Image Models for Wavelet Domain Statistics

by

Seyedeh-Zohreh Azimifar

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2005

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.
I understand that my thesis may be made electronically available to the public.

S. Zohreh Azimifar

Abstract

Statistical models for the joint statistics of image pixels are of central importance in many image processing applications. However the high dimensionality stemming from large problem size and the long-range spatial interactions make statistical image modeling particularly challenging. Commonly this modeling is simplified by a change of basis, mostly using a wavelet transform. Indeed, the wavelet transform has widely been used as an approximate whitener of statistical time series. It has, however, long been recognized that the wavelet coefficients are neither Gaussian, in terms of the marginal statistics, nor white, in terms of the joint statistics.

The question of wavelet joint models is complicated and admits for possibilities, with statistical structures within subbands, across orientations, and scales. Although a variety of joint models have been proposed and tested, few models appear to be directly based on empirical studies of wavelet coefficient cross-statistics. Rather, they are based on intuitive or heuristic notions of wavelet neighborhood structures. Without an examination of the underlying statistics, such heuristic approaches necessarily leave unanswered questions of neighborhood sufficiency and necessity.

This thesis presents an empirical study of joint wavelet statistics for textures and other imagery including dependencies across scale, space, and orientation. There is a growing realization that modeling wavelet coefficients as independent, or at best correlated only across scales, may be a poor assumption. While recent developments in wavelet-domain Hidden Markov Models (notably HMT-3S) account for within-scale dependencies, we find that wavelet spatial statistics are strongly orientation dependent, structures which are surprisingly not considered by state-of-the-art wavelet modeling techniques.

To demonstrate the effectiveness of the studied wavelet correlation models a novel nonlinear correlated empirical Bayesian shrinkage algorithm based on the wavelet joint statistics is proposed. In comparison with popular nonlinear shrinkage algorithms, it improves the denoising results.

Acknowledgements

As a humble being, I praise Almighty God, the Source of my spirituality, inspiration, and guidance that has enabled me to complete my work. This thesis is the result of several years of study, during which many people have walked by my side. At last, I have the pleasant task of expressing my gratitude to all of them.

My Ph.D. studies were supervised by “two great minds” at the University of Waterloo, Professor Ed Jernigan and Professor Paul Fieguth. Not only have they been enthusiastic and innovative teachers, but also inspiring and dedicated mentors. From one day to the next, they have supported my ambitious plans and ideas, had confidence in me when I doubted myself, and were generous when my life outside school reshaped my life in school. Although I have been anticipating the opportunity to thank them, I fear that I do not have the words nor the skills to sincerely express my gratitude. The intellectual influence of both Dr. Jernigan and Dr. Fieguth has had a profound and powerful influence not just on my research, but also on all aspects of my life.

I wish to thank Professor Jernigan for inviting me to work with his vision and image processing research group, for providing me with exciting opportunities and passionate encouragement throughout my Ph.D. program, and for introducing me to Professor Fieguth. His lectures were one of the two activities at the University of Waterloo that I attended eagerly and enjoyed immensely.

I am also grateful to Professor Fieguth, my thesis supervisor. I have been deeply impressed with his dedication to high-quality research and no less. He has been an active listener who also takes the time to give advice. He has taught me how to think philosophically, to ask meaningful questions, to understand difficult concepts, to approach and identify a research problem, and to precisely carry out a research plan to accomplish any goal. Our weekly meetings was the other activity that I was keen to attend. Without his enthusiasm, guidance, and compassion, I could not have completed my dissertation.

My appreciation is also extended to the following people at the university for the stimulating discussions we had: Professor Ed Vrscay, Dr Simon Alexander, Professor Fakhry Khellah, and Dr Fu Jin. What I have learned in the last five years will benefit me for the next fifty years.

I am very grateful to my husband Professor Reza Karimi, for encouraging and mentoring me. One of the best experiences of our partnership throughout this period was the birth of our son Ali, who with his brother, Amir, adds a joyful dimension to our mission in life.

I feel a deep gratitude to my parents who have nurtured my self-confidence, rendered the good things that really matter in my life, and taught me the value of a positive perspective. Their love and loyalty has been unconditional and borderless, as my interests have led me to a language and geography that are very different from what they know. They still provide a persistent voice of inspiration for my journey in this life, and continue to impart that each of their seven children is special.

At last but not the least, there are those who reminded me that a student should have a life outside school. Thank you, Zohreh Abdekhodaie, Mahdieh-Sadat Emrani, Mahtab Kamali, Maryam Malekzadeh, Atefeh Mashatan, Homeyra Pourmohammadali, Betty Pries, Bita Roushanaei, Bing Yue, and all active members of our religious gathering on Friday evenings.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

"As long as a branch of knowledge offers an abundance of problems, it is full of vitality" -David Hilbert [48]

The visual representation of a physical object or scene produced by an optical instrument is called an *image*, which is considered to be the most informative and comprehensive among all kinds of data representations perceived by our senses. A sequence of analysis, manipulation, storage, and display of digitized images by computer from sources such as photographs, drawings, and video is called *digital image processing*.

In the era of multimedia, powerful computers, and modern digital technology in general, digital image processing is at the forefront of information technology. It stands as the basis for a growing variety of applications including medical diagnosis, remote sensing, geophysical prospecting, space exploration, molecular biology, microscopy and machine vision. Classic areas of interest in these fields include the enhancement of such signals, their compression, analysis and synthesis, and many, many more applications.

There are two main streams in the research area of image processing; one attempts to fundamentally analyze and theoretically model different classes of images, another includes the broad spectrum of applications, which greatly benefits from the achievements of the first category. This thesis mainly focuses on the first area, *i.e.*, image modeling.

As digital images become more widely used, digital image analysis must find more tools to work on them. It is very crucial to develop mathematical tools, such as the *wavelet transform*, which help in improving the efficiency of the multiscale-based image processing algorithms. In this work, we are particularly interested in statistical image modeling conducted in the wavelet domain, *i.e.*, *a study of wavelet joint statistics*.

1.1 Statistical Image Processing

Statistical models, in particular, prior probability models for the underlying spatial pixels, are of central importance in many image processing applications. Since the 1950s [54], when television engineers studied the auto-covariance functions of images, a great deal of attention has been devoted to modeling image statistics.

The probabilistic models attempt to characterize the key properties of an image, based on which imaging problem can be described, formulated and resolved. For example, the goal of image restoration is to enhance and to improve the appearance of an image by estimating the original pixel values from the distorted observation. A prior model describing statistics of both the noise and the uncorrupted image plays an essential role in this application.

Due to high-dimensionality (long-range) in spatial interactions, however, modeling the statistics of images is a challenging task. The first step in reducing dimensionality is to make some simplifying assumptions regarding the pixel interrelationships:

- *locality:* given the statistics of its predefined neighbors, the probability density function (pdf) of a pixel is conditionally independent of the pixels beyond that neighborhood;

- *homogeneity:* statistics describing a neighborhood are invariant to changes in spatial positions, *i.e.*, are the same for all such neighborhoods, regardless of their spatial location;

- *Gaussianity:* marginal or joint distributions are assumed Gaussian.

The first assumption leads to a Markov model. Markov Random Field (MRF) models [41, 58, 81] can be simplified by adding the next two assumptions, though in many cases the local statistics are highly non-Gaussian.

The second step in reducing dimensionality, while still improving the capability of statistical image models, is to decompose the spatial-domain pixels into a set of multiscale-multichannel frequency subbands, such as the *wavelet domain* [28, 90]. This linear transformation is not only very effective in reducing long-range dependencies, but also it has a multiresolution structure which allows one to zoom in to the local signal to analyze its details, or zoom out to get a global view of the signal. Studies of the human visual system [21, 76] support this multiscale image analysis approach, since it has been discovered that the human visual cortex can be modeled as a set of independent channels, each with a particular orientation and spatial frequency tuning. That is why wavelet transforms are found useful and significantly effective to the field of multidimensional signal processing [64].

1.2 Multiresolution Image Analysis

The past decade has seen increasing research in the development of new mathematical tools for multiscale image representation, analysis and modeling. Multiscale image representations provide us with a natural framework to describe mathematical phenomena and signals at different levels of resolution. The resolution of a signal is a qualitative description associated with its frequency content, e.g., a low-frequency signal has coarser resolution. The signal is studied at a coarse resolution to get an overall picture, and at higher and higher resolutions to see increasingly fine details. Images are presented at multiple resolutions due to three primary considerations:

1. Such representations support highly efficient algorithms based on the divide and conquer principle.

2. The hierarchy of resolutions provides a smooth transition between local features and global features, an illustration of seeing the forest from the trees.

3. The multiresolution framework provides a model for certain types of early processing in natural human vision.

Motivated by such facts, a number of multiresolution algorithms have been proposed and used for a variety of applications [56, 63, 100].

Methods of multiscale stochastic modeling [29, 55, 100] have led to a variety of efficient signal and image processing algorithms. Furthermore, the powerful theory of wavelets [64] has also brought much attention to applications of multiscale analysis, because wavelet basis functions are well suited to analyze a nested sequence of resolutions. This remarkable marriage of multiscale framework and wavelet theory, the great accomplishment of pioneer works done by Meyer [68], Mallat [63] and Daubechies [28], plays a particularly effective role in many disciplines such as interpolation [99], estimation [26], compression [95], and denoising [78], which are simplified in the wavelet domain because of the energy compaction and localization properties of the wavelet transform.

The significance of the wavelet transform is that it compresses the energy of the signal in a relatively small number of big coefficients [90], which represent the signal's high frequency components, such as edges. This acts as the motivating property behind the wavelet *shrinkage* methods [2, 33, 40]. Wavelet shrinkage is a widely-used and effective non-parametric coefficient thresholding approach to solve many inverse problems, such as regression or denoising, where the wavelet coefficients are subject to non-linear or Bayesian rules [2, 15, 96]. The primary idea is to exploit the localization property of the wavelet transform to develop efficient adaptive algorithms in order to minimize loss of important image features while eliminating noise. To propose an optimally efficient shrinkage estimator, detailed attention needs to be paid to the statistics of wavelet coefficients.

1.3 Motivation and Principal Focus of this Research

The primary interest of this research was to obtain a clear understanding of the image pixel characteristics when projected into the wavelet domain. It was the desire to know how much efficiency could be achieved in principle, when images are processed in the wavelet domain. Some important questions originally motivated this project:

- How efficient are the wavelet-domain algorithms?
- How do the image features and pixels connectivity change in the wavelet domain?
- Have the statistics of image data projected into the wavelet domain been fully characterized, so far?
- How efficient are the current wavelet models in describing the wavelet joint statistics?

Discussions on fundamental issues regarding wavelet image modeling are the focal point of this study.

The principle motivation for this study is to construct statistical models that are rich enough to capture wavelet domain correlation structure. We emphasize that image processing is not the immediate objective, rather *image modeling*. It is not the intent of this work to model different phenomena at multiple resolutions. Indeed, the fusion of multiresolution analysis (MRA) and random fields ideas and the resulting drawbacks have been investigated in the past [44, 56, 60]. Lakshmanan and Derin [56] presented an illustration and treatment of Gaussian Markov random fields (GMRFs) at multiple scales. They developed a multiresolution framework for MRFs and provided consistent model descriptions for GMRFs at multiple resolutions. Our objective is, rather, to fit wavelet domain statistics into the MRA and random fields concepts, particularly the class of MRFs as a prior belief regarding the connectivity of wavelet coefficients across scales and frequency channels.

The main theme of this thesis is, then, to test several hypotheses for wavelet statistical models, assessing model variations from wavelet complete independence to full dependency. The presence of across-scale correlations motivated the development of a wavelet-based multiscale model. Furthermore, the presence of significant within-scale relationships led us to propose a hierarchy of MRFs, capable of capturing coefficient statistics within and across subbands and scales in a sparse structure.

We have seen that the study of independent models has been thorough, whereas the complementary study, the development of Gaussian joint models, is much less thorough, and forms the focus of this thesis (shown in Figure 1.1). The goal, of course, is the merging of these two fields: that is, the development of non-Gaussian joint models with non-trivial neighborhood[1]. However for the purpose of starting the work, we limit ourselves

[1] The recent works such as Gaussian Scale Mixtures (GSM) and steerable pyramids [79] and Multivariate generalized Gaussian distributions [17] are indeed heuristic Gaussian joint models which do not carry out an empirical study of wavelet statistics.

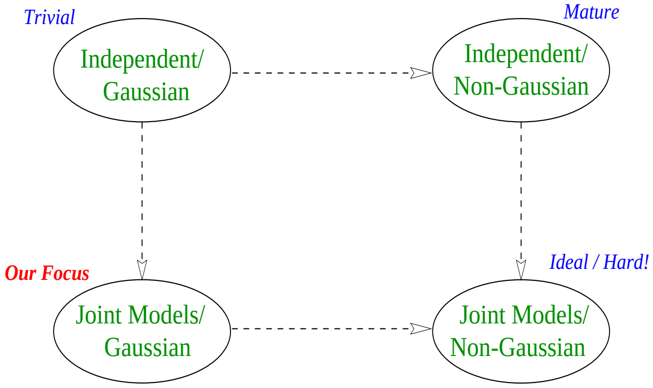


Figure 1.1: Focus of this research: The development of joint Gaussian models with inter-coefficient dependencies.

to simplifying marginal assumptions (Gaussianity) which we know to be incorrect (as we shall discuss later in this thesis), but which allow us to undertake a correspondingly more sophisticated study of joint models. The main novelty is the systematic approach we have taken to define a wavelet-based neighborhood system consisting of 1) inter-scale dependency, 2) within-scale clustering, and 3) across-orientation (geometrical constraints) activities. This probabilistic modeling is directly applied to the coefficient values.

To study the wavelet dependencies, perhaps, the most direct approach is to examine the joint histograms (joint histograms between only two coefficients are tested, due to the visualization limitations), conditional distributions, covariance, and correlation functions.

1.4 Thesis Outline

This dissertation includes four main categories which are shown in Table 1.1.

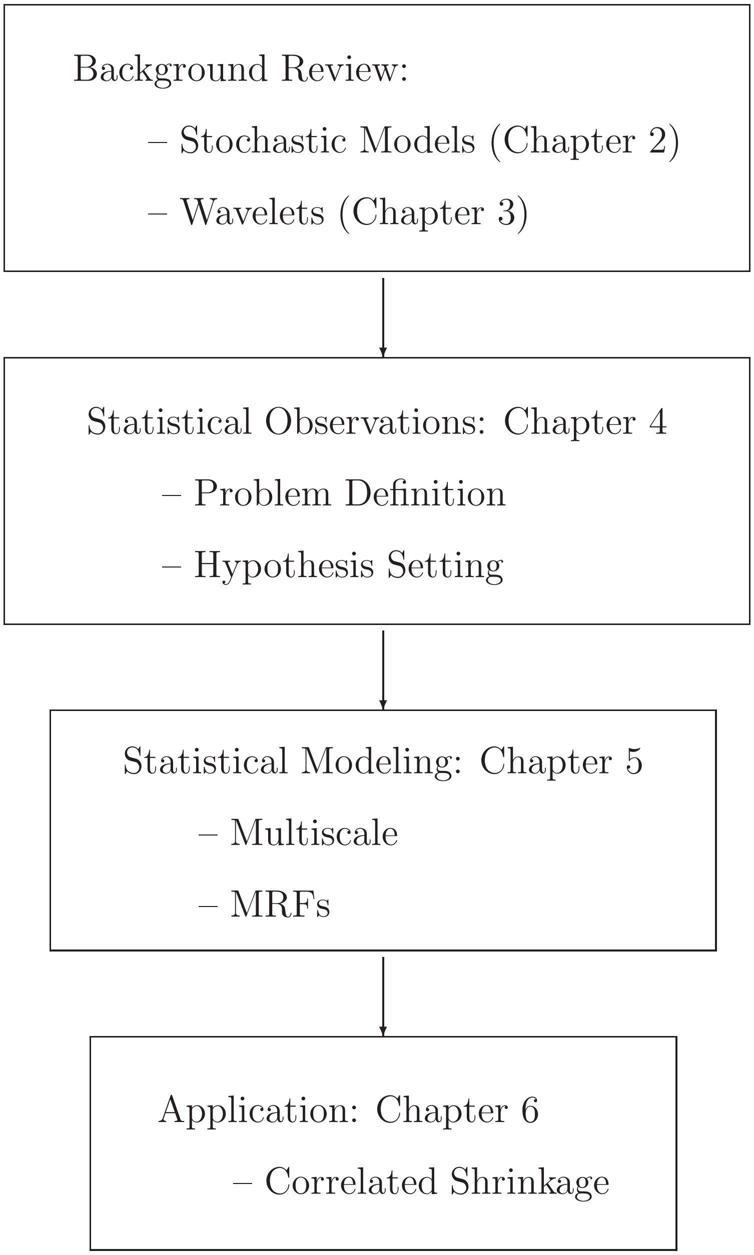


Table 1.1: Thesis Organization

A detail description of each individual chapter:

Chapter 2: Introduces the statistical modeling frameworks including multiscale and Markov random fields. It emphasizes models structured on a hierarchy which will be used to describe the statistics of image elements in the wavelet domain.

Chapter 3: The fundamental aspects of wavelet theory and principles of 2-D wavelet transforms and statistics are reviewed in this chapter. The state-of-the-art models of the wavelet-domain statistics which provide us with insights into the subject are presented. The issues and limitations of the current models that inspired this study to propose more powerful models are presented and discussed.

Chapter 4: This chapter presents an empirical study of joint wavelet statistics for a large range of natural images and random fields. In this chapter we describe possible choices of wavelet statistical interactions by examining the wavelet domain covariance, joint-histograms, conditional distributions, correlation coefficients, and the significance of coefficient relationships. An efficient and fast strategy to demonstrate the wavelet correlation maps and the significance of those inter-relationships will be introduced. This simulation will help us to propose a hypothesis of wavelet joint statistics.

Chapter 5: In this chapter our hypothesis of wavelet domain dependencies is tested using multiscale and Markov random fields models. A detailed description of these two probabilistic approaches in approximating the interrelationships among wavelet coefficients is given. The essential goal is to obtain a well-structured and sparse model absorbing the most striking local correlations including wavelet spatial and interscale dependencies.

Chapter 6: The primary goal of this chapter is to demonstrate effectiveness of the wavelet correlation models in the applied world. A novel non-linear correlated empirical Bayesian shrinkage algorithm based on wavelet joint statistics is proposed and compared with the popular nonlinear shrinkage algorithms.

Chapter 7: Finally, conclusions with perspectives on this contribution are presented and future directions and improvements of this research are discussed.

Chapter 2

Models of Stochastic Processes

> *"It has long been recognized in the field of image processing that the design of processing operations should be based on a model for the ensemble of images to be processed. Unfortunately, it is difficult to formulate realistic models for real-world classes of images; but progress is being made on a number of fronts, including models based on Markov processes, random fields, multiresolution methods, among others."* – Azriel Rosenfeld [83]

This chapter focuses primarily on stochastic models acting as the baseline of an increasing number of signal and image processing algorithms. The multiscale and Markov random fields frameworks that are reviewed in this chapter were principally developed to devise efficient linear estimation techniques. To motivate these methodologies and to provide additional insight into the notion of model-based signal processing, an overview of linear least squares estimation is given. Then more structured models that lead to computationally efficient estimators are discussed. These probabilistic models will be employed in later chapters to characterize the statistics of the wavelet coefficients.

2.1 Linear Least Square Estimation

This section considers a finite-dimensional linear least square estimation problem. A more detailed discussion on this topic can be found in [38].

A linear estimation problem is that of predicting a vector $\underline{\hat{x}}$ (mostly assumed zero-mean) of unknowns $\underline{x}$ with a linear function of an observed vector $\underline{y}$:

$$\begin{aligned} \underline{\hat{x}} &= P_{xy}P_y^{-1}\underline{y} && (2.1)\\ \widetilde{P} &= E[(\underline{x}-\underline{\hat{x}})(\underline{x}-\underline{\hat{x}})^T] && (2.2)\\ &= P_x - P_{xy}P_y^{-1}P_{xy}^T && (2.3)\end{aligned}$$

where $E[.]$ is the expectation operator, P_x the prior covariance of the random vector $\underline{x}$, P_y covariance of the random vector $\underline{y}$, $P_{x,y} = E[\underline{xy}^T] - E[\underline{x}]E[\underline{y}]^T$ the cross-covariance matrix for $\underline{x}$ and $\underline{y}$, and $\widetilde{P}$ the estimation error covariance.

The linear estimator that minimizes the mean square estimation error is called the Linear Least Square (LLS) estimator. Clearly, the above LLS estimator and the error covariance $\widetilde{P}$ depend on the second-order joint statistics of the measurement and the original data.

It is of significance to notice that this estimation counts on a predefined prior on $\underline{x}$. Given the prior

$$\underline{x} \sim (\mu_x, P_x)$$

and the measurement

$$\underline{y} = H\underline{x} + \underline{\nu} \qquad\qquad \underline{\nu} \sim (0, R) \tag{2.4}$$

where $\underline{\nu}$ is assumed additive noise uncorrelated with $\underline{x}$, then the LLS estimator and error covariance have the form

$$\begin{aligned}\underline{\hat{x}} &= \mu_x + P_x H^T \left(H P_x H^T + R\right)^{-1} (\underline{y} - H\mu_x) \\ \widetilde{P}_{LLS} &= P_{x|y} = P_x - P_x H^T (H P_x H^T + R)^{-1} H P_x \end{aligned} \tag{2.5}$$

The computational complexity of this estimator is a cubic function of the data vector $\underline{y}$: $\mathcal{O}(|\underline{y}|^3)$, which grows dramatically and becomes prohibitively burdensome as the data size increases.

This computational complexity can be significantly improved if some additional structures present in the prior model or measurement model matrices are taken into consideration. For example, consider the case in which the measurement matrix H is assumed an identity matrix and the unknown data $\underline{x}$ is toroidally (boundaries are circular) stationary. The Fast Fourier Transform (FFT) algorithm can diagonalize these matrices resulting in computationally more efficient estimation algorithms [38].

Another alternative which leads to computational savings occurs when the underlying phenomenon obeys principles of Markov random fields (MRF) or multiscale (MS) structures. There exist various iterative procedures that can solve the MRF- or MS-based problems with sparse systems of equations [58, 100].

2.2 Random Fields

Herein, a statistical approach of modeling dependency phenomena amongst image pixels is introduced. This section deals with a description of fundamentals of Markov Random Fields from a theoretical and practical point of view. Bayesian estimation theory is presented and some standard notations such as random fields are defined. The discussions here play an important role in the establishment of wavelet hierarchical Markov and non-Markov random fields presented in Ch. 5.

2.2.1 Markov Random Fields

One of the main tasks in statistical image processing is to construct stochastic models for observed images, especially for textures. The pixel values $\{x_i, i = 0, 1, \cdots, n-1\}$ are represented as realizations of random variables and the probability measure representing the joint distribution of all pixel values in an image grid is called a *random field*. An image is often modeled as a sample of a random field process for which the correlation between pixels is proportional to their geometric separations. In real scenes, neighboring pixels usually have similar intensities. In a probabilistic framework, such regularities are well expressed mathematically by Markov random fields. The Markov process, here a two dimensional random process, is basically motivated by the idea that the probability density of a pixel within an image, when conditioned on a set of pixels in a small spatial neighborhood, is independent of the pixels beyond that neighborhood. This notion of decoupling and sparse representation highlights the Markovianity as a focal point in many probabilistic frameworks [56].

In the 1920's, mostly inspired by the Ising model [52], MRFs as a new type of stochastic process appeared in the theory of probability and rapidly became a broadly used tool in a variety of problems not only in statistical mechanics. Its use in image processing became popular with the famous paper of Geman and Geman [42] but its first use in this domain dates back in the early 1970's, when Hassner and Sklansky [46] introduced MRFs to image analysis. Since then MRFs have been used extensively as representations of visual phenomena. For more thorough expositions on MRFs principles and applications the reader is referred to [41, 58, 81].

The most natural way to define MRFs related to image models is to define them on a lattice. However, here MRFs are defined more generally on graphs. It will be useful in Ch. 5 where the wavelet hierarchical MRF models are studied. Let $\mathcal{G} = (\mathcal{V}; \mathcal{E})$ be a undirected graph where $\mathcal{V} = \{v_1, v_2, \cdots, v_n\}$ is a set of vertices (or sites) and $\mathcal{E}$ is the set of edges.

Definition 2.2.1 (Neighbors): Two sites t and r are neighbors if there is an edge $e \in \mathcal{E}$ connecting them. The set of points which are neighbors of a site t, *i.e.*, the neighborhood of t, is denoted by $\mathcal{N}_t$.

Definition 2.2.2 (Neighborhood system): $\mathcal{N} = \{\mathcal{N}_t, t \in \mathcal{V}\}$ is a collection of subsets of $\mathcal{V}$ for which

1. $t \notin \mathcal{N}_t$
2. $r \in \mathcal{N}_t \Leftrightarrow t \in \mathcal{N}_r$.

Each site of the graph, is assigned a label from an infinite set of labels Λ. Such an assignment is called a configuration ω having some probability $P(\omega)$. The set of all possible configurations on $\mathcal{V}$ is called Ω.

Definition 2.2.3 (Markov Random Field): X is a Markov Random Field (MRF) with respect to the neighborhood system $\mathcal{N}$ *iff*

1. $P(X = \omega) > 0$ for all $\omega \in \Omega$,
2. $P(X_t = \omega_t \mid X_r = \omega_r, r \neq t) = P(X_t = \omega_t \mid X_r = \omega_r, r \in \mathcal{N}_t)$ for all $t \in \mathcal{V}$ and $\omega \in \Omega$.

The notion of *cliques* will be very useful in a discussion about probability measures on Ω:

Definition 2.2.4 (Clique): A clique C is a subset of $\mathcal{V}$ for which every pair of sites are neighbors. Single pixels are also considered cliques. The set of all cliques on a grid is called $\mathcal{C}$.

5	4	3	4	5
4	2	1	2	4
3	1	▨	1	3
4	2	1	2	4
5	4	3	4	5

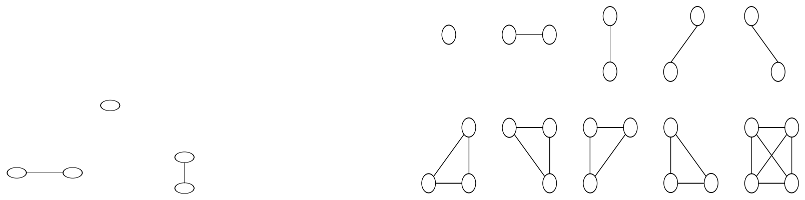
Figure 2.1: Order coding of neighborhood structure. The n-order neighborhood of the center pixel (shaded) contains the pixels with numbers less than or equal to n.

The structure of the neighborhood system determines the order of the MRF. For a first-order MRF the neighborhood of a pixel consists of its four nearest neighbors. The order coding of the neighborhood up to order five is shown in Figure 2.1. In a second-order MRF the neighborhood consists of the eight nearest neighbors. The clique structures are illustrated in Figure 2.2 for a first-order MRF and a second-order MRF.

It is seen that the above discussion is based on joint probability density $P(X)$ which is a computationally hard problem, *e.g.*, if the grid size is n with each pixel representing eight-bit gray values, then there are 256^n different configurations to estimate! To overcome this practical issue the following two alternatives are commonly considered:

1. Gauss-Markov Random Fields (GMRFs): In this case the random field X is Gaussian, a simplifying assumption which characterizes the field in term of expectation instead of the probability function [38, 58]. Since GMRFs are of interest in this work, the following section describes detail characteristics of these random fields and their use in model inference and Bayesian estimation.

2. Gibbs Random Fields (GRFs): Originally used in statistical physics to study the



(a) first-order

(b) second-order

Figure 2.2: Cliques for first- and second-order neighborhood structures.

characteristics of particles interaction, GRFs were introduced to the image processing by Besag [11] and became popular in Bayesian image modeling and inference. A Gibbs distribution is a probability measure π on Ω with the following representation:

$$\pi(\omega) = \frac{1}{Z}\exp(-U(\omega)) \tag{2.6}$$

where Z is the normalizing constant or partition function

$$Z = \sum_{\omega}\exp(-U(\omega))$$

and $U(\omega)$ is the energy function

$$U(\omega) = \sum_{c\in\mathcal{C}} V(\{\omega_i; i \in c\}), \tag{2.7}$$

with c denoting a clique, $\mathcal{C}$ the set of all possible cliques, and V the clique potential. The next famous theorem establishes the equivalence between Gibbs measures and MRFs [11]:

> **Theorem 2.2.1 (Hammersley-Clifford)**: X is an MRF with respect to the neighborhood system $\mathcal{N}$ if and only if X is a Gibbs distribution with respect to $\mathcal{C}$, where $\mathcal{C}$ is the set of cliques based on the neighbourhood system $\mathcal{N}$.

The main benefit of this equivalence is that it provides a simple way to specify MRFs, namely specifying potentials instead of local parameters, which is usually very difficult. Details about GRFs can be found in [58].

2.2.2 Gaussian Markov Random Fields & Inference

The GMRF model is frequently used to describe continuous phenomena. The conditional density is given by the expression

$$P(x_t|x_r, r \in \mathcal{N}_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left[x_t - \sum_{r\in\mathcal{N}_t} g_{t,r}x_r\right]^2\right\} \tag{2.8}$$

where σ is variance of the zero-mean process x_t and $\{g_{t,r}\}$ are the model parameters that will be defined shortly.

This model is also known as the conditional autoregressive (AR) model. More detail can be found in [58]. The very significant advantage of GMRF is the equivalence between decorrelatedness and independence in the case of Gaussian: decorrelation implies independence. Thus, given the GMRF X with the neighborhood system $\mathcal{N}$, the Bayesian estimate

$$E[x_s \mid x_r, r \neq s] = E[x_s \mid x_r, r \in \mathcal{N}_s] \tag{2.9}$$

becomes a linear expectation.

In the case of Gaussianity the random field X can be an AR process with respect to the neighborhood $\mathcal{N}$

$$x_s = \sum_{r\in\mathcal{N}_s} g_{s,r}x_r + \xi_s \qquad\qquad E[x_r\xi_s] = 0 \quad \forall s \neq r \tag{2.10}$$

where $\{g_{s,r}\}$ are model parameters and ξ_s is estimation error with zero mean and variance σ_s^2 for site s.

In this particular case of Gaussianity, (2.9) and (2.10) are identical, implying an equivalence between the decoupling property of (2.9) and the linear estimate in (2.10).

As stated above, $\{g_{s,r}\}$ are the model parameters, which are easily calculated by the LLSE of (2.9). Herein, the impact of $\{g_{s,r}\}$ on the statistical structure of the random field X and driven noise ξ is given. Eq. (2.10) implies that

$$\sum_{r\in\mathcal{N}_s} \bar{g}_{s,r} x_r = \xi_s \qquad\qquad \bar{g}_{s,r} = \begin{cases} -g_{s,r} & r \in \mathcal{N}_s \\ 1 & r = s \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

On the other side, for every two sites $s \neq r$

$$\begin{aligned} E[\xi_s \xi_r] &= E[\xi_s \sum_{t\in\mathcal{N}_r} \bar{g}_{r,t} x_t] \\ &= \bar{g}_{r,s} E[x_s \xi_s] \\ &= \bar{g}_{r,s} E[\xi_s \xi_s] \\ &= \bar{g}_{r,s} \sigma_s^2 \\ &\equiv \bar{g}_{s,r} \sigma_r^2 \end{aligned} \tag{2.12}$$

It is observed that the noise $\underline{\xi}$ is *not* white, with the correlation structure given by $\bar{g}$. To explore the locality and parsimony of the correlation structure of the GMRF X, (2.11) and (2.12) are re-written in vector form

$$\begin{aligned} cov(\underline{\xi}) &= \bar{G}\Sigma \\ \bar{G}\underline{x} &= \underline{\xi} \\ \underline{x} &= \bar{G}^{-1}\underline{\xi} \\ P_x = cov(\underline{x}) = \bar{G}^{-1}\bar{G}\Sigma\bar{G}^{-1} &= \Sigma\bar{G}^{-1} \end{aligned} \tag{2.13}$$

where the matrix $\bar{G}$ contains all the model parameters estimated for the field and Σ is the diagonal matrix of the noise variance σ_s^2.

Since the sparsity of matrix $\bar{G}$ is of interest here, the inverse covariance of the random field, *i.e.*, $cov^{-1}(\underline{x}) = \Sigma^{-1}\bar{G}$ is considered. While the covariance itself represents the correlation between the random field elements, the covariance inverse shows the estimated model parameters. This is a significant observation which characterizes the MRF prior $P^{-1} = \Sigma^{-1}\bar{G}$ used in the Bayesian estimation process, described below.

Assuming the measurement

$$\underline{y} = H\underline{x} + \underline{\nu} \qquad \underline{\nu} \sim \mathcal{N}(\underline{0}, R), \tag{2.14}$$

and the zero-mean prior model

$$\underline{x} \sim \mathcal{N}(0, P_x = \Sigma\bar{G}^{-1}) \tag{2.15}$$

then the linear estimate for the MRF $\underline{x}$ is

$$\begin{aligned} \hat{\underline{x}} &= (H^T R^{-1} H + P_x^{-1})^{-1} H^T R^{-1} \underline{y} \\ &= (H^T R^{-1} H + \Sigma^{-1}\bar{G})^{-1} H^T R^{-1} \underline{y} \end{aligned} \tag{2.16}$$

which is another form of the expression derived in (2.5) but with a new prior describing statistics of $\underline{x}$.[1] For further information on various GMRFs the reader is referred to [58] and citations therein.

