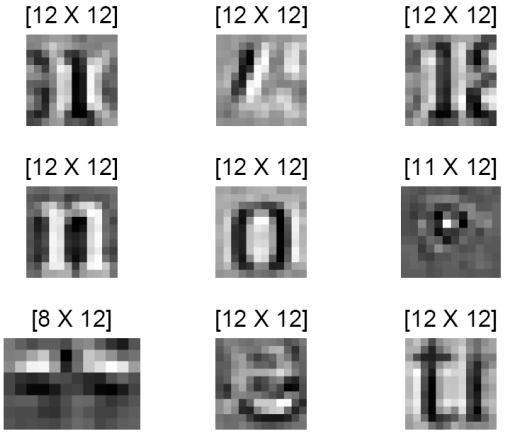
[3]We find it useful to point out that the 1D correspondent is a block-row matrix with Toeplitz blocks (also known as Toeplitz-striped matrix).



(a) Natural image. (b) "Natural image" kernels (300 iter.)

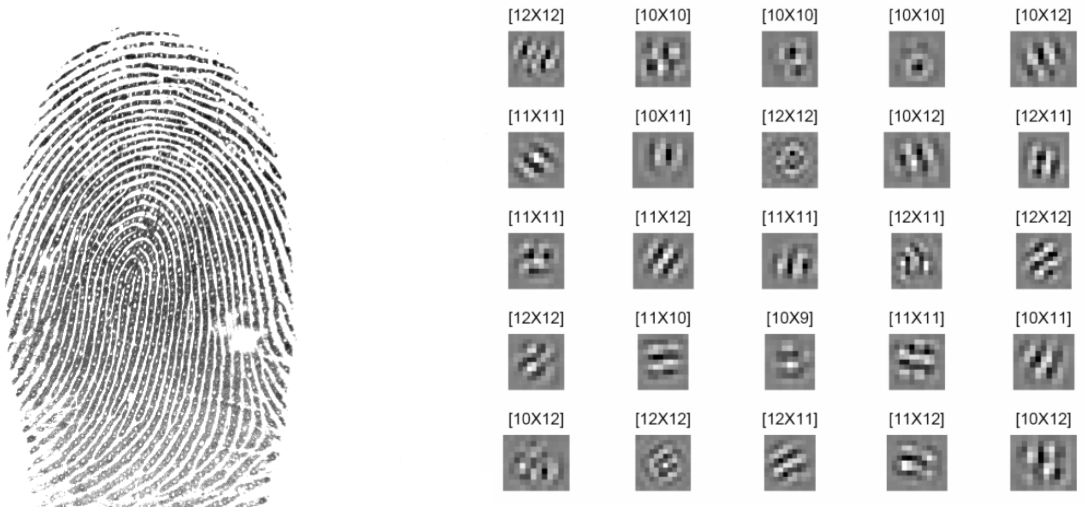
Figure 4.2: Results of applying the Point Coding method to natural images in the Kyoto database. The dictionary was initialized with $K = 25$ random kernels of size 10×10. Kernels are up-scaled for a better visualization; actual pixel size is displayed above each kernel subplot.


ng to representatives of under-
ted minority groups. "It's inex-
ible," said Eric Rodriguez, sen-
olicy analyst with the National
ncil of La Raza, a Hispanic advo-
organization. "I don't under-
d it. We had an undercount of 5
ent of the Latino population in
. If this ruling is upheld we have
eason to expect anything but a 5
ent undercount this time."
ecause of large concentrations of
ks in inner cities and increased
bers of immigrants, mainly in
Southwest — two groups that
e notoriously low response rates
nailed census questionnaires —
1990 census was the first in mod-
history to miss more people than
census that preceded it.
oday's ruling turned on an inter-
tation of two seemingly contra-
ory provisions in the Census Acts
957 and 1976. The 1957 act said
"except for the determination of
ulation for apportionment pur-
tively removed the proh
against it for reapportionmen
poses. Further they argued t
1976 Congress added amendm
the act that directed the Com
Secretary to conduct a census
10 years "in such form and c
as he may determine, includi
use of sampling procedures a
cial surveys."
In today's ruling, written by
Lamberth, the panel disagree
the Justice Department's rea
"Though defendants' inte
tion of the except/shall se
structure is proper in so
stances, the court finds it
strained and incorrect when
to the amended section 19
wrote.
Judge Lamberth stressed
portance of the census to the
of power in Congress betwe
ties and between states, and
was hard to imagine that C
would change the rules with
plicitly saying so. He quote



(a) "Newspaper" image. (b) Newspaper kernels (55 iter.)

Figure 4.3: Results of applying the Point Coding method to newspaper images. The dictionary was initialized with $K = 40$ random kernels (only 9 are shown) of size 8×8. Kernels are up-scaled for a better visualization; actual pixel size is displayed above each kernel subplot.



(a) Fingerprint image. (b) "Fingerprint" kernels (31 iter.)

Figure 4.4: Results of applying the Point Coding method to fingerprint images. The dictionary was initialized with $K = 25$ random kernels of size 10×10. Kernels are up-scaled for a better visualization; actual pixel size is displayed above each kernel subplot.

Natural Images. We first apply the method presented here to images from the Kyoto natural image database [45]. The dictionary started out as a set of 25 random 10×10 patches, and evolved as some of the kernels grew or shrank. The entire set of kernels is displayed in Figure 4.2(b); the learned kernels display the expected aspects for such a dataset (namely edges, ridges, cross patterns) but they also include other shapes (*e.g.*, "round" edges). One interesting aspect is that the kernels in the final dictionary (here, after 300 iterations) do not seem extremely sensitive to the starting point.

Newspaper Images. A different type of signal, with significantly distinct statistical structure, is the class of scanned newspaper images. It is a main characteristic of this class that fewer predominant orientations are present. Figure 4.3(b) exhibits kernels adapted to this class, after 55 iterations and starting from random. As can be observed easily, kernels tend to capture mainly printed symbols; if a large enough set of kernels is used (*e.g.*, 40), they tend to stabilize to individual characters, or pairs thereof.

Fingerprint Images. Finally, we apply our method to another highly distinct class, namely fingerprint images. Figure 4.4(b) displays a set of kernels learned from images in the Cross Match Verifier 300 sample fingerprint database [1]. Although initialized randomly, after only 31 iterations the 25 kernels already localize in frequency and in orientation, although not also in space (easily explained by the structure of the signals).

The results presented in this section have been learned from a training set ranging from 10 images (newspaper) to 50 images (natural). The dictionary size was hand-picked to the reported values in order to avoid redundancy (*e.g.*, several kernels being copies or shifted versions of each other). We currently are working on designing an automatic mechanism to control the number

of “sufficient” kernels.

4.6 Conclusion

We proposed an approach to deriving adaptive shift-invariant image representation. Our method is computationally very efficient and eliminates kernel size constraints, which can lead to a general adaptive multiscale dictionary. In the kernel update step, we focused on what we believe to be an under-explored family of algorithms, that exploit the structure of the least-squares optimization problem.

Chapter 5

Robust Coding

5.1 Introduction

Reliable communication over noisy channels is the fundamental problem of information theory. The abundance of practical modern applications, such as the communication, compression, or storage of data, has produced many variations on the theme. In this chapter, we focus on the problem of finding linear representations that optimally preserve information in the transmitted signals when the representation has limited precision. Introduced by E. Doi and coll. in [43] and further analyzed in [44], the so-called Robust Coding scheme makes use of arbitrarily many coding units to minimize reconstruction error, by explicitly introducing redundancy in the code to compensate for channel noise.

The above problem was pointed out to be of particular relevance to a new and exciting area: mathematical modeling of neural representations. This is not at all surprising; cells can be regarded as communication channels for traveling neural spikes and their coding precision is limited by intrinsic biological constraints (see for instance [8, 16, 43, 46]). By identifying the short time activity of a neuron with a real value, the limited information capacity of the encoding unit can be modeled effectively by additive Gaussian noise. This abstraction has been observed to be better suited for neural modeling than are noise-free representations, employed by existing standard linear models like PCA or ICA; nevertheless, in this chapter we intend to focus on theoretical optimality, rather than on biological plausibility issues.

Many problems tightly related to Robust Coding have been studied in the literature. In spite of the high conceptual similarity of these problems, optimal solutions depend on the various objectives and, just as importantly, on the particular assumptions or constraints involved. These factors, as it turns out, determine both the structural properties of the solutions and the computational cost of obtaining them. One instance of such problems appears in the context of communicating real-valued signals over parallel independent Gaussian channels ([33, ch. 9]). There, the objective is optimal power allocation for maximizing mutual information among the altered components of the signal, subject to an average power constraint. The solution essentially depends on the covariance of the original signal, by the so called "water-filling" procedure. This problem addresses only one aspect of the coding (power allocation), which restricts the so-

lutions to the set of diagonal matrices with nonnegative entries and bounded trace. Instead, it would be more useful to investigate more general linear transforms, notably those implemented by non-square dense matrices.

The idea of improving robustness to additive noise via redundant linear transformations has been addressed in the context of frames by Daubechies [34], where the optimality criterion was mean squared error (MSE). Several classes of frames have been identified as optimal solutions of other problems involving the design of linear representations that are resilient to various types of coefficient alterations (quantization noise [58], erasures [57, 97]), or that have optimal numerical stability of reconstruction [34]. (For a recent and thorough review of frame properties and applications, we recommend Kovačević and Chebira [72, 73]). In general, the frame design problem can become rather complex and often times algebraic methods, usually employed for structural characterization, must be complemented by numerical methods, as argued by Dhillon and coll. [40, 41]. Several insightful techniques have been developed in search for accurate characterization and interpretation of optimal frame representations which reduce the numerical optimization burden considerably (see Casazza and coll. [26]).

To formally define our problem, let us consider our signal to be made of samples from an N-dimensional zero-mean data distribution, with known full-rank covariance matrix $\mathbf{\Sigma_x}$. For a given number M of communication channels[1], we shall search for *analysis matrix* $\mathbf{W} \in \mathbb{R}^{M\times N}$ and *synthesis matrix* $\mathbf{A} \in \mathbb{R}^{N\times M}$ that maximally reduce the effect of additive Gaussian noise, independent of the signal and having the same power σ_δ^2 on each channel. More precisely, our objective is to minimize the reconstruction MSE[2]:

$$\mathbf{tr}\left\{(\mathbf{I}_N - \mathbf{AW})\mathbf{\Sigma_x}(\mathbf{I}_N - \mathbf{AW})^T\right\} + \sigma_\delta^2 \mathbf{tr}\left\{\mathbf{AA}^T\right\} \tag{5.1}$$

The signal power on each of the M channels also may be assumed to be identical (call it σ_u^2), via a simple rescaling of $\mathbf{W}$'s rows. This implies the existence of a common signal-to-noise ratio (SNR) parameter $\gamma^2 = \sigma_u^2/\sigma_\delta^2 > 0$ to characterize the precision, and consequently the information capacity of each channel[3]. In the following we will try to simplify the optimization problem, by reducing it to a more convenient form, a first step of which is to eliminate the correlation in the signal. By the eigendecomposition of the covariance matrix $\mathbf{\Sigma_x}$, we obtain a diagonal spectrum matrix $\mathbf{S}^2$ with positive diagonal entries, and an orthogonal matrix $\mathbf{E} \in \mathbb{R}^{N\times N}$, such that $\mathbf{\Sigma_x} = \mathbf{ES}^2\mathbf{E}^T$. By the changes of variable $\mathbf{T} = \mathbf{WES}/\sigma_u$, and $\mathbf{B} = \mathbf{E}^T\mathbf{A}$ we can formulate our optimization problem as:

$$\left\{\begin{array}{ll} \min\limits_{\mathbf{B},\mathbf{T}} & ||\mathbf{BT} - \mathbf{S}||_F^2 + \frac{1}{\gamma^2}||\mathbf{B}||_F^2 \\ s.t. & \mathrm{diag}(\mathbf{TT}^T) = \mathbf{1}_{M,1} \end{array}\right. \tag{RC$_{\text{simple}}$}$$

[1]No assumption about the relation between M and N is explicitly made. We shall refer the case $M < N$ as *undercomplete*, and the opposite one as *overcomplete*.

[2]Here we take the average over the data, as well as over the noise.

[3]In Cover and Thomas [33], information capacity is defined as $\frac{1}{2}\log(\gamma^2 + 1)$, which is a function of the channel SNR γ^2.

The new variables $\mathbf{B}$ and $\mathbf{T}$ also represent synthesis and analysis matrices, while the capacity constraint simply translates in the rows of $\mathbf{T}$ being unit-length vectors in $\mathbb{R}^N$.

When $N = 1$ and $N = 2$, a complete characterization of the optimal encoding/decoding pair for the Robust Coding problem, based on exhaustive case analysis, was presented in [43, 44]. For general dimension N, numerical solutions were obtained via gradient-based optimization with Lagrange multipliers. Although such solutions are perfectly legitimate for many optimization problems, we could not provide optimality guarantees as precise as in the one- and two-dimensional cases. A conjecture on the error function lower bound was given in [44], yet even though the result was strongly supported by numerical results, a proof remained out of reach.

The present chapter addresses this issue by extending the previous analysis to the most general case (any N and M). Namely, we characterize the algebraic structure of the solutions via singular value decomposition (SVD; see [56, 110]). We also prove an exact formula for the MSE lower bound, as a function on the covariance spectrum, channel SNR γ^2, and on the dimensions N and M. This exact, therefore tight, bound not only enables us to easily verify optimality of a given encoding/decoding pair, but also gives insight into how to manipulate the parameters (*e.g.*, increase the number of units M) to minimize reconstruction error. Last but not least, structural characterization leads to fast and direct algorithms for obtaining the optimal solutions, which confers an immense advantage over our previous approach. Thus, the computation becomes independent on the notorious step-size sensitivity of gradient methods, both from numerical stability and from computational complexity points of view.

The chapter is organized as follows. In subsection 5.2, we identify necessary conditions on the optimal $\mathbf{T}$, via constraints on the components of its singular value decomposition (SVD), and compute the optimal singular values. Also, the structure of the right singular matrix $\mathbf{V}$ is described, and a generic procedure for computing the left singular matrix $\mathbf{U}$ is analyzed. Also, we derive the closed-form expression of the lower bound, interpret its dependence on the problem parameters, and illustrate with several interesting cases. Section 5.4 concludes the chapter by summarizing the results, by comparing our study with related approaches, and by revealing several directions we plan to explore in the future.

5.2 Robust Coding Solutions

In section 5.1 we provided an extended motivation for the Robust Coding framework. Namely, it models the problem of reliable communication on parallel Gaussian channels having identical signal-to-noise ratios. Here, we address the properties of optimal solutions and propose efficient ways of computing them.

First, we shall introduce several notions that we intend to use throughout the rest of this chapter. We will show how to effectively reduce the parameter search space to the set of unit-row, M-by-N matrices $\mathbf{T}$. Then, we identify necessary conditions for the optimality of such matrices via constraints on its singular vectors, and compute the singular values exactly. This will enable us to exactly describe the structure of the right singular matrix $\mathbf{V}$, and identify several algorithms for computing the left singular matrix $\mathbf{U}$. Next, we shall derive the closed-form expression of

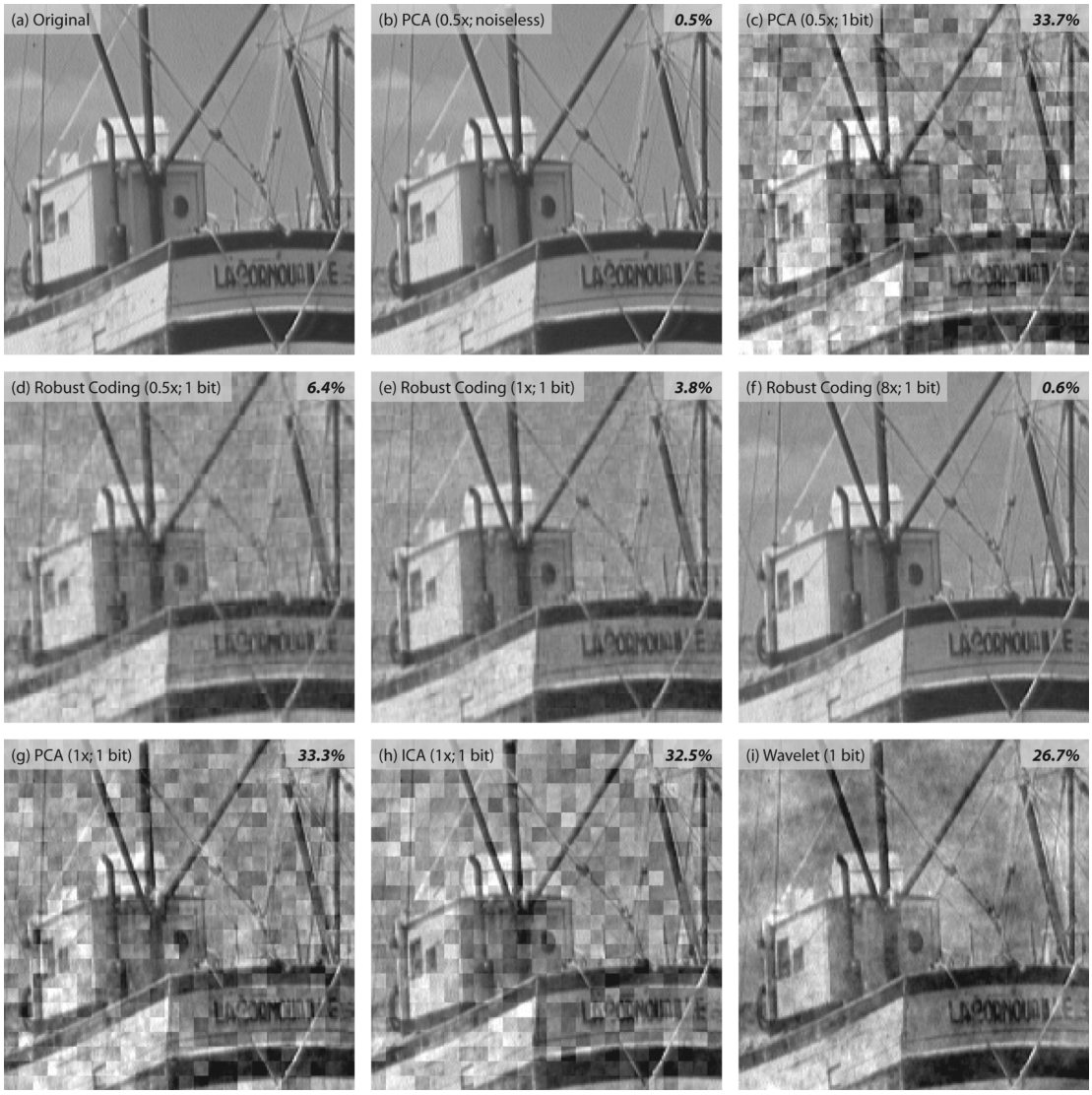


Figure 5.1: Image coding under the presence of channel noise. For each reconstruction its percent error is indicated. (a) Original image. (b) PCA (M=32) with noiseless representation. (c) PCA (M=32) with 1-bit precision code. (d) Robust coding (M=32) with 1-bit precision code. (e) Robust coding (M=64) with 1-bit precision code. (f) Robust coding (M=512) with 1-bit precision code. (g) PCA (M=64) with 1-bit precision code. (h) ICA (M=64) with 1-bit precision code. (i) Daubechies 9/7 wavelet with 1-bit precision code.

the lower bound, interpret its dependence on the problem parameters, and illustrate with several interesting cases.

Definitions and notations

Let us define several notions we shall use throughout the chapter.

Definition 2. *For any $p \in \mathbb{N}^*$, let us denote $\mathcal{O}_p(\mathbb{R})$* the orthogonal group of index p, i.e.*, the space of all $p \times p$ orthogonal matrices.*

Definition 3. *For any $p \in \mathbb{N}^*$ and $r > 0$, let $\mathcal{B}_p(0;r) = \{x | x \in \mathbb{R}^p, \|x\|_2 \leq r\}$ the* p-dimensional zero-centered unit ball *and $\mathcal{S}_p(0;r) = \partial\mathcal{B}_p(0;r)$ its* surface.

Definition 4. *For any two matrices $\mathbf{M}_1$ and $\mathbf{M}_2$, we define the* direct sum *of $\mathbf{M}_1$ and $\mathbf{M}_2$ as the block-diagonal matrix*

$$\mathbf{M}_1 \oplus \mathbf{M}_2 = \begin{pmatrix} \mathbf{M}_1 & \\ & \mathbf{M}_2 \end{pmatrix}. \tag{5.2}$$

Definition 5. *For any two matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ of the same size, we will denote by $\mathbf{M}_1 \circ \mathbf{M}_2$ their* Hadamard*, or* entry-wise product*.*

Definition 6. *Let n be a positive integer. For any permutation τ of $\{1, 2, ..., n\}$, its* corresponding permutation matrix *is defined as*

$$\mathbf{P}(\tau) \equiv \left(\delta_{\tau(i),j}\right)_{i,j} \tag{5.3}$$

where δ is Kronecker's symbol: $\delta_{\alpha,\beta} = 1$ if $\alpha = \beta$ and 0 otherwise.

Remark. *Permutation matrices are both orthogonal and doubly-stochastic [62].*

We will now show how to further simplify the Robust Coding optimization problem, as well as to reduce the parameter search space. For the sake of completeness, let us review the assumptions and conditions imposed on the parameters appearing in the simplified form ($\mathrm{RC}_{\mathrm{simple}}$).

Let $M, N \in \mathbb{N}^*$, $\gamma > 0$, and $\mathbf{S} = \mathrm{diag}(s_1, ..., s_N)$, where $s_1 \geq \ldots \geq s_N > 0$ without loss of generality. As explained in section 5.1, we want to find $\mathbf{B} \in \mathbb{R}^{N\times M}$, $\mathbf{T} \in \mathbb{R}^{M\times N}$ that minimize the cost function

$$\mathcal{E}(\mathbf{B}, \mathbf{T}) = ||\mathbf{BT} - \mathbf{S}||_F^2 + \frac{1}{\gamma^2}||\mathbf{B}||_F^2, \tag{5.4}$$

subject to $\mathrm{diag}(\mathbf{TT}^T) = \mathbf{1}_{M,1}$. We can regard $\mathcal{E}$ as a function defined on $\mathbb{R}^{MN} \times \mathcal{S}_N(0;1)^M$. This is possible, because the constraint $\mathrm{diag}(\mathbf{TT}^T) = \mathbf{1}_{M,1}$ means that the rows of $\mathbf{T}$ are unit-length, N-dimensional vectors and therefore we can identify the set of feasible matrices $\mathbf{T}$, with $\mathcal{S}_N(0;1)^M \equiv \mathcal{S}_N(0;1) \times \ldots \times \mathcal{S}_N(0;1)$. We observe that for any fixed $\mathbf{T} \in \mathcal{S}_N(0;1)^M$, function $\mathcal{E}_\mathbf{T}(\cdot) \equiv \mathcal{E}(\cdot, \mathbf{T})$ is a quadratic (therefore continuously differentiable) function of the entries of $\mathbf{B}$, and so a necessary condition on $\mathbf{B}$ to minimize $\mathcal{E}_\mathbf{T}$ can be expressed in matrix form as:

$$\mathbf{0}_{M,N} = \frac{\partial}{\partial\mathbf{B}}\mathcal{E}_\mathbf{T}(\mathbf{B}) = \frac{\partial}{\partial\mathbf{B}}\mathcal{E}(\mathbf{B}, \mathbf{T}) = 2\left((\mathbf{BT} - \mathbf{S})\,\mathbf{T}^T + \frac{1}{\gamma^2}\mathbf{B}\right) \tag{5.5}$$

which further implies that

$$\mathbf{B}\left(\frac{1}{\gamma^2}\mathbf{I}_M + \mathbf{TT}^T\right) = \mathbf{ST}^T. \tag{5.6}$$

Remark. *Matrices $\frac{1}{\gamma^2}\mathbf{I}_M + \mathbf{TT}^T$ and $\frac{1}{\gamma^2}\mathbf{I}_N + \mathbf{T}^T\mathbf{T}$ are positive definite matrices for any $\gamma > 0$ and $\mathbf{T} \in \mathbb{R}^{M\times N}$. As such, they are invertible and $\det(\frac{1}{\gamma^2}\mathbf{I}_M + \mathbf{TT}^T) > 0$, $\det(\frac{1}{\gamma^2}\mathbf{I}_N + \mathbf{T}^T\mathbf{T}) > 0$. Consequently, function $\mathcal{E}_\mathbf{T}$ has a unique point of extremum, namely*

$$\mathbf{B_T} = \mathbf{S}\mathbf{T}^T \left(\frac{1}{\gamma^2}\mathbf{I}_M + \mathbf{T}\mathbf{T}^T \right)^{-1}. \tag{5.7}$$

Since $\mathcal{E}_{\mathbf{T}} : \mathbb{R}^{MN} \to \mathbb{R}$ is bounded below by 0, and obviously unbounded above, this extremum can only be a minimum. Therefore eq. (5.7) *represents a necessary and sufficient condition for* $\mathbf{B}$ *to minimize* $\mathcal{E}_{\mathbf{T}}$.

From the observation above, we can infer (see details in Appendix 5.5):

$$\mathcal{E}(\mathbf{B}, \mathbf{T}) \geq \mathcal{E}(\mathbf{B_T}, \mathbf{T}) = \mathbf{tr}\left(\mathbf{S}^2(\mathbf{I}_N - \mathbf{T}^T(\frac{1}{\gamma^2}\mathbf{I}_M + \mathbf{T}\mathbf{T}^T)^{-1}\mathbf{T}) \right) \tag{5.8}$$

with equality if and only if $\mathbf{B} = \mathbf{B_T}$. This allows us to conclude that $(\mathbf{B}_{min}, \mathbf{T}_{min})$ is a minimizing pair for $\mathcal{E}$ if and only if $\mathbf{B}_{min} = \mathbf{B}_{\mathbf{T}_{min}}$ and $\mathbf{T}_{min}$ is minimizing the cost function $\mathcal{F} : \mathcal{B}_N(0;1)^M \to \mathbb{R}$,

$$\mathcal{F}(\mathbf{T}) = \mathbf{tr}\left(\mathbf{S}^2(\mathbf{I}_N - \mathbf{T}^T(\frac{1}{\gamma^2}\mathbf{I}_M + \mathbf{T}\mathbf{T}^T)^{-1}\mathbf{T}) \right) = \mathbf{tr}\left(\mathbf{S}^2(\mathbf{I}_N + \gamma^2\mathbf{T}^T\mathbf{T})^{-1} \right). \tag{5.9}$$

where the last equality follows by applying Sherman-Morrison-Woodbury formula ([56, p.50]; see Appendix 5.5). Since $\mathcal{F}$ is a continuous function defined on a compact set, Weierstrass's theorem implies that it reaches its minimum on its domain, which means that there exists at least one such optimal matrix $\mathbf{T}_{min}$ and moreover, $\min_{\mathbf{T}} \mathcal{F} = \min_{\mathbf{B},\mathbf{T}} \mathcal{E}$.

Let us notice that matrix $\mathbf{T}$ shall serve now as a parametric descriptor of our optimal analysis/synthesis Robust Coding pair. We have succeeded in effectively reducing the search space to the set of feasible matrices $\mathbf{T}$. Next, we need to describe the structure of function $\mathcal{F}$'s minima.

The SVD structure of $\mathbf{T}_{min}$

In this section, we shall identify necessary conditions on the minimizers of $\mathcal{F}$, after which we shall point out which of these are also sufficient. For this purpose, we employ the singular value decomposition (SVD) of $\mathbf{T}$. Thus, if we denote $K \equiv \min(M, N)$, there exists a decomposition:

$$\mathbf{T} = \mathbf{U} \cdot \mathbf{\Sigma} \cdot \mathbf{V}^T \tag{5.10}$$

where $\mathbf{U} \in \mathcal{O}_M(\mathbb{R})$, $\mathbf{V} \in \mathcal{O}_N(\mathbb{R})$, and $\mathbf{\Sigma} \in \mathbb{R}^{M \times N}$ a diagonal matrix whose diagonal entries are, without loss of generality, sorted in decreasing order: $\sigma_1 \geq \ldots \geq \sigma_K \geq 0$. Let us denote by $\mathbf{\Sigma}_K = \mathrm{diag}(\sigma_1, \ldots, \sigma_K)$, the "reduced", square version of $\mathbf{\Sigma}$.