2.3 Multiscale

In the study of the wavelet joint statistics, there are several reasons to adopt a multiscale (MS) modeling framework [100]:

[1] The zero-mean assumption is made for simplicity and is easily relaxed by adding a deterministic term to (2.16).

- Many signals, especially images, have a sparse multiscale representation and consequently useful Bayesian priors are easily specified in the multiscale analysis domain.
- The multiresolution property of the wavelet transform implies coarse-to-fine dependencies among the coefficients, which is a significant characteristic of the MS framework.
- The associated coarse-scale estimators are capable of providing information to improve fine-scale estimators.

This section defines the general class of MS models organized on branches of a tree. The multiscale (MS) algorithm is introduced followed by a brief discussion on the MS-based estimator.

2.3.1 The Multiscale Algorithm

The class of multiscale (MS) random processes introduced in [20] is indexed by the nodes that make different tree layers as different scales. The coarsest scale is called the root node, while the finest scale is indexed by the set of leaf nodes. For example, the multiscale process defined on the binary tree shown by Figure 2.3 consists of a set of random processes $z(w)$ for each node on the tree. The scale of node w shows the distance between node w and the leaf nodes (finest resolution) of the tree, *e.g.*, the root node's scale is denoted by J showing the tree's maximum depth. Denote the parent of any node w to be p_w and the children of w to be $\{c_{1w}, c_{2w}\}$ as illustrated in Figure 2.3. The class of multiscale processes generally satisfies the following autoregression (AR) across scales

$$z(w) = A_w z(p_w) + B_w \nu_w \qquad \nu_w \sim \mathcal{N}(0, I) \tag{2.17}$$

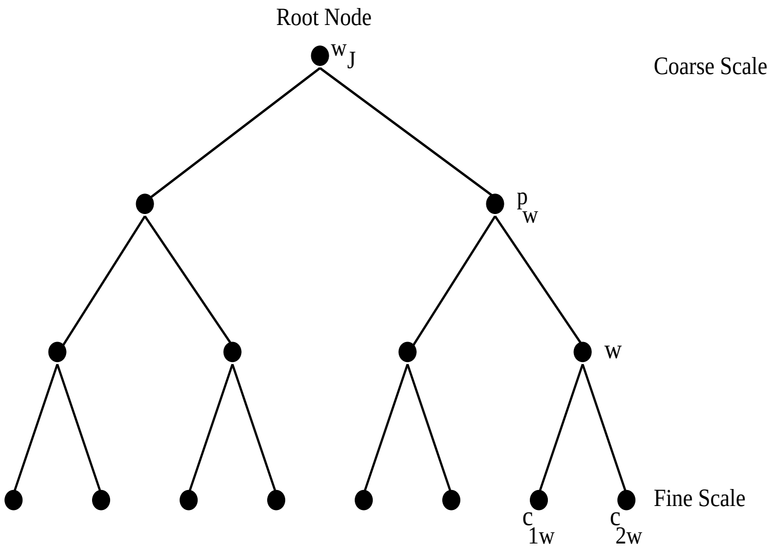


Figure 2.3: A binary tree of random processes at multiple resolutions. For each node w, p_w shows its parent at the coarser scale and $\{c_{iw}; i = 1, 2\}$ denote its children at the finer scale.

where $z(w)$ is the process value at node w and $z(p_w)$ is the process value at node p_w. Eq. (2.17) defines a stochastic dynamic from coarse to fine scale, with ν_w as its white process noise. This AR process is initialized at the root node w_J by

$$z(w_J) \sim \mathcal{N}(0, P_J). \tag{2.18}$$

The whiteness of the process noise implies that the multiscale model can be completely characterized by P_J (the root node covariance) and the AR parameters A_w and B_w. The white assumption for the process noise adds Markovian properties to the AR process driven by white noise [100]. Let q_w denote the number of children of node w, meaning that the node w partitions the tree into $q_w + 1$ subtrees. According to the Markov property of multiscale tree processes, the q_w+1 subsets of states partitioned by node w are conditionally independent, given the state $z(w)$, *i.e.*,

$$p\left(z(c_{1w}), z(c_{2w})|z(w)\right) = p\left(z(c_{1w})|z(w)\right) p\left(z(c_{2w})|z(w)\right) \tag{2.19}$$

for all nodes $c_{1w} \neq c_{2w}$ belonging to the distinct descendent subtrees of the node w. More formal derivations are given in Ch. 5, where this modeling framework is used to describe the statistical characteristics of wavelet coefficients.

2.3.2 Multiscale Model Inference

The Markovian property of the MS-based processes leads to algorithms that can efficiently estimate the original data at every node w on the tree based upon measurement, which is a noise-corrupted observation at every individual node of the tree, *i.e.*,

$$y(w) = H_w z(w) + v(w), \qquad v(w) \sim \mathcal{N}(0, R) \tag{2.20}$$

where v is a white noise uncorrelated with all the node processes. Measurements at coarse-scale nodes are assumed equivalent to measurements of coarse-resolution or nonlocal functions of the finest-scale process [27]. The multiscale estimation algorithm provided in [20] includes two steps. The first sweep of the estimator is a recursion from fine to coarse scale, followed by a recursion from coarse to fine scale. The result is that the linear least-squared error (LLSE) estimate (2.5) of the state $\hat{z}(w)$ at every node in the tree is computed in $\mathcal{O}(d^3n)$ computations for a tree which has n nodes of maximum state dimension d. Thus the efficiency of the estimator depends primarily upon whether a tree model can be realized with manageable state dimension.

2.4 Chapter Summary

This chapter started with an emphasis on the need for describing any meaningful collection of information within the framework of stochastic calculus, among which two efficient stochastic models – MRFs and MS – were introduced and their advantages were discussed.

These two are the most effective and frequently used probabilistic models which cover a broad range of multidimensional signal analysis and estimation.

Before leaving this chapter, recall that the ultimate goal of this thesis is to use the probabilistic models to describe the existence and the structures of the joint statistics for image elements when projected into another domain called wavelet domain. Before making any use of these probabilistic models, the theory of wavelets and the principles of wavelet transforms are reviewed. The material covered by this chapter will be revisited in Ch. 5 where wavelet joint correlations will be investigated.

Chapter 3

Wavelets

> *"If you painted a picture with a sky, clouds, trees, and flowers, you would use different size brush depending on the size of the features. Wavelets are like those brushes."* – Ingrid Daubechies [28]

This scientific journey starts with learning the basics of wavelet theory. Wavelets are introduced from a historical perspective, as a theoretical concept and within the context of image processing.

This chapter begins by exploring what is known about wavelet transforms and the characteristics of wavelet coefficients. The primary purpose of the discussion here is to introduce notation and to provide a suitable background on the wavelet transform as a mathematical and microscopic tool for image representation and analysis. The wavelet domain's most effective estimation technique known as wavelet shrinkage is introduced and in particular some of the well-known shrinkage algorithms are reviewed. Then, some of the most effective probabilistic models describing the wavelet statistics and the associated issues will be addressed.

At the end, I define the motivating points and sketch the plan for this presented research work which essentially is to study the wavelet coefficients relationship and their efficacy in a statistical modeling framework in order to capture the key dependencies amongst the wavelet coefficients.

3.1 Wavelet Transforms

From a historical point of view, wavelet analysis is a new method, though its mathematical concepts date back to the work of Joseph Fourier in the nineteenth century. Fourier laid the foundations with his theory of frequency analysis, which proved to be enormously powerful and important [91]. However, the attention of researchers has gradually turned from frequency-based analysis to scale-based analysis since it started to become clear that an approach measuring average fluctuations at different scales might prove less sensitive to noise. The first recorded citation of what we now call a *"wavelet"* seems to be in 1909, in a thesis by Alfred Haar [45]. In the late nineteen-eighties, when Daubechies [28], Mallat [63] and Meyer [67] first explored and applied the ideas of wavelet transforms, there was a great amount of literature addressing the wavelet related signal processing techniques such as shrinkage and compression. Most of the early contributions utilized the wavelet transform as a black box without doing an in-depth study of the wavelet coefficients characteristics. Fortunately, there is an increasing interest in describing the wavelet statistics, which has had an influential impact on the application of model-based wavelet methods, such as Bayesian estimation. Having said that, the rest of this chapter, after introducing the mathematics of the wavelet transform and the principles of wavelet shrinkage, is devoted to summarize the literature achievements on the probabilistic models of the wavelet coefficients statistics, in particular addressing their associated problems which has motivated this research.

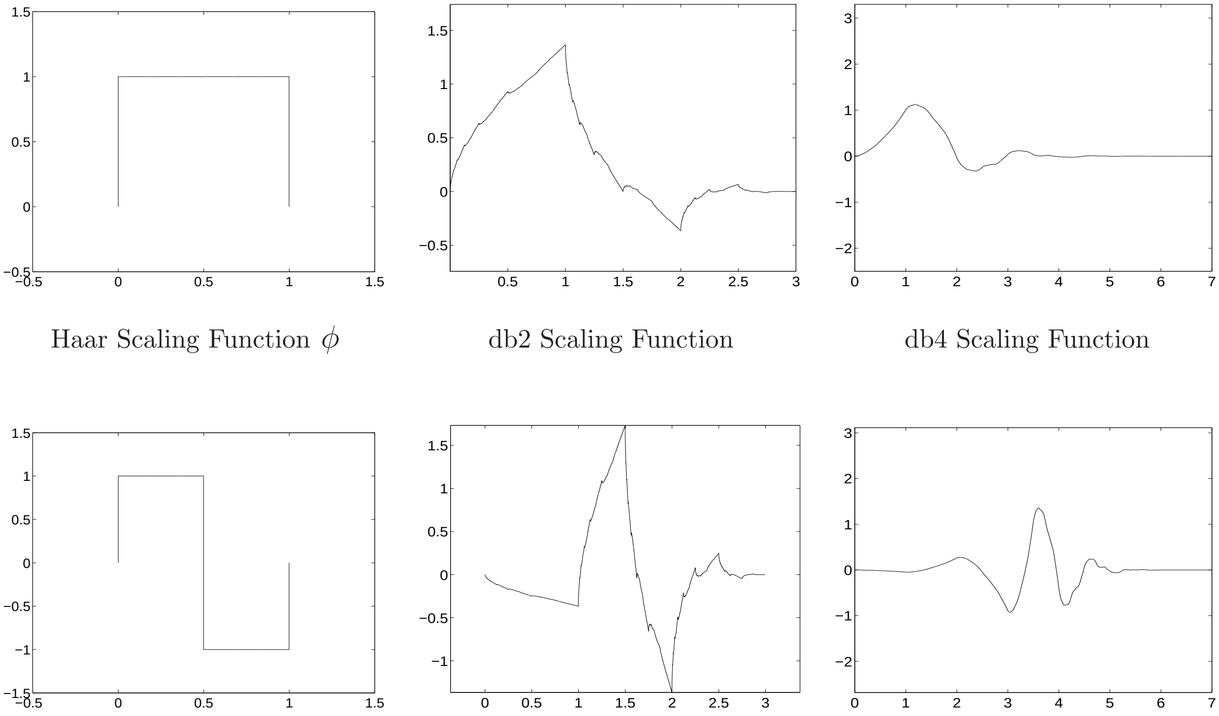
A Wavelet Transform (WT) of a function is a decomposition of that function into a weighted sum of a particular family of functions generating from a *mother wavelet* and forming a basis for $\mathcal{L}^2(\mathbb{R})$. Wavelets are functions that satisfy certain mathematical demands in multiresolution analysis. The name *wavelet* comes from the requirement that 1) the funtion magnitude should integrate to zero [72] and 2) the function has to be well localized [72]. Figure 3.1 shows some of the commonly used orthogonal wavelet functions and their corresponding scaling function $\phi(x)$.

It is important to notice the significant differences between Fourier analysis and wavelet analysis [91]. Fourier basis functions are localized in frequency but not in time. Small frequency changes in the Fourier domain will produce changes everywhere in the time domain. Wavelets are, however, local in both frequency (scale) and time. This localization is a major advantage of the WT. Another important feature is that a large class of functions can be represented by wavelets in a compact mode. For example, functions with discontinuities or with sharp transitions usually take substantially fewer wavelet basis functions than sine-cosine basis functions to obtain a comparable approximation. Furthermore, large data sets can be easily and quickly transformed by the WT. Indeed, the word *"fast"* for the fast Fourier transform (FFT) can be replaced by *"faster"* for wavelets. It is well known that the computational complexity of the FFT is $O(n \log_2 n)$, while that of the fast WT reduces to $O(n)$ [91].

The Continuous WT (CWT) is simply the correlation of an input function $f(x) \in \mathcal{L}^2(\mathbb{R})$ with a family (in particular, an orthogonal family) of wavelet functions $\psi_{a,b}(x) = 2^{-a/2}\psi(2^{-a}x - b)$, $a, b \in \mathbb{R}$ and $a \neq 0$ generated by scaling (dilating or compressing) and shifting a single mother wavelet $\psi(x) \in \mathcal{L}^2(\mathbb{R})$

$$(\mathcal{W}f)(a,b) = \langle f, \psi_{a,b} \rangle = \int f(x)|a|^{-\frac{1}{2}}\overline{\psi(\frac{x-b}{a})}dx \tag{3.1}$$

The result of these correlations are referred to as *wavelet coefficients.* When two signals



Haar Scaling Function ϕ

db2 Scaling Function

db4 Scaling Function

Haar Wavelet Function ψ

db2 Wavelet Function

db4 Wavelet Function

Figure 3.1: Examples of some common wavelet basis functions.

are correlated with each other, a measure of similarity is obtained between the two signals. Thus, when the WT is computed at a scale such that the wavelet is compressed, a measure of similarity between the signal and the high-frequency wavelet is obtained. Likewise, when the wavelet function is dilated, a measure of how similar the input signal is to the low-frequency wavelet is obtained. In other words, the WT can be interpreted as frequency decomposition with corresponding coefficients which provide information about the frequency contributions of the original signal, as well as their spatial position [49]. This kind of analysis is also referred to as *multiresolution analysis*. For a comprehensive review of the WT theory and applications, see [28, 64, 72, 95].

As a matter of fact, it is possible to compute the transform $\mathcal{W}f(a, b)$ where only discrete values for a and b are used. A common choice is to use the dyadic numbers, *i.e.*, to let $a = 2^{-j}$ and $b/a = k$ with $j, k \in \mathbb{Z}$. The transform which uses only the dyadic values of a and b is called the *Discrete Wavelet Transform* (DWT). This term is also used to denote the transform from the sequence of scaling function coefficients to its wavelet coefficients.

After this general introduction to the wavelet transform, here, I explain how this abstract theory of WT can be practically implemented. To bridge the gap between theory and practice, first, I start by defining the multiresolution analysis as a microscope that allows looking at a function at different scales. Then it will be seen how the scaling and wavelet functions are dedicated to the multiresolution analysis of functions and the construction of orthogonal wavelets.

Definition 3.1 A *multiresolution analysis* of $\mathcal{L}^2(\mathbb{R})$ is defined as a sequence of closed subspaces $V_j \subset \mathcal{L}^2(\mathbb{R}),\ j \in \mathbb{Z}$, with the following properties [63]:

1. $V_{j+1} \subset V_j \qquad \forall j \in \mathbb{Z}$,
2. $f \in V_0 \Leftrightarrow f(2^{-1}\cdot) \in V_1$,
3. $f \in V_0 \Rightarrow f(\cdot - k) \in V_0, \qquad \forall k \in \mathbb{Z}$,
4. $\lim_{j \to -\infty} V_j = closure(\bigcup_{j=-\infty}^{\infty} V_j) = \mathcal{L}^2(\mathbb{R})$ is dense in $\mathcal{L}^2(\mathbb{R})$. That is, as the resolution increases, the approximated signal converges to the original function in the $\mathcal{L}^2(\mathbb{R})$,
5. $\lim_{j \to +\infty} V_j = \bigcap_{j=-\infty}^{\infty} V_j = \{0\}$. This implies that at the limit where the resolution approaches zero, the approximation function contains less and less information until it converges to zero. These two properties can be written as:

$$\{0\} \cdots \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \subset \cdots \subset \mathcal{L}^2(\mathbb{R})$$

6. A unique *scaling function* $\phi(x) \in \mathcal{L}^2(\mathbb{R})$, with a non-vanishing integral $\int_{-\infty}^{\infty} \phi(x)dx = 1$ exists such that for each $j \in \mathbb{Z}$ the set of functions $\{\phi_{j,k}(x) = 2^{-j/2}\phi(2^{-j}x - k) \mid k \in \mathbb{Z}\}$ forms an orthogonal basis of V_j. For a sequence of coefficients $\{h_k\}$ the scaling function satisfies

$$\phi(x) = \sum_k h_k \sqrt{2}\phi(2x - k) \tag{3.2}$$

where the sequence $\{h_k\}$ constitutes the low-pass filter bank coefficients used in the transformation.

The spaces V_j are defined as the approximate spaces for a general function. This is done by defining appropriate projections of the function onto these spaces. Since the union of all the V_j is dense in $\mathcal{L}^2(\mathbb{R})$, any given function belonging to $\mathcal{L}^2$ can be approximated by such projections.

Having defined the approximate subspaces, W_j is defined to denote a subspace complementing V_j in V_{j-1}, *i.e.*, a space that satisfies

$$V_{j-1} = V_j \bigoplus W_j$$

where $\bigoplus$ is defined as a direct sum of the two vector spaces [63]. The subspace W_j contains the detail information needed to go from one approximation at resolution j to an approximation at resolution $j-1$, *i.e.*, $\bigoplus_j W_j = \mathcal{L}^2(\mathbb{R})$. A function ψ is a *wavelet function* if the collection of functions $\{\psi_{j,k}(x) = 2^{-j/2}\psi(2^{-j}x - k) \mid k \in \mathbb{Z}\}$ forms an orthogonal basis of W_j. The wavelet function ψ provides a way of characterizing the information that is lost when a function is approximated at lower resolutions. Similar to the scaling function, a sequence of coefficients $\{g_k\}$ exists so that the wavelet function satisfies

$$\psi(x) = \sum_k g_k \sqrt{2}\phi(2x - k) \tag{3.3}$$

where the sequence $\{g_k\}$ constitutes the high-pass filter bank coefficients used in the transformation.

For the particular case of orthogonal multiresolution analysis, the wavelet spaces W_j are defined as the orthogonal complement of V_j in V_{j-1}, *i.e.*, $W_j \perp V_j$. This is a class of orthogonal wavelets, which is a powerful representation in signal processing [63]. These lower resolution approximations can be interpreted as a removal of detail information between consecutive levels V_j and V_{j-1}. Detail information present in V_{j-1} that is missing from V_j is captured in W_j.

Since the coarsest approximation level is $\{0\}$, as the space of functions with no detail, any function $f \in V_j$ can be built up to any level $j = J$ simply by adding detail back into the approximation:

$$V_J = \bigoplus_{j=J+1}^{+\infty} W_j$$

Therefore, any $f(x) \in \mathcal{L}^2(\mathbb{R})$ has its discrete wavelet representation and $\forall J \in \mathbb{Z}$, as

$$f(x) = \sum_{k=-\infty}^{\infty} a_{J,k}\phi_{J,k}(x) \ + \sum_{j=J}^{+\infty}\sum_{k=-\infty}^{\infty} w_{j,k}\psi_{j,k}(x) \tag{3.4}$$

where the scaling coefficients $a_{J,k}$ and the wavelet coefficients $w_{j,k}$ are computed by inner products $a_{J,k} = \langle f, \phi_{J,k}\rangle$ and $w_{j,k} = \langle f, \psi_{j,k}\rangle$, respectively.

The above framework leads to wavelet designs which are particularly useful in approximating the class of signals with only a few non-zero coefficients. Two important design criteria that have the greatest effect on the number of significant wavelet coefficients are:

1. The number of vanishing moments of the wavelet function ψ:

   The wavelet function ψ is said to have p vanishing moments if

$$\int_{-\infty}^{\infty} x^n \psi(x)dx = 0 \quad for \ \ 0 \le n < p$$

It has been shown in [49] that the order of vanishing moments limits the order of smoothness of the signal f that can be characterized by the wavelet.

2. The compact support of ψ:

   For both the scaling and wavelet functions it is very important to have compact support. This is equivalent to the fact that the filter coefficients of $\{h_k\}$ and $\{g_k\}$ in the equations (3.2) and (3.3) are finite. The size of the support of ψ affects the amplitude of wavelet coefficients around discontinuities. If f has a discontinuity, such as an edge, at x_0, and x_0 is in the support of $\psi_{j,k}(x)$, then $|w_{j,k}|$ will be large. Minimizing the support of ψ will ensure that few wavelets $\psi_{j,k}$ contain the discontinuity, hence minimizing the number of high amplitude wavelet coefficients. This is an important consideration in image analysis, since much of the energy of an image is concentrated in its edges.

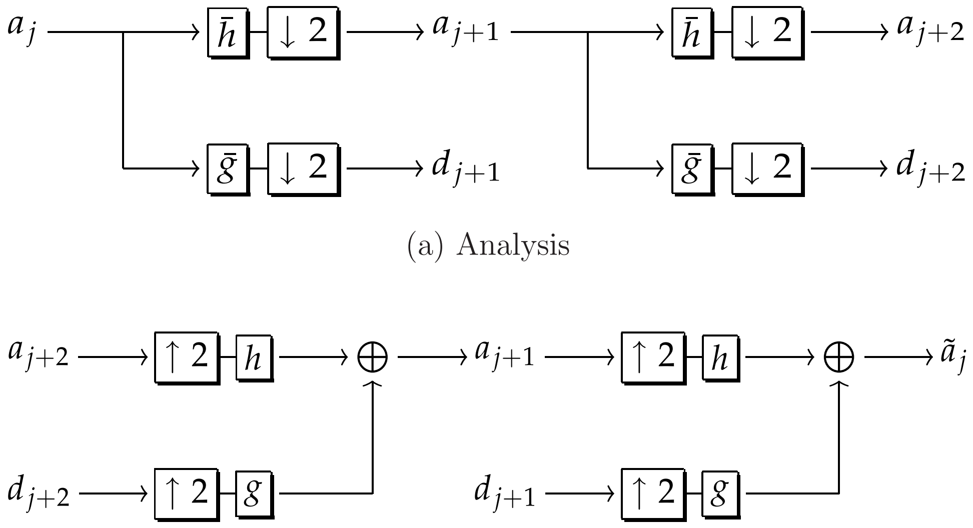
The oldest and the simplest known orthonormal wavelet is *Haar* wavelet [45], derived from the constant scaling function $\phi_{Haar}(x)$ (Figure 3.1(a))

$$\phi_{Haar}(x) = \begin{cases} 1 & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

with an associated mother wavelet

$$\psi_{Haar}(x) = \begin{cases} 1 & \text{if } x \in [0,\frac{1}{2}) \\ -1 & \text{if } x \in [\frac{1}{2},1] \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

This is one of the few wavelets that can be written in closed form, and it has the shortest support among all orthogonal wavelets. However, the main drawback of the Haar wavelet is that it is not well adapted to approximating functions with discontinuities because of its single vanishing moment [63].



(a) Analysis

(b) Synthesis

Figure 3.2: Filter bank implementation of the Mallat algorithm [63] for 1-D signals.

Probably the most frequently used orthogonal wavelets are the *Daubechies* wavelets [49] (Figure 3.1(b-c)). They are a family of orthogonal wavelets indexed by $N \in \mathbb{N}$, where N is the number of vanishing moments. They are supported on an interval of length $2N - 1$, *i.e.*, the Daubechies wavelets are compactly supported wavelets with a maximum number of vanishing moments for a fixed support width:

$$\int_{-\infty}^{\infty} x^n \psi(x) dx = 0 \qquad 0 \leq n < N \tag{3.7}$$

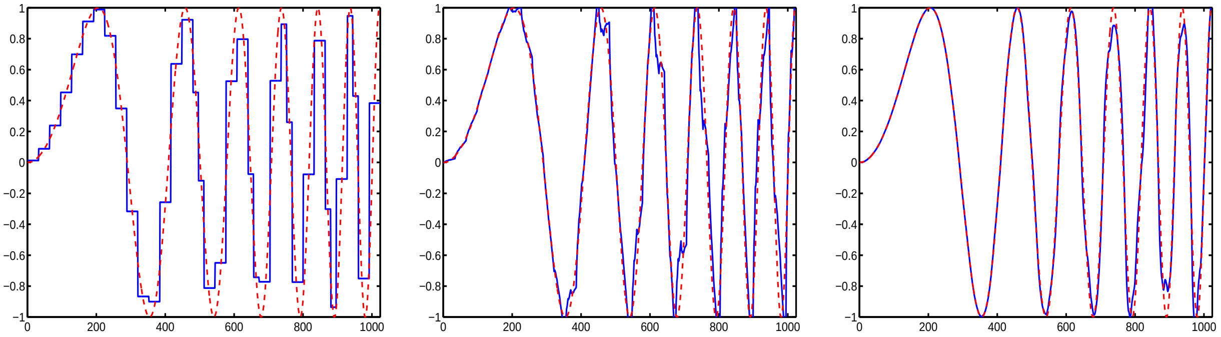
An important issue after selecting a compactly supported wavelet is the implementation of the orthogonal WT of discrete signals. In practice, signal f has only a finite number of samples, which are filtered with a cascade of discrete low pass and high pass filters, as illustrated in Figure 3.2 [63]. In the multiresolution analysis the set of coefficients $\{h_k\}$ and $\{g_k\}$ denote the low pass and high pass filters, respectively. They are related as $g(n) = (-1)^n h(1-n)$ to form a *quadrature mirror pair* [63]. This derivation of quadrature mirror filters from the scaling and wavelet functions implies that one can compute the approximate coefficients $\{a_J\}$ by convolving the discrete signal f with the filter $\{h_k\}$ and retaining every other sample of the output. Consequently, the detail coefficients $\{w_j\},\ 1 \leq j < J$ are obtained by convolving f with the filter $\{g_k\}$ followed by down sampling. This pyramidal approach, known as the Mallat algorithm is shown in Figure 3.2, and continues recursively from the approximate coefficients into lower resolutions.

As is depicted in Figure 3.3, one of the primary and key characteristics of the WT is its sparse representation of a signal's energy content. In other words, only a handful of wavelet coefficients carry a significant energy content of many complicated signals (also referred as the energy compaction property). In the following sections, it is explained how this parsinomous representation of the WT plays a crucial role in many signal processing applications, such as wavelet shrinkage and compression.

Other significant characteristics of the WT are its simultaneous localization in both time and frequency domains (wavelets are localized in frequency as well as in space, *i.e.*, their rate of variation is restricted), and its multiresolution representation, *i.e.*, analysis of a nested sequence of fine to coarse resolution, is a primary tool of the WT, which allows efficient analysis of the small details and the big picture: "Seeing the forest and the trees! [61]"

3.2 Wavelets and Images

The previous section focused on the fundamentals of the DWT of functions of one variable, *i.e.*, the 1-D WT. There are also wavelets in higher dimensions such as in 2-D. In this section, a method to extend 1-D DWT to the two dimensional case is introduced and applications of wavelet methods to images will be discussed.



Haar Daubechies2 Daubechies4

Figure 3.3: Approximations to $\sin(x^2)$ on $[0, 2\pi]$ resulting from keeping only the largest (in absolute value) 25% of the wavelet coefficients.

As defined in Definition 3.1, suppose V_j is a multiresolution subspace of $\mathcal{L}^2(\mathbb{R})$ and consider the tensor product space

$$\mathbf{V}_j = V_j \otimes V_j$$

which forms a multiresolution in $\mathcal{L}^2(\mathbb{R}^2)$ [64]. Following the discussion given in § 3.1, the orthogonal component of $\mathbf{V}_j$ is assumed to be $\mathbf{W}_j$:

$$\mathbf{V}_{j-1} = \mathbf{V}_j \bigoplus \mathbf{W}_j$$

where

$$\mathbf{W}_j = (V_j \otimes W_j) \oplus (W_j \otimes V_j) \oplus (W_j \otimes W_j) = \mathbf{W}_j^h \oplus \mathbf{W}_j^v \oplus \mathbf{W}_j^d$$

where h, v, and d stand for horizontal, vertical, and diagonal components respectively. Thus, the complete decomposition is obtained by

$$\begin{aligned} \mathbf{V}_j &= \mathbf{V}_J \oplus \bigoplus_{k=0}^{J-j-1} \mathbf{W}_{J-k}, \qquad j < J \\ &= \mathbf{V}_J \oplus \bigoplus_{k=0}^{J-j-1} (\mathbf{W}_{J-k}^h \oplus \mathbf{W}_{J-k}^v \oplus \mathbf{W}_{J-k}^d) \end{aligned} \tag{3.8}$$

Accordingly, the two dimensional DWT of a fine scale image I_0 at finest scale $j = 0$ is a process in which low and high frequency components of I_0 are represented by separate sets of coefficients, namely the approximation $\underline{a}_J$ and the detail $\underline{w}_j, 1 \leq j \leq J$, with J denoting the coarsest resolution. Following the above discussion, define the linear operators H_j and G_j (in matrix form, with different sizes at different resolutions) as high- and low-pass filters respectively, then the scaling and wavelet coefficients are recursively computed as

$$\begin{aligned}
\underline{a}^1 &= H_0 H_0 I_0 \\
\underline{a}^{j+1} &= H_j H_j \underline{a}^j \\
\underline{w}_h^{j+1} &= G_j H_j \underline{a}^j \\
\underline{w}_v^{j+1} &= H_j G_j \underline{a}^j \\
\underline{w}_d^{j+1} &= G_j G_j \underline{a}^j
\end{aligned} \tag{3.9}$$

with $\underline{w}_h^j, \underline{w}_v^j$, and $\underline{w}_d^j$ denoting, respectively, the horizontal, vertical, and diagonal subbands at scale j. Each resulting frequency channel is decimated by suppression of three samples out of four. The three high frequency subbands are left and the process recursively continues with decomposition of the low frequency channel. The maximum decomposition level for a discrete image with size $n = N \times N$, would be $J = log_2 N$, with $n/4^j$ detail coefficients in every subband at scale j [64].

To simplify the notation, all the scaled versions of the linear operators H and G are grouped into one linear wavelet kernel $\mathcal{W}$ reforming 2-D wavelet decomposition (3.9) into

$$\mathcal{W} I_0 = \{\underline{w}^1, \underline{w}^2, \cdots, \underline{w}^J, \underline{a}^J\} \tag{3.10}$$

where $\underline{w}^j$ contains the three orientation subbands $\underline{w}_h^j, \underline{w}_v^j$, and $\underline{w}_d^j$ at scale j, and $\underline{a}^J$ represents the scaling coefficients at the coarsest scale J.

Figure 3.4 illustrates the simple recursive algorithm of the 2-D DWT. In practice, the rows are first passed through the high-pass and low-pass filters, followed by down sampling,

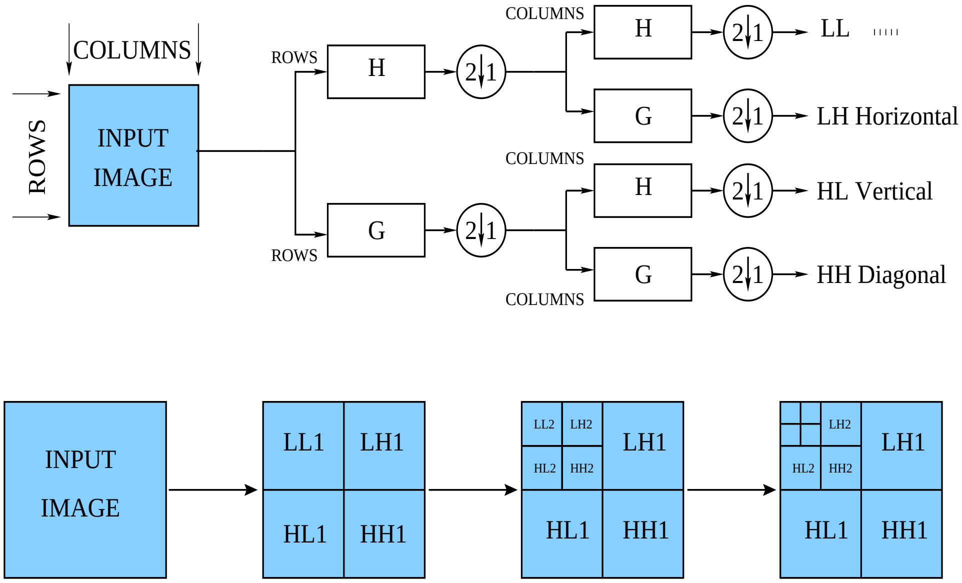


Figure 3.4: Implementation of the DWT algorithm for 2-D signals.

with the process then applied to the columns. The total number of the coefficients in this representation is identical to the number of pixels in the original image. This is due to the orthogonality of the WT representation.[1]

All of the primary properties of the 1-D WT discussed in § 3.1 are still applicable to the

[1] For simplicity, this statistical study is based on the class of orthogonal DWT. There exist some practical weaknesses associated with the 2-D DWT when used in image processing, such as shift invariance, aliasing artifacts when approximating coefficients, ambiguous edge directions in three subbands, etc. The past few years have seen many alternatives and extensions to the basic DWT which address some of these issues, such as shift-invariant WT [22, 23], biorthogonal WT [64], overcomplete steerable pyramids [87], complex WT [53], ridgelets [12] and curvelets [89] transforms, which combine ideas of multiscale analysis that results in improved edge orientations. Although it is important to be aware of the above limitations, the primary interest of this research is to study statistical properties of the wavelet coefficients. Those newly proposed transformations can also be modeled through extensions of the modeling framework in conjunction with some explicit geometrical constraints.