Remark. *Matrices $\mathbf{T}^T\mathbf{T} \in \mathbb{R}^{N \times N}$ and $\mathbf{T}\mathbf{T}^T \in \mathbb{R}^{M \times M}$ are symmetric, and their SVD (the same as their eigenvalue decomposition) is*

$$\mathbf{T}^T\mathbf{T} = \mathbf{V}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{V}^T = \mathbf{V}\left(\mathbf{\Sigma}_K^2 \oplus \mathbf{0}_{N-K}\right)\mathbf{V}^T, \tag{5.11}$$

respectively

$$\mathbf{T}\mathbf{T}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^T\mathbf{U}^T = \mathbf{U}\left(\mathbf{\Sigma}_K^2 \oplus \mathbf{0}_{M-K}\right)\mathbf{U}^T. \tag{5.12}$$

Remark. *Since* $\mathbf{T} \in \mathcal{B}_N(0;1)^M$*, a necessary condition on its singular values is:*

$$M = \mathbf{tr}\left(\mathbf{T}\mathbf{T}^T\right) = \mathbf{tr}\left(\mathbf{\Sigma}_K^2 \oplus \mathbf{0}_{M-K}\right) = \mathbf{tr}\,\mathbf{\Sigma}_K^2 = \sum_{i=1}^{K} \sigma_i^2 \tag{5.13}$$

Let us examine now the cost function. By algebraic manipulations, from eq. (5.9) we obtain (see Appendix 5.5):

$$\mathcal{F}(\mathbf{T}) = \mathbf{tr}\left(\mathbf{S}^2\mathbf{V}\left(\mathrm{diag}(\frac{1}{1+\gamma^2\sigma_1^2},\ldots,\frac{1}{1+\gamma^2\sigma_K^2}) \oplus \mathbf{I}_{N-K}\right)\mathbf{V}^T\right). \tag{5.14}$$

As function $\mathcal{F}$ above depends only on $\mathbf{\Sigma}_K$ and $\mathbf{V}$, we choose to refer it (by a slight abuse of notation) as $\mathcal{F}(\mathbf{\Sigma}_K, \mathbf{V})$. Our goal now became to characterize the minimizing pairs $(\mathbf{\Sigma}_{K,min}, \mathbf{V}_{min})$ of this function. First of all, let us observe that for any $\mathbf{\Sigma}_K$, there exists[4]:

$$\mathcal{G}(\mathbf{\Sigma}_K) \equiv \min_{\mathbf{V}\in\mathcal{O}_N(\mathbb{R})} \mathcal{F}(\mathbf{\Sigma}_K, \mathbf{V}) \tag{5.15}$$

We can show that function $\mathcal{G}$ is invariant to the ordering of the entries on the diagonal of its argument (for proof, see Appendix).

Lemma 1. *For any permutation matrix* $\mathbf{P} \in \mathbb{R}^{K\times K}$*, we have* $\mathcal{G}(\mathbf{\Sigma}_K) = \mathcal{G}(\mathbf{P}\mathbf{\Sigma}_K\mathbf{P}^T)$.

This lemma guarantees that condition $\sigma_1 \geq \ldots \geq \sigma_K$ imposed on the diagonal elements of $\mathbf{\Sigma}_K$ does not restrict the generality in any way.

Remark. *From the conditions imposed so far, we have*

$$0 < \frac{1}{1+\gamma^2\sigma_1^2} \leq \frac{1}{1+\gamma^2\sigma_2^2} \leq \ldots \leq \frac{1}{1+\gamma^2\sigma_K^2} \leq 1 \tag{5.16}$$

$$s_1^2 \geq s_2^2 \geq \ldots \geq s_N^2 > 0. \tag{5.17}$$

This implies that the two diagonal matrices appearing in (5.14) have their (positive) diagonal elements ordered differently. The following general Lemma will help us find the exact expression of $\mathcal{G}$, and moreover characterize all the orthogonal matrices $\mathbf{V}$ for which $\mathcal{G}(\mathbf{\Sigma}_K) = \mathcal{F}(\mathbf{\Sigma}_K, \mathbf{V})$.

Lemma 2. *Let* $n \in \mathbb{N}^*$ *a positive integer, and* $\mathbf{D}_A = \mathrm{diag}\,(a_1, a_2, \ldots, a_n)$, $\mathbf{D}_B = \mathrm{diag}\,(b_1, b_2, \ldots, b_n)$, *two diagonal matrices such that* $a_1 \geq \ldots \geq a_n > 0$ *and* $b_n \geq \ldots \geq b_1 > 0$.

a) Then, $\min_{\mathbf{V}\in\mathcal{O}_n(\mathbb{R})} \mathbf{tr}\left(\mathbf{D}_A\mathbf{V}\mathbf{D}_B\mathbf{V}^T\right) = \mathbf{tr}\left(\mathbf{D}_A\mathbf{D}_B\right) = \sum_{i=1}^{n} a_i b_i$.

b) Let $\tau_1, ..., \tau_m$ *be all the permutations of* $\{1, ..., n\}$*, for which* $\sum_{i=1}^{n} a_i b_{\tau_k(i)} = \sum_{i=1}^{n} a_i b_i$*,* $1 \leq k \leq m$*. Then* $\forall\, \mathbf{V} \in \mathcal{O}_n(\mathbb{R})$*,* $\mathbf{V}$ *is a minimizer of* $\mathbf{tr}\left(\mathbf{D}_A\mathbf{V}\mathbf{D}_B\mathbf{V}^T\right)$ *if and only if the entry-wise product* $\mathbf{V} \circ \mathbf{V}$ *is a convex combination of the permutation matrices* $\mathbf{P}(\tau_k)$*,* $1 \leq k \leq m$.

Proof. *See Appendix.*

[4]Function $\mathcal{F}(\mathbf{\Sigma}_K, \cdot) : \mathcal{O}_N \to \mathbb{R}$ is continuous, and defined on a compact set.

The above lemma guarantees that the minimum of $\mathbf{tr}\left(\mathbf{D}_A\mathbf{V}\mathbf{D}_B\mathbf{V}^T\right)$ over the set of orthogonal matrices, is reached for the identity matrix. This minimizer may not be unique, as this property should depend on the diagonal values as well, and not only on their relative order. If we substitute $\mathbf{D}_A$ by $\mathbf{S}^2$, and $\mathbf{D}_B$ by $(\mathbf{I}_K + \gamma^2\mathbf{\Sigma}_K^2)^{-1} \oplus \mathbf{I}_{N-K}$, we obtain:

$$\begin{aligned}
\mathcal{G}(\mathbf{\Sigma}_K) &= \min_{\mathbf{V}\in\mathcal{O}_N(\mathbb{R})} \mathcal{F}(\mathbf{\Sigma}_K, \mathbf{V}) = \mathcal{F}(\mathbf{\Sigma}_K, \mathbf{I}_N) && (5.18)\\
&= \mathbf{tr}\left(\mathrm{diag}(s_1^2, \ldots, s_K^2) \cdot (\mathbf{I}_K + \gamma^2\mathbf{\Sigma}_K^2)^{-1}\right) + \sum_{i=K+1}^{N} s_i^2 && (5.19)\\
&= \sum_{i=1}^{K} \frac{s_i^2}{1+\gamma^2\sigma_i^2} + \sum_{i=K+1}^{N} s_i^2 && (5.20)
\end{aligned}$$

Since function $\mathcal{G}$ is continuous, we can guarantee that there exists a minimizer for $\mathcal{G}$ on the compact set of all diagonal matrices of trace M, with nonnegative diagonal entries. Consequently, $(\mathbf{\Sigma}_{K,min}, \mathbf{V}_{min})$ is a minimizing pair for $\mathcal{F}$ if and only if $\mathbf{\Sigma}_{K,min}$ is minimizing $\mathcal{G}$, and moreover, $\min_{\mathbf{\Sigma}_K,\mathbf{V}} \mathcal{F} = \min_{\mathbf{\Sigma}_K} \mathcal{G}$.

From eq. (5.20), we observe that an immediate lower bound for $\mathcal{G}(\mathbf{\Sigma}_K)$ is the sum of squares of the smallest $N-K$ diagonal entries of $\mathbf{S}$. Moreover, optimizing $\mathcal{G}$ is equivalent to solving the following optimization problem:

$$\left\{\begin{array}{ll} \min & \mathcal{G}_K(\sigma_1, \ldots, \sigma_K) \\ s.t. & \sum_{i=1}^{K} \sigma_i^2 = M \\ & \sigma_1 \geq \ldots \geq \sigma_K \geq 0 \end{array}\right. \tag{P_1}$$

where we denoted

$$\mathcal{G}_K(\sigma_1, \ldots, \sigma_K) = \sum_{i=1}^{K} \frac{s_i^2}{1+\gamma^2\sigma_i^2} \tag{5.21}$$

Fortunately, this problem has a closed-form solution.

Theorem 5.2.1. *There exists $1 \leq R \leq K$, such that for any index $j \leq K$,*

$$s_j > \frac{\sum_{i=1}^{j} s_i}{j + \gamma^2 M} \Leftrightarrow j \leq R. \tag{5.22}$$

Then, problem (P_1) *has the unique solution*

$$\sigma_i = \left\{\begin{array}{ll} \dfrac{1}{\gamma}\sqrt{\dfrac{s_i}{\sum_{j=1}^{R} s_j}(R+\gamma^2 M) - 1}, & \text{if } 1 \leq i \leq R \\ 0, & \text{if } R+1 \leq i \leq K. \end{array}\right. \tag{5.23}$$

Proof. *See Appendix 5.5*

The above theorem provides the general form of the singular values of $\mathbf{T}_{min}$. At a closer examination of the result, we observe that the optimal singular values depend not only on the dimensions and the SNR level, but also on the concentration of the eigenvalues of the covariance matrix. The threshold R, the index of the smallest nonzero singular value, and consequently the rank of the optimal encoding matrix $\mathbf{T}$ (or $\mathbf{W}$, for that matter), directly reflects a degree of spectral concentration. The necessity of identifying this degree would easily become apparent in a naive attempt to minimize the sum $\sum_{i=1}^{K} s_i^2/(1+\gamma^2\sigma_i^2)$ s.t. $\sum_{i=1}^{K} \sigma_i^2 = M$, via the Cauchy-Schwartz inequality (see Appendix 5.5). On the other hand, R is an indicator on "how much" and "which part of the data space we can fill" with the available number of noisy coding units. It therefore determines an interesting power allocation scheme, somewhat related to the well-known "waterfilling" scheme. Similarly, the subspaces, or directions, lacking significant energy are sacrificed and the resources are devoted to the more important ones.

Finding $\mathbf{V}$, the right singular vector matrix of $\mathbf{T}_{min}$

In the proof of Lemma 2 we used Birkhoff's theorem [62, p. 527] stating that for any $\mathbf{V} \in \mathcal{O}_N$, matrix $\mathbf{V} \circ \mathbf{V}$ is doubly stochastic, and therefore it must be a convex combination of permutation matrices. Furthermore, matrix $\mathbf{V}$ is a minimizer for $\mathcal{F}(\Sigma_{K,min}, \cdot)$ if and only if all the permutation matrices just mentioned are also minimizers. In this section, we will address the structural characterization of the matrices minimizing $\mathcal{F}(\Sigma_{K,min}, \cdot)$, which are actually all the possible right singular matrices of $\mathbf{T}_{min}$. As in the previous section, we shall state a slightly more general result (we omit the proof).

Lemma 3. *Let $n \in \mathbb{N}^*$ a positive integer, and $\mathbf{D}_A = \operatorname{diag}(a_1, a_2, \ldots, a_n)$, $\mathbf{D}_B = \operatorname{diag}(b_1, b_2, \ldots, b_n)$, two diagonal matrices such that $a_1 \geq \ldots \geq a_n > 0$ and $b_n \geq \ldots \geq b_1 > 0$. Let us denote by $\bigcup \mathcal{I}_k^a$ the partition of $\{1, ..., n\}$ determined by the values $a_1, a_2, \ldots, a_n$; namely, $\forall\, i \in \mathcal{I}_{k_1}^a,\ j \in \mathcal{I}_{k_2}^a,\ i \leq j$ we have:*

$$(a_i = a_j \Leftrightarrow k_1 = k_2) \qquad \text{and} \qquad (a_i > a_j \Leftrightarrow k_1 < k_2)\,. \tag{5.24}$$

Let $\bigcup \mathcal{I}_l^b$ be the partition of $\{1, ..., n\}$, determined by $b_1, b_2, \ldots, b_n$. (In the definition above, we only need to substitute a's with b's, and $a_i > a_j$ with $b_i < b_j$.) Then all orthogonal matrices $\mathbf{V} = (v_{ij})_{1 \leq i,j \leq n}$ minimizing the function $\mathbf{tr}(\mathbf{D}_A \mathbf{V} \mathbf{D}_B \mathbf{V}^T)$ share the same support structure determined by these partitions:

$$v_{ij} \neq 0 \;\Rightarrow\; i \in \mathcal{I}_k^a,\ \text{and}\ j \in \mathcal{I}_l^b,\ \text{and}\ \mathcal{I}_k^a \bigcap \mathcal{I}_l^b \neq \emptyset. \tag{5.25}$$

Consider sequences $\mathbf{a} = (a_i)_{1 \leq i \leq N}$ and $\mathbf{b} = (b_i)_{1 \leq i \leq N}$, where

$$a_i = s_i^2,\ \forall\, 1 \leq i \leq N, \tag{5.26}$$

$$b_i = \begin{cases} \dfrac{1}{1+\gamma^2\sigma_i^2}, & \text{if } 1 \leq i \leq K \\ 1, & \text{if } K+1 \leq i \leq N. \end{cases} \tag{5.27}$$

As we observed, there is a correspondence between contiguous intervals of optimal singular values σ_i and intervals of variances s_i. This correspondence will directly influence the structure of the $\mathbf{V}$ matrix. We shall start discussing this issue by means of observations on the "optimal permutations" $\mathbf{P}(\tau)$ involved in the convex expansion of $\mathbf{V} \circ \mathbf{V}$. Namely, we notice that such a permutation τ_k is optimal if and only if $\sum_{i=1}^{N} a_i b_{\tau(i)} = \sum_{i=1}^{N} a_i b_i$, which is equivalent to saying that the ordering of the values of $\mathbf{b}$ remains unchanged via permutation τ.

In the following, let us consider the partition $(\mathcal{J}_j)_{1 \leq j \leq \beta+1}$ of $\{1, 2, \ldots, N\}$, where $\mathcal{J}_{\beta+1} = \{K+1, \ldots, N\}$, and the rest of the $\mathcal{J}_j$'s defined as in the previous section. We will denote $\widehat{b}_j$ the (common) value of b_i for $i \in \mathcal{J}_j$. First, let us notice that for $1 \leq j < \beta$, we have $\tau(\mathcal{J}_j) = \mathcal{J}_j$, that is, optimal permutation τ takes interval $\mathcal{J}_j$ onto itself. An immediate consequence of this observation is that all the corresponding permutation matrices $\mathbf{P}(\tau)$ have a block-diagonal structure, of whose first $\beta - 1$ diagonal blocks are of size $|\mathcal{J}_j| \times |\mathcal{J}_j|$. In turn, this will imply that matrix $\mathbf{V} \circ \mathbf{V}$ has this property, and therefore matrix $\mathcal{C}$ itself will have the same structure. Namely, we showed that the first $\beta - 1$ blocks of $\mathcal{C}$ are matrices from $\mathcal{O}_{|\mathcal{J}_j|}$, respectively. What about the rest of $\mathcal{C}$?

If $\widehat{b}_\beta = 1$, since $\widehat{b}_{\beta+1} = 1$ it follows that however we permute the last $|\mathcal{J}_\beta| + |\mathcal{J}_{\beta+1}|$ elements of $\mathbf{b}$, the result will be optimal. To conclude, let us prove that any orthogonal, block-diagonal matrix $\mathcal{C}$ is optimal.

$$\begin{aligned} \mathbf{tr}\left(\mathbf{D}_A \mathbf{V} \mathbf{D}_B \mathbf{V}^T\right) &= \mathbf{tr}\left(\mathbf{D}_{A,1} \oplus \ldots \oplus \mathbf{D}_{A,\beta}\right) \cdot \left(\mathbf{V}_1 \oplus \ldots \oplus \mathbf{V}_\beta\right) & (5.28) \\ &\quad \left(\mathbf{D}_{B,1} \oplus \ldots \oplus \mathbf{D}_{B,\beta}\right) \cdot \left(\mathbf{V}_1^T \oplus \ldots \oplus \mathbf{V}_\beta^T\right)\big) & (5.29) \\ &= \mathbf{tr}\left(\mathbf{D}_{A,1} \mathbf{V}_1 \mathbf{D}_{B,1} \mathbf{V}_1^T \oplus \ldots \oplus \mathbf{D}_{A,\beta} \mathbf{V}_\beta \mathbf{D}_{B,\beta} \mathbf{V}_\beta^T\right) & (5.30) \end{aligned}$$

Since $\mathbf{D}_{B,j} = \widehat{b}_j \cdot \mathbf{I}_{|\mathcal{J}_j|}$, $1 \leq j < \beta$, and $\mathbf{D}_{B,\beta} = \widehat{b}_\beta \cdot \mathbf{I}_{|\mathcal{J}_\beta| + |\mathcal{J}_{\beta+1}|}$, for each j we have

$$\mathbf{D}_{A,j} \mathbf{V}_j \mathbf{D}_{B,j} \mathbf{V}_j^T = \widehat{b}_j \cdot \mathbf{D}_{A,j} \mathbf{V}_j \mathbf{V}_j^T = \widehat{b}_j \cdot \mathbf{D}_{A,j} = \mathbf{D}_{A,j} \mathbf{D}_{B,j} \tag{5.31}$$

and therefore $\mathbf{tr}\left(\mathbf{D}_A \mathbf{V} \mathbf{D}_B \mathbf{V}^T\right) = \mathbf{tr}\left(\mathbf{D}_A \mathbf{D}_B\right)$, which as we know is optimal. Let us observe that the previous analysis works almost unaltered for the case when $N \leq M$ (*i.e.*, $N = K$), regardless of the value of $\widehat{b}_\beta$. (The only difference is that $\mathcal{J}_{\beta+1} = \emptyset$.)

Let us analyze the case when $\widehat{b}_\beta < 1$. If $N > M (= K)$, and $a_M > a_{M+1}$, the analysis goes almost exactly as before, the only difference being in the structure of the optimal $\mathbf{V}$. Namely, $\mathbf{V}$ would be block-diagonal, having $\beta+1$ (orthogonal) blocks of size $|\mathcal{J}_j|$, respectively. An interesting case occurs when $N > M (= K)$ (*i.e.*, we are in the "undercomplete case"), and $a_M = a_{M+1}$ (*i.e.*, all the singular values are non-degenerate and M "splits" a contiguous interval of variance values). We can assume, without losing generality that $\beta = 1$, otherwise the first $\beta - 1$ blocks of $\mathbf{V}$ are orthogonal, as before. In other words, our problem is the following: given $a_1 = \ldots = a_r > \ldots \geq a_N > 0$, and $b_1 = \ldots = b_p < b_{p+1} = \ldots = b_N = 1$, with $r > p$, for which permutations τ of the indices we have $\sum_{i=1}^{N} a_i b_{\tau(i)} = \sum_{i=1}^{N} a_i b_i$? The answer is quite immediate, namely we can only consider permutations for which the smallest p values of $\mathbf{b}$ are paired with some of

the largest values values of $\mathbf{a}$, or more simply stated for which $\{1, 2 \ldots, p\} \subseteq \tau(\{1, 2 \ldots, r\})$. But this condition is equivalent to $\tau(\{r+1, \ldots, N\}) \bigcap \{1, 2 \ldots, p\} = \emptyset$, which means that the lower left submatrix of $\mathbf{P}(\tau)$, corresponding to rows $r+1, \ldots, N$ and columns $1, \ldots, p$ is null. But then $\mathbf{V} \circ \mathbf{V}$ and $\mathbf{V}$ will also have this property! To complete the analysis, let us show that, for any orthogonal matrix $\mathbf{V}$ of the form

$$\mathbf{V} = \begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix} \tag{5.32}$$

we have $\mathbf{tr}\left(\mathbf{D}_A \mathbf{V} \mathbf{D}_B \mathbf{V}^T\right) = \mathbf{tr}\left(\mathbf{D}_A \mathbf{D}_B\right)$ (here $\mathbf{X} \in \mathbb{R}^{r \times p}$). First, let us observe that the orthogonality of $\mathbf{V}$ implies

$$\mathbf{X}^T \mathbf{X} = \mathbf{I}_p, \qquad \mathbf{X}\mathbf{X}^T + \mathbf{Y}\mathbf{Y}^T = \mathbf{I}_r, \quad \text{and} \quad \mathbf{Z}\mathbf{Z}^T = \mathbf{I}_{N-r}. \tag{5.33}$$

We have:

$$\mathbf{tr}\left(\mathbf{D}_A \mathbf{V} \mathbf{D}_B \mathbf{V}^T\right) = \mathbf{tr}\left(\mathbf{D}_A \begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix} \begin{pmatrix} \widehat{b}_\beta \mathbf{I}_p & \\ & \mathbf{I}_{N-p} \end{pmatrix} \begin{pmatrix} \mathbf{X}^T & \mathbf{0} \\ \mathbf{Y}^T & \mathbf{Z}^T \end{pmatrix}\right) \tag{5.34}$$

$$= \mathbf{tr}\left(\mathbf{D}_A \begin{pmatrix} \widehat{b}_\beta \mathbf{X} & \mathbf{Y} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix} \begin{pmatrix} \mathbf{X}^T & \mathbf{0} \\ \mathbf{Y}^T & \mathbf{Z}^T \end{pmatrix}\right) \tag{5.35}$$

$$= \mathbf{tr}\left(\mathbf{D}_A \begin{pmatrix} \widehat{b}_\beta \mathbf{X}\mathbf{X}^T + \mathbf{Y}\mathbf{Y}^T & \mathbf{Y}\mathbf{Z}^T \\ \mathbf{Z}\mathbf{Y}^T & \mathbf{Z}\mathbf{Z}^T \end{pmatrix}\right) \tag{5.36}$$

$$= \mathbf{tr}\left(\mathbf{D}_A \begin{pmatrix} (\widehat{b}_\beta - 1)\mathbf{X}\mathbf{X}^T & \mathbf{Y}\mathbf{Z}^T \\ \mathbf{Z}\mathbf{Y}^T & \mathbf{0} \end{pmatrix} + \mathbf{D}_A\right) \tag{5.37}$$

$$= \mathbf{tr}\, \mathbf{D}_A - (1 - \widehat{b}_\beta)\, \mathbf{tr}\left(\mathbf{D}_{A,\beta} \mathbf{X}\mathbf{X}^T\right) \tag{5.38}$$

Since $a_1 = \ldots = a_r$, and $r > p$, it follows that $\mathbf{D}_{A,\beta} = a_1 \mathbf{I}_r$, and so

$$\begin{aligned}
\mathbf{tr}\left(\mathbf{D}_A \mathbf{V} \mathbf{D}_B \mathbf{V}^T\right) &= \mathbf{tr}\, \mathbf{D}_A - (1 - \widehat{b}_\beta)\, \mathbf{tr}\left(\mathbf{D}_{A,\beta} \mathbf{X}\mathbf{X}^T\right) && (5.39)\\
&= \mathbf{tr}\, \mathbf{D}_A - a_1(1 - \widehat{b}_\beta)\, \mathbf{tr}\left(\mathbf{X}\mathbf{X}^T\right) && (5.40)\\
&= \mathbf{tr}\, \mathbf{D}_A - a_1(1 - \widehat{b}_\beta)\, \mathbf{tr}\left(\mathbf{X}^T\mathbf{X}\right) && (5.41)\\
&= \mathbf{tr}\, \mathbf{D}_A - a_1(1 - \widehat{b}_\beta)\, \mathbf{tr}\left(\mathbf{I}_p\right) && (5.42)\\
&= \sum_{i=1}^{N} a_i - p a_1 (1 - \widehat{b}_\beta) && (5.43)\\
&= \sum_{i=1}^{p} a_i - p a_1 (1 - \widehat{b}_\beta) + \sum_{i=p+1}^{N} a_i && (5.44)\\
&= \sum_{i=1}^{p} a_i \left(1 - (1 - \widehat{b}_\beta)\right) + \sum_{i=p+1}^{N} a_i && (5.45)\\
&= \widehat{b}_\beta \sum_{i=1}^{p} a_i + \sum_{i=p+1}^{N} a_i && (5.46)\\
&= \mathbf{tr}\left(\mathbf{D}_A \mathbf{D}_B\right) && (5.47)
\end{aligned}$$

Algorithm 1 Compute $\mathbf{U}$, the left singular matrix of $\mathbf{T}_{min}$.

Let $\mathbf{U}_0 \in \mathcal{O}_M$ arbitrary.
$\mathbf{Q} \Leftarrow \mathbf{U}_0\ (\Sigma^2_{K,min} \oplus \mathbf{0}_{M-K})\mathbf{U}_0^T$
for $i = 1$ to $M-1$ **do**
  Find indices j, k such that $(1-q_{jj})(1-q_{kk}) < 0$; otherwise, **stop**.
  Let $\theta_i \in [0, 2\pi)$, such that $q'_{jj} = 1$, where $\mathbf{Q}' = \mathbf{G}_{j,k}(\theta_i)\ \mathbf{Q}\ \mathbf{G}_{j,k}(\theta_i)^T$
  $\mathbf{U}_i \Leftarrow \mathbf{G}_{j,k}(\theta_i)\mathbf{U}_{i-1}$
  $\mathbf{Q} \Leftarrow \mathbf{Q}'$
end for
Output $\mathbf{U} = \mathbf{U}_i$

In conclusion, in the case when $N > M$ and $a_M = a_{M+1}$, matrix $\mathbf{V}$ is block-diagonal, with the first $\beta - 1$ blocks orthogonal matrices, and the β^{th} block of the form shown in eq. (5.32). Our attempt to characterize the right singular vector matrices of $\mathbf{T}_{min}$ is now complete.

Finding U, the left singular vector matrix of $\mathbf{T}_{min}$

In this section we shall describe how to compute the left singular matrix of $\mathbf{T}_{min}$ when we know the optimal singular values (i.e., $\mathbf{\Sigma}_{K,min}$). As we noticed already, this is needed only to satisfy the constraint and does not influence the cost function. Namely, our goal will be to search for $\mathbf{U} \in \mathcal{O}_M(\mathbb{R})$, such that $\mathrm{diag}\left(\mathbf{U}\ (\Sigma^2_{K,min} \oplus \mathbf{0}_{M-K})\mathbf{U}^T\right) = \mathbf{1}_{M,1}$.

Remark. *For any $\mathbf{U} \in \mathcal{O}_M(\mathbb{R})$, we have:*

$$\mathbf{tr}\left(\mathbf{U}\ (\mathbf{\Sigma}^2_{K,min} \oplus \mathbf{0}_{M-K})\mathbf{U}^T\right) = \mathbf{tr}\left(\mathbf{T}\mathbf{T}^T\right) = M \tag{5.48}$$

Using this observation, we shall provide a very simple and efficient way to obtain the $\mathbf{U}$ matrix, based on Givens rotations. We illustrate this procedure (hereby referred as Algorithm 5.2), and then prove its correctness.

Lemma 4. *Algorithm 5.2 computes a matrix $\mathbf{U} \in \mathcal{O}_M$, satisfying the constraint condition*

$$\mathrm{diag}\left(\mathbf{U}\ (\Sigma^2_{K,min} \oplus \mathbf{0}_{M-K})\mathbf{U}^T\right) = \mathbf{1}_{M,1}. \tag{5.49}$$

Proof. *See Appendix.*

As it can easily be observed, there may be multiple ways to "fix" the initial matrix $\mathbf{U}_0$ into an acceptable solution. Let us also remark that the procedure just described remains virtually unchanged if we work with the incomplete SVD, rather than with full SVD, in the overcomplete case ($M > N$). The only difference is in the choice of the starting point $\mathbf{U}_0 \in \mathbb{R}^{M \times N}$, having an orthonormal set of columns. Unfortunately, this observation does not essentially reduce the computational complexity of finding matrix $\mathbf{U}$, which is still going to be $O(M^2)$.

Error Bound

It is useful to know what is the theoretical limit of the cost function. To find the exact formula for the lower bound, we first need to identify the rank R of the encoding matrix, that is we need to

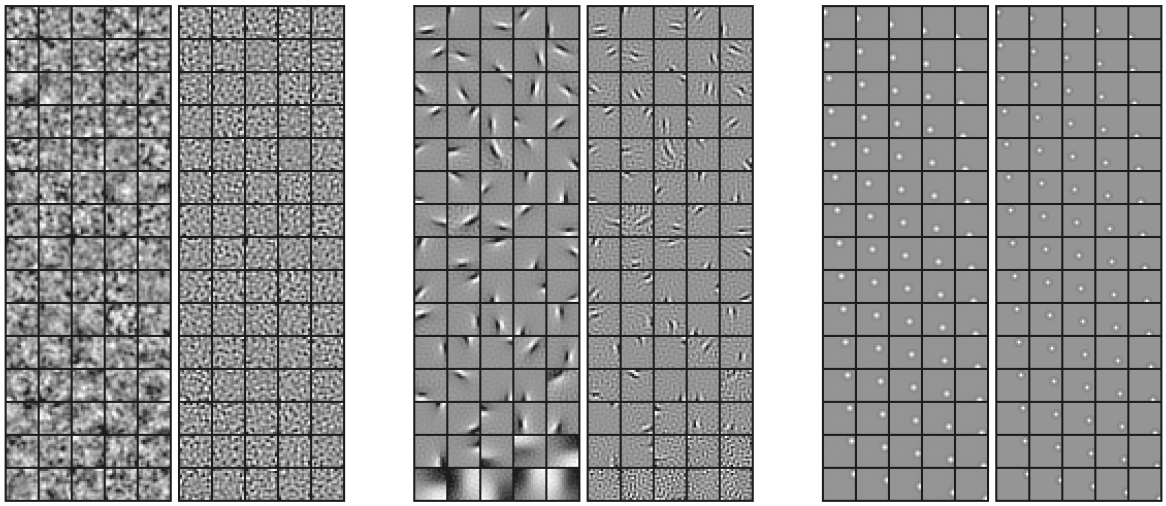


Figure 5.2: Robust Coding solutions: computed without additional constraints (left), with "sparsity" constraints (center), and with "locality" constraints (right). [Courtesy of E. Doi.]

check until when does the non-degeneracy condition hold. To do this, binary search is sufficient, and so we can find R in $O(\log M)$ time. Once it is known, we can express the closed-form lower bound of $\mathcal{E}$, as demonstrated by the following lemma.