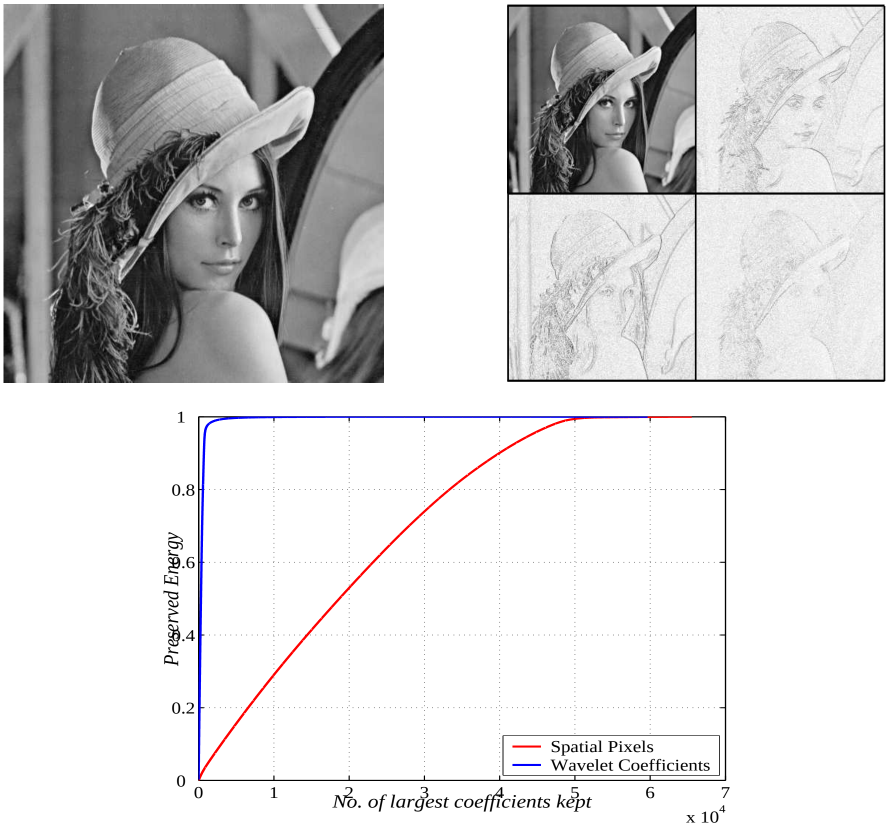


Figure 3.5: Illustration of the WT energy compaction. Top panels: the Lena image and its 2-D WT. Bottom panel: cumulative plot sketching the amount of the original energy preserved as a function of the number of the contributing pixels or coefficients. Only a very small fraction of the wavelet coefficients is needed to represent the signal energy.

two dimensional case. For instance, the bottom plot in Figure 3.5 illustrates the energy compaction of the 2-D WT. It is interesting that how a small number of wavelet coefficients present a large amount of information about the original image. [2]

In addition to the primary properties of the WT, one can examine the characteristics

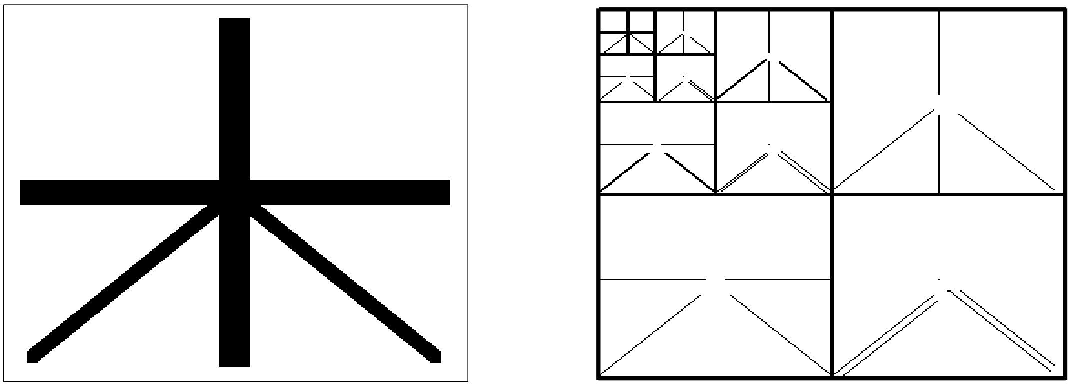
[2] As an aside note: it is observed that the wavelet coefficients have a heavy-tailed distribution. Detail discussion is given in § 3.4.

of wavelet coefficients, which basically represent the coefficients interrelationships and are known as the secondary properties [26]:

- ***clustering***: if the magnitude of a particular wavelet coefficient is small/large, then its adjacent coefficients magnitudes are very likely to also be small/large [75].
- ***persistence***: Small/large magnitude of wavelet coefficients tend to propagate across scales [66, 65].

See [26] and [82] for a detailed discussion on these properties.

These secondary properties, displayed by Figure 3.6, more or less depend on the characteristics of the original image and on the choice of wavelet. Romberg *et al.* [82] called these properties persistency and exponential decay across scales. In addition to the discussion in [82], the connectivity of wavelet coefficients is simultaneously spatial and scale dependent. Spatial dependency states that a large (small) valued coefficient has very likely large (small) valued siblings within the subband that it belongs to and siblings across other two subband counterparts. To define persistency across scales, one, however, needs to be cautious! If a coefficient value is large – meaning that an edge was included inside support of the basis function – it is highly possible that this large magnitude propagates through its children. However, small magnitude of a coefficient does not necessarily represent a smooth region of the original image. It may be a cancellation result of two or more edges covered by the support of the basis function. The second property, *i.e.*, the rapid decay of coefficients variance toward finer scales, is mostly a consequence result of the self-similar characteristic of real images [34, 39]. The decaying property of the wavelet coefficients has also been studied for both stationary and non-stationary stochastic processes. Tewfik and Kim [93] proved that the correlation between coefficients of the class of fractional Brownian motion, as a particular case of non-stationary signals, decreases exponentially fast across



(a) original image (b) wavelet transformed image

Figure 3.6: Orthogonal three-level discrete WT of of an artificial binary image.

scales and hyperbolically fast through time.

In summary, the DWT is very attractive because it tends to represent signals and images sparsely, with a few large scaling and detail coefficients. Indeed, the WT is intended to compress real-world signals, which is due to the vanishing moments of the wavelet functions. As mentioned earlier, this parsimonious property, along with the localized support of wavelets, implies that most signals or images would have a very sparse representation in the wavelet domain. The non-zero coefficients are, however, crucial in reconstructing important details such as edges.

In the section that follows the classical method of wavelet thresholding will be addressed. It will be seen that the key to effective wavelet domain filtering is to determine which wavelet coefficients do not have significant signal energy and can be discarded without critical signal loss.

3.3 Wavelet Shrinkage

The WT enables the representation of signals with a large degree of sparsity (Figure 3.5). This is a key property: for most signals a large fraction of signal energy is captured by very few wavelet coefficients. Motivated by and capitalizing on this property, *wavelet shrinkage* [32, 33] is a widely-used and effective non-parametric coefficient thresholding approach to solve many inverse problems, such as regression or denoising. In this technique, the wavelet coefficients are subject to non-linear or Bayesian rules [2, 15, 40, 96] that suppress the small coefficients, dominated by noise, and that retain high-magnitude ones. The main idea is to exploit the localization property of the WT to develop efficient adaptive algorithms in order to minimize the loss of important image features while eliminating noise. Since the parsimony of wavelet coefficients ensures that the signal of interest can be well described by a relatively small number of large coefficients, wavelet shrinkage tends to keep these large valued coefficients while discarding the negligible ones.

Many data operations can now be done by processing the corresponding wavelet coefficients. In fact, when details are small, they might be omitted without significantly affecting the original image. Thus, the intuition of *shrinking* the wavelet coefficients is a way of cleaning out unimportant detail coefficients considered to be noise. Wavelet shrinkage denoising should not be confused with smoothing. Whereas smoothing removes high frequencies and retains low frequency ones, denoising attempts to remove whatever noise is present and retain whatever signal is present regardless of the signal's frequency content. Wavelet shrinkage denoising does involve denoising in the wavelet domain and consists of three steps: a linear forward WT, a non-linear shrinkage denoising by comparing the wavelet coefficients with a predefined threshold value, and a linear inverse WT as is displayed by Figure 3.7. The non-linear shrinkage of the coefficients in the transform domain distinguishes this procedure from entirely linear denoising methods. Furthermore,

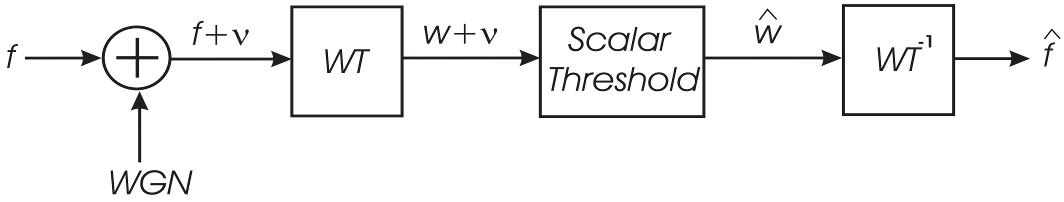


Figure 3.7: Block diagram for standard wavelet thresholding. The signal f plus additive white Gaussian noise ν is wavelet transformed, passed through nonlinear shrinkage, and inverse transformed to get the denoised signal $\hat{f}$.

the procedure exploits the fact that the WT maps white noise in the signal domain to white noise in the transform domain. Thus, although signal energy becomes more concentrated into fewer coefficients in the wavelet domain, noise energy does not. It is this important principle that enables the separation of signal from noise.

For a precise explanation of wavelet shrinkage, assume that the observed data is transformed into the wavelet domain and

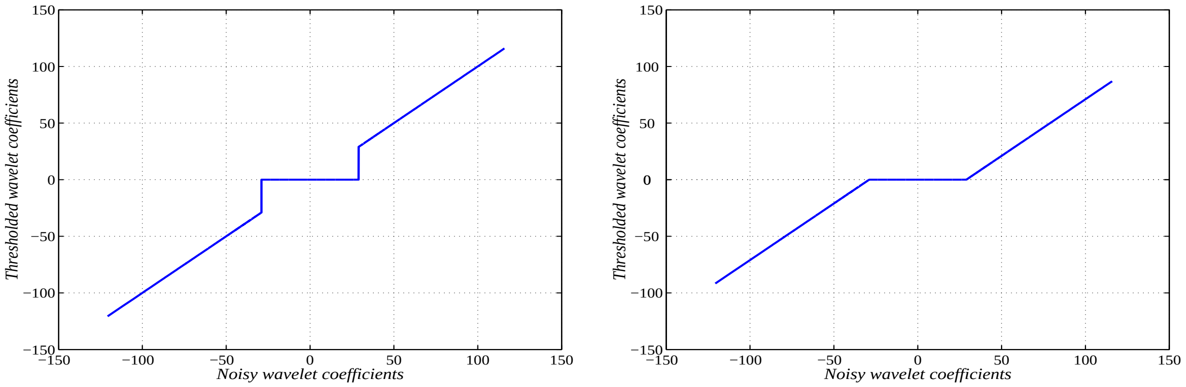
$$\underline{y} = \underline{w} + \underline{\nu} \tag{3.11}$$

where $\underline{w}$ is the original wavelet coefficients corrupted with the Gaussian additive white noise $\nu_i \sim \mathcal{N}(0, \sigma_\nu^2)$.

The process of thresholding wavelet coefficients is divided into two major steps. The first step is the choice of the threshold function $\mathcal{T}$. Two standard choices are *hard* and *soft* thresholding [33]:

$$\mathcal{T}^{hard}(w, \lambda) = \begin{cases} w & \text{if } |w| > \lambda \\ 0 & \text{otherwise} \end{cases} \tag{3.12}$$

$$\mathcal{T}^{soft}(w, \lambda) = \begin{cases} (w - sgn(w)\lambda) & \text{if } |w| > \lambda \\ 0 & \text{otherwise} \end{cases} \tag{3.13}$$



(a) Hard Thresholding

(b) Soft Thresholding

Figure 3.8: Hard and soft thresholding transfer functions. $\lambda = 25\%$ of max $|\underline{w}|$.

where w is the wavelet coefficient of interest and λ is the threshold level to be determined.

Hard thresholding is a “keep or kill” procedure which is intuitively appealing. Its transfer function is shown in Figure 3.8(a). The alternative, soft thresholding, whose transfer function is shown in Figure 3.8(b), shrinks coefficients above the threshold in absolute value. While at the first sight hard thresholding may seem to be natural, the continuity of soft thresholding has some advantages. It makes algorithms mathematically more tractable [30]. Sometimes, pure noise coefficients may pass the hard threshold and appear as annoying artifacts in the output. Soft thresholding shrinks these false structures. Soft thresholding has also been shown to achieve near minimax rate over a large number of Besov spaces [30].

Threshold determination, the choice of λ, is a very important question and a crucial step in wavelet shrinkage. A small threshold may preserve too much noise, whereas a large threshold leads to a loss of signal, causing artifacts and blurred edges.

For the rest of this section, several thresholding techniques such as VisuShrink [32], SUREShrink [104] and BayesShrink [2, 13, 96, 97] are explored. Further, Gaussian-based shrinkage techniques for natural images are discussed and their performances are compared.[3]

The idea of an MMSE estimate was first applied to the wavelet representation of signals and images by Donoho and Johnstone [32, 33]. They proposed a universal threshold λ as

$$\lambda_{universal} = \sqrt{2\log n}\ \hat{\sigma}_\nu$$

based on treating the wavelet coefficients as *i.i.d.* random variables, with n being the signal length and $\hat{\sigma}_\nu$ being an estimate for the additive white noise standard deviation. It is the optimal threshold in the asymptotic sense and in minimizing the cost function of the difference between the original wavelet coefficients and the soft thresholded version in the $\mathcal{L}^2$ norm sense, *i.e.*, it minimizes $E[(\underline{\hat{w}} - \underline{w})^2]$. VISUShrink is the thresholding which applies the universal threshold. Although, this method ensures, with high probability, that no noise (of size $\hat{\sigma}_\nu$) appears in the image after thresholding, its global adaptation to the signal to noise ratio (SNR) creates unpleasant visual artifacts especially in the vicinity of edges. VISUShrink is found to yield an overly smoothed estimate, because the universal threshold is derived under the constraint that with high probability the estimate should be at least as smooth as the signal. The threshold value, therefore, tends to be high for large values of n, killing many signal coefficients along with the noise. Thus, the threshold does not adapt well to discontinuities in the signal. A qualitative comparison of VISUShrink with two other wavelet shrinkage algorithms for the a corrupted image (512 × 512 pixels) is illustrated by Figure 3.10.

To overcome the problems of universal thresholding, adaptive denoising based on minimizing Stein's Unbiased Risk Estimator (SUREShrink) was proposed [40, 104]. SUREShrink is a scale dependent thresholding scheme which combines the universal threshold method with a scale-dependent adaptive selecting scheme and provides better visually

[3] In Ch. 6, we will revisit this section when formulating our model-based shrinkage rule.

appearing estimated results. This method estimates the loss $E[(\underline{\hat{w}} - \underline{w})^2]$ in an unbiased fashion:

$$SURE(\lambda; \underline{y}) = \underline{y} - 2 \cdot |\{i : |y_i| < \lambda\}| + \sum_{i=1}^{d} min(|y_i|, \lambda)^2 \tag{3.14}$$

where $|.|$ shows the number of elements is a set.

For an observed vector $\underline{y}$ (the set of noisy wavelet coefficients in a subband), find the threshold λ_{SURE} that minimizes $SURE(\lambda; \underline{y})$, *i.e.*, $\lambda_{SURE} = argmin_\lambda \; SURE(\lambda; \underline{y})$. The above optimization problem is computationally straightforward. This technique performs different global operations across scales. However, no spatial adaptation is assumed within each scale or each orientation, as seen in Figure 3.10, which shows SUREShrink denoising results for the "Lena" image. Clearly, the results are much better than VISUShrink. The sharp features of the image are retained. This is because SUREShrink is subband adaptive, *i.e.*, a separate threshold is computed for each detail subband.

An even more precise approach is spatially adaptive wavelet shrinkage [13, 70], called BayesShrink. In this method the coefficients in each subband are modeled as realizations of the class of Generalized Gaussian Distributed (GGD) [13] and independent random variables with different unknown parameters estimated on a pixel level using context modeling. Indeed it has been shown [13, 47]that for a large class of images, the coefficients of each detail subband form a symmetric distribution that is sharply peaked at zero, which is well described by the zero-mean GGD as is shown in Figure 3.9. The GGD is given by [13]

$$GGD_{\sigma_w,\beta}(w) = C(\sigma_w, \beta) e^{-[\alpha(\sigma_w,\beta)|w|]^\beta} \qquad -\infty < w < +\infty, \;\; \beta > 0, \;\; \sigma_w > 0 \tag{3.15}$$

where

$$\alpha(\sigma_w, \beta) = \sigma_w^{-1} \left[\frac{\Gamma(3/\beta)}{\Gamma(1/\beta)}\right]^{1/2}$$

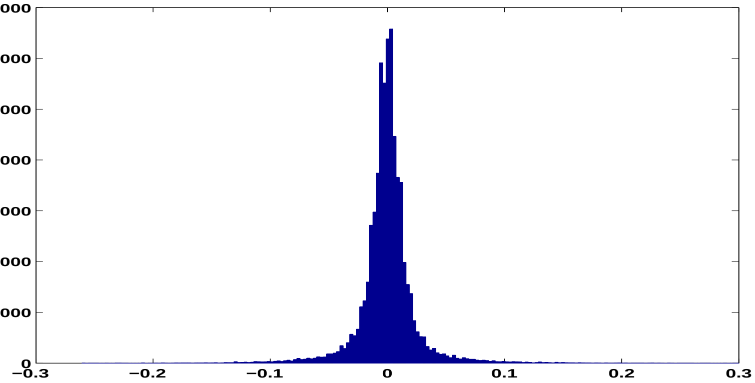


Figure 3.9: Histogram of the horizontal-band wavelet coefficients for a typical image.

and

$$C(\sigma_w, \beta) = \frac{\beta.\alpha(\sigma_w, \beta)}{2\Gamma(1/\beta)}$$

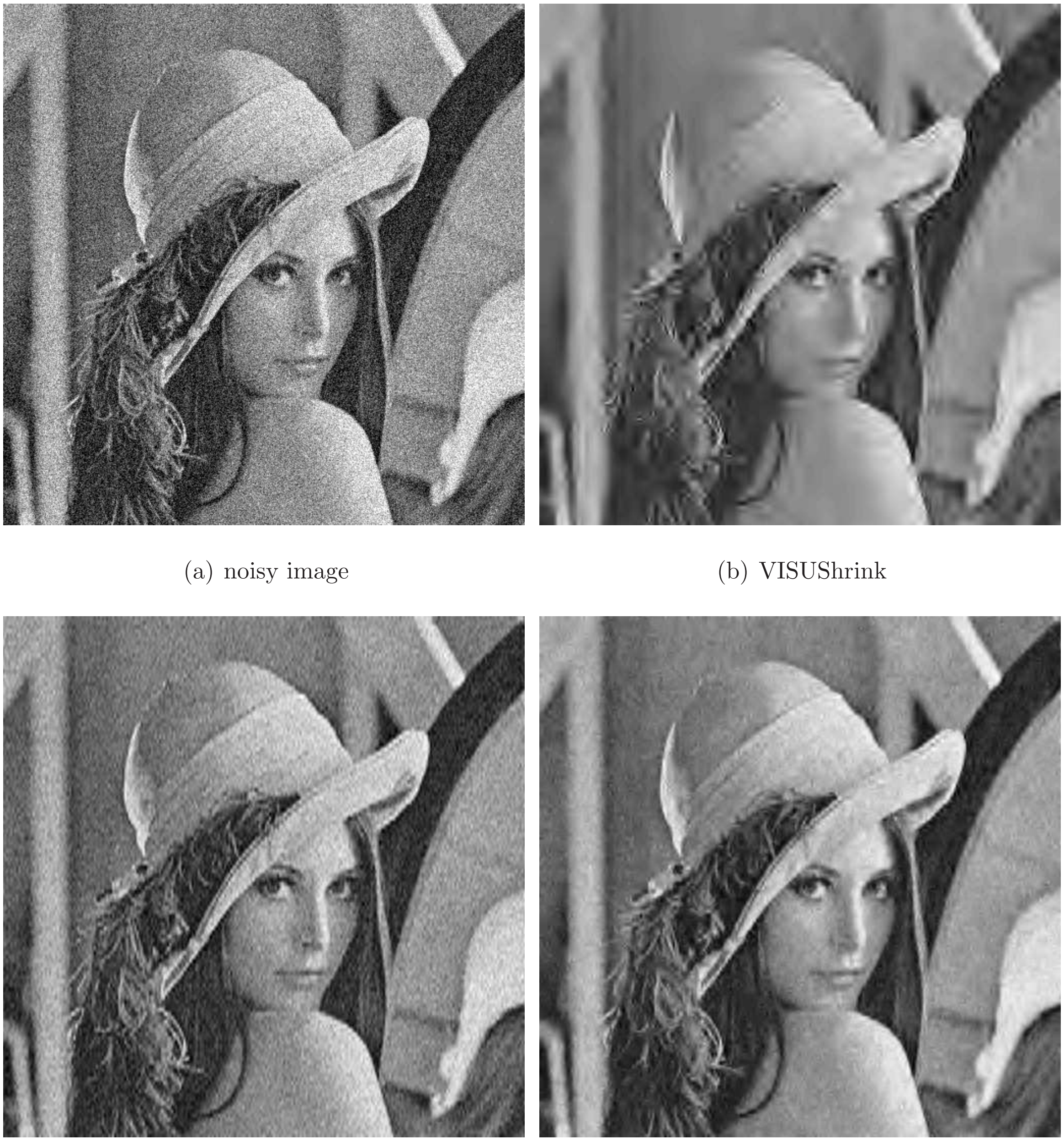
and $\Gamma(t) = \int_0^\infty e^{-u} u^{t-1} du$.

The parameter σ_w is the coefficient standard deviation and β is the shape parameter. It has been observed [13] that a shape parameter β ranging from 1 to 2 (Laplacian to Gaussian distribution), can describe the the distribution of coefficients in a subband for a large set of natural images. Assuming such a distribution for the wavelet coefficients, estimate β and σ_w for each subband and try to find threshold λ_{Bayes} which minimizes the Bayesian Risk, *i.e.*, the expected value of the mean square error

$$E[(\underline{\hat{w}} - \underline{w})^2] = E_W E_{Y|W}[(\underline{\hat{w}} - \underline{w})^2] \tag{3.16}$$

where $\underline{w} \sim GGD_{\sigma_w,\beta}$

Since there is no closed form solution for λ_{Bayes}, a numerical calculation is used to find its value. It was observed by Chang *et al.* [14] that the threshold value set by $\lambda_{Bayes} = \frac{\sigma_\nu^2}{\sigma_w}$ is very close to the optimum threshold values. The estimated threshold value is not only nearly optimal but also has an intuitive appeal. The normalized threshold, $\lambda_{Bayes}/\sigma_\nu$ is



(a) noisy image

(b) VISUShrink

(c) SUREShrink

(d) BayesShrink

Figure 3.10: Wavelet shrinkage using VISUShrink, SUREShrink and BayesShrink techniques, $\sigma_\nu = 0.2$. The db4 wavelet was used with four levels of decompositions.

inversely proportional to σ_w, the standard deviation of $\underline{w}$, and proportional to σ_ν, the noise standard deviation. When $\sigma_\nu/\sigma_w \ll 1$, the signal is much stronger than the noise, $\lambda_{Bayes}/\sigma_\nu$ is chosen to be small in order to preserve most of the signal and remove some of the noise; when $\sigma_\nu/\sigma_w \gg 1$, the noise dominates and the normalized threshold is chosen to be large to remove the noise which has overwhelmed the signal. Thus, this threshold choice adapts to both the signal and the noise characteristics as reflected in the parameters σ_ν and σ_w.

In summary, BayesShrink performs soft thresholding, with the data-driven, subband dependent threshold. The results obtained by BayesShrink for the image "Lena", shown in Figure 3.10, looks more appealing than those obtained using VISUShrink and SUREShrink.

At this stage it is unclear how much efficiency could be achieved in principle and how well the wavelet-domain shrinkage will operate, since the statistics of the wavelet coefficients have not been well characterized. Most of the existing thresholding procedures are essentially universal, *i.e.*, they uniformly treat the coefficients either at the same scale or across scales. They do not take into account the secondary properties of the wavelet coefficients. A natural way of using the prior information about the unknown image I is via a Bayesian framework, within which a prior distribution on the wavelet coefficients is specified. Wavelet thresholding via a Bayesian approach has been thoroughly studied during the last few years [2, 15, 97]. The following section summarizes the research achievements on the wavelet-domain statistics used as *a priori* in the Bayesian framework.

3.4 Wavelets and Statistics

This section reviews a series of most cited proposed models of the wavelet statistics. Each model's pros and cons are discussed and the open research questions and motivations for this present work will be highlighted.

Among numerous developed shrinkage methods, there is a common assumption that the wavelet coefficients are marginally Gaussian, and that the WT is a perfect whitener, such that all of the wavelet coefficients are independent. There is, however, a growing recognition [26] that the wavelet coefficients are neither Gaussian (Laplacian [86]), in terms of marginal statistics, nor white in terms of the joint statistics. There have been several recent efforts [26, 47, 77] to study wavelet statistics; most of these focus on the individual (marginal) statistics, only recent literature has addressed providing models representing the interrelationship (joint) statistics. Table 3.1 summarizes a comprehensive list of the probabilistic models proposed to address the wavelet domain statistics.

3.4.1 Wavelet Marginal Models

1. Double-exponential distribution [47]

$$w \sim \frac{1}{2\sqrt{2}\sigma} e^{-2\sqrt{2}\sigma|w|}$$

2. Generalized Laplacian (stretched exponential) [86]

$$w \sim \frac{e^{-|w/s|^p}}{2\frac{s}{p}\Gamma(1/p)}$$

where $p \in [0.5, 0.8]$ and s varies with the scales variances.

3. Mixture of Gaussians [15, 26, 47] (Figure 3.11).

4. Mixture of a Gaussian and a point mass function [2, 71]

$$w \sim \alpha \left[\frac{1}{\sqrt{2\pi}\sigma_w} e^{-(w-\mu)^2/(2\sigma_w^2)} \right] + (1-\alpha)\delta(0), \qquad 0 \le \alpha \le 1$$

where $\delta(0)$ is a point mass function at zero. For all coefficients at a given scale the same prior parameters α and σ_w^2 are assumed.

Wavelet Statistics	**Marginal**	**Joint**
Chipman *et al.* [15, 16]	two Gaussian dist.	—
Leporini *et al.* [55]	non-Gaussian	—
Simoncelli [86]	generalized Laplacian	—
Moulin and Liu [70]	generalized Gaussian	—
Huang and Mumford [47]	non-Gaussian, heavy-tailed	—
Abramovich [2]	a point mass f. & a Gaussian	—
Donoho and Johnstone [33]	non-Gaussian	—
Vidakovich [96]	a point mass f. & a Gaussian	—
Crouse *et al.* [26]	—	hidden states
Romberg *et al.* [82]	—	hidden states (Bayes Ets.)
Choi and Baraniuk [19]	—	hidden states (Segmentation)
Nowak [73, 74]	—	hidden states (Bayes Ana.)
Xu *et al.* [101]	—	heuristic
Portilla *et al.*[79, 98]	—	Gaussian scale mixtures
Strela *et al.* [92]	—	Gaussian scale mixtures
Fan and Xia [36]	—	hidden states
Mihcak *et al.* [69]	—	adaptive
Simoncelli [78, 86]	—	local coeff. energy
Yoo *et al.* [103]	—	adaptive
Chang and Vetterli [14, 13]	—	adaptive
Crouse and Barabiuk[25]	—	hidden states of local NBHD
Fan and Xia [35]	—	hidden states of local NBHD
Maifait and Roose [62]	—	MRFs
Pizurica *et al.* [77]	—	MRFs
Fan and Xia [37]	—	hidden states on a hybrid quad-tree
Liu and Moulin [59]	—	mutual information, MRFs
Azimifar *et al.* [6, 7, 8]	—	MS, MRFs

Table 3.1: A list of the probabilistic models proposed to describe wavelet domain statistics.

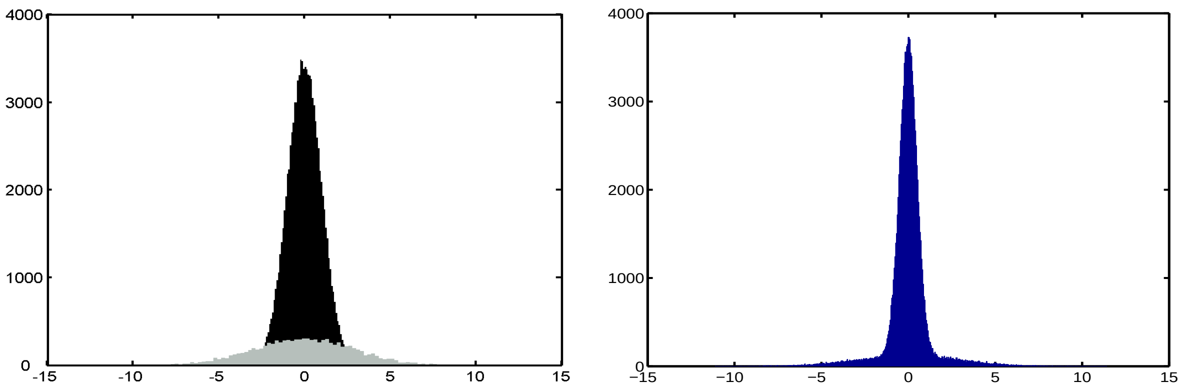


Figure 3.11: Two-state, zero-mean Gaussian mixture model for the wavelet coefficients. Left: Two normal density functions with different variances; Right: mixture model of both Gaussian distributions.

5. Generalized Gaussian distribution (GGD) [70, 78]

$$w \sim C(\sigma_w, \beta) \exp\left\{ - \left[\alpha(\sigma_w, \beta)|w|\right]^{\beta} \right\}$$

   where the parameters are defined in (3.15).

6. Explicit modeling of edges and boundary-like features [33, 98].

7. Bessel functions [88].

All of these models propose heavy-tailed distributions for the individual wavelet coefficients. The heavy-tailed behaviour is supported by empirical tests, and is due to the energy compaction property: the coefficient statistics are a mixture of low-variance (noise) and a few high-variance (signal) elements.

Virtually all marginal models currently being used in wavelet shrinkage [78], assume the coefficients to be decorrelated and treated individually; *i.e.*, only the diagonal elements of the wavelet covariance are considered. Although such independent models result in simple nonlinear operations on individual coefficients, the approach is suboptimal because the WT is not a perfect whitening process.

3.4.2 Wavelet Joint Models

As opposed to the marginal models, the question of joint models is much more complicated and admits for more possibilities, with structures possible across subbands, orientations, and scales. Since Shapiro [85] proposed zerotree coding for image compression there have been many efforts to model joint structures. Researchers have proposed a variety of wavelet dependency models including

a. Hidden Markov models (HMMs) [26, 37, 82],

b. Markov random field priors (MRFs) [62, 77], and

c. Gaussian Scale Mixtures (GSMs) [79, 98].

These models are defined on the basis of the observed characteristics of wavelet coefficients: across-scale persistence, within-scale clustering [82], and sparse representation. In general, these models investigate a combination of these three main categories:

1- interscale [26, 82, 101],

2- intrascale (spatial) [35, 69, 74, 86], and

3- combined intra- and interscale [14, 25, 37, 62, 70, 77, 94] dependencies.

A brief discussion on some of these models follows.

In a similar Bayesian fashion but independent investigations, Malfait and Roose [62] and Pizurica *et al.* [77] examined non-decimated wavelet-domain joint within- and across-scale dependencies by assigning a significance mask and a binary label to the individual coefficients. The significance map is defined based on an estimation of the local Lipschitz regularity and coefficient evolution and inside a local cone of influence. On a similar track and motivated by the empirical Bayes estimator of Lee [57], a number of authors have estimated the local variance from a collection of wavelet coefficients at nearby positions and scales and used these estimated variances in order to denoise the coefficients [1, 13, 69]. Surprisingly, none of the aforementioned modeling procedures has investigated orientation-dependent priors. Indeed, only very little literature has studied models to describe across-orientation correlations.

A multivariate normal prior for the wavelet coefficients has been proposed by Vannucci and Corradi [94]. By taking into account the correlation between the wavelet coefficients, they showed that the wavelet prior P_w (wavelet covariance) is a J-diagonal matrix, where J is the number of scales. The covariance structure is considered for coefficients within the same resolution level as well as across different scales. The covariance matrix is expressed as a diagonal band outside which the correlation is assumed zero and within the band, the largest correlations happen between coefficients that are close in the same location and scale. A thorough investigation of the wavelet domain covariance structure will be discussed in Ch. 4.

Portilla *et al.* [79, 98] developed a Gaussian Scale Mixture (GSM) model to describe the kurtotic behavior of marginal distributions as well as the pairwise joint distributions and used Bayesian least square to estimate the coefficients. GSMs model the neighborhood of coefficients at adjacent positions and scales as the product of two independent random variables: a Gaussian vector and a hidden scaler multiplier [3]. This multiplier plays a crucial rule: the key property of the GSM model is that the density of coefficient w is Gaussian when conditioned on the multiplier value. A GSM assumes that the local variance is governed by a continuous multiplier variable, which is an extension to the two-state hidden multiplier variable used by Romberg *et al.* [82] to characterize the two-mode behavior of large and small valued coefficients.