Lemma 5. *The minimal value of the cost function $\mathcal{E}$ is*

$$\min \mathcal{E} = \frac{\left(\sum_{j=1}^{R} s_j\right)^2}{R + \gamma^2 M} + \sum_{i=R+1}^{N} s_i^2. \tag{5.50}$$

Proof. *See Appendix.*

This result proves, as well as improves our conjecture in [44].

5.3 Robust Coding Algorithms

We have described the optimal solutions of the Robust Coding problem when the only constraint was that the SNR of the channels was identical, arguing that this should lead to a more realistic model for neural encoding. As we pointed out, any solution of the form described in section 5.2 is equally good, although it is not obvious what kind of image structure do these optimal solutions represent (see figure 5.2).

The most natural explanation for this lack of structure is the fact that we only focused on one biological constraint that accompany neural encoding. Other constraints, such as the sparsity of the distribution of encoding coefficients, known to hold in real biological systems, should also be taken into account.

This motivation has lead to more structured optimal RC solutions in [47]. There, by extending the RC model to account for optical blur, as well as for sparsity of the neural activity, they managed to produce optimal solutions closely resembling receptive fields of retinal ganglion cells. For the computation, they used gradient descent with Lagrange multipliers, thus grouping all the constraints and the objective function. Although the optimization converged, the number of iterations was considerable, even for a relatively moderate-size problem. We will focus more on the computational side of their problem. Namely, we are interested in deriving algorithms that take advantage of the known structure of the Robust Coding optimal solutions to accommodate additional constraints. The obvious advantage of such an approach is the significant reduction of the (structured) search space.

In our case, since the only set of parameters needed to differentiate among all RC solutions is the set of left singular vectors (an orthonormal matrix), we propose using optimization algorithms on Stiefel and Grassmann manifolds (for a comprehensive description and categorization of most widely used algorithms of this kind, see [7]). The idea of using a "natural gradient" (as it is sometimes referred) has lead to what is currently state-of-the-art stochastic gradient approach to ICA ([30, 66]), but has been applied to many other problems (see [7] for a thorough review, and [2] for a very good monograph).

We recently found that variants of the (unconstrained) Robust Coding problem have been known already in the information theory literature. The most similar to our analysis is the one presented in [77, 78]. Our structural characterization of the optimal encoder matrices is (in our opinion) more explicit and elegant. Regarding the algorithm for computation of the left singular matrix $\mathbf{U}$, we learned that it has been rediscovered at least two more times since [78], being included in [15] and [27]. Fortunately, we can now take advantage of recent mathematical advances such as optimization algorithms on manifolds to identify algorithms with better computational and numerical properties. Applicability of the general theory of optimization on Stiefel and Grassmann manifolds was eased significantly by the implementation and release of the `sg` MATLAB package (see [82]). We feel that by adapting such power to the specifics of Robust Coding (and its generalizations) we can advance our understanding about this fundamental problem.

5.4 Discussion

We have provided a theoretical analysis of Robust Coding solutions in the general case: arbitrary dimension, arbitrary number of encoding units, arbitrary precision, and spectrum of the data covariance matrix.

A consequence of our result is that we can manipulate the parameters to arbitrarily reduce the error. For instance, one possible intuitive interpretation of the formula is that by increasing the redundancy (*i.e.*, M), we can compensate for the limited coding precision and thus we can overcome the effect of noise in the representation. Alternatively, for a fixed number of encoding units a higher capacity will result in a lower error[5].

[5]We should emphasize that the "break" index J is dependent of both M and γ. However its effect does not affect

By keeping our model flexible enough to handle arbitrary (*i.e.*, non-isotropic) covariance also implies that Robust Coding optimal encoder/decoder are not necessarily tight frames. Indeed, because of the way index R is defined we notice that the optimal matrices may not have full rank. This is to say that we first select a subspace *most relevant* for the data distribution, and then allocate all the representation resources (encoding vectors) to span this subspace as well as possible. Such an interpretation is related to the one in [26], however it is more general.

5.5 Appendix

Reformulate cost function $\mathcal{E}$. We show how the cost function is transformed when we plug in the expression of the optimal decoding matrix ($\mathbf{B_T}$ from eq. (5.7)) into eq. (5.4):

$$\begin{aligned}
\mathcal{E}(\mathbf{B_T}, \mathbf{T}) &= ||\mathbf{B_T T} - \mathbf{S}||_F^2 + \frac{1}{\gamma^2}||\mathbf{B_T}||_F^2 && \text{(A-1)}\\
&= \mathbf{tr}\left((\mathbf{B_T T} - \mathbf{S})(\mathbf{B_T T} - \mathbf{S})^T\right) + \frac{1}{\gamma^2}\mathbf{tr}\left(\mathbf{B_T B_T^T}\right) && \text{(A-2)}\\
&= \mathbf{tr}\left(\mathbf{B_T T T}^T\mathbf{B_T^T} + \mathbf{S}^2 - \mathbf{S T}^T\mathbf{B_T^T} - \mathbf{B_T T S} + \frac{1}{\gamma^2}\mathbf{B_T B_T^T}\right) && \text{(A-3)}\\
&= \mathbf{tr}\left(\mathbf{B_T}(\frac{1}{\gamma^2}\mathbf{I}_M + \mathbf{T T}^T)\mathbf{B_T^T} + \mathbf{S}^2 - \mathbf{S T}^T\mathbf{B_T^T} - \mathbf{B_T T S}\right) && \text{(A-4)}\\
&= \mathbf{tr}\left(\mathbf{S T}^T\mathbf{B_T^T} + \mathbf{S}^2 - \mathbf{S T}^T\mathbf{B_T^T} - \mathbf{B_T T S}\right) && \text{(A-5)}\\
&= \mathbf{tr}\left((\mathbf{S} - \mathbf{B_T T})\mathbf{S}\right) && \text{(A-6)}\\
&= \mathbf{tr}\left(\mathbf{S}(\mathbf{S} - \mathbf{B_T T})\right) && \text{(A-7)}\\
&= \mathbf{tr}\left(\mathbf{S}(\mathbf{S} - \mathbf{S T}^T(\frac{1}{\gamma^2}\mathbf{I}_M + \mathbf{T T}^T)^{-1}\mathbf{T})\right) && \text{(A-8)}\\
&= \mathbf{tr}\left(\mathbf{S}^2(\mathbf{I}_N - \mathbf{T}^T(\frac{1}{\gamma^2}\mathbf{I}_M + \mathbf{T T}^T)^{-1}\mathbf{T})\right). && \text{(A-9)}
\end{aligned}$$

Sherman-Morrison-Woodbury formula.

Proposition 1 ([56]). *Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{N\times M}$, $\mathbf{C} \in \mathbb{R}^N$, such that both $\mathbf{C}$ and $\mathbf{I}_M + \mathbf{Y}^T\mathbf{C}^{-1}\mathbf{X}$ are nonsingular. Then $\left(\mathbf{C} + \mathbf{X Y}^T\right)^{-1}$ is also nonsingular and*

$$\left(\mathbf{C} + \mathbf{X Y}^T\right)^{-1} = \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{X}\left(\mathbf{I}_M + \mathbf{Y}^T\mathbf{C}^{-1}\mathbf{X}\right)^{-1}\mathbf{Y}^T\mathbf{C}^{-1}. \quad \text{(A-10)}$$

By plugging $\mathbf{C} = \mathbf{I}_N$, and $\mathbf{X} = \mathbf{Y}^T = \gamma\mathbf{T}^T$ into the identity above, we obtain:

$$(\mathbf{I}_N + \gamma^2\mathbf{T}^T\mathbf{T})^{-1} = \mathbf{I}_N - \mathbf{T}^T(\frac{1}{\gamma^2}\mathbf{I}_M + \mathbf{T T}^T)^{-1}\mathbf{T}. \quad \text{(A-11)}$$

consistency of the interpretations.

Reformulate cost function $\mathcal{F}$.

$$\mathcal{F}(\mathbf{T}) = \mathbf{tr}\left(\mathbf{S}^2(\mathbf{I}_N + \gamma^2\mathbf{T}^T\mathbf{T})^{-1}\right) \tag{A-12}$$
$$= \mathbf{tr}\left(\mathbf{S}^2(\mathbf{V}\mathbf{V}^T + \gamma^2\mathbf{V}\left(\mathbf{\Sigma}_K^2 \oplus \mathbf{0}_{N-K}\right)\mathbf{V}^T)^{-1}\right) \tag{A-13}$$
$$= \mathbf{tr}\left(\mathbf{S}^2\mathbf{V}(\mathbf{I}_N + \gamma^2(\mathbf{\Sigma}_K^2 \oplus \mathbf{0}_{N-K})))^{-1}\mathbf{V}^T\right) \tag{A-14}$$
$$= \mathbf{tr}\left(\mathbf{S}^2\mathbf{V}\left((\mathbf{I}_K + \gamma^2\mathbf{\Sigma}_K^2)^{-1} \oplus \mathbf{I}_{N-K}\right)\mathbf{V}^T\right) \tag{A-15}$$
$$= \mathbf{tr}\left(\mathbf{S}^2\mathbf{V}\left(\mathrm{diag}(\frac{1}{1+\gamma^2\sigma_1^2}, \ldots, \frac{1}{1+\gamma^2\sigma_K^2}) \oplus \mathbf{I}_{N-K}\right)\mathbf{V}^T\right) \tag{A-16}$$

Proof of Lemma 1. Fix arbitrary permutation matrix $\mathbf{P}$, and define matrix $\mathbf{P}_N = \mathbf{P} \oplus \mathbf{I}_{N-K}$. Then,

$$\mathcal{G}(\mathbf{P}\mathbf{\Sigma}_K\mathbf{P}^T) = \min_{\mathbf{V}\in\mathcal{O}_N(\mathbb{R})} \mathcal{F}(\mathbf{P}\mathbf{\Sigma}_K\mathbf{P}^T, \mathbf{V}) = \min_{\mathbf{V}\in\mathcal{O}_N(\mathbb{R})} \mathcal{F}(\mathbf{\Sigma}_K, \mathbf{V}\mathbf{P}_N) = \min_{\mathbf{V}'\in\mathcal{O}_N(\mathbb{R})} \mathcal{F}(\mathbf{\Sigma}_K, \mathbf{V}') = \mathcal{G}(\mathbf{\Sigma}_K). \tag{A-17}$$

Proof of Lemma 2. Due to the inverse ordering of elements on the two diagonals, by Hardy-Littlewood-Polya rearrangement lemma we know that for any permutation π of $\{1, 2, \ldots, n\}$ we have:

$$\sum_{i=1}^{n} a_i b_{\pi(i)} \geq \sum_{i=1}^{n} a_i b_i. \tag{A-18}$$

Let us inspect the function we intend to optimize:

$$\mathbf{tr}\left(\mathbf{D}_A\mathbf{V}\mathbf{D}_B\mathbf{V}^T\right) = \mathbf{tr}\left(\mathbf{D}_A^{1/2}\mathbf{V}\mathbf{D}_B\mathbf{V}^T\mathbf{D}_A^{1/2}\right) = \|\mathbf{D}_A^{1/2}\mathbf{V}\mathbf{D}_B^{1/2}\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i c_{ij}^2 b_j$$
$$= \mathbf{a}^T(\mathbf{V}\circ\mathbf{V})\mathbf{b}$$

where $\mathbf{a} = (a_1, \ldots, a_n)^T$, $\mathbf{b} = (b_1, \ldots, b_n)^T$.

Remark. *For any orthogonal matrix $\mathbf{V} = (v_{ij}) \in \mathcal{O}_n(\mathbb{R})$, the Hadamard product $\mathbf{V}\circ\mathbf{V} = \left(v_{ij}^2\right)$ is a doubly stochastic matrix. This property is simply a restatement of the fact that the rows and columns of $\mathbf{V}$ are unit length vectors.*

Birkhoff's theorem [62, p. 527] states that for any $n \times n$ doubly stochastic matrix $\mathbf{Q}$ (in particular, for $\mathbf{Q} = \mathbf{V}\circ\mathbf{V}$), we can write $\mathbf{Q}$ as convex combination of permutation matrices. In other words, there exist $m \in \mathbb{N}^*$, nonnegative coefficients $\alpha_1, \ldots, \alpha_m$ with $\sum_{k=1}^{m} \alpha_k = 1$, and permutations $\tau_1, \ldots, \tau_m$ of $\{1, 2, \ldots, n\}$ such that $\mathbf{Q} = \sum_{k=1}^{m} \alpha_k \mathbf{P}(\tau_k)$. Consequently,

$$\mathbf{tr}\left(\mathbf{D}_A\mathbf{V}\mathbf{D}_B\mathbf{V}^T\right) = \mathbf{a}^T\left(\sum_{k=1}^{m}\alpha_k\mathbf{P}(\tau_k)\right)\mathbf{b} = \sum_{k=1}^{m}\left[\alpha_k\left(\mathbf{a}^T\mathbf{P}(\tau_k)\mathbf{b}\right)\right] = \sum_{k=1}^{m}\left[\alpha_k\left(\sum_{i=1}^{n} a_i b_{\tau_k(i)}\right)\right]$$
$$\geq \sum_{k=1}^{m}\left[\alpha_k\left(\sum_{i=1}^{n} a_i b_i\right)\right] = \left(\sum_{i=1}^{n} a_i b_i\right)\cdot\sum_{k=1}^{m}\alpha_k = \sum_{i=1}^{n} a_i b_i = \mathbf{tr}\left(\mathbf{D}_A\mathbf{D}_B\right).$$

The cost is minimized for those and only for those orthogonal matrices $\mathbf{V}$ such that $\mathbf{V} \circ \mathbf{V}$ is a convex combination of matrices corresponding to permutations τ for which $\sum_{i=1}^{n} a_i b_{\tau(i)} = \sum_{i=1}^{n} a_i b_i$. Consequently, we obtain the desired result.

Proof of Theorem 5.2.1. Let $J = K$. Let us characterize the ensembles $(\sigma_i)_{1 \leq i \leq J}$ that minimize

$$\mathcal{G}_J(\sigma_1, \ldots, \sigma_J) = \sum_{i=1}^{J} \frac{s_i^2}{1 + \gamma^2 \sigma_i^2} \tag{A-19}$$

subject to $\sum_{i=1}^{J} \sigma_i^2 = M$, and $\sigma_1 \geq \ldots \geq \sigma_J \geq 0$.