Xu *et al.* [101] used the scale-dependent consistency between the wavelet coefficients for the denoising process. In separate works by Simoncelli [86], Strela *et al.* [92], and Crouse *et al.* [26] probabilistic models that capture wavelet coefficient dependencies, essentially across scales, were studied. Crouse *et al.* [26] considered hidden states describing each coefficient's significance. Instead of the coefficients values, they proposed statistical models for coefficient's hidden state dependencies. Normally an assumption is present that the correlation between coefficients' states does not exceed the parent-child dependencies, *e.g.*, given the state of parent, the child is decoupled from the tree on the other side of its parent. The wavelet-based HMMs [74], in particular, have been thoroughly studied and successfully outperform many wavelet-based techniques in Bayesian denoising, estimation, texture analysis, synthesis and segmentation. For the rest of this section I introduce the HMMs as the most influential wavelet joint models, then explain how the issues assigned to these models can direct and motivate this work.

3.4.3 Wavelet HMM Joint Models

As a finite state machine, an HMM is basically a Markov chain process characterized by its state transition probabilities [80]. In an HMM, at every time interval an observation is made from the current state according to a pdf depending only on that state. In contrast to a Markov chain, it is not possible to determine the current state by simply examining the current observation. Thus, the state of an HMM is hidden to the observer (empty circles shown in Figure 3.12).

Although HMMs have been successfully applied in 1-D signal processing (*e.g.*, speech processing), it is hard to directly adopt them in the spatial domain image modeling due to large extent of spatial correlations which result in a large number of states. On the other hand, the reported characteristics of the wavelet coefficients including across-scale persistence and within-scale clustering [82], in addition to the sparse representation property of the WT which results in reducing the number of states, has inspired researchers to propose HMMs to describe the structure of wavelet local dependencies. The wavelet-domain HMMs were developed and successfully applied in many image processing tasks, such as estimation [26], Bayesian image analysis and denoising [73, 74], image segmentation [18, 82] and texture analysis, synthesis and classification [37]. Since Nowak and his colleague [26] introduced these models, significant work has been done to improve performance of the HMMs, either in model accuracy or in parameter estimation. In general, these models adopt a probabilistic graph in which every wavelet coefficient w_i is associated with a discrete hidden state $s_i \in \{0, 1, \ldots, M-1\}$ thus modeling w_i as an M-state Gaussian mixture, conditionally independent of all of the other coefficients $p(w_i, w_j|s_i) = p(w_j|s_i)p(w_i|s_i) \quad \forall i \neq j$. A binary state, $M = 2$, is particularly common, used to specify a low/high variance of w_i. Figure 3.12 sketches a simple HMM on the 1-D wavelet tree. Clearly the tree branches are formed by connecting the hidden states. The simplest case is when the coefficiets are assumed maginally, *i.e.*, the dashed lines are disappeared.

The wavelet parsimony representation indicates that the majority of the coefficients are small and only a few coefficients are large in magnitude. As mentioned earlier, Chipman *et al.* [15] showed that the heavy-tailed non-Gaussian marginal pdf, $f_W(w)$ of the wavelet coefficient w can be well approximated by GMM (Figure 3.11). Accordingly, wavelet non-linear shrinkage, such as Bayesian estimation has been achieved with these non-Gaussian priors, which consider the kurtotic behavior of the wavelet coefficients (§ 3.3). At this point the marginal modeling of Chipman *et al.*is compared to the HMMs of Crouse *et al.*. The main distinction between these two approaches is that

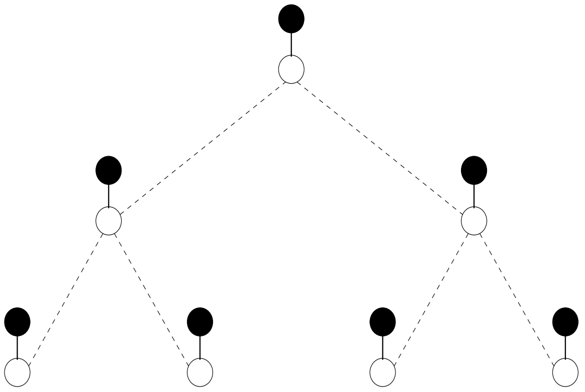


Figure 3.12: The 1-D wavelet hidden Markov model. The statistics of each coefficient (filled circle) are described by its associated hidden state (empty circle) value.

Chipman *et al.* [15] consider independence in the prior, while Crouse *et al.* [26] introduce dependent priors on the mixture parameters. Crouse *et al.* [26] consider a Hidden Markov Model (HMM) for the dependencies among the wavelet coefficients, in which whether or not a specific coefficient is non-zero will depend on the state of its immediate neighbors. This neighborhood can be at the same resolution but locations $k \pm 1$, or at an analogous location but across resolutions $j \pm 1$, *i.e.*, parent and children. In fact, if a certain coefficient is significant, then its neighbors are likely to be significant.

The GMM observation of Chipman *et al.*, in addition to the association of the hidden states to each w (*i.e.*, the hidden state indicates that the associated node is either negligible or significant in magnitude), has directed the researchers to assume M-state GMM with conditional pdf $f_{W|S}(w|s = m, m = 0, 1, \ldots, M-1)$ with mean μ_m and variance σ_m^2 for every w, given its hidden state value. The overall pdf of w is defined to be

$$f_W(w) = \sum_{m=0}^{M-1} p_S(m) f_{W|S}(w|s = m), \tag{3.17}$$

where

$$f_{W|S}(w|s=m) = \mathcal{N}(w;\mu_m,\sigma_m^2)$$

A practical two-state zero-mean GMM with parameter set $\Theta = \{p_S(m), \sigma_m^2 | m = 0, 1\}$ is shown in Figure 3.11.

Although, the GMM can necessarily describe the wavelet-domain marginal statistics, it can not be a sufficient tool to characterize the joint statistics of the wavelet domain. Therefore, different neighborhood structures (mostly Markov) are adopted to describe the joint statistics of the hidden states. Indeed, the HMMs are multidimensional GMMs in which hidden states have a Markovian dependency structure. Note that all different kinds of HMMs assume that the wavelet coefficients have the same statistics, regardless of their spatial position [37]. The class of HMMs includes

1- GMM: Gaussian Mixture Model [15],

2- IMM: Independent Mixture Model [26],

3- HMT: Hidden Markov Tree [26, 73],

4- HMT-2: Improved Hidden Markov Tree [36],

5- CHMM: Contextual Hidden Markov Model [25, 35], and

6- HMM-3S: Hidden Markov Model-Three Subbands [37].

A brief definition of each model follows.

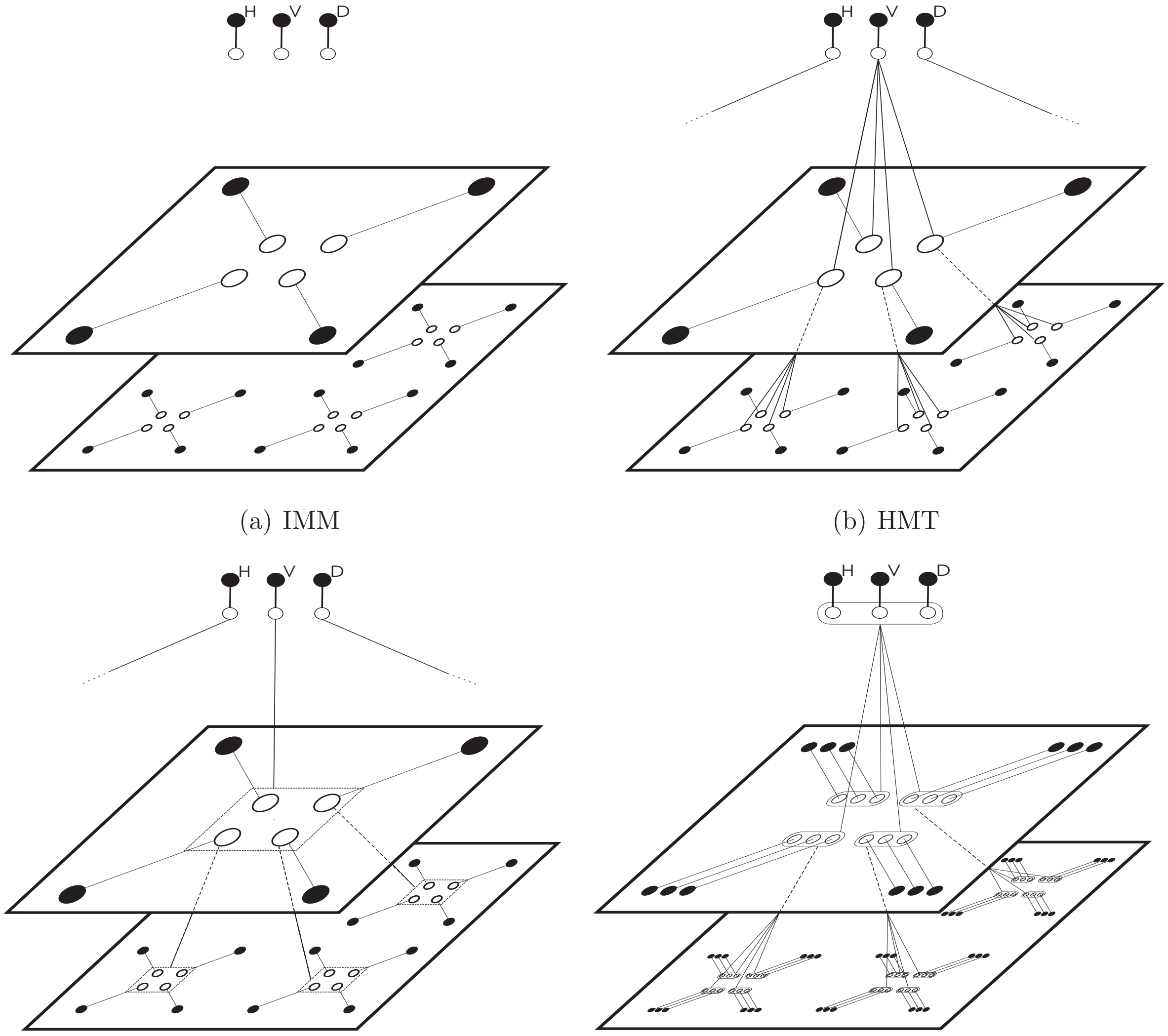
Independent Mixture Model (IMM): The simplest HMM, the IMM [26] (Figure 3.13(a)) models the hidden states as independent GMMs, motivated by the observation that the WT is an approximate whitening process such as Karhunen-Loeve transform, making the coefficients nearly decorrelated.

Although, the IMM can describe the wavelet marginal statistics well as a Gaussian mixture, it does not characterize the remaining wavelet-domain joint statistics. More advanced models in which hidden states have a Markovian dependency structure are discussed below.

Hidden Markov Tree (HMT): More sophisticated approaches sought to model the local wavelet statistics by introducing Markovian dependencies between the hidden state variables across scales and orientations. Crouse *et al.* [26] observed the persistence and clustering properties of wavelet coefficients and introduced the HMT (Figure 3.13(b)) which captures wavelet interscale dependencies by imposing a tree structure on the hidden states across scales, while assuming independence within and across the three subbands. The statistic of hidden state s_i is a function of its parent, $s_{\rho(i)}$, based on a transition probability $p_{S_i|S_{\rho(i)}}(s_i = m|s_{\rho(i)} = n)$. It was shown in [82] that the HMT algorithm outperforms traditional wavelet-based techniques.

Improved Hidden Markov Tree (HMT-2): A generalization, to capture additional correlations between scales, is the HMT-2 [36], where the state s_j depends on the state of its parent $s_{\rho(j)}$, as before, but also on siblings of its parent (Figure 3.13(c)). The approach is motivated by the correlation of the wavelet bases in two adjacent scales and the long length of the filters used in the decomposition process. It, however, leads to higher-order hidden states. Unlike the HMT, in the HMT-2 each node is associated with a vector of four hidden states (within-scale), *i.e.*, if $M = 4$ is assumed, there are sixteen different possibilities for each node. Since a vector of coefficients grouped in each node with two different GMMs, the two-dimensional Gaussian mixture field is considered in the training process of the HMT-2 model. HMT-2 empirical results show some improvement in signal denoising [36].

Contextual Hidden Markov Model (CHMM): The HMT models focus on the vertical interscale dependencies by imposing a tree structure in the wavelet domain. To support



(a) IMM

(b) HMT

(c) HMT-2

(d) HMT-3S

Figure 3.13: Illustration of hidden Markov models. Empty circles denote hidden states and filled circles the coefficient values. (a) Independent hidden states; (b,c) Interscale dependencies; (d) Three subbands (H, V, D) integrated into one hybrid HMT.

additional connectivity the CHMM was developed [25], which adds a context structure to model both interscale and intrascale dependencies. The basic idea of the CHMM is to define contexts, as a function of the wavelet coefficient w_j and its local neighbors, to capture the spatial dependencies such that given the context the coefficients are treated as independent. The CHMM has many potential advantages [25] over traditional HMMs in exploiting the wavelet correlation structure, offering similar denoising performance with reduced computational complexity compared with the HMT.

The lack of spatial adaptability of CHHM [35][69] may limit its advantages in image processing tasks, so a Local CHMM was proposed [35]. This model exploits both interscale correlations and local wavelet coefficients statistics.

Hidden Markov Model-Three Subbands (HMM-3S): The above joint models assume that the three high-frequency subbands are independent, an almost universal assumption. Although this simplifies wavelet image modeling, for natural textures the regular spatial patterns may result in noticeable dependencies across subbands [37]. The HMT-3S (Figure 3.13(d)) includes the joint interscale statistics captured by the HMT, but adds the dependencies across subbands by integrating the three corresponding coefficients across the three orientations. The HMT-3S was successfully applied in texture analysis and synthesis with improved performance over the HMT models.

3.5 Chapter Summary

The literature of modeling wavelet statistics has been thoroughly reviewed. It is clear that a significant body of research addresses wavelet joint statistics and modeling. The more general the model, the broader the range of included correlations, and the better the results.

This chapter answered several significant questions:

- Why are the wavelet coefficients assumed independent and non-Gaussian?

Because of the WT primary properties (locality, multiresolution and compression) and interpretation of the WT as a whitening process which attempts to make the coefficients statistically independent.

- Why are the coefficients assumed jointly Gaussian?

  Because this assumption helps to captures the linear correlation between the coefficients.

- Why are the coefficients assumed jointly non-Gaussian?

  Because of the secondary properties of the wavelet coefficients (within-scale clustering and interscale persistence).

- What sub-groups of the coefficients are described by the wavelet joint models?

  Clusters of coefficients across scales, across subbands, or within subbands. The last two groups are sometimes referred as intra-scale or within-scale statistics, but often confusing!

Some important concepts still remain to be thoroughly understood:

- What should be measured in the wavelet domain?
- What statistics should be studied?
- What stochastic model (such as Markov random fields) can model these statistics?
- What are the best parameters describing the wavelet joint model?

The main theme of the next two chapters is to elaborate on the above issues. Ch. 4 will establish a framework to study the existence and the characteristics of the wavelet joint correlations, followed by Ch. 5 which will concentrate on bridging between the wavelet joint statistics and the probabilistic models introduced in Ch. 2.

Chapter 4

Wavelet Correlation Structures

This chapter presents an empirical study of the joint wavelet statistics for a large range of natural images and random fields. The study of wavelet statistics includes results of straightforward Monte-Carlo simulations as well as the *exact* statistical analysis when the given image is a sample of a Gaussian random field or a real seen picture. This sample statistical study highlights, albeit only approximately, the significant residual correlations between coefficients within and across scales. While recent developments in wavelet-domain Hidden Markov Models (notably HMT-3S) [26, 37] account for within-scale dependencies, empirically, wavelet spatial statistics are strongly orientation dependent, structures which are surprisingly not considered by state-of-the-art wavelet modeling techniques.

This chapter describes possible choices of wavelet statistical interactions by examining the wavelet domain covariance, joint-histograms, conditional distributions, correlation coefficients, and the significance of coefficient relationships. An efficient and fast strategy which describes the wavelet-based statistical correlations and their significance of inter-relationships will be demonstrated.

4.1 Empirical Correlations: a Monte-Carlo Study

This statistical study of the wavelet coefficients starts by generating an ensemble of parameterized random fields with a spatially stationary assumption (only an assumption to simplify the inference). Later observations of the real image wavelet statistics will show that the structure of wavelet-domain joint correlations is not a direct consequence of stationarity assumption.

Because of the stationarity, the correlation structure of any random field $X \in \mathbb{R}^{n \times n}$ is invariant to spatial location, *i.e.*,

$$E[x_{i,j}x_{i+\Delta i,j+\Delta j}] = E[x_{0,0}x_{\Delta i,\Delta j}], \quad 1 \leq i,j \leq n \tag{4.1}$$

resulting in an autocorrelation structure

$$\Pi_{i,j} = E[x_{0,0}x_{i,j}], \quad 1 \leq i,j \leq n \tag{4.2}$$

a circulant matrix, which can be diagonalized by the 2-D Fast Fourier Transform (FFT) [38]. Then a toroidally stationary random field covariance matrix $P \in \mathbb{R}^{n^2 \times n^2}$ which corresponds to Π can be formed. A Gaussian random field $X \sim \mathcal{N}(\underline{0}, P)$ is then synthesized as

$$X = FFT^{-1}\{\sqrt{FFT\{\Pi\}} \cdot FFT\{Q\}\} \tag{4.3}$$

where $\sqrt{\cdot}$ and $\cdot$ are element-by-element operations and $Q \sim \mathcal{N}(\underline{0}, I)$ is a matrix of unit variance Gaussian random variables. Because of the Gaussian assumption in prior P this model is called a Gauss-Markov Random Field (GMRF).

The prior covariance structure P plays a critical role in the sample generation, *i.e.*, the first step in sampling is to have a prior model. In the absence of a particular prior correlation structure of the field X, one can impose a smoothness constraint between the field elements. This subject is known as data regularization [38]. Two regularization methods which are commonly used in surface interpolation problems, asserting local constraints between the spatially close pixels, are reviewed here.[1] If the first-order derivative is considered, then

$$\begin{aligned} H_{membrane}(X) &= \int\int \left|\frac{\partial X}{\partial x}\right|^2 + \left|\frac{\partial X}{\partial y}\right|^2 dxdy \\ &= \sum\sum \left|\mathcal{L}_x * X\right|^2 + \left|\mathcal{L}_y * X\right|^2 \qquad (4.4) \end{aligned}$$

where $*$ is a convolutional operator and $\mathcal{L}_x$ and $\mathcal{L}_y$ are defined to be the first-order convolutional kernels

$$\mathcal{L}_x = [-1 \quad 1] \qquad \mathcal{L}_y = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \qquad (4.5)$$

which form the associated *membrane* model [38]

$$\begin{aligned} \mathcal{L}_{membrane} &= \mathcal{L}_x * \mathcal{L}_x + \mathcal{L}_y * \mathcal{L}_y \\ &= \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \qquad (4.6) \end{aligned}$$

The smoothness constraint on second-order derivatives is called a *thin-plate* model:

$$\begin{aligned} H_{thin-plate}(X) &= \int\int \left|\frac{\partial^2 X}{\partial x^2}\right|^2 + 2\left|\frac{\partial^2 X}{\partial x \partial y}\right|^2 + \left|\frac{\partial^2 X}{\partial y^2}\right|^2 dxdy \\ &= \sum\sum \left|\mathcal{L}_{xx} * X\right|^2 + 2\left|\mathcal{L}_{xy} * X\right|^2 + \left|\mathcal{L}_{yy} * X\right|^2 \qquad (4.7) \end{aligned}$$

[1] In solving for an interpolation problem if the measurement alone is considered, normally the problem is not well-posed and a unique optimum solution is not guaranteed. However, by asserting certain prior knowledge such as these smoothness constraints, the problem may become well-posed, in which case a single estimate, *i.e.*, least square is found [38].

where $\mathcal{L}_{xx}$ and $\mathcal{L}_{yy}$ and $\mathcal{L}_{xy}$ are defined to be the second-order convolutional kernels

$$\mathcal{L}_{xx} = [-1 \quad 2 \quad 1] \qquad \mathcal{L}_{yy} = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix} \qquad \mathcal{L}_{xy} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \tag{4.8}$$

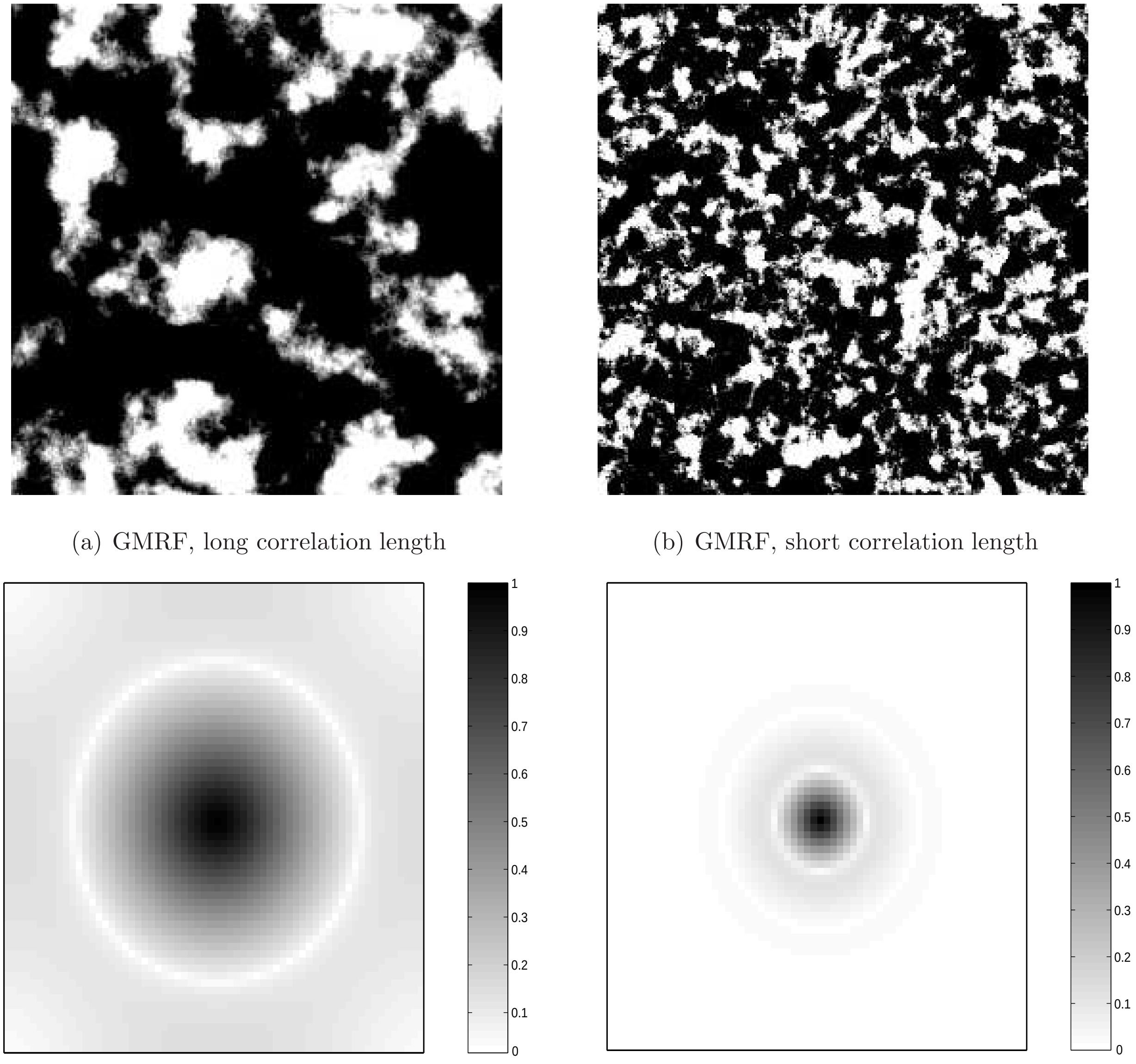
with the associated discrete constraint model

$$\begin{aligned} \mathcal{L}_{thin-plate} &= \mathcal{L}_{xx} * \mathcal{L}_{xx} + \mathcal{L}_{yy} * \mathcal{L}_{yy} + 2\mathcal{L}_{xy} * \mathcal{L}_{xy} \\ &= \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 2 & -8 & 2 & 0 \\ 1 & -8 & 20 & -8 & 1 \\ 0 & 2 & -8 & 2 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \end{aligned} \tag{4.9}$$

being employed in the linear estimation process (2.5).

Comparison of (4.6) and (4.9) with Figure 2.1 indicates that membrane is a first-order Markov process and that thin-plate, which imposes a larger neighborhood, is third-order Markov.

Figure 4.1 illustrates two different GMRF realizations based on (4.3) and a given thin-plate autocorrelation Π_x and covariance P_x structures with different correlation lengths. Correlation length is defined to be average distance between any pixel and its farthest neighbor whose correlation is at least half of the maximum correlation. Figure 4.3 shows the covariance matrix for two fields displayed in Figure 4.1. Note that P_x shown in this figure illustrates the correlation structure for the field elements when they are lexicographically re-ordered into a vector format (*e.g.*, column-wise) as is shown in Figure 4.2. In other words, each row of the field covariance matrix P_x is obtained by circularly shifting the vector-formatted autocorrelation map Π_x.



(a) GMRF, long correlation length

(b) GMRF, short correlation length

(c) $|\Pi_x|$

(d) $|\Pi_x|$

Figure 4.1: Gaussian random fields generated based on the FFT algorithm given in (4.3). *Top panels*: the thin-plate random fields X, *Bottom panels*: absolute values of the associated autocorrelation map Π_x, where zero-lag position is assumed in the middle of the image.

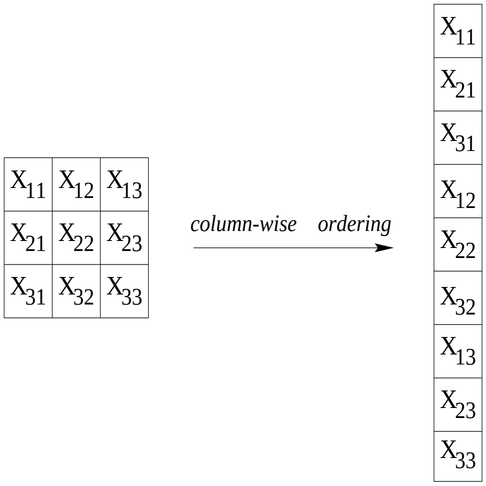


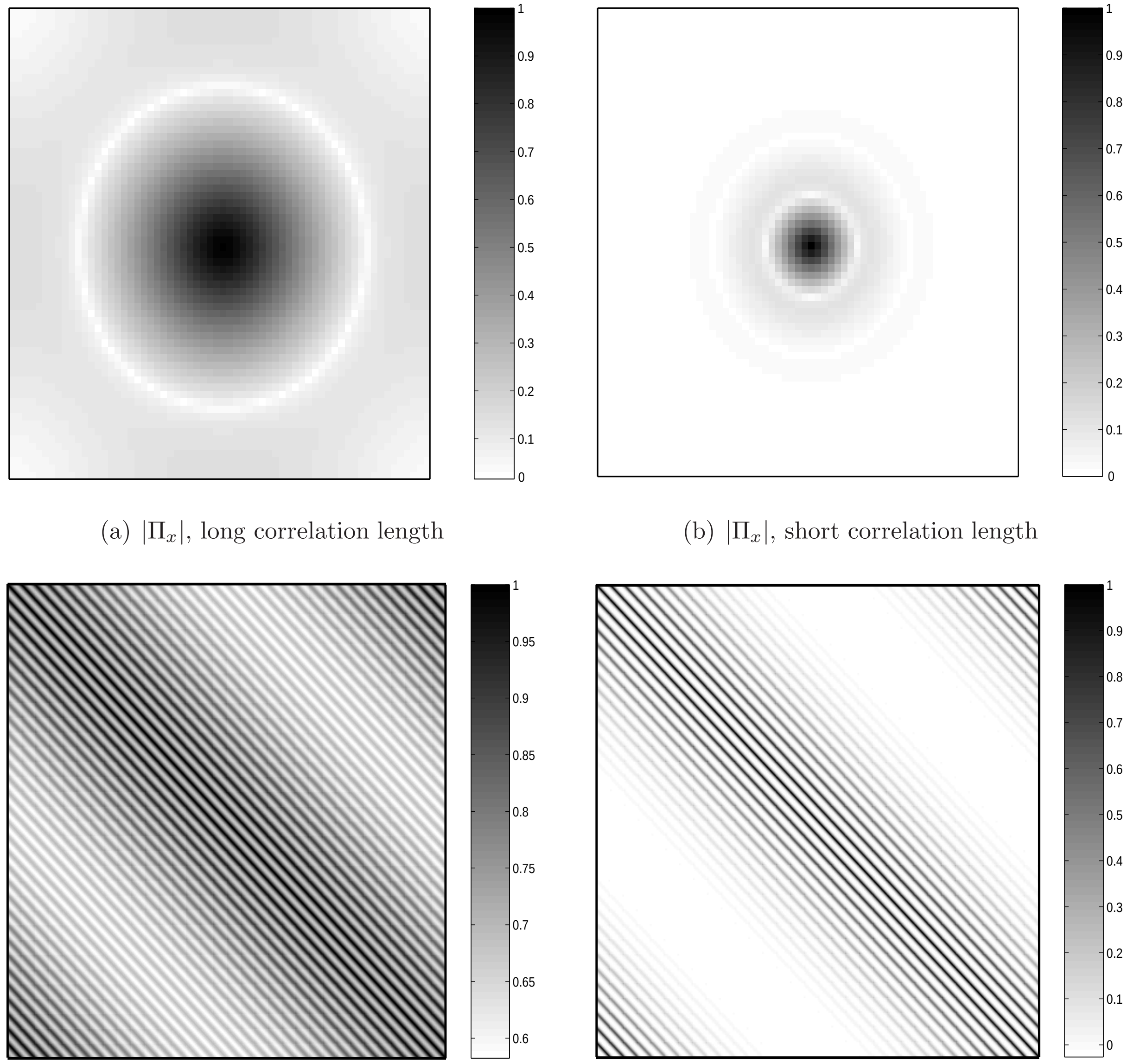
Figure 4.2: Lexicographical stacking a 2-D matrix into a 1-D vector.

Having defined the above framework to generate a number of model-based fields, a Monte-Carlo study was done. An ensemble (~ 500 iterations) of sample paths (thin-plate prior) with various correlation lengths were transformed into the wavelet domain. For each case the within scale sample correlation coefficients were calculated for a local spatial neighborhood at every orientation, *i.e.*, horizontal, vertical and diagonal directions. For convenience in understanding the results, the resulting variances were normalized, so that the inter-coefficient relationships are measured as a correlation coefficient

$$\rho_{i,j} = \frac{E[(w_i - \mu_{w_i})(w_j - \mu_{w_j})]}{\sigma_{w_i}\sigma_{w_j}}, \qquad -1 \leq \rho_{i,j} \leq 1 \tag{4.10}$$

where w_i and w_j are two wavelet coefficients, with mean and standard deviation μ_{w_i}, σ_{w_i} and μ_{w_j}, σ_{w_j}, respectively. Indeed, $|\rho_{i,j}| = 1$ shows that the coefficients are deterministically related, $\rho = 0$ indicates total uncorrelatedness between two wavelet coefficients and $|\rho_{i,j}| = 0.5$ was considered the threshold in measuring the correlation length. In other words

$$\mathcal{P}_{w_{i,j}} = \frac{P_{w_{i,j}}}{\sqrt{P_{w_{i,i}} P_{w_{j,j}}}} \tag{4.11}$$



(a) $|\Pi_x|$, long correlation length

(b) $|\Pi_x|$, short correlation length

(c) $|P_x|$

(d) $|P_x|$

Figure 4.3: Absolute value of the covariance matrices $P_x \in \mathbb{R}^{n^2 \times n^2}$ of two Gauss-Markov random fields obtained according to the autocorrelation structures $\Pi_x \in \mathbb{R}^{n \times n}$ of a thin-plate model with two different correlation lengths.

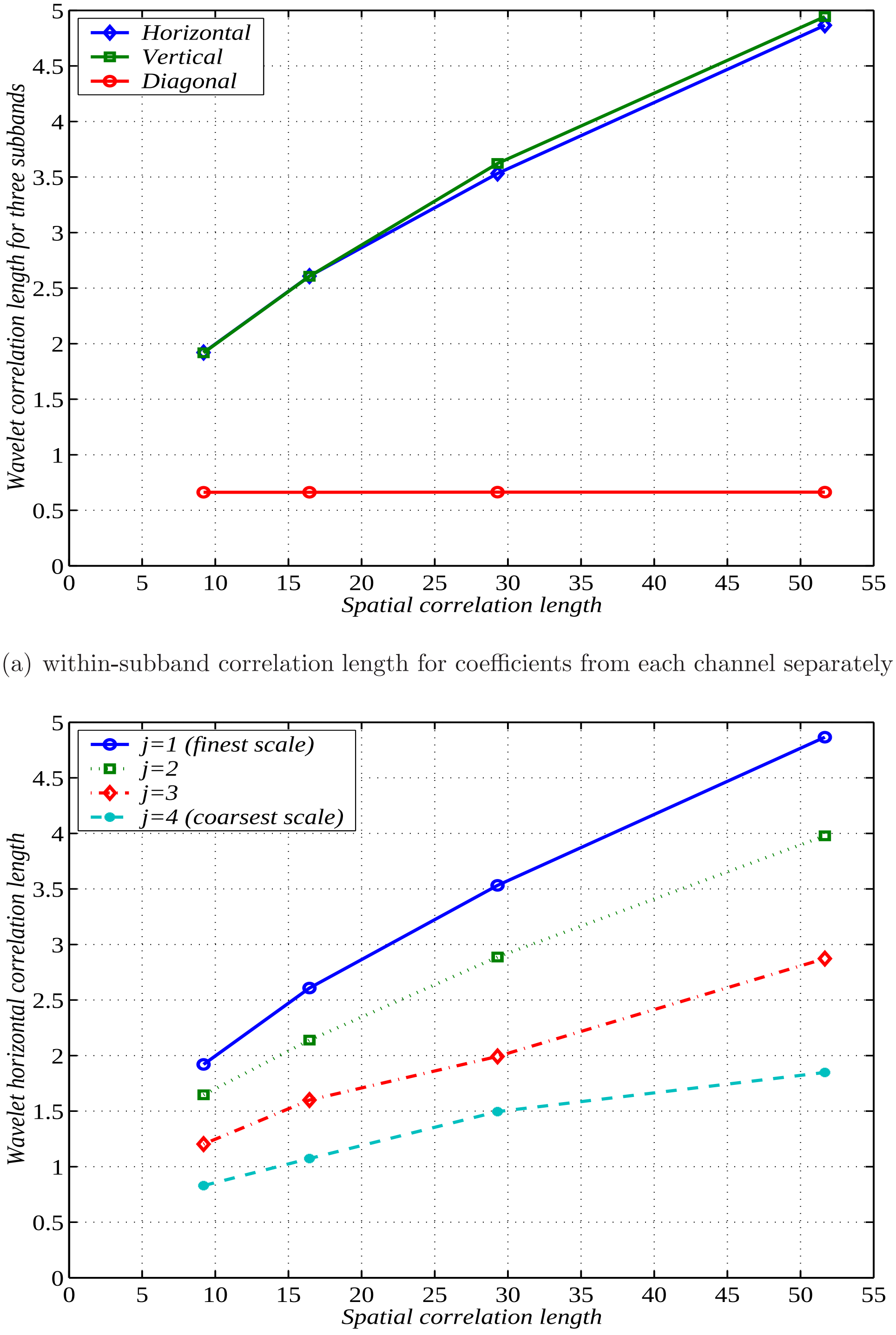
where $\mathcal{P}_w$ shows matrix of correlation coefficient values.

For a fixed spatial domain correlation length it was observed that

- the wavelet domain correlation length is much smaller than the spatial domain one, but is **not** zero,
- the wavelet domain correlation strength reduces from fine to coarse scales,
- the extent of correlation is identical in both vertical and horizontal subbands (because the spatial prior is identical in vertical and horizontal directions),
- correlations in the diagonal subband are weaker than in the other two orientations.

Figure 4.4(a) plots within-subband correlation length in the Haar wavelet domain for all three orientation channels as a function of the spatial correlation length. In this experiment the thin-plate model as a third-order neighborhood smoothness constraint was considered. Although the whitening effect of the wavelet transform is quite clear for the diagonal subband coefficients, the coefficients at the horizontal and vertical channels exhibit a residual inter-relation along their orientation which grows, albeit slowly, with increasing correlation length in the spatial domain. Figure 4.4(b) highlights the correlations between pairs of horizontal subband coefficients within four different scales. The increasing trend of correlation between the coefficients up to five pixels apart is quite obvious in this plot.

This wavelet correlation study was started with the above Monte-Carlo examination of the spatial and wavelet correlation lengths, and observing the wavelet residual sample statistics (non-whitening effect of the wavelet transform) [5]. It is continued by investigating exact statistics of the wavelet domain.



(a) within-subband correlation length for coefficients from each channel separately

(b) within-subband correlation length for horizontal coefficients at different scales

Figure 4.4: Wavelet (Haar) within-subband correlation lengths for wavelet coefficients are plotted as a function of the spatial correlation lengths.

4.2 Empirical Correlations: A Statistical Approach

To analyze the statistical behavior of any random process, one alternative is to evaluate a process whose statistical correlation is known, *i.e.*, the exact covariance structure is given, such that one can directly evaluate the actual correlation structure instead of manipulating the sample statistics of data. In other words, since the objective is to assess any change in correlation structure, by having a mathematical description of this structure it is possible to directly work with the wavelet coefficients of synthetic or real-world data. Instead, one can explicitly evaluate the wavelet transformation of the predefined spatially correlation structure, explained as follows.

4.2.1 Examining one-dimensional signals

It is assumed that low and high frequency components of the 1-D signal $\underline{x}$ projected into the wavelet domain, *i.e.*, $\mathcal{W}\underline{x}$, are represented by separate sets of coefficients, namely the approximation $\underline{a}_J$ and the detail $\{\underline{w}_j\}, 1 \leq j \leq J$ coefficients. If, as before (Eq. (3.9)), the linear operators H and G are defined as low-pass and high-pass filters respectively, then clearly the coefficient vectors may be recursively computed in scale as

$$\begin{aligned} \underline{a}_{j+1} &= H_j \underline{a}_j \\ \underline{w}_{j+1} &= G_j \underline{a}_j \end{aligned} \tag{4.12}$$

Having defined the vectors of $\underline{a}_j$ and $\underline{w}_j$ coefficients, one can recursively calculate the within- and across-scale auto- and cross-covariances from the covariance $P_{\underline{a}_j,\underline{a}_j}$ at the finest scale $j = 1$ as follows:

$$\begin{aligned} P_{\underline{w}_{j+1},\underline{w}_{j+1}} &= G_j P_{\underline{a}_j,\underline{a}_j} G_j^T \\ P_{\underline{a}_{j+1},\underline{a}_{j+1}} &= H_j P_{\underline{a}_j,\underline{a}_j} H_j^T \\ P_{\underline{a}_{j+1},\underline{w}_{j+1}} &= H_j P_{\underline{a}_j,\underline{a}_j} G_j^T \\ P_{\underline{a}_{j+2},\underline{w}_{j+1}} &= H_{j+1} H_j P_{\underline{a}_j,\underline{a}_j} G_j^T \\ P_{\underline{w}_{j+2},\underline{w}_{j+1}} &= G_{j+1} H_j P_{\underline{a}_j,\underline{a}_j} G_j^T \end{aligned} \tag{4.13}$$

The extension of this process to the covariance structure for 2-D wavelet coefficients needs to repeat the above processes for each row or each column of the wavelet coefficient matrix.

Having this well-defined variance-covariance structure of the wavelet coefficients, one can exactly assess the extent of correlation between the coefficients at the same scale or different resolutions. The very first observation one can make from (4.13) is that the appearance of H and G matrices in the recursive computing of each wavelet local covariance indicates connectivity of all wavelet coefficients through these two operators.

A time domain model of correlation should be assumed to be re-assessed in the wavelet domain. An exponential correlation structure is common for real signals and remotely-sensed fields [50], so it is assumed that the second-order statistics of the finest-scale signal $\underline{x}$ is given by $\underline{x} \sim (\underline{0}, P_x)$, that is, $\underline{x}$ has zero mean and covariance structure

$$P_{x_i,x_j} = cov(x_i, x_j) = \sigma_x^2 \exp\,(-\frac{|i-j|}{\tau}) \tag{4.14}$$

with parameter τ controlling the correlation length between two pixels.[2] The chosen distribution has constant correlation length and is spatially stationary; this assumption is for convenience only and is not fundamental to the analysis.

[2] The 1-D membrane GMRF model of (4.6) which is known to be exponentially distributed [38] can also be considered, whose respective covariance has periodic boundaries more consistent with structures of the GMRFs examined in § 4.2.2.

With the covariance structure P_x determined, one can transform it into the wavelet domain by computing the wavelet operator $\mathcal{W}$, containing all translated and dilated versions of the selected wavelet basis functions, *e.g.*,

$$\begin{bmatrix} \underline{a}^J \\ \underline{w}^J \\ \vdots \\ \underline{w}^1 \end{bmatrix} = \mathcal{W}\underline{x} \tag{4.15}$$

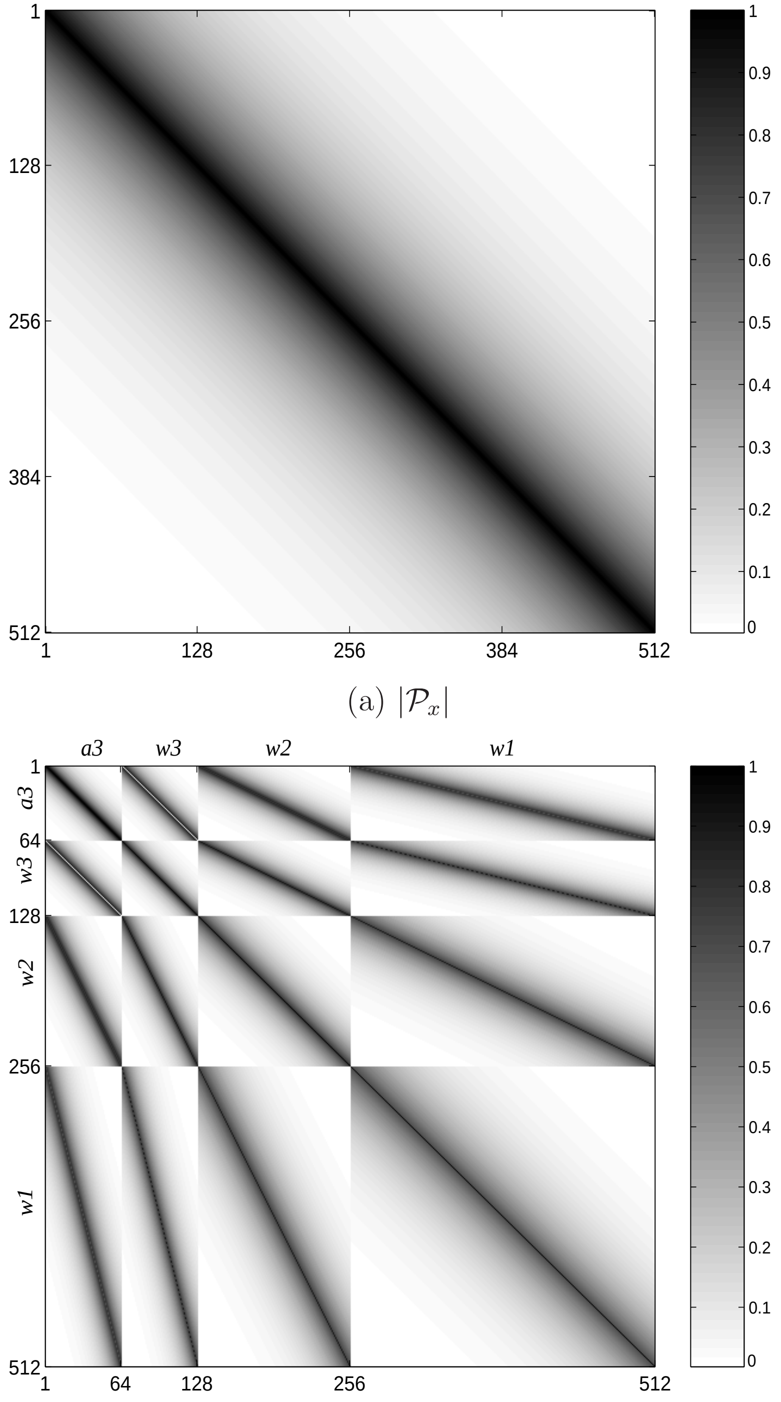
The covariance structure of the wavelet-decomposed signal is then

$$P_w = \mathcal{W} P_x \mathcal{W}^T \tag{4.16}$$

The covariance matrix P_w is normalized to $\mathcal{P}_w$ by (4.11) so that the auto- and cross-correlations of the coefficients at different scales are obtained. Figure 4.5 displays correlation structure for a 1-D signal with exponential joint statistics (4.14) in both time domain: $\mathcal{P}_x$, and "db1" wavelet domain: $\mathcal{P}_w$.[3] The resulting wavelet correlation is a block matrix, with the block diagonals showing the within-scale autocorrelations and off-diagonal blocks presenting the cross-correlations between $\underline{a}^J$ and $\{\underline{w}^j\}, 1 \leq j \leq J$, at different resolutions.

Figure 4.6 zooms into a small fraction (structure) of the correlation matrix display in Figure 4.5(b) and summarizes the magnitudes of the covariance values between a typical detail coefficient w and its spatially local neighbors, both within the same scale and across scales. It is seen most clearly that the within-scale correlations tend to decay very quickly, while the dependencies across different resolutions surprisingly remain strong, even for the

[3] For the purpose of illustrations in this thesis the results of simulation with the piecewise linear family of the first three members of Daubechies wavelet family are displayed. Simulations with other commonly used basis functions such as longer Daubechies wavelet bases, and more regular wavelets of Coiflet [49] and Meyer [67], exhibit a stronger decorrelation effect within scale, nevertheless the qualitative structure is similar, and the across-scale correlations are no less significant.



(a) $|\mathcal{P}_x|$

(b) $|\mathcal{P}_w|$

Figure 4.5: Time- and wavelet-domain correlation structures for exponentially correlated signals. $\mathcal{P}_w$ was obtained from (4.16) by using a three-level wavelet operator $\mathcal{W}$ of Haar basis function. The resulted P_w was normalized to correlation coefficient matrix $\mathcal{P}_w$ and its absolute value is shown above.

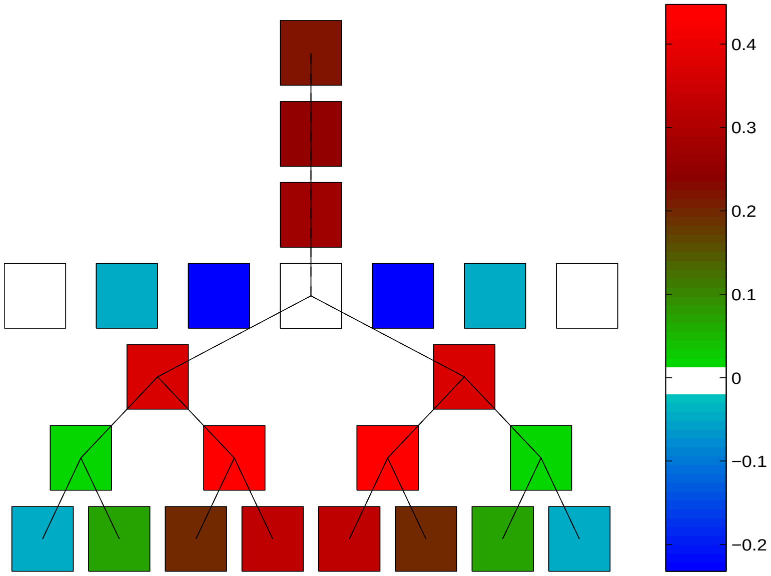


Figure 4.6: The extent of correlation between a typical 1-D wavelet coefficient at scale j and its adjacent coefficients within the same scale and across several resolutions towards both parents and children. For the purpose of demonstration the centered coefficient variance value was set to zero.

coefficients located several scales apart. This result confirms that: *the wavelet coefficients are uncorrelated and in some cases their correlation can be quite significant!*

This statistical result of 1-D wavelet coefficient inter-correlations is considered as a motivating point to investigate wavelet correlation structures for higher dimensional data.

4.2.2 Examining two-dimensional signals

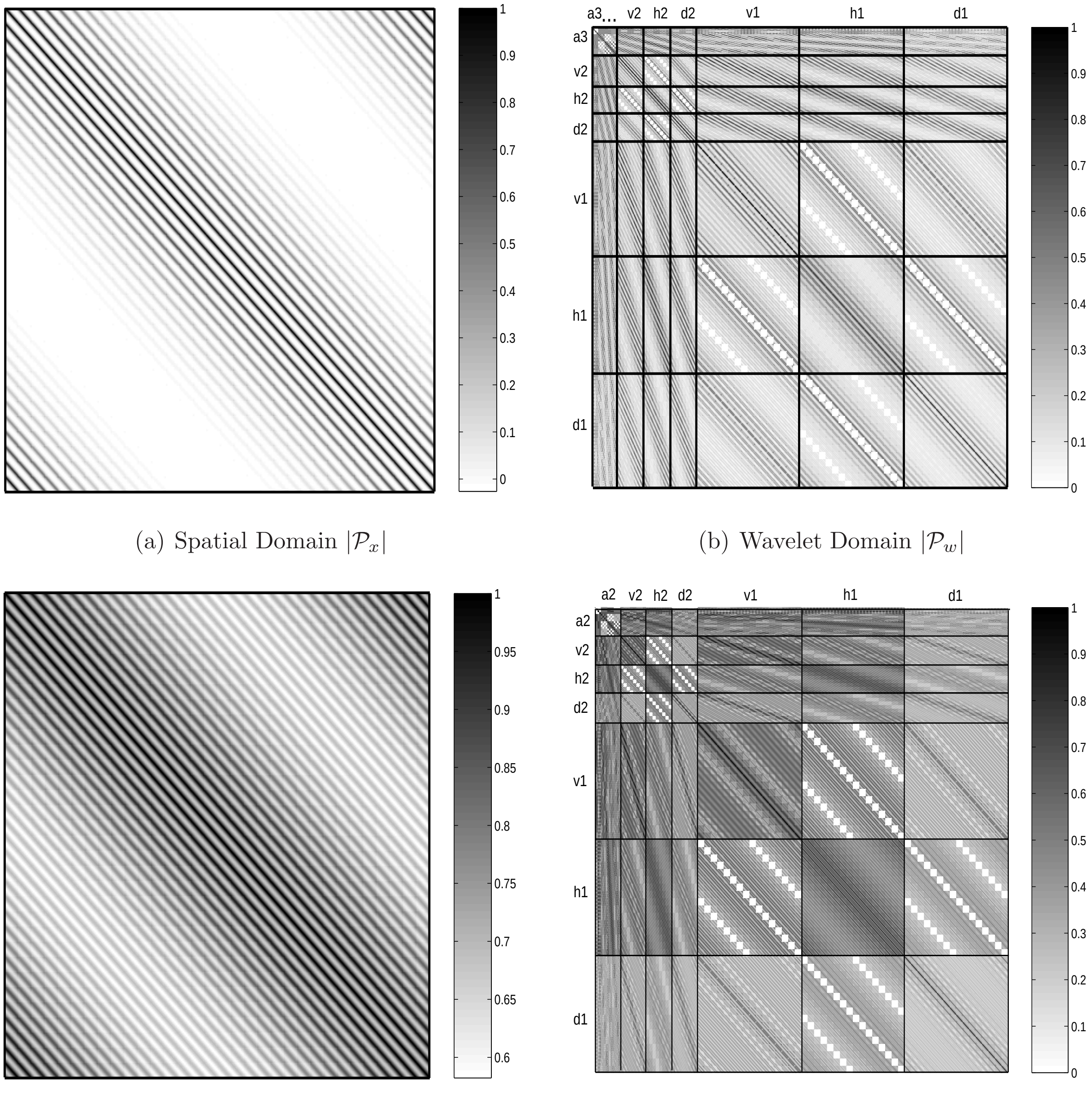
The joint-statistical study of 1-D wavelet coefficients can be extended to that of the 2-D wavelet transform. To perform an empirical study of wavelet statistics with an assumed spatial prior, one needs to examine the covariance matrix for the given random field. Any $n \times n$-sized image has an $n^2 \times n^2$-sized covariance matrix. Consequently, the limitations caused by huge covariance matrices in terms of computational time and space can affect the problem size. Due to the dramatic increase in covariance matrix size, the empirical results discussed herein were obtained by examining the correlation structure of small sized images (*e.g.*, 32×32 pixels) with reasonably sized covariance matrices (*e.g.*, $32^2 \times 32^2$).

The wavelet operator $\mathcal{W}$ is defined by, first, lexicographically re-ordering the 2-D image is into a 1-D vector format (*e.g.*, column-wise) as is shown in Figure 4.2. This pixel arrangement keeps the vertically adjacent image pixels next to each other, while horizontal neighbors are located n pixels apart. This phenomenon (the vicinity of vertical elements and the segregation of horizontal pixels) still remains clearly visible when the associated P_x is projected into the wavelet domain, *i.e.*, P_w, by (4.16).

Figure 4.7 illustrates the spatial (identical to those shown in Figure 4.3) and wavelet domain correlation structures of a thin-plate GMRF with two different correlation lengths. The main diagonal blocks in $\mathcal{P}_w$ show autocorrelation of coefficients correspond to the same scale and orientation, whereas off-diagonal blocks illustrate cross-correlations across orientations or across scales. To zoom in some detail coefficients inter-relationships given by matrix P_w, Figure 4.8 presents the two-dimensional parallel of Figure 4.6, showing the correlation pattern for a typical horizontal detail coefficient.

The wavelet covariance P_w, shown above, is not a diagonal matrix, indicating that the wavelet coefficients are *not*, in fact, independent. Indeed, it is well known that localized image structures, such as edges, tend to have substantial power across many scales. We have observed [7] that, although the majority of correlations are very close to zero, a relatively significant percentage (10%) of the coefficients are strongly correlated across several scales or across orientations.

Since the wavelet statistical results illustrated by Figure 4.7 and 4.8 play a critical role in this research study which focuses on developing wavelet joint probabilistic models, the next section is devoted to performing a comprehensive examination and understanding of



(a) Spatial Domain $|\mathcal{P}_x|$

(b) Wavelet Domain $|\mathcal{P}_w|$

(c) Spatial Domain $|\mathcal{P}_x|$

(d) Wavelet Domain $|\mathcal{P}_w|$

Figure 4.7: Correlation coefficient absolute values of the thin-plate GMRF model in the spatial and wavelet domains. Top panels: $|\mathcal{P}_x|$ and its associated $|\mathcal{P}_w|$ for the prior with short correlation length. Bottom panels: $|\mathcal{P}_x|$ and $|\mathcal{P}_w|$ for the prior with long correlation length.

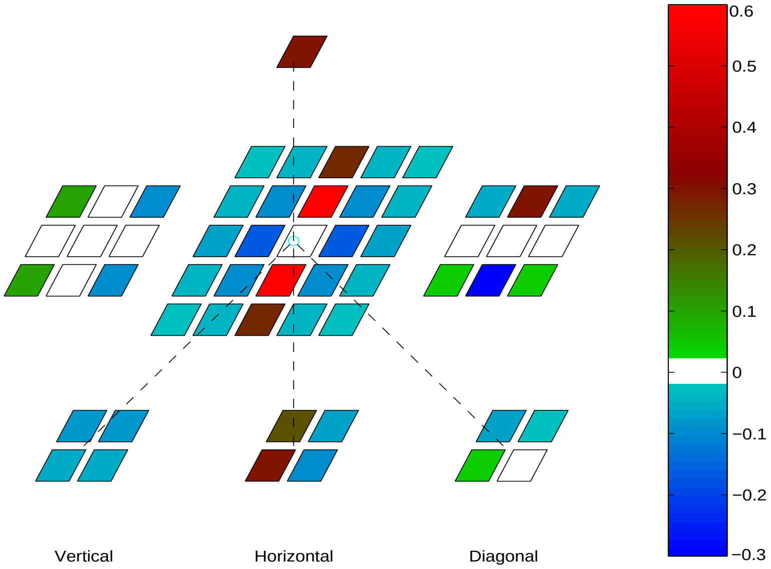


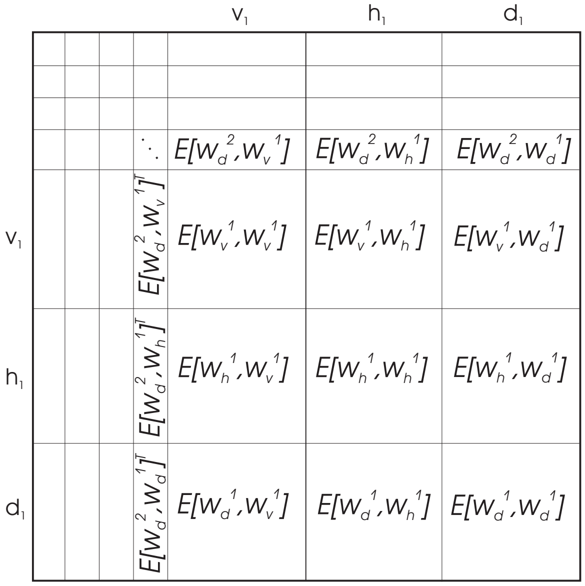
Figure 4.8: Summary of correlation between a horizontal coefficient and its spatially local neighbors at the same scale, but different orientations and across scales but the same orientation. For the purpose of demonstration the centered coefficient variance value was set to zero.

wavelet correlations.

4.3 Wavelet Covariance Matrix

In order to study wavelet correlations exactly various statistical GMRF textures $\underline{x}$ with known covariance P_x, as shown in Figure 4.10 are considered. The fine scale texture $\underline{x} \equiv \underline{x}^0$ has a 2-D wavelet decomposition

$$\mathcal{W}\underline{x}^0 = \begin{bmatrix} \underline{a}^J \\ \underline{w}^J \\ \vdots \\ \underline{w}^1 \end{bmatrix} \tag{4.17}$$



					v_1	h_1	d_1
				$\ddots$	$E[w_d^2, w_v^1]$	$E[w_d^2, w_h^1]$	$E[w_d^2, w_d^1]$
v_1				$E[w_d^2, w_v^1]^T$	$E[w_v^1, w_v^1]$	$E[w_v^1, w_h^1]$	$E[w_v^1, w_d^1]$
h_1				$E[w_d^2, w_h^1]^T$	$E[w_h^1, w_v^1]$	$E[w_h^1, w_h^1]$	$E[w_h^1, w_d^1]$
d_1				$E[w_d^2, w_d^1]^T$	$E[w_d^1, w_v^1]$	$E[w_d^1, w_h^1]$	$E[w_d^1, w_d^1]$

Figure 4.9: Schematic plot of the wavelet covariance matrix P_w. The main diagonal blocks show autocorrelation of coefficients correspond to the same scale and orientation, whereas off-diagonal blocks illustrate cross-correlations across orientations or across scales.

where

$$\underline{w}^j = \begin{bmatrix} \underline{w}_h^j \\ \underline{w}_v^j \\ \underline{w}_d^j \end{bmatrix} \tag{4.18}$$

contains the three orientation subbands at scale j, and where $\underline{a}^J$ represents the scaling coefficients at the coarsest scale J. The wavelet operator $\mathcal{W}$ is linear, therefore the chosen spatial domain image $\underline{x}$ and its covariance structure P_x can be projected into the wavelet domain via (4.16). As before, and because it is so widely used in wavelet HMMs, the Daubechies class of wavelets is considered.

Figure 4.9 plots a schematic view of P_w by labelling each block as a sub-covariance showing expectation between two subbands. Figure 4.11 shows correlation structures for two textures displayed in Figure 4.10 in the spatial domain as well as in the wavelet domain. The clear block structure governing the wavelet covariance P_w is due to various subbands carrying high frequency components, *i.e.*, edges with different orientations. Blocks representing within-scale correlations (auto- or cross-) are square in shape, due to equality between within-scale subband sizes. However, cross-correlations across scales with subbands of different sizes result in rectangular sub-covariances.

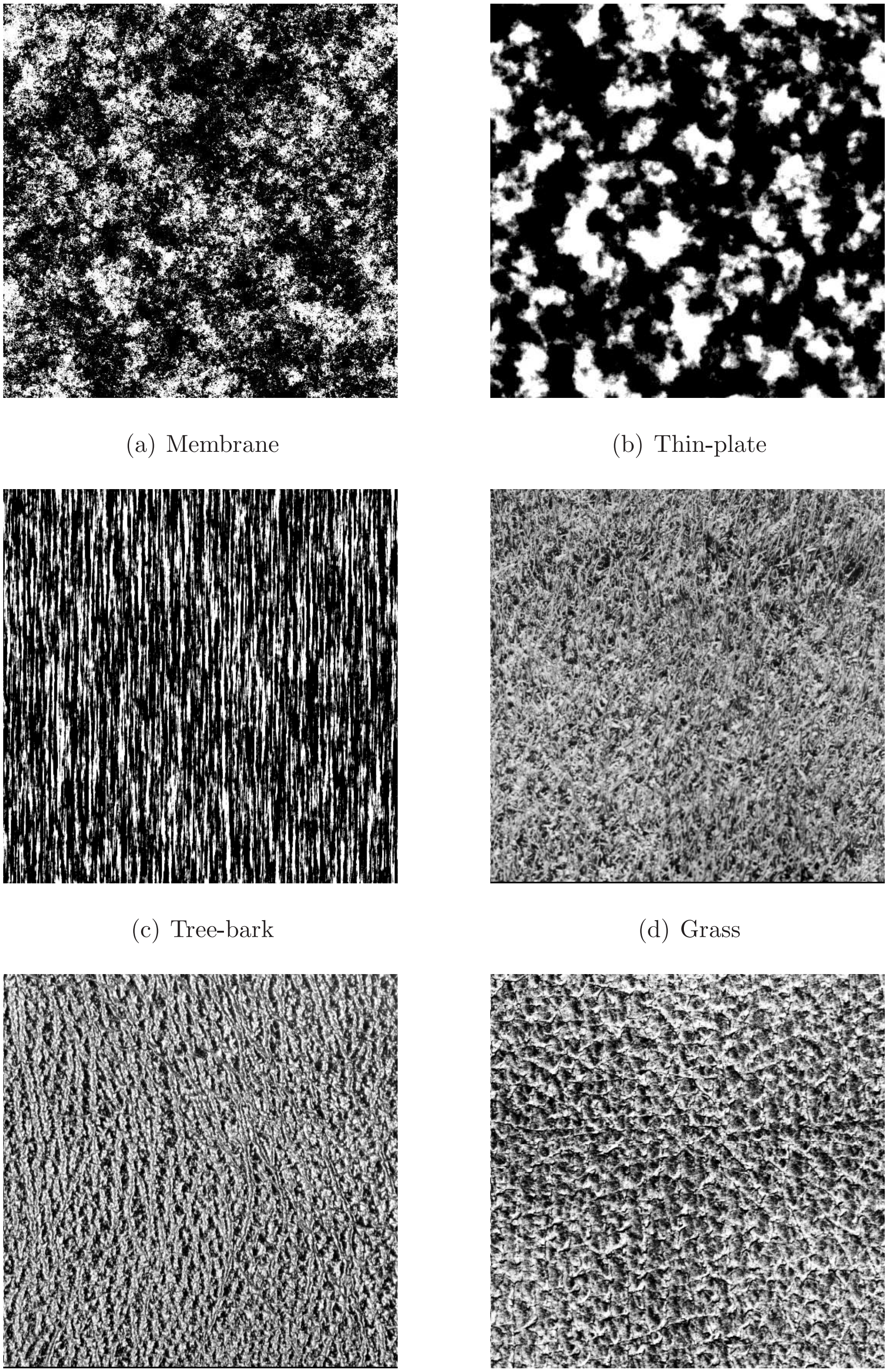
The above illustrations of the wavelet domain covariances lead to the following observations:

- Diagonal sub-covariances in P_w: The diagonal blocks show the auto-correlation of within-scale coefficients located at the horizontal, vertical and diagonal orientations, respectively. Due to the column-wise 2-D to 1-D data stacking (Figure 4.2), large magnitude auto-correlations of the vertical coefficients tend to concentrate near the main diagonal, whereas those of the horizontal coefficients are distributed on the diagonals n pixels apart.

- Within-subband correlations: It is significant to notice that regardless of the orientation, the large magnitude correlations at any subband are basically arranged in lines following the orientation of the subband. The bottom panel in Figure 4.11 shows P_x and P_w for tree-bark texture (shown in Figure 4.10) with strong spatial domain correlation among vertical neighbors. Consequently, the strong wavelet correlation among vertical subband coefficients is apparent.

- Across-subband correlations: The off-diagonal square blocks represent within-scale correlation between coefficients at the same spatial positions but from horizontal, vertical, or diagonal directions (*i.e.*, cousins).