If $J = 1$, then the solution is obvious: there is only one nonnegative value (namely $\sigma_1 = \sqrt{M}$) satisfying the constraint and therefore it will also be optimal. If $J > 1$, we need a more elaborate way of analysis and for this purpose we shall use an auxiliary lemma which will allow us to relax the constraints, without critically influencing the optimal solutions.

Lemma. *With the above notation, we have:*

$$\min_{\substack{\sum_{i=1}^{J} \sigma_i^2 = M \\ \sigma_1 \geq \ldots \geq \sigma_J \geq 0}} \mathcal{G}_J(\sigma_1, \ldots, \sigma_J) = \min_{\sum_{i=1}^{J} \sigma_i^2 = M} \mathcal{G}_J(\sigma_1, \ldots, \sigma_J) \tag{A-20}$$

Proof. *It is sufficient to prove that the minimum value on the left-hand side (i.e., on the more restricted domain) is no larger than the one on the right-hand side. First, let us observe that $\mathcal{G}_J$ is a continuous function. If we define this function on the compact $\mathcal{S}_J(0; \sqrt{M})$ (the surface of a ball), $\mathcal{G}_J$ reaches its minimum at a point $(x_1, \ldots, x_J)^T \in \mathcal{S}_J(0; \sqrt{M})$. But since $\mathcal{G}_J$ is even in each of its arguments, we have*

$$\mathcal{G}_J(x_1, \ldots, x_J) = \mathcal{G}_J(|x_1|, \ldots, |x_J|) \tag{A-21}$$

and so we can assume without losing generality that $x_i \geq 0,\ 1 \leq i \leq J$.

Moreover, $\mathcal{G}_J(x_1, \ldots, x_J) \geq \mathcal{G}_J(x_{\tau(1)}, \ldots, x_{\tau(J)})$ with equality if $x_{\tau(1)} \geq \ldots \geq x_{\tau(J)}$. This means that we can also assume without losing generality that $x_1 \geq \ldots \geq x_J \geq 0$. Thus the lemma is proved.

We transformed our problem into minimizing a continuous function on a less restricted domain, which will help us describe the optimal ensembles in a simple fashion. We will employ a slightly different parametrization of $\mathcal{G}_J$, which not only embeds the constraint, but also allows us to optimize a continuously differentiable function defined on a closed ball (as opposed to the surface of such a ball).

Namely, let $\mathcal{H}_J : \mathcal{B}_{J-1}(0; \sqrt{M}) \to (0, \infty)$,

$$\mathcal{H}_J(\sigma_1, ..., \sigma_{J-1}) = \sum_{i=1}^{J-1} \frac{s_i^2}{1 + \gamma^2 \sigma_i^2} + \frac{s_J^2}{1 + \gamma^2 \left(M - \sum_{i=1}^{J-1} \sigma_i^2 \right)} \tag{A-22}$$

which is a continuous function on the compact $\mathcal{B}_{J-1}(0;\sqrt{M})$, and therefore has at least a minimum point $(\overline{\sigma}_i)_{1\leq i\leq J}$ on this set. Such a point can lie either on the border of the ball, or on the interior. The first case implies that $\sum_{i=1}^{J-1}\overline{\sigma}_i^2 = M$, and so $\overline{\sigma}_J = 0$. The problem is then reduced to finding the minimum of the analogously defined cost function $\mathcal{G}_{J-1}$, subject to $\sum_{i=1}^{J-1}\sigma_i^2 = M$.

As the sum of $\overline{\sigma}_i^2$'s is M, there must exist an index i for which $\overline{\sigma}_i \neq 0$. We can assume without losing generality that J is the largest such index. As noticed before, if $J = 1$, we are done. Otherwise, the minimum point $(\overline{\sigma}_1,\ldots,\overline{\sigma}_{J-1})$ of $\mathcal{H}_J$ lies on the interior of $\mathcal{B}_{J-1}(0;\sqrt{M})$. From Lemma 5.5, we can assume that the optimal ensemble satisfies $\overline{\sigma}_1 \geq \ldots \geq \overline{\sigma}_J > 0$.

Since $\mathcal{H}_J$ is differentiable on the interior of its domain, it follows that for all $1 \leq i < J$:

$$\frac{\partial \mathcal{H}_J}{\partial \sigma_i} = \frac{-2\overline{\sigma}_i\gamma^2 s_i^2}{\left[1+\gamma^2\overline{\sigma}_i^2\right]^2} + \frac{2\overline{\sigma}_i\gamma^2 s_J^2}{\left[1+\gamma^2\left(M-\sum_{j=1}^{J-1}\overline{\sigma}_j^2\right)\right]^2} \tag{A-23}$$

$$= -2\overline{\sigma}_i\gamma^2\left(\frac{s_i^2}{\left[1+\gamma^2\overline{\sigma}_i^2\right]^2} - \frac{s_J^2}{\left[1+\gamma^2\left(M-\sum_{j=1}^{J-1}\overline{\sigma}_j^2\right)\right]^2}\right) \tag{A-24}$$

and the gradient is null in $(\overline{\sigma}_1,\ldots,\overline{\sigma}_{J-1})$. We obtain

$$\frac{s_i^2}{\left[1+\gamma^2\overline{\sigma}_i^2\right]^2} = \frac{s_J^2}{\left[1+\gamma^2\left(M-\sum_{j=1}^{J-1}\overline{\sigma}_j^2\right)\right]^2} = \frac{s_J^2}{\left[1+\gamma^2\overline{\sigma}_J^2\right]^2},\ \forall\, 1 \leq i < J \tag{A-25}$$

and consequently

$$\frac{s_i}{1+\gamma^2\overline{\sigma}_i^2} = \frac{\sum_{i=1}^{J} s_i}{J+\gamma^2\sum_{i=1}^{J}\overline{\sigma}_i^2} = \frac{\sum_{i=1}^{J} s_i}{J+\gamma^2 M},\ \forall\, 1 \leq i \leq J \tag{A-26}$$

$$\gamma^2\overline{\sigma}_i^2 = \frac{s_i}{\sum_{j=1}^{J} s_j}\left(J+\gamma^2 M\right) - 1,\ \forall\, 1 \leq i \leq J \tag{A-27}$$

We observe that the existence of an interior minimum point depends on how $\frac{s_J}{\sum_{i=1}^{J} s_i}\left(J+\gamma^2 M\right)$ compares to 1. This is the same as saying that, in order for such a solution to exist, not even the smallest among the s_i should be too much smaller than their "average"[6].

[6] We used quotes here because the lower bound is strictly smaller than the actual average.

Let us denote by R the largest index $J \leq K$ such that

$$\frac{s_J}{\sum_{i=1}^{J} s_i} \left(J + \gamma^2 M\right) > 1. \tag{A-28}$$

Such an index indeed exists, as for $J = 1$ the above condition is satisfied, and the set $\{1, \ldots, K\}$ is finite. We shall prove now that for any index $J \leq R$, it too satisfies condition (A-28).

Denote $t_j = \sum_{i=1}^{j} s_i$, $1 \leq j \leq K$. As observed before, the case $J = 1$ is clear. Assume $J > 1$. Then condition (A-28) is equivalent to

$$s_J > \frac{t_J}{J + \gamma^2 M} \Leftrightarrow t_J - t_{J-1} > \frac{t_J}{J + \gamma^2 M} \Leftrightarrow (t_J - t_{J-1})(J + \gamma^2 M) > t_J$$
$$\Leftrightarrow\ t_J(J - 1 + \gamma^2 M) > t_{J-1}(J + \gamma^2 M) \Leftrightarrow \frac{t_J}{J + \gamma^2 M} > \frac{t_{J-1}}{J - 1 + \gamma^2 M}.$$

But then, since the s_j values were ordered in decreasing order, we have:

$$s_{J-1} \geq s_J > \frac{t_J}{J + \gamma^2 M} > \frac{t_{J-1}}{J - 1 + \gamma^2 M}$$

which proves that if J satisfies (A-28), then so does $J - 1$. Thus, we showed the existence of an index R as in the statement of the theorem.

It follows immediately that the optimal values $\overline{\sigma}_i$ are

$$\overline{\sigma}_i = \begin{cases} \frac{1}{\gamma}\sqrt{\frac{s_i}{t_R}\left(R + \gamma^2 M\right) - 1}, & \text{if } 1 \leq i \leq R \\ 0, & \text{if } R + 1 \leq i \leq K. \end{cases} \tag{A-29}$$

We can easily verify that both constraints on the σ_i are verified by the values above, and thus the theorem is completely proved.

Cauchy-Schwartz fails. Let us try to minimize $\sum_{i=1}^{K} s_i^2/(1 + \gamma^2\sigma_i^2)$ s.t. $\sum_{i=1}^{K} \sigma_i^2 = M$, using the Cauchy-Schwartz inequality:

$$\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) \geq \left(\sum_{i=1}^{n} a_i b_i\right)^2 \tag{A-30}$$

with equality if and only if either all b_i's are zero, or if there exists a real constant ρ, such that $a_i = \rho b_i,\ \forall 1 \leq i \leq n$. In our case, this becomes

$$\left(\sum_{i=1}^{K} \frac{s_i^2}{1 + \gamma^2\sigma_i^2}\right)\left(\sum_{i=1}^{K} (1 + \gamma^2\sigma_i^2)\right) \geq \left(\sum_{i=1}^{K} s_i\right)^2 \tag{A-31}$$

$$\Leftrightarrow\ \sum_{i=1}^{K} \frac{s_i^2}{1 + \gamma^2\sigma_i^2} \geq \frac{\sum_{i=1}^{K} (1 + \gamma^2\sigma_i^2)}{\left(\sum_{i=1}^{K} s_i\right)^2} = \frac{K + \gamma^2 M}{\left(\sum_{i=1}^{K} s_i\right)^2}. \tag{A-32}$$

We may be tempted to say that we have equality in the relation above, if and only if:

$$\frac{s_i}{\sqrt{1+\gamma^2\sigma_i^2}} = ct., \ \forall\, 1 \leq i \leq K. \tag{A-33}$$

Unfortunately, this is *not always* true, as this condition would necessarily imply the condition in eq. (A-27). Nor should it be so, since we are overconstraining the quantities involved in the Cauchy-Schwartz inequality (namely, we impose that all $|b_i|$ be supraunitary).

Proof of Correctness for Algorithm 5.2

First, we need to prove that it is indeed possible to construct the matrices $\mathbf{U}_i$ mentioned above. Assume we are at the i^{th} step in the loop ($1 \leq i \leq M-1$). Namely, we know that the first $i-1$ diagonal entries of matrix $\mathbf{Q}_{i-1} = \mathbf{U}_{i-1}\ldots\mathbf{U}_0(\Sigma^2_{K,min} \oplus \mathbf{0}_{M-K})\mathbf{U}_0^T\ldots\mathbf{U}_{i-1}^T$ are equal to 1. We will search for the orthogonal matrix $\mathbf{U}_i$ such that the first $i-1$ diagonal entries of $\mathbf{Q}_i = \mathbf{U}_i\mathbf{Q}_{i-1}\mathbf{U}_i^T$ remain unchanged, while the i^{th} entry becomes 1. We will restrict our search to orthogonal matrices of a particular form, namely to Givens rotations:

$$\mathbf{G}_{\alpha,\beta}(\theta) = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & \cos\theta & \cdots & \sin\theta & & \\ & & \vdots & \ddots & \vdots & & \\ & & -\sin\theta & \cdots & \cos\theta & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix}. \tag{A-34}$$

where $1 \leq \alpha < \beta \leq M$, and $\theta \in [0, \pi)$. In our particular case, we shall take $\mathbf{U}_i = \mathbf{G}_{i,\beta_i}(\theta_i)$ and we show how to find parameters β_i and θ_i such that the invariant is satisfied.

To simplify the description, let us denote $\mathbf{Q}_l = \left(q_{jk}^{(l)}\right)_{j,k=\overline{1,M}}$. Then, by assumption we have $q_{jj}^{(i-1)} = 1, \forall\, 1 \leq j < i$. Let us observe that due to the particular form of $\mathbf{U}_i$, we have $q_{jj}^{(i)} = q_{jj}^{(i-1)}, \forall\, 1 \leq j \leq M, i \neq j \neq \beta_i$. In the simplest case (namely if $q_{ii}^{(i-1)} = 1$), we take $\theta_i = 0$, $\alpha = i$ and $\beta = i+1$, which gives $\mathbf{U}_i = \mathbf{I}_M$. Then $q_{jj}^{(i)} = 1, \forall\, 1 \leq j \leq i$. Now, assume without loss of generality that $q_{ii}^{(i-1)} < 1$. Since $M = \mathbf{tr}\ \mathbf{Q}_{i-1} = \sum_{j=1}^{M} q_{jj}^{(i-1)}$, it follows that there exists an index β_i, $i+1 \leq \beta_i \leq M$, such that $q_{\beta_i\beta_i}^{(i-1)} > 1$. We will search now for θ_i such that $q_{ii}^{(i)} = 1$. Let $\mathbf{g} \in \mathbb{R}^M$ be defined by

$$g_j = \begin{cases} \cos\theta_i, & \text{if } j = i \\ \sin\theta_i, & \text{if } j = \beta_i \\ 0, & \text{otherwise.} \end{cases} \tag{A-35}$$

(In other words, vector $\mathbf{g}$ is the transposed of the i^{th} row of matrix $\mathbf{U}_i$.) Then:

$$\begin{aligned} q_{ii}^{(i)} &= \mathbf{g}^T \mathbf{Q}_{i-1} \mathbf{g} && \text{(A-36)} \\ &= \begin{pmatrix} \cos\theta_i, & \sin\theta_i \end{pmatrix} \begin{pmatrix} q_{ii}^{(i-1)} & q_{\beta_i i}^{(i-1)} \\ q_{i\beta_i}^{(i-1)} & q_{\beta_i\beta_i}^{(i-1)} \end{pmatrix} \begin{pmatrix} \cos\theta_i \\ \sin\theta_i \end{pmatrix} && \text{(A-37)} \\ &= \begin{pmatrix} \cos\theta_i, & \sin\theta_i \end{pmatrix} \begin{pmatrix} q_{ii}^{(i-1)}\cos\theta_i + q_{\beta_i i}^{(i-1)}\sin\theta_i \\ q_{i\beta_i}^{(i-1)}\cos\theta_i + q_{\beta_i\beta_i}^{(i-1)}\sin\theta_i \end{pmatrix} && \text{(A-38)} \\ &= q_{ii}^{(i-1)}\cos^2\theta_i + 2q_{i\beta_i}^{(i-1)}\cos\theta_i\sin\theta_i + q_{\beta_i\beta_i}^{(i-1)}\sin^2\theta_i && \text{(A-39)} \end{aligned}$$

We notice that $\sin\theta_i = 0$ will not give an acceptable solution, since then $q_{ii}^{(i)} = q_{ii}^{(i-1)} < 1$. Therefore we can assume $\sin\theta_i \neq 0$. Then

$$\begin{aligned} \cot^2\theta_i + 1 &= \frac{1}{\sin^2\theta_i} = \frac{q_{ii}^{(i)}}{\sin^2\theta_i} && \text{(A-40)} \\ &= \frac{q_{ii}^{(i-1)}\cos^2\theta_i + 2q_{i\beta_i}^{(i-1)}\cos\theta_i\sin\theta_i + q_{\beta_i\beta_i}^{(i-1)}\sin^2\theta_i}{\sin^2\theta_i} && \text{(A-41)} \\ &= q_{ii}^{(i-1)}\frac{\cos^2\theta_i}{\sin^2\theta_i} + 2q_{i\beta_i}^{(i-1)}\frac{\cos\theta_i}{\sin\theta_i} + q_{\beta_i\beta_i}^{(i-1)} && \text{(A-42)} \\ &= q_{ii}^{(i-1)}\cot^2\theta_i + 2q_{i\beta_i}^{(i-1)}\cot\theta_i + q_{\beta_i\beta_i}^{(i-1)} && \text{(A-43)} \end{aligned}$$

Now, all need to do is to solve a quadratic equation in $t = \cot\theta_i$:

$$(1 - q_{ii}^{(i-1)}t^2 - 2q_{i\beta_i}^{(i-1)}t + 1 - q_{\beta_i\beta_i}^{(i-1)} = 0 \qquad \text{(A-44)}$$

Since we assumed $1 - q_{ii}^{(i-1)} > 0$ and $1 - q_{\beta_i\beta_i}^{(i-1)} < 0$, it follows that the discriminant is positive:

$$\Delta = 4(q_{i\beta_i}^{(i-1)})^2 - 4(1 - q_{ii}^{(i-1)})(1 - q_{\beta_i\beta_i}^{(i-1)}) > 0 \qquad \text{(A-45)}$$

and therefore we can take $\theta_i = \cot^{-1}\left(\frac{q_{i\beta_i}^{(i-1)} \pm \sqrt{\Delta/4}}{1 - q_{ii}^{(i-1)}}\right)$.

Proof of Lemma 5 (Error lower-bound). Let us recall that

$$\min MSE = \min_{\mathbf{B},\mathbf{T}} \mathcal{E} = \min_{\mathbf{T}} \mathcal{F} = \min_{\mathbf{\Sigma}_K,\mathbf{V}} \mathcal{F} = \min_{\mathbf{\Sigma}_K} \mathcal{G} \qquad \text{(A-46)}$$

where in each case we assumed the appropriate constraint. Due to eq. (5.20) and to the formula (5.23) on the optimal singular values, we have:

$$\begin{aligned} \min\mathcal{E} &= \sum_{i=1}^{R} \frac{s_i^2}{1+\gamma^2\sigma_i^2} + \sum_{i=R+1}^{N} s_i^2 = \sum_{i=1}^{R} \frac{s_i^2}{\frac{s_i}{\sum_{j=1}^{R} s_j}(R+\gamma^2 M)} + \sum_{i=R+1}^{N} s_i^2 && \text{(A-47)} \\ &= \frac{\sum_{j=1}^{R} s_j}{R+\gamma^2 M} \sum_{i=1}^{R} \frac{s_i^2}{s_i} + \sum_{i=R+1}^{N} s_i^2 = \frac{\left(\sum_{j=1}^{R} s_j\right)^2}{R+\gamma^2 M} + \sum_{i=R+1}^{N} s_i^2. && \text{(A-48)} \end{aligned}$$

This gives an exact formula for the lower bound of the error function.

Chapter 6

Conclusions

We have studied several ways in which related existing frameworks for signal representations, particularly those pertaining to visual signal encoding, can be extended and improved. Since the description of a signal or of an entire class of signals is intrinsically suboptimal if it does not account for statistical properties of that class, we embrace the adaptive encoding point of view as essential to an optimal design. The main challenge is that in so many cases adaptivity is too impractical to employ, and therefore trading optimality for computation might not necessarily be worth it. We beg to differ. As such, we suggested several ways in which the particularity of the problem can be exploited to allow for practical adaptive solutions. In each of the instances hereby studied, we follow two main goals: to clearly identify the theoretical principles which govern the representation's optimality, and to convey the most efficient algorithm to compute it.

Multiresolution ICA. We designed and implemented a hybrid multiresolution adaptive method (MrICA) for image encoding. We demonstrated that it combines the advantages of multiresolution methods (representational power, and computational efficiency) and of adaptive methods (statistical optimality), thus improving over both classes of representations. We illustrated the practical merits of MrICA (specifically, its coding efficiency) by direct comparison with the current image coding standard for images JPEG2000. The new method demonstrated that for a large range of encoding rates, the average quality of the reconstruction (both perceptual, and measured by SNR) is significantly better than that of JPEG2000. This strongly supports the idea of using adaptivity as a source of practical improvement for modern image coders.

Point Coding. Existing approaches for the adaptive sparse encoding of large signals are known to involve significant computational costs. A particularly useful approach in this respect is representing the signal via a set of adaptive variable-size shiftable kernels (much smaller in size than the signal itself). We studied the particularities of applying such an approach to images. The most important merit of our method (called Point Coding) is that it produces a very efficient adaptive code, by what can be considered a direct approach towards an approximately shift-invariant representation. This is especially desirable in modeling natural or artificial encoding systems necessarily robust to signal shifts, such as the visual sensory system. A significant contribution of our implementation, is that both the encoding and the learning steps are performed

using computationally efficient (*e.g.*, fast and superfast) algorithms, thus allowing a practical way to attain optimality.

Robust Coding. Finally, we provide a detailed mathematical study of Robust Coding - the problem of optimal linear coding with limited precision units. We show how to characterize optimal encoding solutions in the case of Gaussian channel noise and arbitrarily many encoding units, and derive efficient and stable algorithms for their computation. By conveniently expressing the limit of optimization as the closed-form bound, we formally explain the intuition that noisy encoding units can preserve signal information if sufficiently many are used - a case very relevant to modeling neural encoding systems such as the retina.

Future Research Directions

The research topics I have shortly presented here describe various situations when existing signal representations are improved by exploiting either the theoretical properties, or the statistical structure of the signals, or both. These ideas can be extended and refined, sometimes with far-reaching consequences.

Sparse ICA. Multiresolution ICA is limited in that the representation depends highly on the associated multiscale transform; moreover, the correspondence between the subband ICA bases and the full-scale optimal basis is not straightforward. A direct approach to deriving efficient adaptive representations for large images can exploit an observed by-product of ICA for images: namely, basis elements look like localized oriented edge features [14], which implies that the basis matrix is sparse. By restricting the optimization to matrices satisfying some (fixed) sparsity pattern, the search space reduces considerably. In addition to estimating fewer parameters, requiring less memory, and exploiting fast sparse-matrix algorithms, this method should likely be very efficient, comparable to unconstrained ICA in this respect. The idea of speculating the structure of optimal solutions suggests a general and simple recipe for designing adaptive representations for large signals.

Algebraic Signal Processing Theory. Rigorously classifying signal transforms and their algorithms is a primary goal of algebraic signal processing theory. The DFT, the sixteen trigonometric transforms, and many other linear transforms [98] proved to be particular cases of polynomial transforms fit by the general theory via algebraic matrix structure. An exciting direction is reconciling the two apparently divergent views – deterministic and stochastic – on signal processing. The first steps were made by establishing, for instance, conditions for the equivalence of Gauss-Markov random fields and algebraic signal models, which connects the concept of Karhunen-Loève transform to the Fourier transform. We plan to push these ideas further, by investigating what are the correspondents of other adaptive linear transforms – Robust Coding, ICA – in the algebraic theory. The immediate outcome would be our better understanding the structure of these transforms. On the long term, this could lead to discovering more efficient algorithms, which would make adaptive representations suitable for general-purpose hardware implementation.

Online Matching Pursuit. A direction of particular interest to me is designing efficient algorithms for nonlinear, greedy approximations of time-varying signals (sound, video). Matching Pursuit has proven to be a very practical choice for spike extraction, in the case of shifted-kernel dictionaries [107]. To apply it properly though, the entire signal must be available ahead of time; however, in some situations (*e.g.*, recording speech, music, or video) the encoding must be performed in real-time and in and online manner. The main bottleneck in Matching Pursuit is not as much updating the residual, as it is selecting the next atom in the representation. In the case of a small set of kernels, the first issue is easily solvable. As for the second, a solution is to only focus on a small, "sliding window" area of the signal and thus process the signal in a quasi-online fashion. The ability of such an approach in producing a sparse set of atoms remains open for now, yet by inspecting partial results it seems comparable to that its offline counterpart. This approach to online greedy approximation has a potentially high impact not only within signal processing, but also in other research fields. For example, Spike Coding [107] has been shown to produce a representation that is relevant to modeling the auditory nerve. However, the spikes in the representation are computed in an offline fashion, unlike the real neural spikes. We expect that an online greedy approach will fix this shortcoming, and thus improve the biological relevance of Spike Coding. This promising idea could be of great potential help in manufacturing better prosthetic hearing devices.

Bibliography

[1] ***. Cross Match Verifier 300 sample fingerprint database. http://neurotechnology.com/download/CrossMatch_Sample_DB.zip. 4.5

[2] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, January 2008. 2.2, 5.3

[3] M. Aharon and M. Elad. Sparse and redundant modeling of image content using an image-signature-dictionary. *SIAM J. Imaging Sci.*, 1(3):228–247, 2008. 4.1

[4] M. Andrle and L. Rebollo-Neira. Biorthogonal techniques for optimal signal representation. http://www.ncrg.aston.ac.uk/Projects/BiOrthog/index.html. 2.3

[5] M. Andrle and L. Rebollo-Neira. A swapping-based refinement of orthogonal matching pursuit strategies. *Signal Processing, sp. iss. on Sparse Approximations in Signal and Image Processing*, 86:480–495, 2006. 2.3

[6] M. Andrle, L. Rebollo-Neira, and E. Sagianos. Backward-Optimized Orthogonal Matching Pursuit Approach. *IEEE Signal Processing Letters*, 11:705–708, 2004. 2.3

[7] T. Arias, A. Edelman, and S. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20:303–353, 1998. 5.3

[8] J. J. Atick and A. N. Redlich. What does the retina know about natural scenes? *Neural Computation*, 4:196–210, 1992. 5.1

[9] D. C. Balcan and J. Rosca. Independent Component Analysis for Speech Enhancement with Missing TF Content. In *Proc. Intl. Conf. on ICA*, pages 552–560, Charleston, SC, USA, 2006. 1.2

[10] D. C. Balcan, A. Sandryhaila, J. Gross, and M. Püschel. Alternatives to the Discrete Fourier Transform. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, pages 3537–3540, 2008. Las Vegas, NV. 1.2

[11] D.C. Balcan and M.S. Lewicki. Adaptive coding of images via Multiresolution ICA. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, pages 1021–1024, 2009. Taipei, Taiwan. 1

[12] D.C. Balcan and M.S. Lewicki. Point Coding: Sparse Image Representation with Adaptive Shiftable-Kernel Dictionaries. In *SPARS Workshop*, Saint Malo, France, 2009. 2

[13] A. J. Bell and T. J. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Comput.*, 7:1129–1159, 1995. 2.2

[14] A. J. Bell and T. J. Sejnowski. The independent components of natural scenes are edge filters. *Vision Res.*, 37:3327–3338, 1997. 3.1, 3.4, 6

[15] R. B. Bendel and M. R. Mickey. Population correlation matrices for sampling experiments. *Commun. Statist. Simul. Comp.*, B7:163–182, 1978. 5.3

[16] A. Borst and F. E. Theunissen. Information theory and neural coding. *Nature Neuroscience*, 2:947–957, 1999. 2.2, 5.1

[17] C. Brislawn. Classification of nonexpansive symmetric extension transforms for multirate filter banks. *Appl. Comput. Harm. Anal.*, 3:337–357, 1996. 3.1

[18] C. Brislawn, J. Bradley, R. Onyshczak, and T. Hopper. The FBI compression standard for digitized fingerprint images. In *Proc. SPIE*, pages 344–355, 1996. 3.1

[19] R. W. Buccigrossi and E. P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Trans. Image Proc.*, 8(12):1688–1701, 1999. 3.1, 4.1

[20] E. Candès and D. Donoho. Ridgelets: a key to higher-dimensional intermittency? *Phil. Trans. R. Soc. Lond. A.*, 357:2495–2509, 1999. 4.1

[21] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying. Fast discrete curvelet transforms. *Multiscale Model. Simul.*, 5:861–899, 2005. 4.1

[22] E. J. Candès and D. L. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise $\mathcal{C}^2$ singularities. *Comm. Pure Appl. Math.*, 57(2):219–266, 2004. 3.1

[23] J.-F. Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Sig. Proc. Let.*, 4:109–111, 1997. 2.2, 2.2

[24] J.-F. Cardoso. High-Order Constrasts for Independent Component Analysis. *Neural Computation*, 11(1):157–192, 1999. 2.2

[25] J.-F. Cardoso and B. Laheld. Equivariant adaptive source separation. *IEEE Trans. Signal Proc.*, 44(12):3017–3030, 1996. 2.2

[26] P. G. Casazza, M. Fickus, J. Kovačević, M. Leon, and J. Tremain. *Harmonic Analysis and Applications*, chapter A Physical Interpretation for Finite Tight Frames, pages 51–76. Birkhäuser, Boston, MA, 2006. 5.1, 5.4

[27] N. N. Chan and K.-H. Li. Diagonal elements and eigenvalues of a real symmetric matrix. *J. Math. Anal. Appl.*, 91:562–566, 1983. 5.3

[28] S. S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM journal on Scientific Computing*, 20(1):33–61, 1998. 2.3, 2.3, 4.1

[29] H. Choi and S. Choi. A relative trust-region algorithm for independent component analysis. *Neurocomputing*, 70(7–9):1502–1510, 2007. 2.2, 3.4

[30] A. Cichocki and S. i. Amari. *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. John Wiley & Sons, 2002. 2.2, 5.3

[31] P. Comon. Independent component analysis, a new concept? *Signal Proc.*, 36:287–314, 1994. 2.2, 2.2

[32] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, second edition, 2001. 4.3

[33] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006. 5.1, 3

[34] I. Daubechies. *Ten Lectures on Wavelets*. Number 61 in CBMS/NSF Series in Applied Math. SIAM, 1992. 2, 3.1, 3.2, 5.1