- Across-scale correlations: In addition to the square blocks (representing within-scale



(a) Membrane

(b) Thin-plate

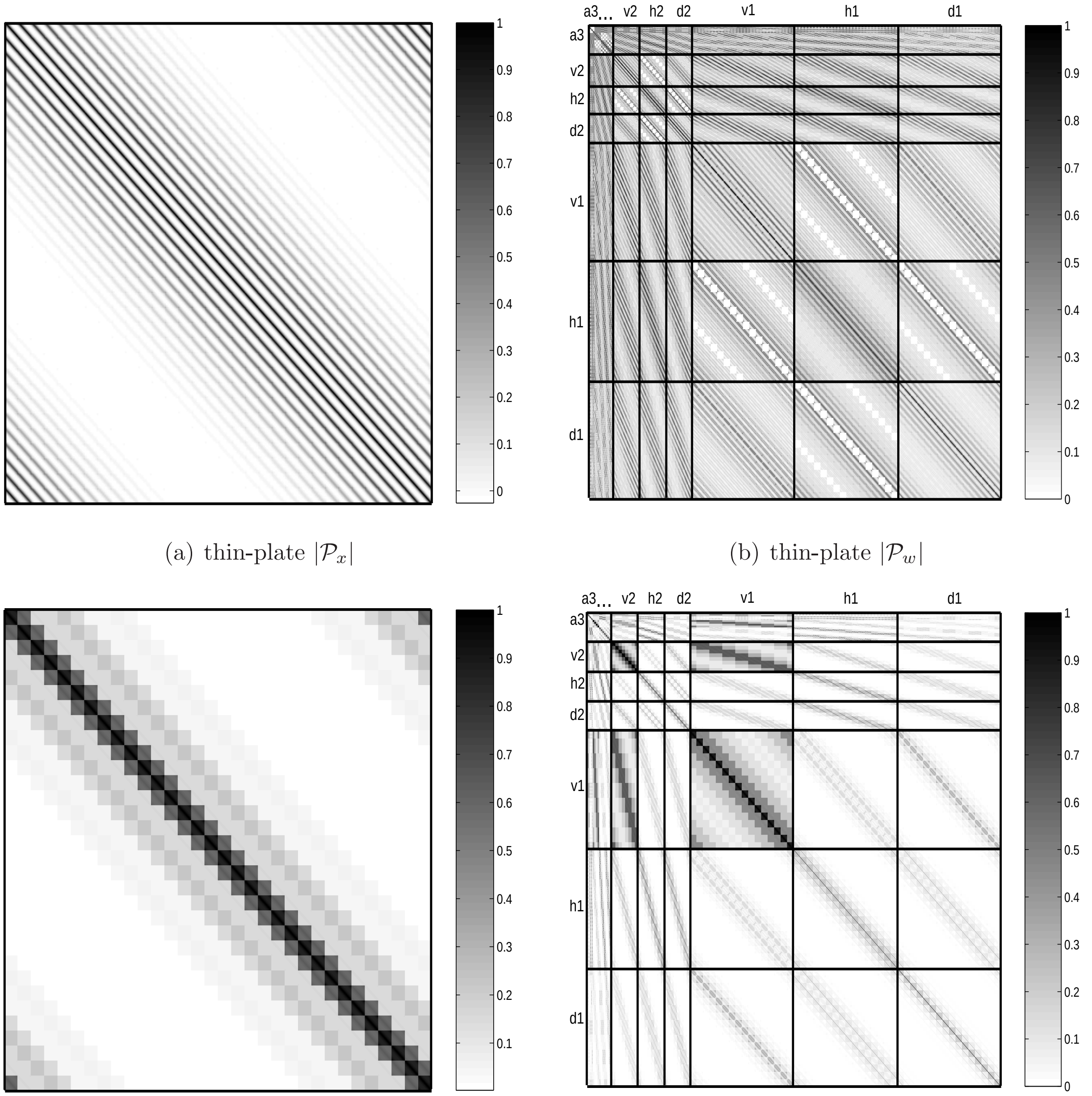
(c) Tree-bark

(d) Grass

(e) Calf leather

(f) Pigskin

Figure 4.10: Six GMRF textures used to generate wavelet statistics.



(a) thin-plate $|\mathcal{P}_x|$

(b) thin-plate $|\mathcal{P}_w|$

(c) tree-bark $|\mathcal{P}_x|$

(d) tree-bark $|\mathcal{P}_w|$

Figure 4.11: Correlation coefficient magnitudes of the thin-plate and tree-bark models in the spatial and db2 wavelet domains. The main diagonal blocks in $\mathcal{P}_w$ show autocorrelation of coefficients correspond to the same scale and orientation, whereas off-diagonal blocks illustrate cross-correlations across orientations or across scales.

localities), the rectangular blocks exhibit significant correlations between subbands across different resolutions, even for subbands located at several scales apart.

As is obvious from these results, there is a clear locality to the correlation structures both within and across scales, and so the wavelet coefficients are proposed to be modeled not as independent, but as governed by a random field. Just as with the HMM methods, described in § 3.4, the neighborhood structures exhibited by the wavelet statistics must be assessed.

4.3.1 Numerical Experiments with the Wavelet Correlations

So far, the first goal – to emphasize the non-whitening property of the WT – has been met. This important observation and the block structure of the wavelet domain prior P_w led this work toward a series of numerical studies which are discussed here. In this section, several exhaustive tests on P_w are performed to characterize the underlying significant correlation structure and to make its content values and relationship more meaningful. It is generally infeasible to directly utilize the huge covariance matrix P_w in an estimation process, due to space and time complexities. The goal is to study the properties of P_w in order to deduce a simple, but still accurate, representation of the underlying correlation model; that is, to construct a new sparse covariance matrix, which contains the most significant information from the prior model. Of course, the study of large covariance matrices is for diagnostic and research purposes; ultimately any practical estimation algorithm will be based on some implicit sparse model of the statistics.

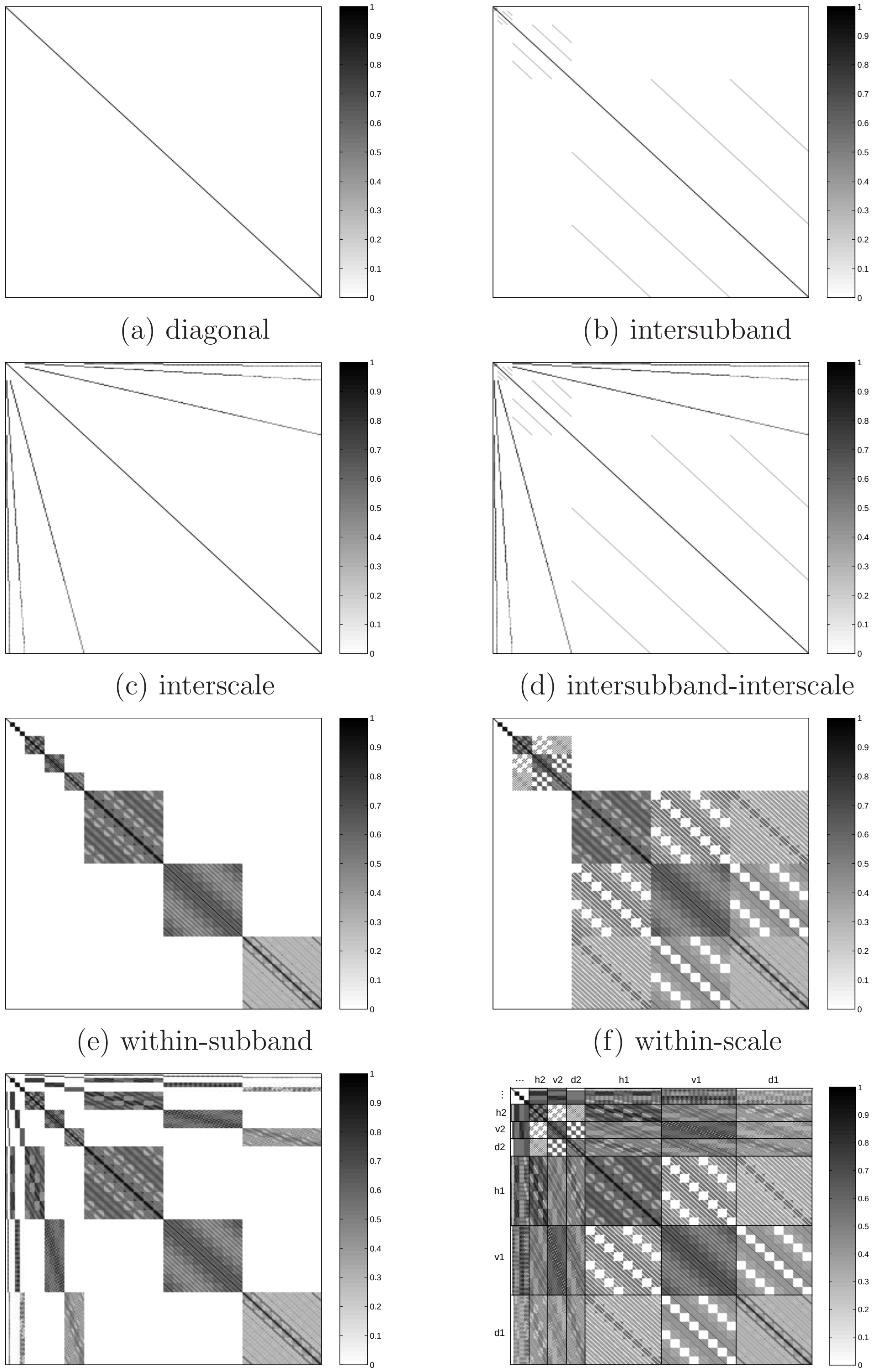
In these experiments the wavelet coefficients are treated in various ways, from complete independence to full dependency among all coefficients over the entire wavelet tree. As is shown in Table 4.1, eight different (amongst many) cases of adding more features to the covariance matrix may be considered. For each case, except the diagonal case, at least one

Figure 4.12 panels	**within subband**	**across subbands**	**across scales**
(a) diagonal	no	no	no
(b) intersubband	no	yes	no
(c) interscale	no	no	yes
(d) intersubband-interscale	no	yes	yes
(e) within-subband	yes	no	no
(f) within-scale	yes	yes	no
(g) intrasubband-interscale	yes	no	yes
(h) full	yes	yes	yes

Table 4.1: Eight different ways to obtain a new wavelet-based covariance structure which contains a combination of three important neighborhood correlation factors, namely intra-orientation, intra-scale, and inter-scale.

of the three important neighborhood correlation maps – within-subband, within-scale, and across-scale – is considered.

Figure 4.12 visualizes all eight structures obtained from the original correlation matrix P_w in Figure 4.7. The test is started with the simplest case, in which all the wavelet coefficients are treated as being independent. Theoretically, this means that all off-diagonal entities of P_w are zero as is depicted in Figure 4.12(a). Note that simple wavelet-based algorithms, such as point-wise shrinkage [33], in which the coefficients are treated as statistically decorrelated, only consider the diagonal entries of the covariance matrix. This is obviously a weak assumption which ignores the fact that significant interactions remain between the coefficients. Hence, one needs to seek a structure which carries the most significant correlations between wavelet coefficients and in a spare representation. It is important to notice that the cross-correlation of any two subbands can be eliminated by simply re-



(a) diagonal

(b) intersubband

(c) interscale

(d) intersubband-interscale

(e) within-subband

(f) within-scale

(g) intrasubband-interscale

(h) Full

Figure 4.12: The correlation structures from Table 4.1. As within-scale dependencies are considered (e), the structural density increases dramatically. The across scale correlations (g) add significant information, but have less impact on density increment.

placing their corresponding correlation block in P_w by zero. The structures displayed by Figure 4.12(f) and (g) indicate that adding intra-scale correlations increases the structure's density (Figure 4.12(f)) much more than the inter-scale dependencies (Figure 4.12(g)).

It is important to notice that there are many possible combinations of correlation structures that one could examine. However, we are limited with the positive definiteness of the resulted matrices. Among the structures tested here, Figure 4.12(c) and (d) are not positive definite, thus are not used in estimator error evaluations discussed next.

As is illustrated in Figure 4.13, each variation shown in Figure 4.12 clearly will differ in its complexity (matrix density) and statistical accuracy, a comparison which is discussed below for the standard image denoising problem.

4.3.2 Bayesian Estimate: A Quantitative Evaluation

A simple estimation algorithm is adopted here to evaluate and compare the achieved various wavelet statistical structures. To exploit the above dependency maps we implement a method that estimates the original coefficients by explicit use of wavelet covariance structure.

Define the noisy observation $\underline{y}$ as

$$\underline{y} = \underline{x} + \underline{\nu}, \qquad \underline{x} \sim (\underline{0}, P_x), \qquad \underline{\nu} \sim \mathcal{N}(\underline{0}, R), \tag{4.19}$$

where its wavelet counterpart is

$$\underline{w}_y = \underline{w}_x + \underline{w}_\nu, \qquad \underline{w}_x \sim (\underline{0}, P_w), \qquad \underline{w}_\nu \sim \mathcal{N}(\underline{0}, R). \tag{4.20}$$

The Bayesian Least Square (BLS) method which directly takes into account the covariance structure is

$$\begin{aligned} \underline{\hat{x}} &= \arg_{\underline{\hat{x}}} \min\{E[(\underline{x} - \underline{\hat{x}})(\underline{x} - \underline{\hat{x}})^T | \underline{y}]\} \\ \underline{\hat{x}} &= P_x(P_x + R)^{-1}\underline{y} \end{aligned} \tag{4.21}$$

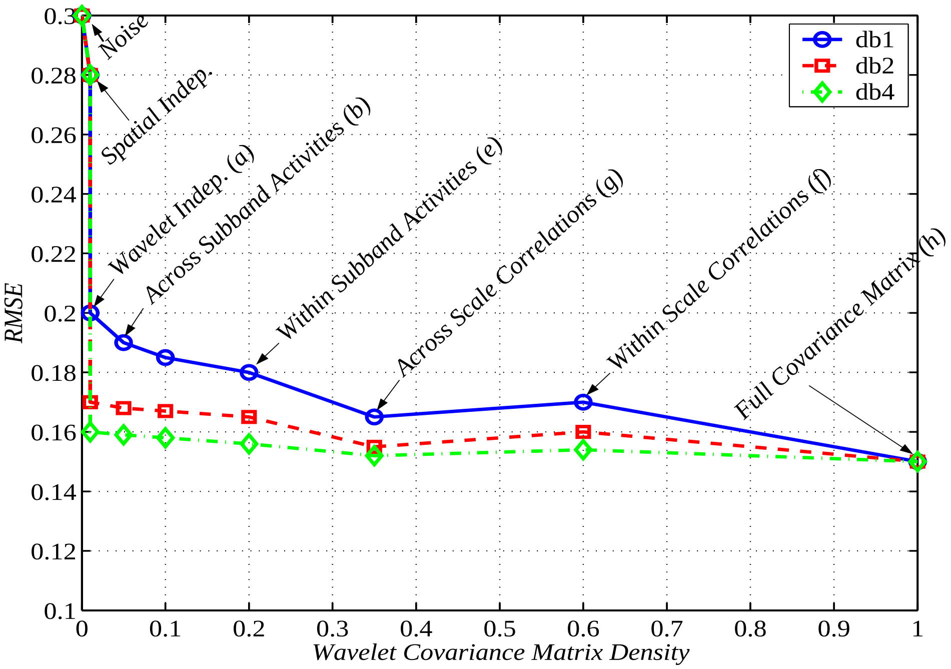


Figure 4.13: RMSE plot as a function of covariance density. It is evident how a tiny fraction of coefficients already provides the majority of the improvement. Labels match the panels in Figure 4.12.

The goal is to estimate $\underline{w}_x$ from noisy observation $\underline{w}_y$, where the additive noise $\underline{w}_\nu$ is decorrelated with the original data $\underline{w}_x$. Therefore the BLS may be applied in the wavelet domain. Because of the linearity and orthogonality of the WT, it is necessary to substitute (4.16) into (4.21). Then the orthogonal wavelet transform of the BLS method is obtained as

$$\begin{aligned} \underline{\hat{w}}_x &= \mathcal{W}P_x\mathcal{W}^T(\mathcal{W}P_x\mathcal{W}^T + \mathcal{W}R\mathcal{W}^T)^{-1}\mathcal{W}\underline{y} \\ \underline{\hat{w}}_x &= P_w(P_w + R)^{-1}\underline{w}_y \\ \therefore\ \underline{\hat{x}} &= \mathcal{W}^{-1}[P_w(P_w + R)^{-1}\underline{w}_y] \end{aligned}$$

In order to perform appropriate comparisons all structures of P_w illustrated in Figure 4.12 are considered in the BLS framework, except those shown in Figure 4.12(c),(d), due to not being positive definite. The resulting estimation error is obtained as

$$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n^2} \tilde{P}_w(i,i)} \qquad (4.22)$$

where

$$\tilde{P}_w = (P_w^{-1} + R^{-1})^{-1}$$

is the estimation error covariance [38].

Figure 4.13 displays the RMSE noise reduction achieved as more correlations are taken into the estimation process:

- The RMSE performance is not necessarily monotonic in matrix density!
- The vast bulk of the benefit is to be gained from relatively few coefficients.
- There is a significant RMSE reduction when across-scale correlations are considered.
- Larger extent of intra-scale dependencies, however, does not lead to significant RMSE reduction. This fact confirms the earlier discussion of reducing the within-scale neighborhood dependency in the model.

It is obvious that modeling the complete joint probability density function of wavelet coefficients $f(\underline{w})$ is computationally intractable. On the other hand, the statistical independence assumption, *i.e.*, $f(\underline{w}) = \prod_i f(w_i)$ ignores the coefficients interconnections. The need to assume dependencies among the coefficients is, thus, obvious and the extent of wavelet dependencies can be deduced from the above numerical simulations, which indicate that, surprisingly, a large fraction of intra-scale correlation values are very close to zero and there are only few significant within-scale correlations. This fact reveals the importance of taking into consideration a small within-subband correlation range, *e.g.*, 3×3 spatially located coefficients, along with a large extent of across-scale dependencies.

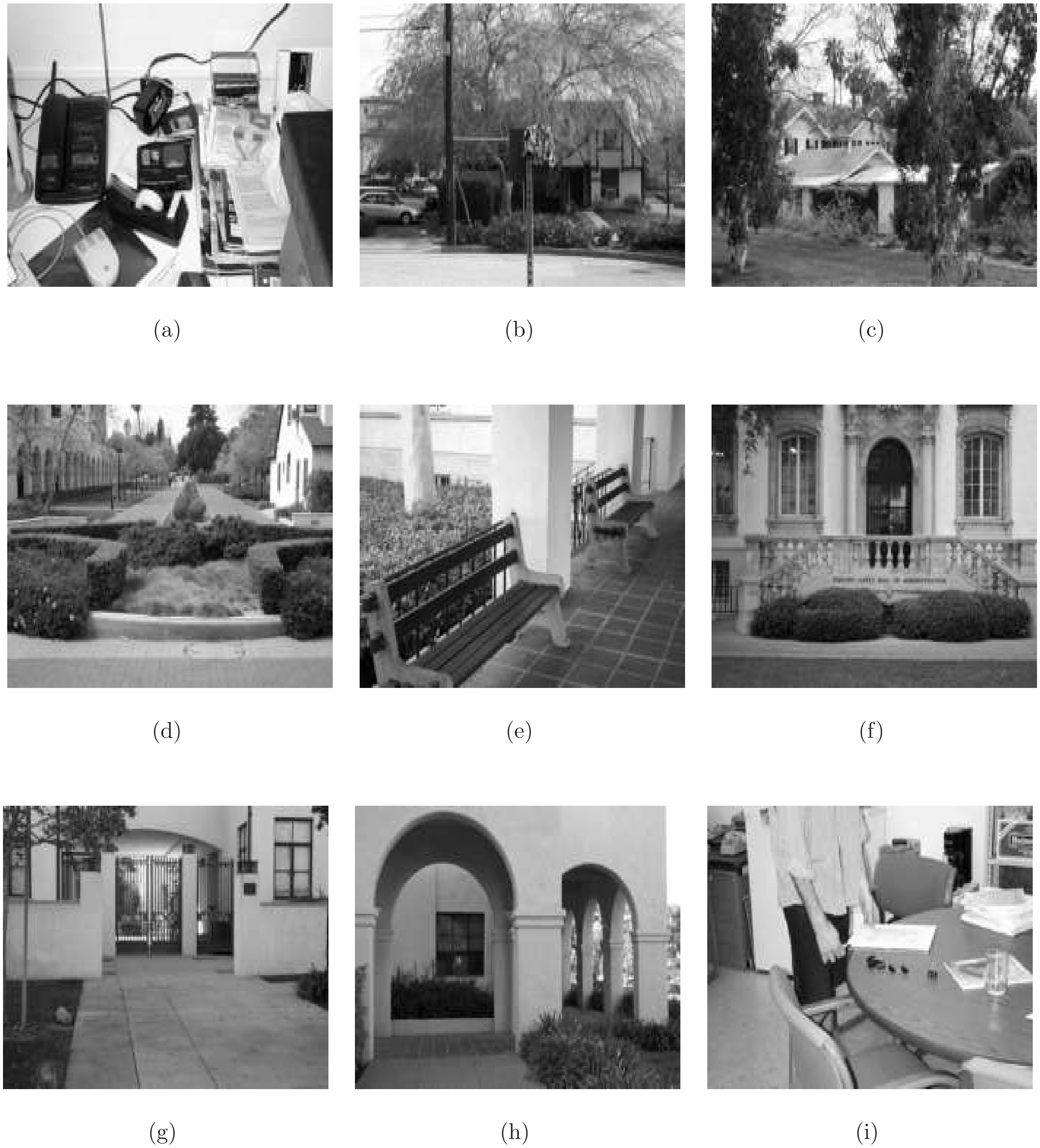
In summary, the goal is revisited: a study of probabilistic models which describe the wavelet random field with a small fraction of coefficients, but which accurately absorb each pixel's dependency on the rest of the wavelet tree. Therefore, the above empirical examinations will continue to highlight the most significant coefficient inter-relationships, which will lead to insights regarding hierarchical correlation models describing the wavelet statistics.

4.4 Wavelet Domain Joint Histograms

Because correlation can be a misleading indicator of statistical relationship (correlation may not be a good measure of degree of dependence, and uncorrelation does not mean independence), before making any further conclusions or drawing any map of wavelet statistics, examine the joint probability densities of pairs of coefficients, illustrated via joint histograms and wavelet dependencies via conditional histograms.

To obtain wavelet joint and conditional histograms a collection of 500 real-world images has been used.[4] Figure 4.14 displays sample images of this image set. The images were projected into the wavelet domain with Figures 4.15 and 4.16 demonstrating the joint histograms of a horizontal $w_h^j(x, y)$ or vertical $w_v^j(x, y)$ coefficient, respectively, with a chosen set of other significantly related coefficients. Figures 4.17 and 4.18 illustrate the conditional histograms of, respectively, an horizontal or vertical coefficient, given the magnitude of a coefficient within subband, across subbands and across scales. Each individual plot

[4] California Institute of Technology CVI Database: www.vision.caltech.edu/html-files/archive.html.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 4.14: Samples of real images used in the study of wavelet joint and conditional histograms.

corresponds to those shown in Figures 4.15 and 4.16.

The conditional distribution were also calculated for an ensemble of GMRF textures displayed in Figure 4.10 and are depicted in Figures 4.19 and 4.20. Each panel in these plots is parallel to its counterpart in Figures 4.17 and 4.18.

These plots highlight the following important aspects:

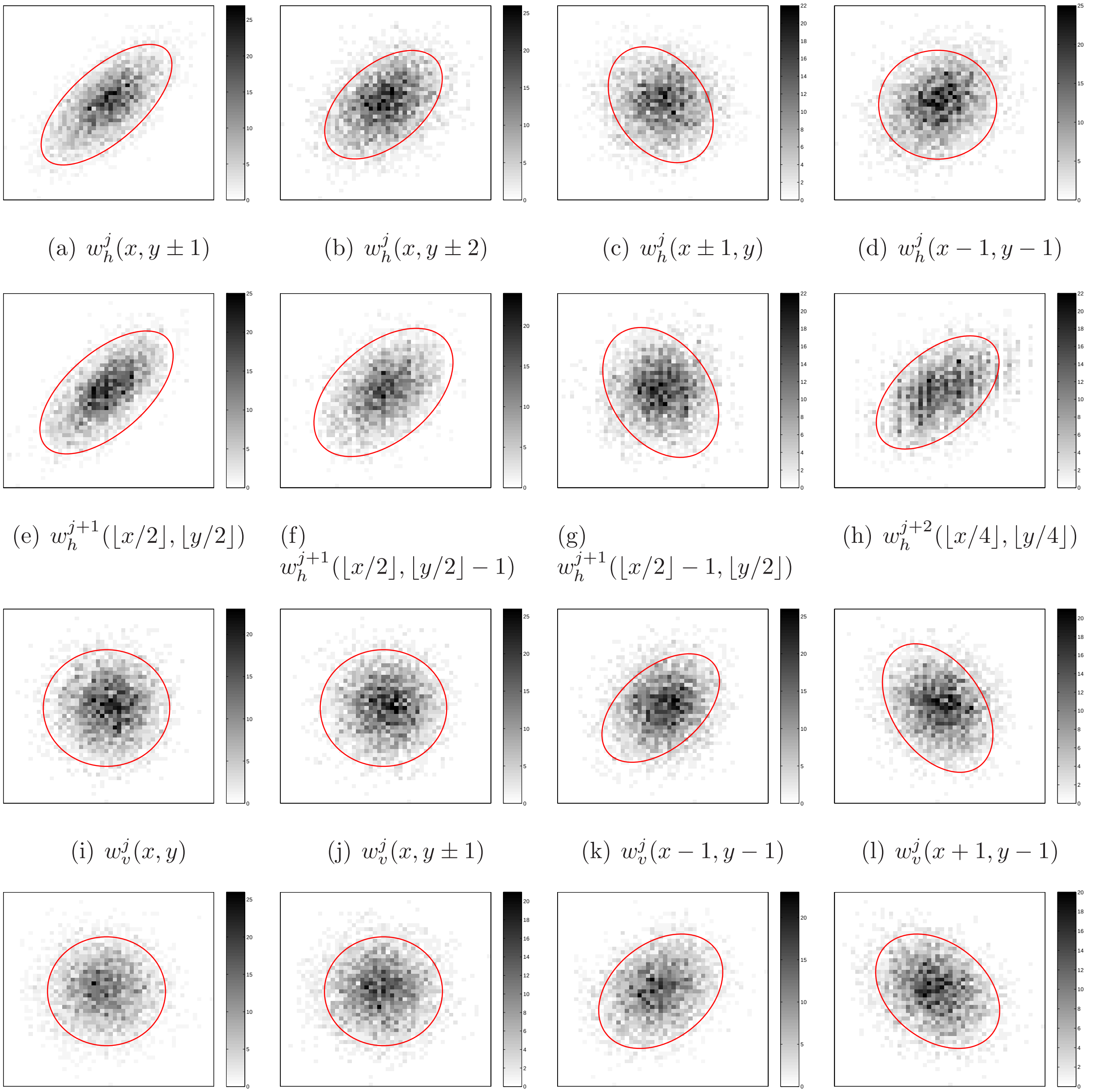
- Panels (a-d): the correlation direction of two spatially adjacent coefficients are a function of subband: within its subband, a horizontal coefficient is more correlated with its vertical neighbors than its horizontal ones. This observation is intuitive: the row elements in the horizontal channel result from the application of a high-pass filter, and are thus more decorrelated than the column elements which result from low-pass filtering. Similarly, an analogous vertical coefficient is more correlated with its horizontal neighbors. Surprisingly, the second-order neighbors are almost uncorrelated (panel d).

- Panels (e-h): a child coefficient strongly depends not only on its parent (a fact observed by many other researchers) but also on its parent's adjacent neighbors, vertically in Figure 4.15. By symmetry, in Figure 4.16 a vertical coefficient depends on its parent's horizontal neighbors.

- Panels (i & m): coefficients at the same location but from different orientations are essentially independent (panel i), directly at odds with most inter-orientation models!

- Panels (j-l) & (n-p): there is, however, inter-orientation correlation, but with pixels at other locations, dependent on the direction of the associated subband.

Thus, the expected child-parent relationships are confirmed, together with a strong subband dependence in the spatial correlations.



(a) $w_h^j(x, y \pm 1)$

(b) $w_h^j(x, y \pm 2)$

(c) $w_h^j(x \pm 1, y)$

(d) $w_h^j(x - 1, y - 1)$

(e) $w_h^{j+1}(\lfloor x/2 \rfloor, \lfloor y/2 \rfloor)$

(f) $w_h^{j+1}(\lfloor x/2 \rfloor, \lfloor y/2 \rfloor - 1)$

(g) $w_h^{j+1}(\lfloor x/2 \rfloor - 1, \lfloor y/2 \rfloor)$

(h) $w_h^{j+2}(\lfloor x/4 \rfloor, \lfloor y/4 \rfloor)$

(i) $w_v^j(x, y)$

(j) $w_v^j(x, y \pm 1)$

(k) $w_v^j(x - 1, y - 1)$

(l) $w_v^j(x + 1, y - 1)$

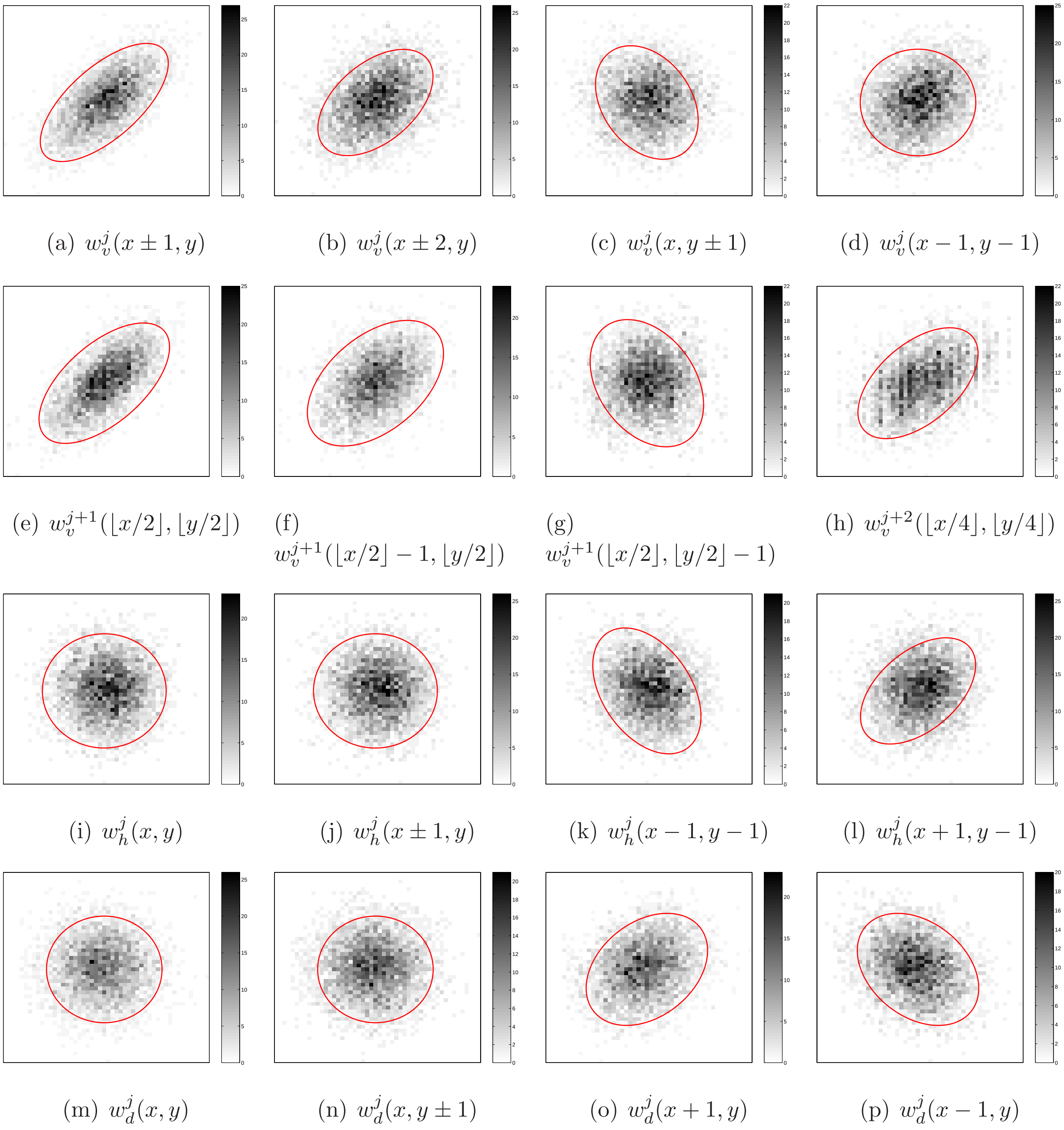
(m) $w_d^j(x, y)$

(n) $w_d^j(x \pm 1, y)$

(o) $w_d^j(x, y + 1)$

(p) $w_d^j(x, y - 1)$

Figure 4.15: Empirical joint histograms of a db2 horizontal coefficient of the real images and at position $w_h^j(x, y)$ associated with coefficients at the same scale and orientation (a-d), at the same orientation but adjacent scales (e-h), at the same scale but across orientations (i-p). The skewness in the ellipsoid indicates correlation.