[35] J. G. Daugman. Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Trans. Acoust. Speech Signal Proc.*, 36:1169–1179, 1988. 3.1

[36] J. G. Daugman. Entropy reduction and decorrelation in visual coding by oriented neural receptive-fields. *IEEE Trans. Biomed. Eng.*, 36:107–114, 1989. 3.1

[37] G. Davis. *Adaptive Nonlinear Approximations*. PhD thesis, Courant Institute, New York University, 1994. 2.3, 4.2

[38] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *J. of Constr. Approx.*, pages 57–98, 1997. 2.3, 4.2

[39] G. Davis, S. Mallat, and Z. Zhang. Adaptive time-frequency decompositions with Matching Pursuit. *Optical Engineering*, 33(7):2183–2191, July 1994. 2.3, 4.2

[40] I. S. Dhillon, R. W. Heath Jr., T. Strohmer, and J. Tropp. Designing structured tight frames via alternating projection. *IEEE Trans. Inform. Th.*, 51(1):188–209, Jan 2005. 5.1

[41] I. S. Dhillon, R. W. Heath Jr., M. Sustik, and J. Tropp. Generalized finite algorithms for constructing Hermitian matrices with prescribed diagonal and spectrum. *SIAM J. Matrix Anal. Appl.*, 27(1):61–71, June 2005. 5.1

[42] M. N. Do and M. Vetterli. Contourlets. In G.W. Welland, editor, *Beyond Wavelets*. Academic Press, 2003. 3.1, 4.1

[43] E. Doi, D. C. Balcan, and M. S. Lewicki. A theoretical analysis of robust coding over noisy overcomplete channels. In *Advances in Neural Information Processing Systems 18*. MIT Press, 2006. 3, 2.2, 5.1, 5.1

[44] E. Doi, D. C. Balcan, and M. S. Lewicki. Robust coding over noisy overcomplete channels. *IEEE Trans. Image Proc.*, 16(2):442–452, February 2007. 3, 2.2, 5.1, 5.1, 5.2

[45] E. Doi, T. Inui, T.-W. Lee, T. Wachtler, and T. J. Sejnowski. Spatiochromatic receptive field properties derived from information-theoretic analyses of cone mosaic responses to natural scenes. *Neural Comp.*, 15:397–417, 2003. `http://www.cnbc.cmu.edu/cplab/data_kyoto.html`. 4.5

[46] E. Doi and M. S. Lewicki. Sparse coding of natural images using an overcomplete set of limited capacity units. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 377–384, Cambridge, MA, 2005. MIT Press. 5.1

[47] E. Doi and M. S. Lewicki. A theory of retinal population coding. In *Advances in Neural Information Processing Systems 19*. MIT Press, 2007. 5.3

[48] D. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, 59:907–934, 2006. 2.3

[49] D. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59:797–829, 2006. 2.3

[50] D. Donoho, M. Vetterli, R. DeVore, and I. Daubechies. Data compression and harmonic analysis. *IEEE Trans. on Information Theory*, 44:2435–2476, Oct. 1998. 4.1

[51] E. H. Adelson E. P. Simoncelli, W. T. Freeman and D. J. Heeger. Shiftable multiscale transforms. *IEEE Trans. Inform. Th.*, 38:587–607, 1992. 2.4

[52] N. Farvardin and J. Modestino. Optimal quantizer performance for a class of non-Gaussian memoryless sources. *IEEE Trans. Inform. Th.*, 30(3):485–497, 1984. 3.2

[53] A. J. Ferreira and M. A. T. Figueiredo. Class-adapted image compression using Independent Component Analysis. In *Proc. IEEE Int. Conf. Image Proc.*, pages 625–628, 2003. 3.1

[54] D. J. Field. Relations between the statistics of natural images and the response profiles of cortical cells. *J. Opt. Soc. Am. A*, 4:2379–2394, 1987. 3.1

[55] Y. Freund, E. Ettinger, S. Cheamanunkul, and M. Jacobsen. The automatic cameraman. `http://seed.ucsd.edu/mediawiki/index.php/Cameraman_Description`. 2

[56] G. Golub and Ch. L. Van Loan. *Matrix Computations*. Johns Hopkins University Press, London, third edition, 1996. 2.1, 5.1, 5.2, 1

[57] V. K. Goyal, J. Kovačević, and J. A. Kelner. Quantized frame expansions with erasures. *Harmonic Analysis and Applications*, 10(3):203–233, May 2001. 5.1

[58] V. K. Goyal, M. Vetterli, and N. T. Thao. Quantized overcomplete expansions in $\mathbb{R}^N$: Analysis, synthesis and algorithms. *IEEE Transactions on Information Theory*, 44(1): 16–31, January 1998. 5.1

[59] R. Grosse, R. Raina, H. Kwong, and A. Ng. Shift-invariant sparse coding for audio classification. In *Proc. UAI*, 2007. 4.1

[60] G. Heinig. Generalized inverses of Hankel and Toeplitz mosaic matrices. *Lin. Alg. Appl.*, 216:43–59, 1995. 4.1, 4.4, 4.4.1

[61] G. Heinig and K. Rost. *Algebraic Methods for Toeplitz-like Matrices and Operators*. Birkhäuser, 1984. 4.1, 4.4

[62] R. Horn and Ch. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990. 5.2, 5.2, 5.5

[63] A. Hyvärinen. Fast and robust fixed-point algorithms for Independent Component Analysis. *IEEE Trans. Neural Networks*, 10(3):626–634, 1999. 2.2

[64] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001. 2.2, 2.2, 3.1

[65] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for Independent Component Analysis. *Neural Computation*, 9(7):1483–1492, 1997. 2.2

[66] S. i. Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10: 251–276, 1998. 2.2, 5.3

[67] Ph. Jost, P. Vandergheynst, S. Lesage, and R. Gribonval. MoTIF: an efficient algorithm for learning translation invariant dictionaries. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, 2006. 4.1

[68] C. Jutten and J. Herault. Blind separation of sources. 1. An adaptive algorithm based on neuromimetic architecture. *Signal Proc.*, 24:1–10, 1991. 2.2

[69] T. Kailath and J. Chun. Generalized Displacement Structure for Block-Toeplitz, Toeplitz-Block, and Toeplitz-Derived Matrices. *SIAM J. Matrix Anal. Appl.*, 15(1):114–128, 1994. 4.4

[70] T. Kailath and A. H. Sayed, editors. *Fast reliable algorithms for matrices with structure*. SIAM, 1999. 4.1, 4.4

[71] P. Kisilev, M. Zibulevsky, and Y. Y. Zeevi. A multiscale framework for blind separation of linearly mixed signals. *JMLR*, 4:1339–1363, Dec 2003. 3.1

[72] J. Kovačević and A. Chebira. Life beyond bases: The advent of frames (Part I). *IEEE Sig. Proc. Mag.*, 24(4):86–104, Jul. 2007. 5.1

[73] J. Kovačević and A. Chebira. Life beyond bases: The advent of frames (Part II). *IEEE Sig. Proc. Mag.*, To appear 2007. 5.1

[74] S. Krstulović and R. Gribonval. Matching pursuit toolkit. `http://mptk.gforge.inria.fr/`. 2.4, 4.3

[75] S. Krstulović and R. Gribonval. MPTK: Matching pursuit made tractable. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, 2006. 2.4, 4.3

[76] G. Labahn, B. Beckermann, and S. Cabay. Inversion of mosaic Hankel matrices via matrix polynomials. *Lin. Alg. Appl.*, 221:253–279, 1995. 4.1

[77] K.-H. Lee. *Optimal Linear Coding for a Multichannel System*. PhD thesis, University of New Mexico at Albuquerque, 1975. 5.3

[78] K. H. Lee and D. P. Petersen. Optimal linear coding for vector channels. *IEEE Trans. Comm.*, 24:1283–1290, Dec 1976. 5.3

[79] T.-W. Lee. *Independent Component Analysis: Theory and Applications*. Kluwer Academic publishers, 1998. 2.2

[80] M. S. Lewicki and B. A. Olshausen. A probabilistic framework for the adaptation and comparison of image codes. *J. Opt. Soc. Am. A*, 16(7):1587–1601, 1999. 3.1

[81] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000. 2.2

[82] R. Lippert. Stiefel Grassmann optimization (SG_MIN). `http://www-math.mit.edu/~lippert/sgmin.html`. 5.3

[83] R. K. Ward M. D. Adams. JASPER: A portable flexible open source software tool kit

for image coding/processing. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, 2004. `http://www.ece.uvic.ca/~mdadams/jasper/`. 3.4

[84] B. Mailhé, S. Lesage, R. Gribonval, and F. Bimbot. Shift-invariant dictionary learning for sparse representations: extending K-SVD. In *Proc. EUSIPCO*, 2008. 4.1

[85] S. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. and Mach. Int.*, 11(7):674–693, July 1989. 3.1

[86] S. Mallat. *A wavelet tour of signal processing*. Academic Press, 1998. 2.1, 3.1

[87] S. Mallat and Z. Zhang. Matching Pursuits with time-frequency dictionaries. *IEEE-Trans-SP*, 41(12):3397–3415, December 1993. 2.3, 2.4, 4.2, 4.3

[88] G. Monaci, Fr. Sommer, and P. Vandergheynst. Learning sparse generative models of audiovisual signals. In *EUSIPCO'08*, 2008. 4.1

[89] M. Narozny and M. Barret. ICA-based algorithms applied to image coding. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, pages I–1033 – I–1036, 2007. 3.1

[90] B.K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput*, 24(2): 227–234, 1995. 2.3, 4.2, 4.3

[91] B. Olshausen. Sparse codes and spikes. In R.P.N. Rao, B. Olshausen, and M.S. Lewicki, editors, *Probabilistic Models of the Brain: Perception and Neural Function*, pages 257–272. MIT Press, Cambridge, MA, 2002. 2.4, 4.1

[92] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Res.*, 37:3311–3325, 1997. 3.1, 3.1, 4.1

[93] B. A. Olshausen, P. Sallee, and M. S. Lewicki. Learning sparse images codes using a wavelet pyramid architecture. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000. 3.1, 4.4

[94] V.Y. Pan. *Structured Matrices and Polynomials: Unified Superfast Algorithms*. Birkhäuser, 2001. 4.1

[95] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Conf. Rec. Asilomar Conf. on Sig., Sys. & Computers*, volume I, pages 40–44, Nov 1993. 2.3

[96] L. Perrinet, M. Samuelides, and S. Thorpe. Sparse spike coding in an asynchronous feed-forward multi-layer neural network using matching pursuit. *Neurocomputing*, 57:125–134, 2004. 2.4

[97] M. Püschel and J. Kovačević. Real, tight frames with maximal robustness to erasures. In *Proc. IEEE Data Compression Conf.*, pages 63–72, Snowbird, Utah, March 2005. 5.1

[98] M. Püschel and J. M. F. Moura. The algebraic approach to the discrete cosine and sine transforms and their fast algorithms. *SIAM Journal of Computing*, 32(5):1280–1316, 2003. 6

[99] M. Püschel and J. M. F. Moura. Algebraic signal processing theory, 2006. (available at http://arxiv.org/abs/cs.IT/0612077, parts of this manuscript are published as [100]). 1.2, 6

[100] M. Püschel and J. M. F. Moura. Algebraic signal processing theory: Foundation and 1-D

time. *IEEE Trans. Signal Proc.*, 56:3572–3588, 2008. (part of [99]). 6

[101] L. Rebollo-Neira. Backward Adaptive Biorthogonalization. *Int. J. Math. Math. Sci.*, 2004: 1843–1853, 2004. 2.3

[102] L. Rebollo-Neira and D. Lowe. Optimised Orthogonal Matching Pursuit Approach. *IEEE Signal Processing Letters*, 9:137–140, 2002. 2.3

[103] J. Rosca, T. Gerkmann, and D. C. Balcan. Statistical inference of missing speech data in the ICA domain. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, pages V–617 – V–620, Toulouse, France, 2006. 1.2

[104] C. Rozell, D. Johnson, R. Baraniuk, and B. Olshausen. Sparse coding via thresholding and local competition in neural circuits. *Neural Computation*, 20(10):2526–2563, 2008. 2.4

[105] Ph. Sallee. *Statistical Methods for Image and Signal Processing*. PhD thesis, University of California at Davis, 2004. 2.4, 4.3

[106] E. C. Smith. *Efficient auditory coding*. PhD thesis, Carnegie Mellon University, 2006. 2.4, 4.1

[107] E. C. Smith and M. S. Lewicki. Efficient coding of time-relative structure using spikes. *Neural Computation*, 17(1):19–45, 2005. 2.4, 2.4, 4.1, 4.4, 6

[108] E. C. Smith and M. S. Lewicki. Efficient auditory coding. *Nature*, 439(7079), 2006. 2.4, 4.1

[109] D. Taubman and M. Marcellin, editors. *JPEG2000: Image Compression Fundamentals, Standards and Practice*. Springer, 2001. 2.1, 3.1, 3.2, 4.1

[110] L. N. Trefethen and D. Bau III. *Numerical Linear Algebra*. SIAM, 1997. 5.1

[111] J. Tropp. Greed is good: Algorithmic results for sparse approximations. *IEEE Trans. Inform. Th.*, 50(10):2231–2242, Oct 2004. 4.2

[112] J. Tropp. Just relax: Convex programming methods for identifying sparse signals. *IEEE Trans. Inform. Th.*, 52(3):1030–1051, March 2006. 2.3

[113] M. Van Barel, G. Heinig, and P. Kravanja. An algorithm based on orthogonal polynomial vectors for Toeplitz least squares problems. In *Numerical analysis and its applications (NAA 2000)*, pages 27–34, 2001. 4.4

[114] M. Van Barel, G. Heinig, and P. Kravanja. A superfast method for solving Toeplitz linear least squares problems. *Lin. Alg. Appl.*, 366:441–457, 2003. 4.1, 4.4

[115] J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Neural Comput.*, 7:1129–1159, 1998. 3.1

[116] J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. Lond. B*, 265:359–366, 1998. 3.4

[117] M. Vetterli and J. Kovačević. *Wavelets and Subband Coding*. Signal Processing. Prentice Hall, Englewood Cliffs, NJ, 1995. 2.1, 3.1

[118] L. von Ahn, M. Blum, and J. Langford. Telling humans and computers apart automatically. *Commun. ACM*, 47(2):56–60, 2004. 1.1

[119] V. Zarzoso, P. Comon, and M. Kallel. How Fast is FastICA? In *Proc. EUSIPCO*, Florence, Italy, 2006. 2.2, 5

[120] M. Zibulevsky, P. Kisilev, Y.Y. Zeevi, and B.A. Pearlmutter. Blind source separation via multinode sparse representation. In *Advances in Neural Information Processing Systems 13*. MIT Press, 2001. 3.1