(a) $w_v^j(x \pm 1, y)$

(b) $w_v^j(x \pm 2, y)$

(c) $w_v^j(x, y \pm 1)$

(d) $w_v^j(x - 1, y - 1)$

(e) $w_v^{j+1}(\lfloor x/2 \rfloor, \lfloor y/2 \rfloor)$

(f) $w_v^{j+1}(\lfloor x/2 \rfloor - 1, \lfloor y/2 \rfloor)$

(g) $w_v^{j+1}(\lfloor x/2 \rfloor, \lfloor y/2 \rfloor - 1)$

(h) $w_v^{j+2}(\lfloor x/4 \rfloor, \lfloor y/4 \rfloor)$

(i) $w_h^j(x, y)$

(j) $w_h^j(x \pm 1, y)$

(k) $w_h^j(x - 1, y - 1)$

(l) $w_h^j(x + 1, y - 1)$

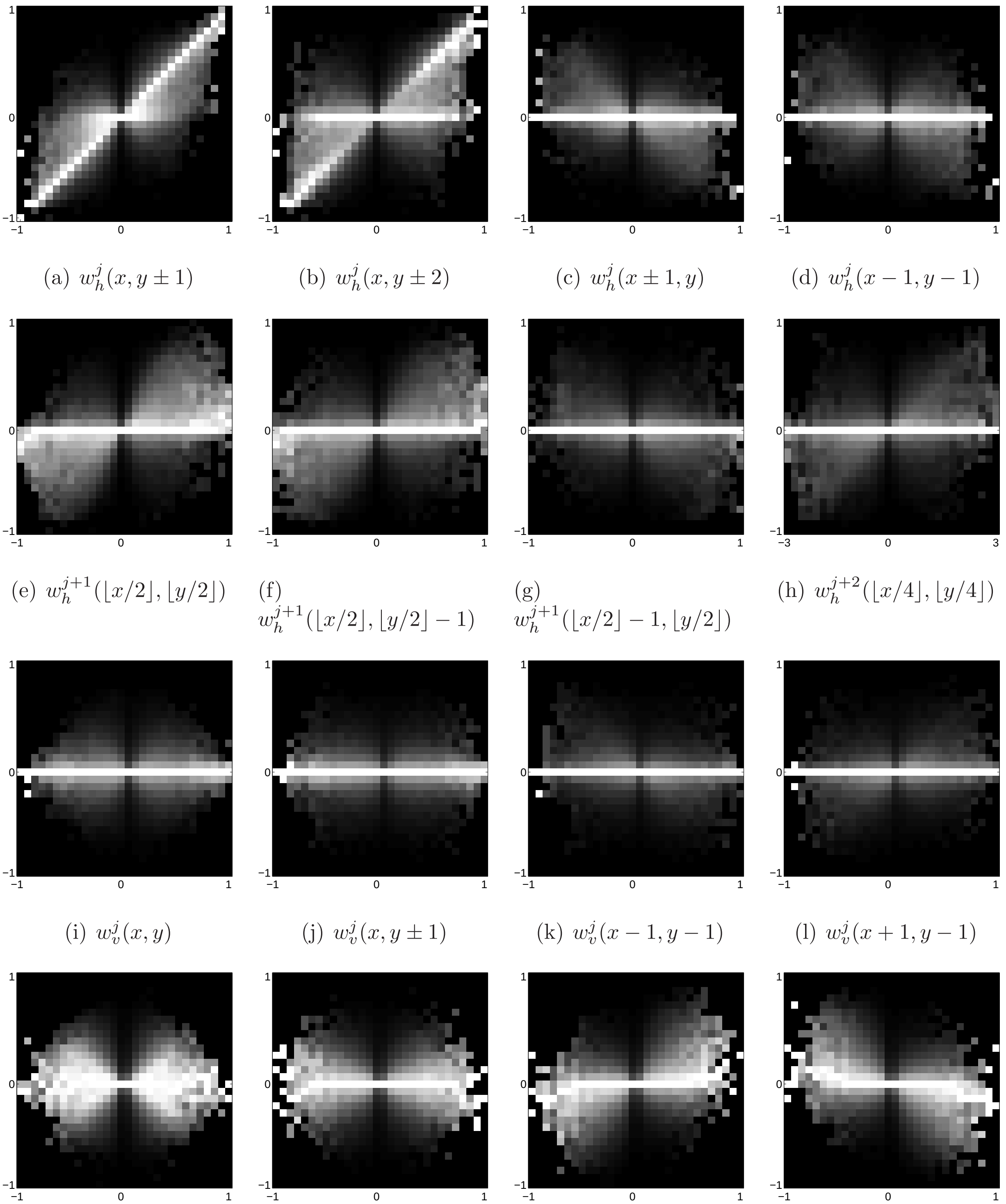
(m) $w_d^j(x, y)$

(n) $w_d^j(x, y \pm 1)$

(o) $w_d^j(x + 1, y)$

(p) $w_d^j(x - 1, y)$

Figure 4.16: Plots parallel to Figure 4.15 showing joint histograms of a db2 vertical coefficient at position $w_v^j(x, y)$. Dependencies for horizontal and vertical coefficients are symmetrically identical.



(a) $w_h^j(x, y \pm 1)$

(b) $w_h^j(x, y \pm 2)$

(c) $w_h^j(x \pm 1, y)$

(d) $w_h^j(x - 1, y - 1)$

(e) $w_h^{j+1}(\lfloor x/2 \rfloor, \lfloor y/2 \rfloor)$

(f) $w_h^{j+1}(\lfloor x/2 \rfloor, \lfloor y/2 \rfloor - 1)$

(g) $w_h^{j+1}(\lfloor x/2 \rfloor - 1, \lfloor y/2 \rfloor)$

(h) $w_h^{j+2}(\lfloor x/4 \rfloor, \lfloor y/4 \rfloor)$

(i) $w_v^j(x, y)$

(j) $w_v^j(x, y \pm 1)$

(k) $w_v^j(x - 1, y - 1)$

(l) $w_v^j(x + 1, y - 1)$

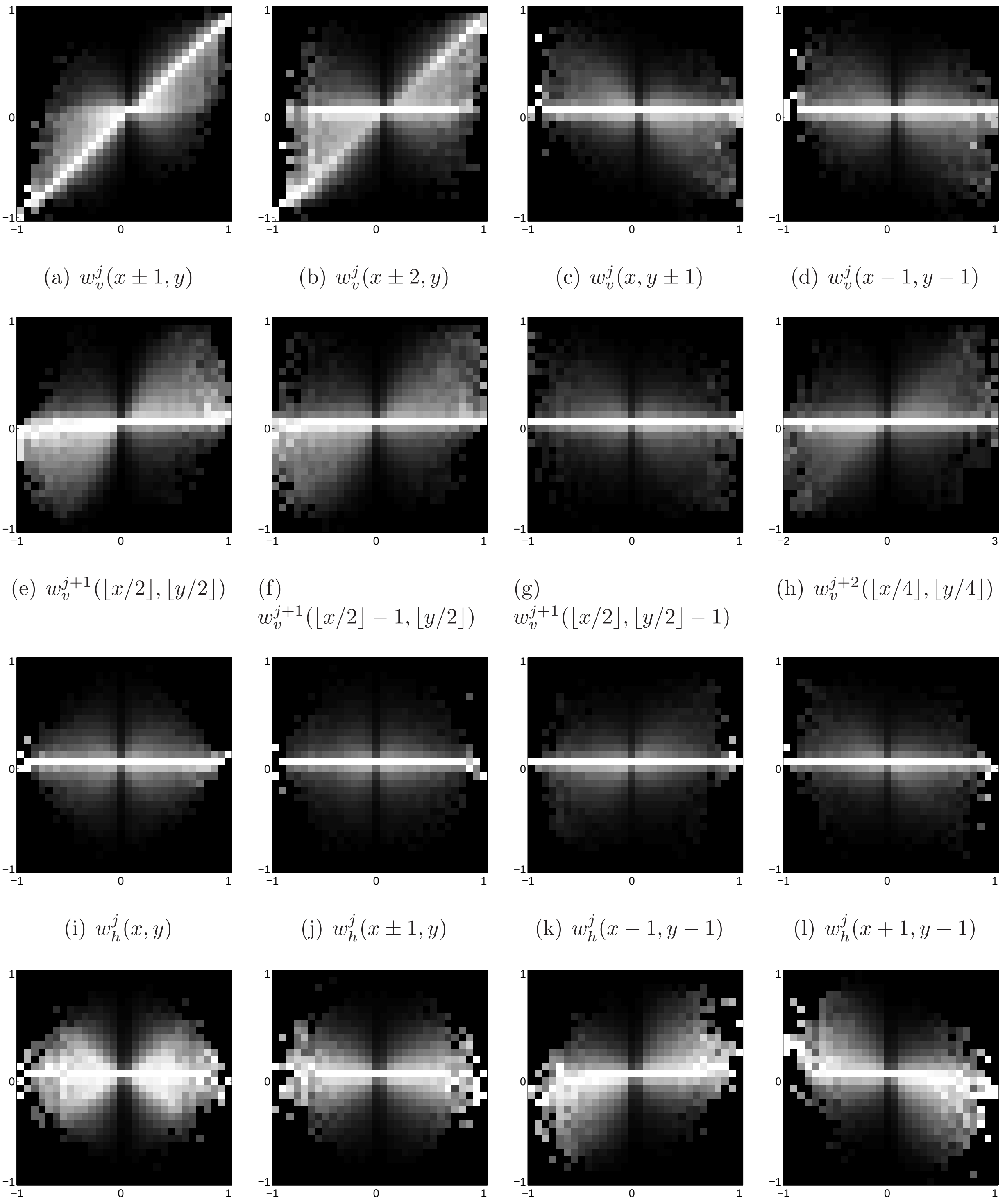
(m) $w_d^j(x, y)$

(n) $w_d^j(x \pm 1, y)$

(o) $w_d^j(x, y + 1)$

(p) $w_d^j(x, y - 1)$

Figure 4.17: Conditional histograms of a db2 horizontal coefficient corresponding to the plots in Figure 4.15. In each plot, brightness indicates probability, with each column being independently rescaled to cover the whole range of intensities.



(a) $w_v^j(x \pm 1, y)$

(b) $w_v^j(x \pm 2, y)$

(c) $w_v^j(x, y \pm 1)$

(d) $w_v^j(x-1, y-1)$

(e) $w_v^{j+1}(\lfloor x/2 \rfloor, \lfloor y/2 \rfloor)$

(f) $w_v^{j+1}(\lfloor x/2 \rfloor - 1, \lfloor y/2 \rfloor)$

(g) $w_v^{j+1}(\lfloor x/2 \rfloor, \lfloor y/2 \rfloor - 1)$

(h) $w_v^{j+2}(\lfloor x/4 \rfloor, \lfloor y/4 \rfloor)$

(i) $w_h^j(x, y)$

(j) $w_h^j(x \pm 1, y)$

(k) $w_h^j(x-1, y-1)$

(l) $w_h^j(x+1, y-1)$

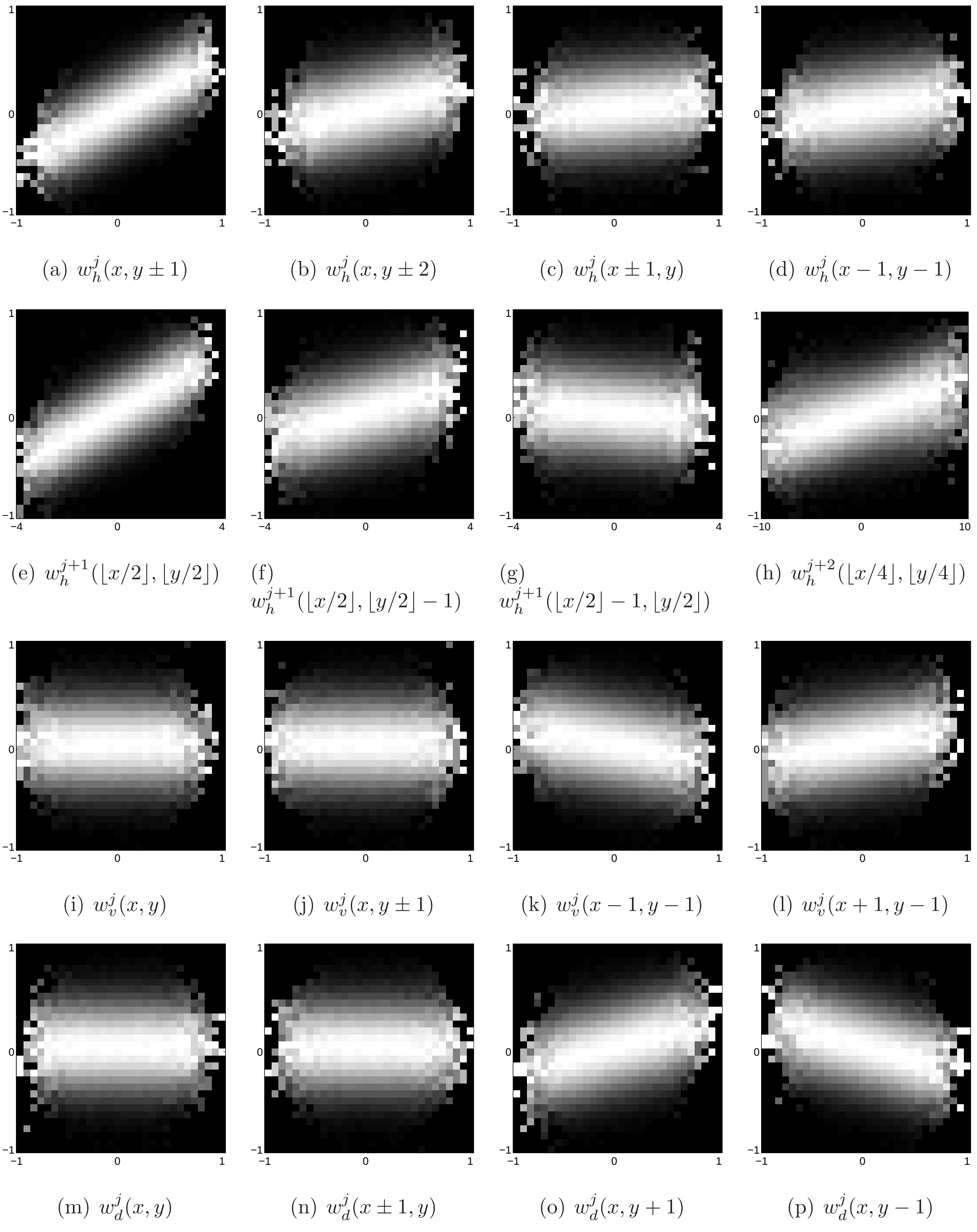
(m) $w_d^j(x, y)$

(n) $w_d^j(x, y \pm 1)$

(o) $w_d^j(x+1, y)$

(p) $w_d^j(x-1, y)$

Figure 4.18: Plots parallel to Figure 4.17 showing conditional densities for a db2 vertical coefficient. The vertical-band dependence map is symmetrically identical to the horizontal-band dependence map.



(a) $w_h^j(x, y \pm 1)$

(b) $w_h^j(x, y \pm 2)$

(c) $w_h^j(x \pm 1, y)$

(d) $w_h^j(x-1, y-1)$

(e) $w_h^{j+1}(\lfloor x/2 \rfloor, \lfloor y/2 \rfloor)$

(f) $w_h^{j+1}(\lfloor x/2 \rfloor, \lfloor y/2 \rfloor - 1)$

(g) $w_h^{j+1}(\lfloor x/2 \rfloor - 1, \lfloor y/2 \rfloor)$

(h) $w_h^{j+2}(\lfloor x/4 \rfloor, \lfloor y/4 \rfloor)$

(i) $w_v^j(x, y)$

(j) $w_v^j(x, y \pm 1)$

(k) $w_v^j(x-1, y-1)$

(l) $w_v^j(x+1, y-1)$

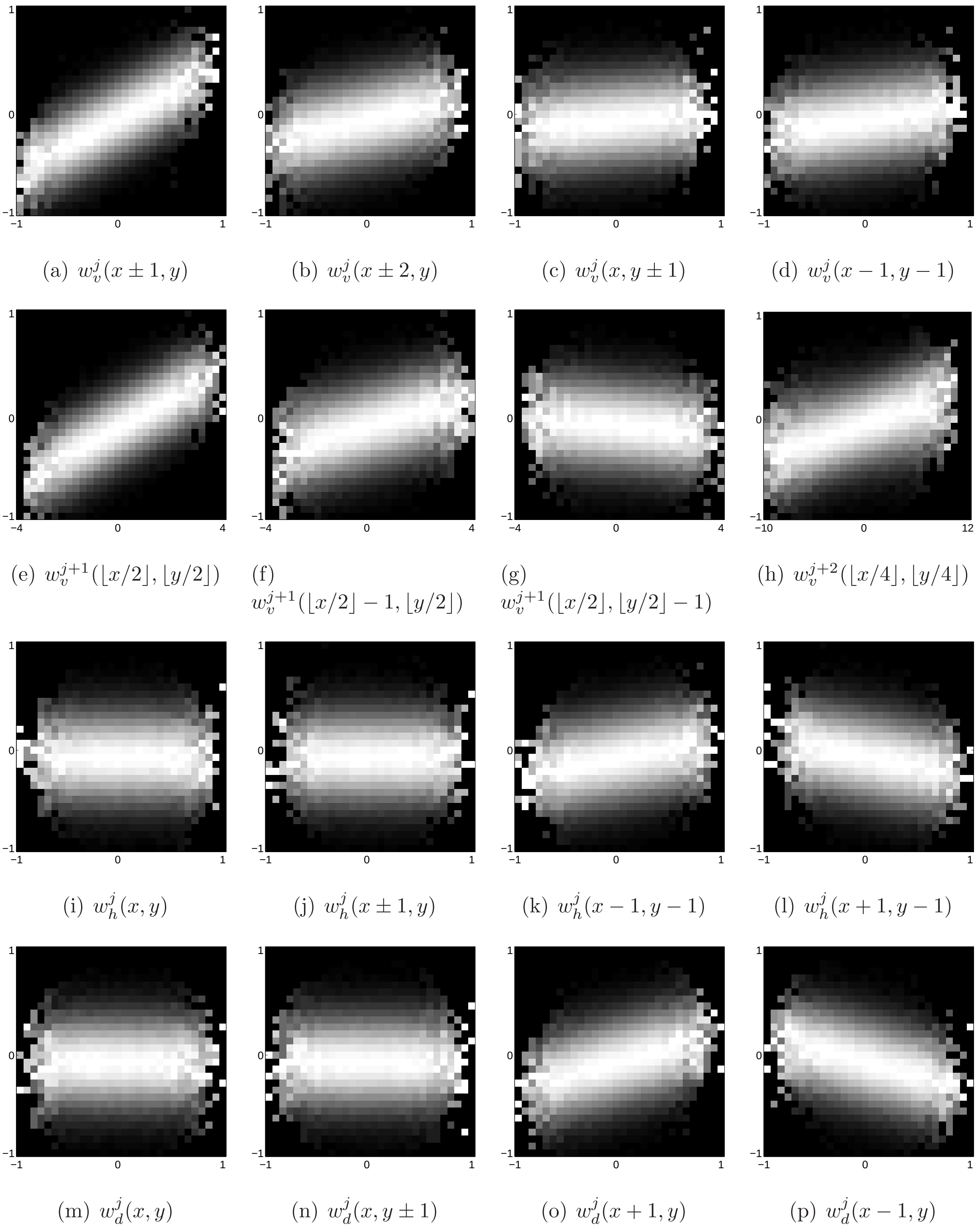
(m) $w_d^j(x, y)$

(n) $w_d^j(x \pm 1, y)$

(o) $w_d^j(x, y+1)$

(p) $w_d^j(x, y-1)$

Figure 4.19: Parallel plots to those displayed in Figure 4.17 for a db2 horizontal coefficient, but for GMRF thin-plate texture of Figure 4.10.



(a) $w_v^j(x \pm 1, y)$ (b) $w_v^j(x \pm 2, y)$ (c) $w_v^j(x, y \pm 1)$ (d) $w_v^j(x-1, y-1)$

(e) $w_v^{j+1}(\lfloor x/2 \rfloor, \lfloor y/2 \rfloor)$ (f) $w_v^{j+1}(\lfloor x/2 \rfloor - 1, \lfloor y/2 \rfloor)$ (g) $w_v^{j+1}(\lfloor x/2 \rfloor, \lfloor y/2 \rfloor - 1)$ (h) $w_v^{j+2}(\lfloor x/4 \rfloor, \lfloor y/4 \rfloor)$

(i) $w_h^j(x, y)$ (j) $w_h^j(x \pm 1, y)$ (k) $w_h^j(x-1, y-1)$ (l) $w_h^j(x+1, y-1)$

(m) $w_d^j(x, y)$ (n) $w_d^j(x, y \pm 1)$ (o) $w_d^j(x+1, y)$ (p) $w_d^j(x-1, y)$

Figure 4.20: Parallel plots to those displayed in Figure 4.18 for a db2 vertical coefficient, but for GMRF thin-plate texture of Figure 4.10.

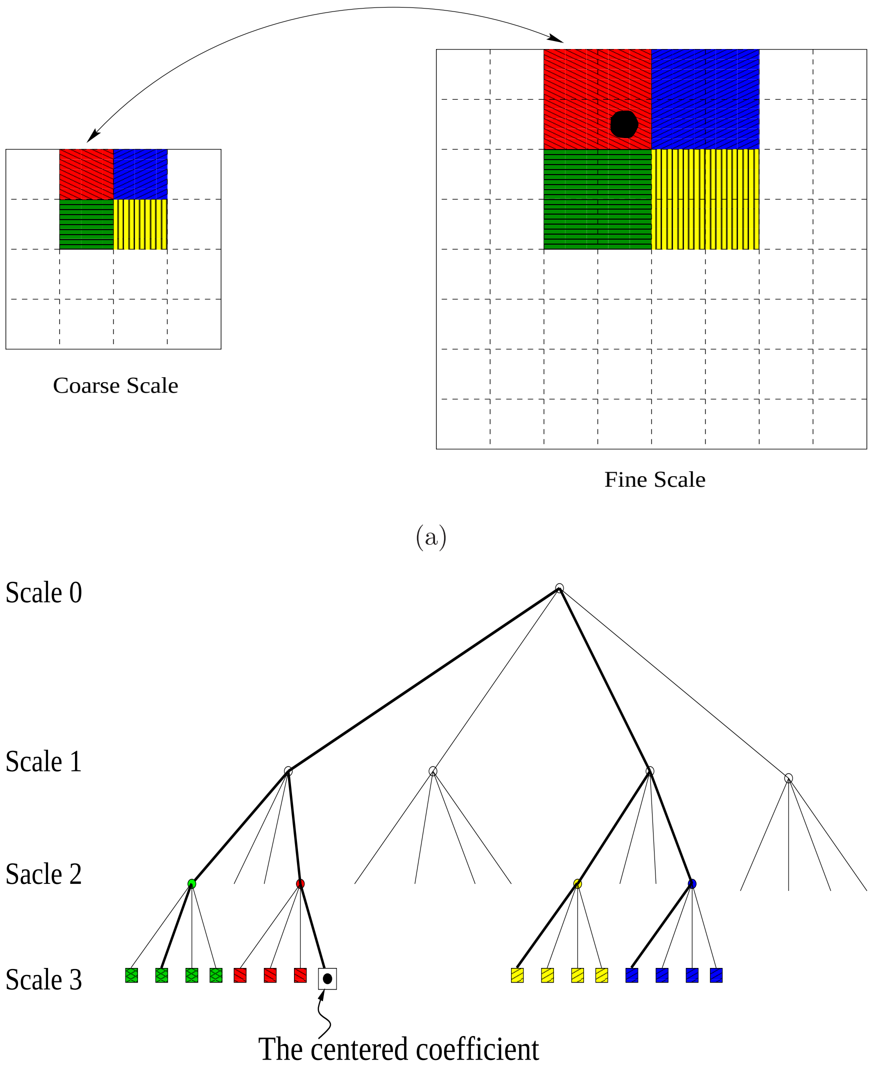
In summary, this section is not to report the striking wavelet correlations exhibited in these empirical observations, as it was seen earlier in this chapter. Rather, it is observed that, surprisingly, the existing wavelet joint models not only consider a *subset* of these interrelationships but also choose to relate some coefficients which are indeed independent, *e.g.*, in HMT-3S three coefficients at the same location from three subbands are grouped into one node, an assumption that is rejected by these plots. With the joint and conditional distributions of Figures 4.15-4.20 by way of introduction, a thorough study of wavelet correlations is considered next.

4.5 Simulation of Wavelet Joint Statistics

Because all of the joint histograms displayed in § 4.4 are well-characterized by their respective correlations (red curves), a more comprehensive study of wavelet correlations is justified.

Figure 4.21(a) illustrates the arrangement of a typical coefficient at a fine scale along with its spatially closed neighbors painted in different colors (right) and the corresponding parents (left) at the coarser scale. Figure 4.21(b) displays the top panel locality in a quad-tree structure. It is immediately obvious that first-order neighbors (siblings) of a pixel are not necessarily spawned from the same parent.

The purpose of these illustrations is to point toward an important issue: although two coefficients may be spatially close, they can be located on distantly separated branches of the wavelet tree. Consequently a standard wavelet quad-tree, modeling only parent-child relationships (as was done mostly and exclusively in the literature), will only poorly represent spatial interrelationships, in those cases where they are found to be significant. A clear neighborhood structure, such as for a Markov random field [41], which is capable



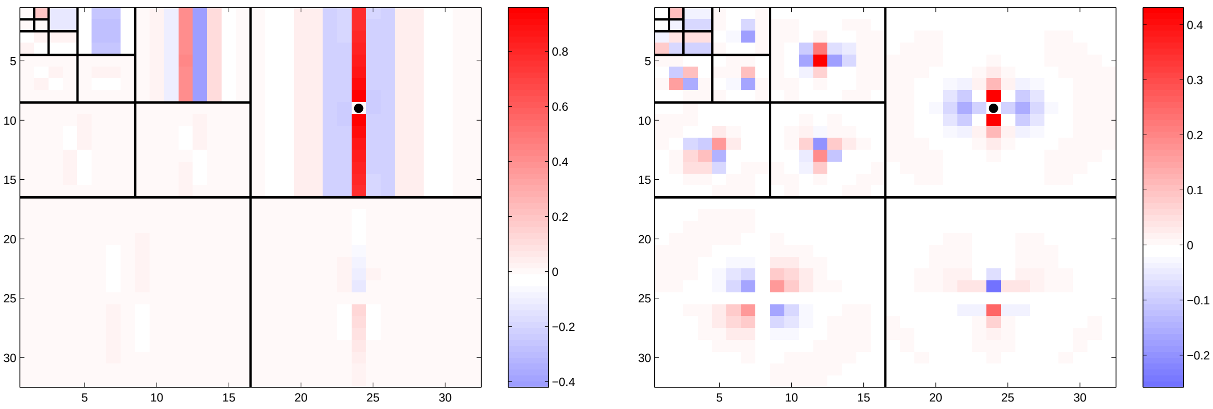
(a)

(b)

Figure 4.21: Illustration of a coefficient in the fine scale (shown by •), whose spatial neighbors come from different parents in the coarser scale.

of describing these statistical interactions of wavelet coefficients, is required.

Answering this question is challenging because of the issues raised in Figure 4.21: the tree-relationship between a pixel and its spatial neighbors is pixel dependent. So notions



(a) (b)

Figure 4.22: The correlation structure of a db2 wavelet coefficient (marked by •) with all other coefficients. (a): Correlation structure of a horizontal coefficient for the tree texture Figure 4.10(c). (b): Correlation structure for a horizontal coefficient of the thin-plate model Figure 4.10(b).

of stationarity, obvious in the spatial domain, become subtle (or completely invalid) in the wavelet domain. In short, does one need to specify a different neighborhood structure for *every* wavelet coefficient (since each coefficient occupies a unique position on the tree), or perhaps one structure for all of the "lower-left" children of parents and another for "upper-right" etc., or is there some degree of uniformity that applies?

4.5.1 2-D Wavelet Diagram of Wavelet Correlations

The problem was studied visually, and without any particular spatial assumptions. For simplicity of interpretation, and as shown in Figure 4.22, a tool which utilizes the traditional 2-D WT structure to display the correlation between any specified coefficient and all other coefficients on the entire wavelet tree, was devised. In this simulation, the wavelet covariance P_w is employed from which the associated correlation row (column) showing the chosen coefficient's covariance with all other wavelet coefficients is extracted and displayed in 2-D WT diagram. For the purpose of illustration, only, the autocorrelation of the selected coefficient is set to zero. Figure 4.22(a) shows the correlation of a typical horizontal coefficient (indicated by •) of the tree texture (Figure 4.10(c)), exhibiting a strong vertical correlation both within and across scales. Similarly Figure 4.22(b) displays correlation structure of a selected horizontal coefficient of the thin-plate model (Figure 4.10(b)).

Evidently the coefficient interactions show a clear preference to locality, as must be expected. This locality increases toward finer scales, which supports the persistency property of wavelet coefficients, the basic attributes which nearly all wavelet models have in common [19]. In particular, the correlation structure is spatially-localized and sparse. The local neighborhood definition for any given pixel does not confine to the pixel's subband: it extends to dependencies across directions and resolutions. Besides the long range across scale correlations, every typical coefficient exhibits strong correlation with its immediate neighbors both within subband and scale. The correlation structure for horizontally and vertically aligned coefficients are almost symmetrically identical. For textures whose edges extend more or less toward one direction (such as tree-bark), this similarity does not hold.

Figure 4.23 shows the correlation coefficients, averaged over the wavelet priors corresponding to the textures of Figure 4.10: the behaviors evidenced in the plots are thus persistent patterns, not the peculiar behavior of a single, particular texture.

There is a very clear consistency between these maps and the conclusions reached from the joint histograms. There are, however, striking patterns which are *not* reflected in other models:

- A given coefficient is not correlated with siblings at other orientations: the hybrid HMT-3S model proposed by Fan and Xia [37] integrates three corresponding siblings across other orientations. These observations, however, indicate that within any par-

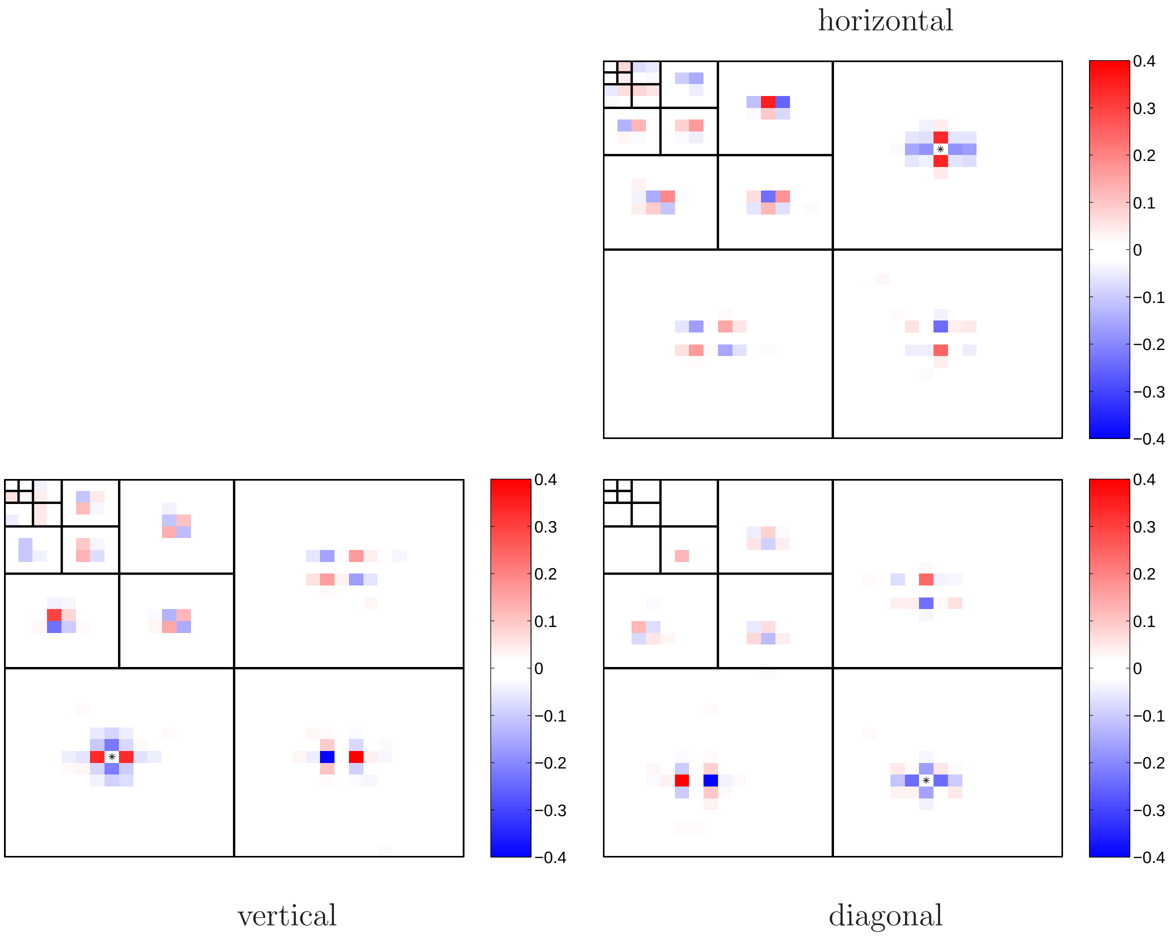


Figure 4.23: Wavelet (db2) correlation structures averaged over the textures displayed in Fig. 4.10, except tree-bark (because of its special structure). Each panel plots the correlation of a selected coefficient (•) with all other coefficients at all orientations and scales.

ticular scale, across-subband siblings are nearly decorrelated, though across-subband neighbours of siblings are related.

- Within-subband correlations are orientation-dependent: horizontal coefficients are vertically correlated, vertical coefficients are horizontally correlated.
- Inter-subband correlations are orientation-dependent: horizontal coefficients are cor-

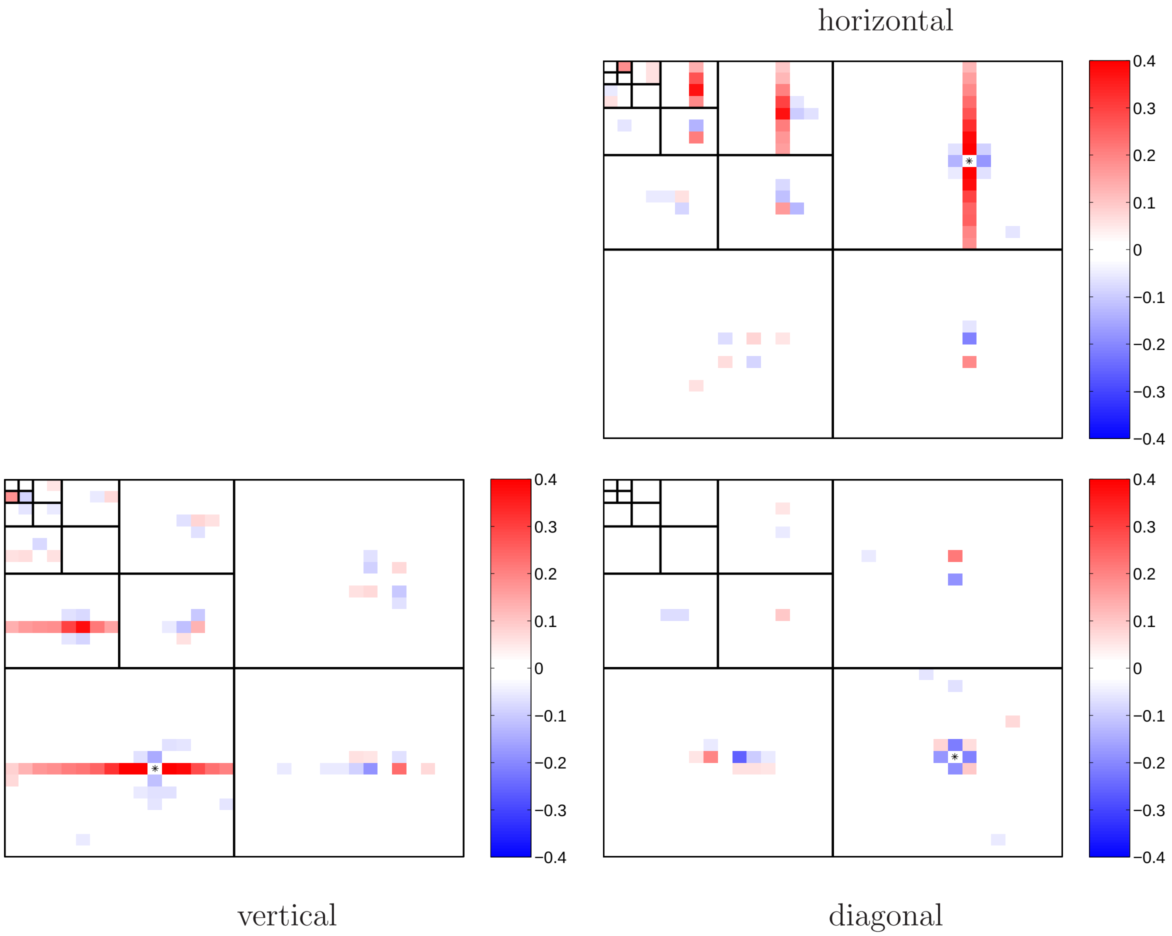


Figure 4.24: Wavelet (db2) correlation structures averaged over a collection of 5000 real images. Each panel contains three plots illustrating the correlations of a given coefficient (•) with its local neighborhoods in the horizontal, vertical, and diagonal subbands.

related with vertical neighbours in the diagonal subband, vertical coefficients are correlated with horizontal neighbours.

- Inter-scale correlations are orientation-dependent: in addition to its parent, a coefficient is correlated with the spatial neighbors of its parent, *e.g.* a horizontal coefficient is more related with vertical neighbors of its parent.

Finally, to confirm that the conclusions are not the result of Markovianity, Gaussianity,

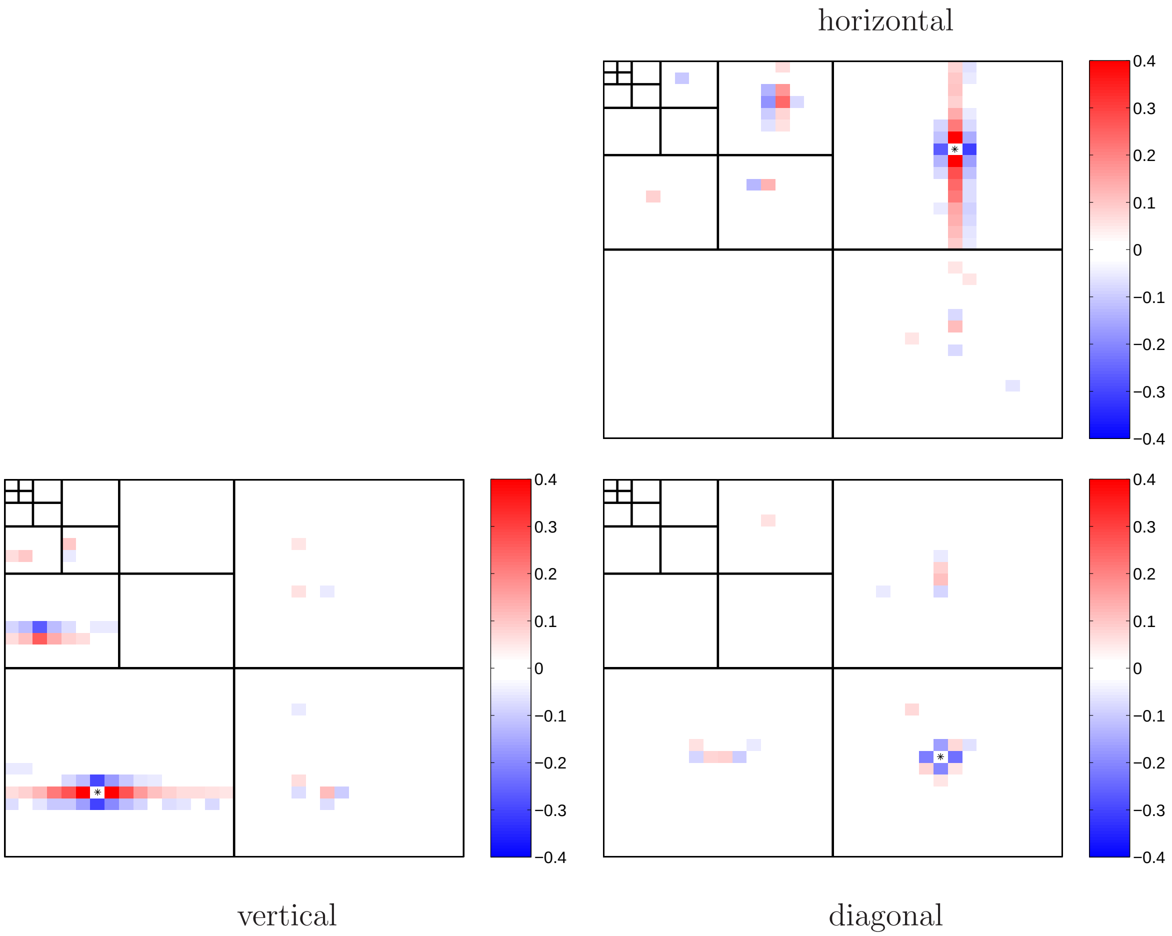


Figure 4.25: Wavelet (db4) correlation structures averaged over a collection of 5000 real images. Each panel contains three plots illustrating the correlations of a given coefficient (•) with its local neighborhoods in the horizontal, vertical, and diagonal subbands.

or other coincidences associated with our choices of textures, Figures 4.24 and 4.25 plot the correlation maps for db2 and db4 wavelets averaged over a collection of 5000 randomly cropped and subsampled real-world images, with example shown in Figure 6.4. The consistency between Figures 4.24 and 4.25 and Figure 4.23 is very clear. Furthermore, all panels of Figures 4.24 and 4.25 support the conclusion of sibling uncorrelatedness and orientation-dependence.

4.5.2 Significance of Wavelet Correlations

Because correlation coefficients can mislead (a high coefficient between two tiny-variance wavelet elements may not be of modeling significance), we propose to measure the correlation *significance* [7] as the reduction in mean-square estimation error induced by including the correlation relationship.

It is not remotely obvious that the correlation coefficients plotted above necessarily quantitatively correspond to importance in considering coefficient interactions. That is, can one more objectively quantify what it means for some correlation to be important or significant?

For small test problems the wavelet-based covariance P_w can be determined exactly. Suppose two coefficients w_1, w_2 are observed in the presence of noise:

$$\begin{bmatrix} w_{n1} \\ w_{n2} \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}, \quad \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix} \sim \mathcal{N}\left(\underline{0}, \begin{bmatrix} r & 0 \\ 0 & r \end{bmatrix}\right) \tag{4.23}$$

Under the standard independence assumption, if only the coefficient variances are kept from the full covariance, then their estimate is

$$\begin{bmatrix} \hat{w}_1 \\ \hat{w}_2 \end{bmatrix} = \left(\begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}^{-1} + \begin{bmatrix} r & 0 \\ 0 & r \end{bmatrix}^{-1} \right)^{-1} \begin{bmatrix} r & 0 \\ 0 & r \end{bmatrix}^{-1} \begin{bmatrix} w_{n1} \\ w_{n2} \end{bmatrix} \tag{4.24}$$

with the associated estimation error

$$\tilde{P}_1 = \left(\begin{bmatrix} \sigma_{w_1}^2 & 0 \\ 0 & \sigma_{w_2}^2 \end{bmatrix}^{-1} + \begin{bmatrix} r & 0 \\ 0 & r \end{bmatrix}^{-1} \right)^{-1} \tag{4.25}$$

On the other hand, if we model the two coefficients with their precise correlation, the estimation error proceeds as

$$\tilde{P}_2 = \left(\begin{bmatrix} \sigma_{w_1}^2 & \lambda_{w_1,w_2} \\ \lambda_{w_2,w_1} & \sigma_{w_2}^2 \end{bmatrix}^{-1} + \begin{bmatrix} r & 0 \\ 0 & r \end{bmatrix}^{-1} \right)^{-1} \tag{4.26}$$

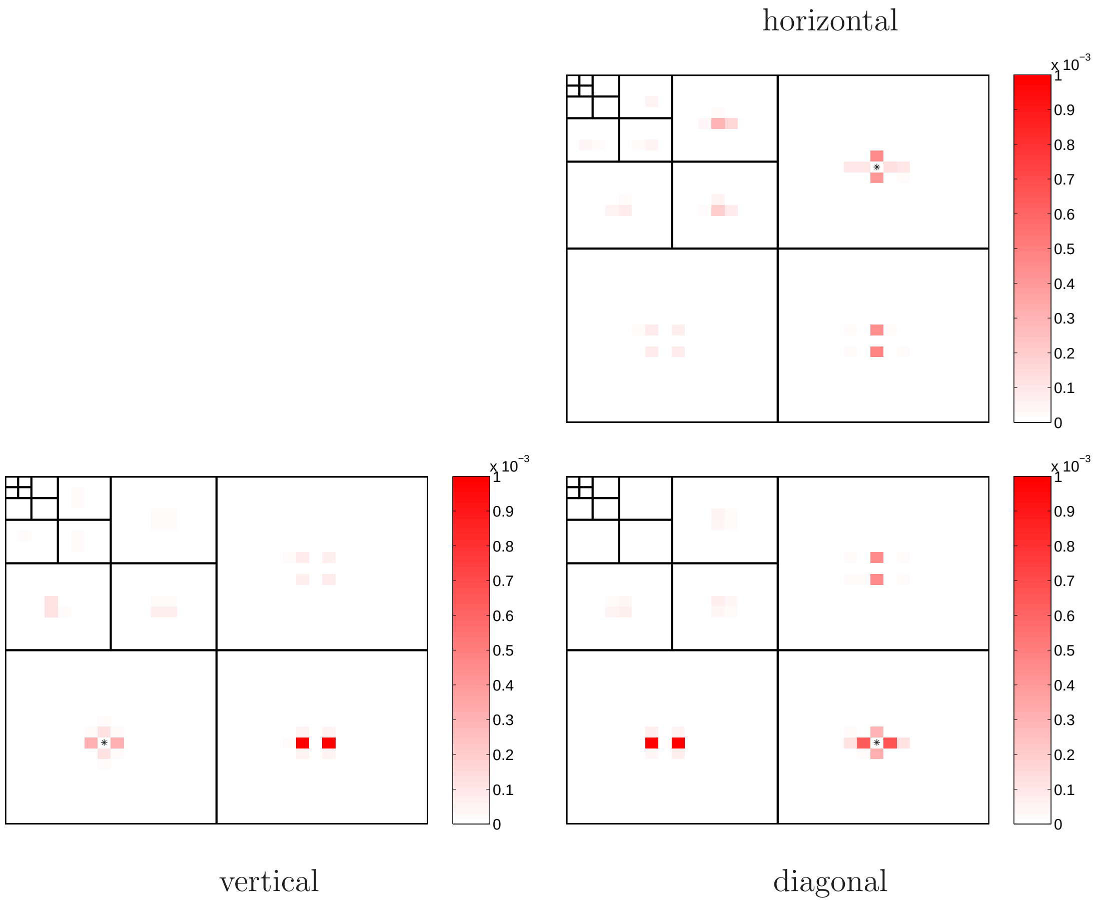


Figure 4.26: Plot of the estimation MSE significance of the db2 wavelet interrelationships depicted in Fig. 4.23.

In other words, the importance of this particular correlation can be quantified as the degree to which it affects the accuracy of the estimation, which although related to the correlation coefficient, is not proportional to it. Define the significance to be the difference of the total MSE under the two approaches:

$$\hat{\lambda}_{w_1,w_2} = tr(\widetilde{P}_1) - tr(\widetilde{P}_2) \tag{4.27}$$

Figure 4.26 shows the significance of correlations for the corresponding coefficients displayed in Figure 4.23. It is evident from these diagrams that within scale dependency range reduces to shorter locality, but across scale activities still present up to several scales. The computation of significant covariances, thus, confirms that the well-structured coefficients dependencies to be hierarchical.

4.6 Chapter Summary

In summary, a thorough study of empirical 2-D wavelet correlation structures has been presented in this chapter. The expected patterns – correlation sparsity and parent-child persistence – are clear, however, there are additional striking relationships which, as yet, are not normally found in wavelet models. Examples of interscale and intra-scale dependencies that are missing in the existing models were discussed. In particular, coefficients are very nearly decorrelated with siblings across orientations, however, there is a very strong orientation-dependence governing correlations within subbands, across orientations, and across scales.

Following the wavelet statistical observations discussed here, there are two primary research directions: (1) proposing models with high and precise capability to describe the wavelet statistics along with utilizing tools to examine the model accuracy by comparing it with the current thresholding methods, in an RMSE sense (Ch. 5), and (2) devising an estimation/ denoising/ shrinkage algorithm, which takes into account the proposed models and results in optimum error and low computational cost (Ch. 6).

Chapter 5

Models of Wavelet Statistics

This chapter presents a detailed description of the statistical modeling approaches that have been taken throughout this research to approximate the interrelationships among wavelet coefficients. The essential goal is to reduce the dimensionality of the wavelet joint statistics (studied in Ch. 4) but to account for the most striking local structures.

As discussed in Ch. 4, the numerical simulations of the wavelet covariance structure have revealed the importance of devising a model which covers a small within-scale locality along with a large extent of across-scale dependencies. In order to meet this objective, there are two alternatives to consider:

1. Imposing models which describe the long range statistical dependencies, such as the full covariance matrix. Such models, however, lead to estimation algorithms that are considerably complex and difficult to implement. If high accuracy is desired, however, one can resort to this approach.

2. Proposing a statistical model which approximates the structural correlations over the entire wavelet tree. The advantage of this approach is in the existence of estimation

techniques which are fast and very easy to implement [100].

Having followed the latter approach, and in addition to the HMM work of others, described in Ch. 3, two models of the wavelet joint statistics are developed in this study:

1. A wavelet multiscale statistical model [7], which captures the parent-child correlations with no assumption of the activities across orientations.

2. An approach to Markov modeling [6, 8] the wavelet across scale, orientation, and space activities.

The theories and methodologies associated with each of these two models are defined and discussed in the current chapter. Just as with the HMM methods, we seek to better assess the neighborhood structures asserted by these probabilistic models, which describe the wavelet random field with a small fraction of coefficients, but which accurately absorb each pixels dependency on the rest of the wavelet tree.

5.1 Multiscale Modeling

This section proposes the MS modeling of the wavelet joint statistics. It explains the results of MS-based approximations of the wavelet domain covariance matrix corresponding to 1-D as well as 2-D signals.

Before getting into the details of the proposed modeling approach, first the general class of MS models organized on a quad-tree (Figure 5.1) is defined as a natural tree of wavelet subbands.[1] In this model, a random variable or random vector $x(w)$ is associated with each node w, representing some information related to the resolution and location

[1] Although methods of multiscale modeling are the topic of § 2.3, the principle intuition is discussed here for the sake of clarity and continuity.

corresponding to that node. In particular, assume w_J denotes the root node, located at the top of the tree, *i.e.*, coarsest resolution, to which a covariance $P(x(w_J))$ is assigned to show the root's marginal distribution. Note that $J = \log_2 n$ (n: data size) depicts the number of resolution levels between the finest scale up to the coarsest one, *i.e.*, tree's depth. To all other nodes w a parent node p_w is connected which is located at the next coarser scale. Accordingly, one can define a coarse-to-fine transition probability density $Pr(x(w)|x(p_w))$. A complete set of these transition probabilities in addition to the root's initial distribution is sufficient to specify the joint pdf of $x(.)$ over the entire MS tree.[2]

A significant assumption in MS modeling is the conditional independence. Given the statistics of a node, all subtrees initiated from that node are conditionally independent, *e.g.*, for the children of node w shown in Figure 5.1:

$$\begin{aligned} Pr(x(c_{1w}), x(c_{2w}), x(c_{3w}), x(c_{4w}) \mid x(w)) = \\ Pr(x(c_{1w})|x(w))\ Pr(x(c_{2w})|x(w))\ Pr(x(c_{3w})|x(w))\ Pr(x(c_{4w})|x(w)) \end{aligned} \tag{5.1}$$

Consider a class of linear MS models, where $P(x(w_J))$ is assumed to be Gaussian along with the following coarse-to-fine recursive stochastic dynamic:

$$x(w) = A_w x(p_w) + B_w \nu_w \tag{5.2}$$

where A_w and B_w are parameter matrices associated with node w and $\nu_w \sim \mathcal{N}(0, I)$ is a Gaussian white noise process.

The connection between the MS modeling framework and the wavelet tree can be intuitively explained as follows. In Figure 5.1, each wavelet coefficient w is shown as a node with p_w as its parent and $\{c_{iw}|i = 1, \ldots, 4\}$ being the set of its four children. As the

[2] Note that when w is root, p_w is empty, implying that the expression $P(x(w)|x(p_w)) \equiv P(x(w))$, the prior probability of w.

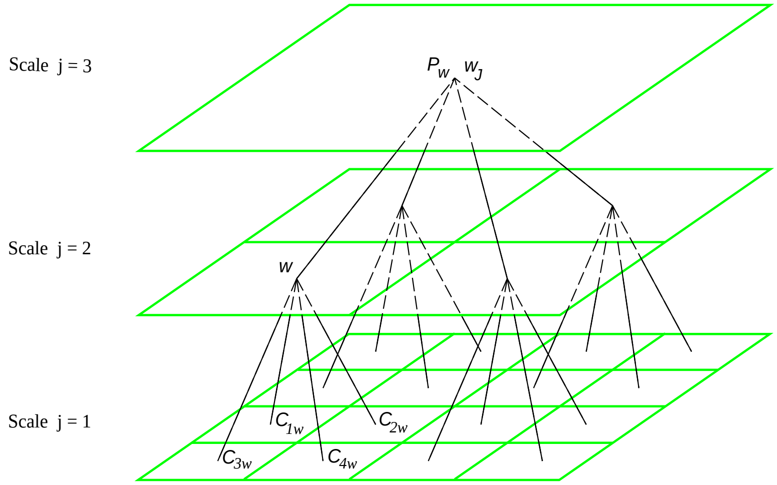


Figure 5.1: Tree-based illustration of a typical coefficient w along with its parent and four children within one wavelet tree subband. At the top: the coarsest resolution. At the bottom: the finest resolution, expressed by the image pixels themselves.

scale j decreases, the children add finer and finer details into the spatial regions occupied by their ancestors [19].

The first-order scalar MS model (5.2) is adopted to devise an approximate model of wavelet coefficient. First-order means that only the parent-child relationship, *i.e.*, the parent-child transition probability $Pr(x(w)|x(p_w))$ is assumed. A scalar MS model means that only one coefficient $d = 1$ is considered per node, *i.e.*, $x(.)$ is a random variable. The discussion on the higher-order MS modeling with random vector processes investigated in the present work, will follow in § 5.1.1.

At the coarsest resolution, the root node, the stochastic process $x(w_J)$ obeys the following statistics:

$$\begin{aligned} E\left[x(w_J)\right] \quad &= \quad 0 \\ E\left[x(w_J)\ x^T(w_J)\right] \quad &= \quad P_J \end{aligned} \tag{5.3}$$

and the cross-correlation of each node w with its parent is computed through the scale-recursive relationship [100]:

$$\begin{aligned} P_{w,p_w} &= E\left[x(w)\; x^T(p_w)\right] \\ &= E\left[\{A_w x(p_w) + B_w \nu_w\}\; x^T(p_w)\right] \\ &= A_w P_{p_w} \end{aligned} \tag{5.4}$$

where P_{p_w} is the auto-covariance of node p_w.

Having defined the initial conditions in (5.3) and across-scale statistics in (5.4), one can easily calculate the system parameters A and B given in (5.2) [100]

$$\begin{aligned} A_w &= P_{w,p_w} P_{p_w}^{-1} \\ B_w B_w^T &= P_w - A_w P_{p_w} A_w^T \end{aligned} \tag{5.5}$$

Following the above introduction to the theory of MS modeling, two basic directions taken to examine the wavelet correlation structure of both 1-D and 2-D signals are discussed next. The objective is to study the capability and accuracy of the MS framework in capturing the significant wavelet correlations.

5.1.1 MS Modeling of a Binary Tree

The very first step is to understand the behavior of the MS model (5.2) applied on a binary tree (Figure 5.2), when the 1-D wavelet domain covariance is given.

Once again an exponentially distributed signal of size $n = 128$ as given in (4.14) is considered with its wavelet covariance P_w and its associated correlation map $\mathcal{P}_w$ depicted in Figure 5.3(a). The first-order scalar MS modeling (5.2) was used with the root's auto-covariance P_J extracted directly from the wavelet covariance P_w and resulted the approximated covariance $\hat{P}_w$. The approximated wavelet correlation map $\hat{\mathcal{P}}_w$ is shown in Figure 5.3(b).

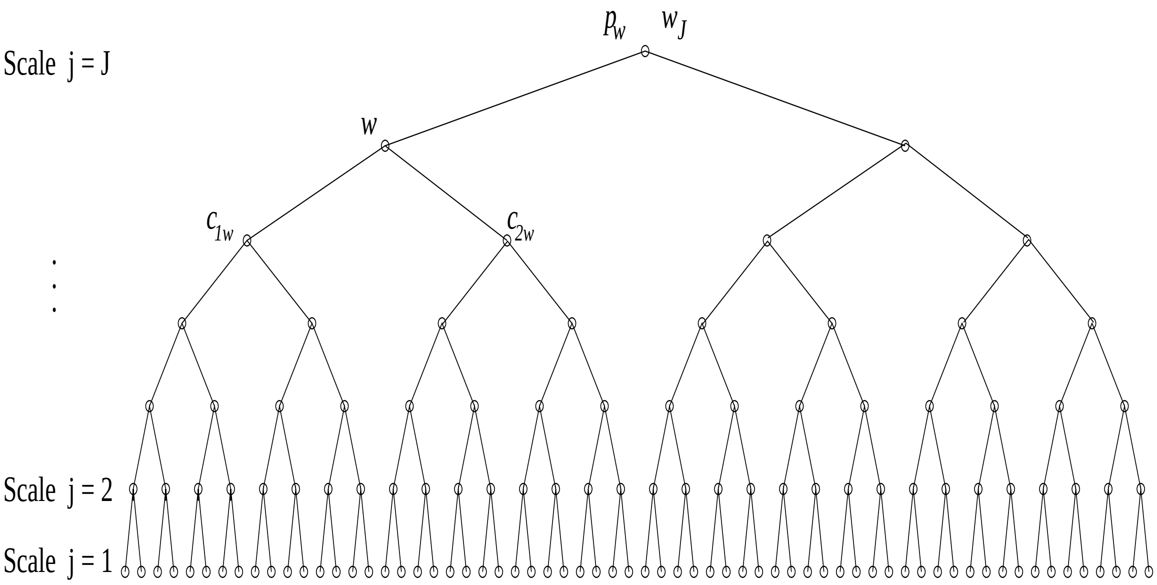
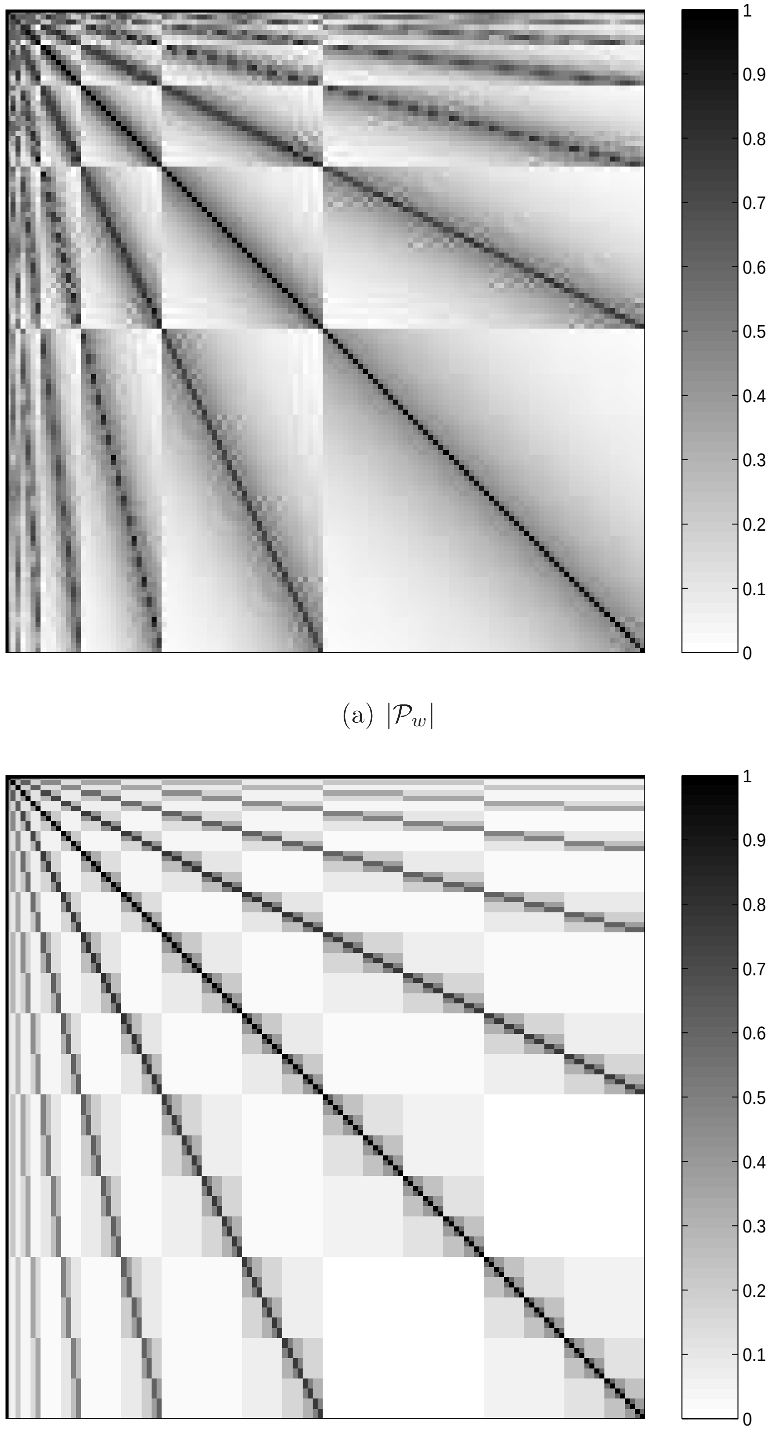


Figure 5.2: A binary tree with seven decomposition levels.

The above methodology leads to the following observations:

- The finger structure of P_w is preserved in $\hat{P}_w$. Note that the number of fingers depends on the number of decomposition levels [43].
- The parent-child dependencies exhibited by $\hat{P}_w$ are quite clear.
- Although the tree-based correlations are captured up to many scales apart, the within-scale stationarity is lost, because the first-order MS model is not accurate enough to assert an efficient within-scale conditional decorrelation, *i.e.*, this model is a relatively poor approximate of the true within-subband model.

To solve the wavelet-spatial modeling issue associated with the MS-based modeling, *i.e.*, to make the conditional decorrelation asserted by the MS model more valid, either (or both) of the following two alternatives can be considered:



(a) $|\mathcal{P}_w|$

(b) $|\hat{\mathcal{P}}_w|$

Figure 5.3: (a) A 1-D wavelet correlation map and (b) its MS-based approximate. $\mathcal{P}_w$, the correlation coefficient of P_w, is displayed.

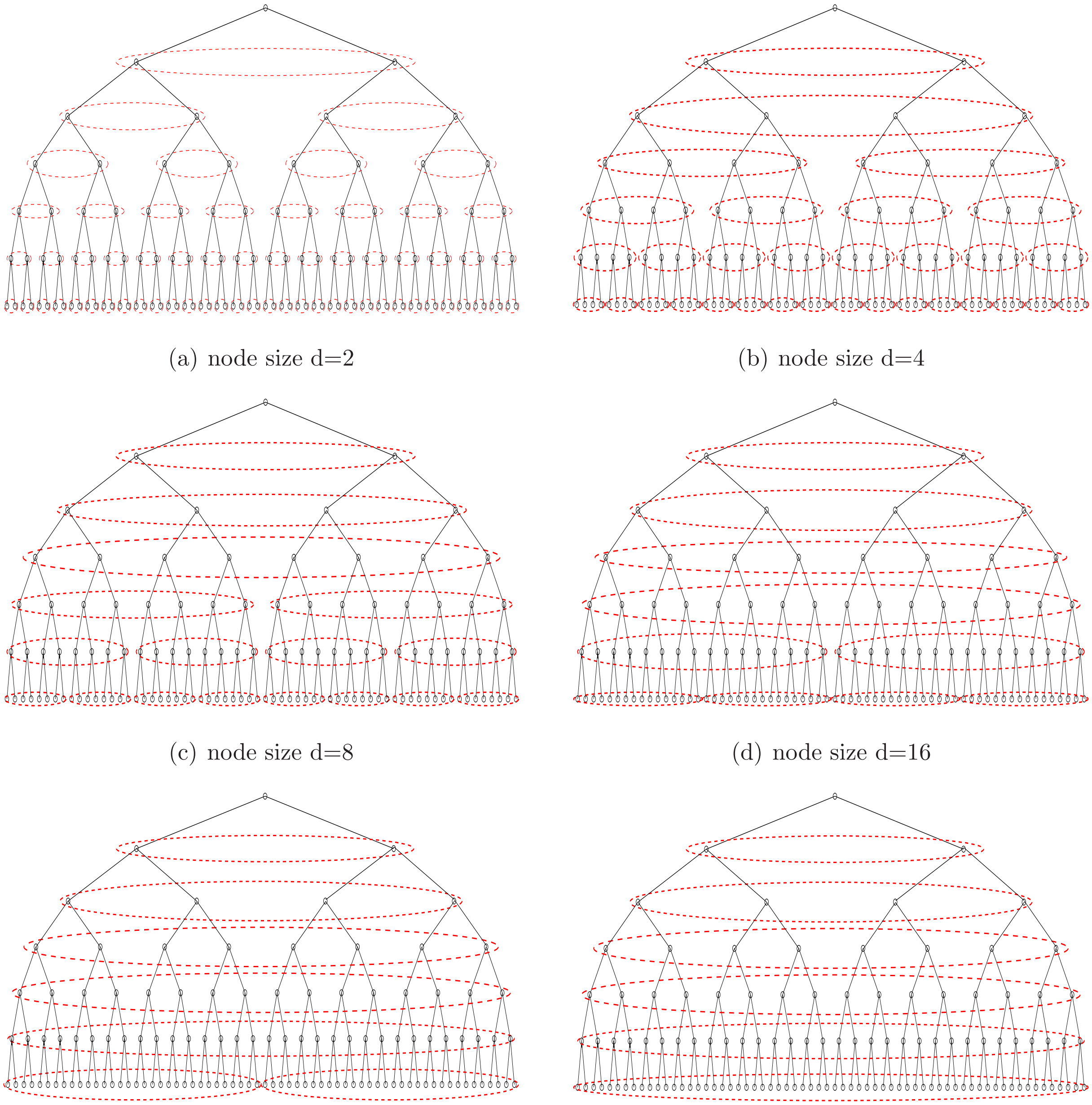
1. Dimension: the number of coefficients that form each node on the tree can vary. A particular node may contain only a single wavelet coefficient, *i.e.*, a scalar, or a finite set of the coefficients, *i.e.*, a vector. Figure 5.4 shows several possible node sizes which could be employed in the MS-based modeling of the typical binary tree of Figure 5.2.

2. Order: the modeling accuracy can be increased from first-order (the state of the parent is sufficient for a child to be decoupled from all other nodes) to second-order (the grand-parent's statistics are also necessary for a node to be independent from the rest of the tree), and even to higher orders. Figure 5.5 illustrates second-order MS modeling with various node sizes applied on a binary tree.

To understand the trade off between the accuracy of the MS model and the computational complexity of the estimation process, a variety of MS models, including first-order to log_2nth-order and scalar to vector of coefficients per node, are examined. In particular, the first-order MS model with all six tree node sizes depicted by Figure 5.4 was used to approximate the entities of the wavelet covariance P_w with the experimental results displayed in the top left panels in Figures 5.6-5.12. The lower left panels of each figure show the difference between each estimate and the original covariance matrix, and the right panels zoom into the finest scale diagonal block entries of the left panels. Clearly, as the model complexity increases the parsimony in the estimated $\hat{P}_w$ decreases.

The purpose of this study is to demonstrate that the MS-based accuracy grows as a function of its complexity. In accordance, Table 5.1 summarizes the computational complexity for each MS model. From top left to bottom right the correlation structure becomes dense while the complexity of even simple estimation algorithms gets harder.

The observations, up until now, indicate that the MS-based approximation is an elegant and reasonable tool to model the 1-D wavelet coefficients locality. The next step is to



(a) node size d=2

(b) node size d=4

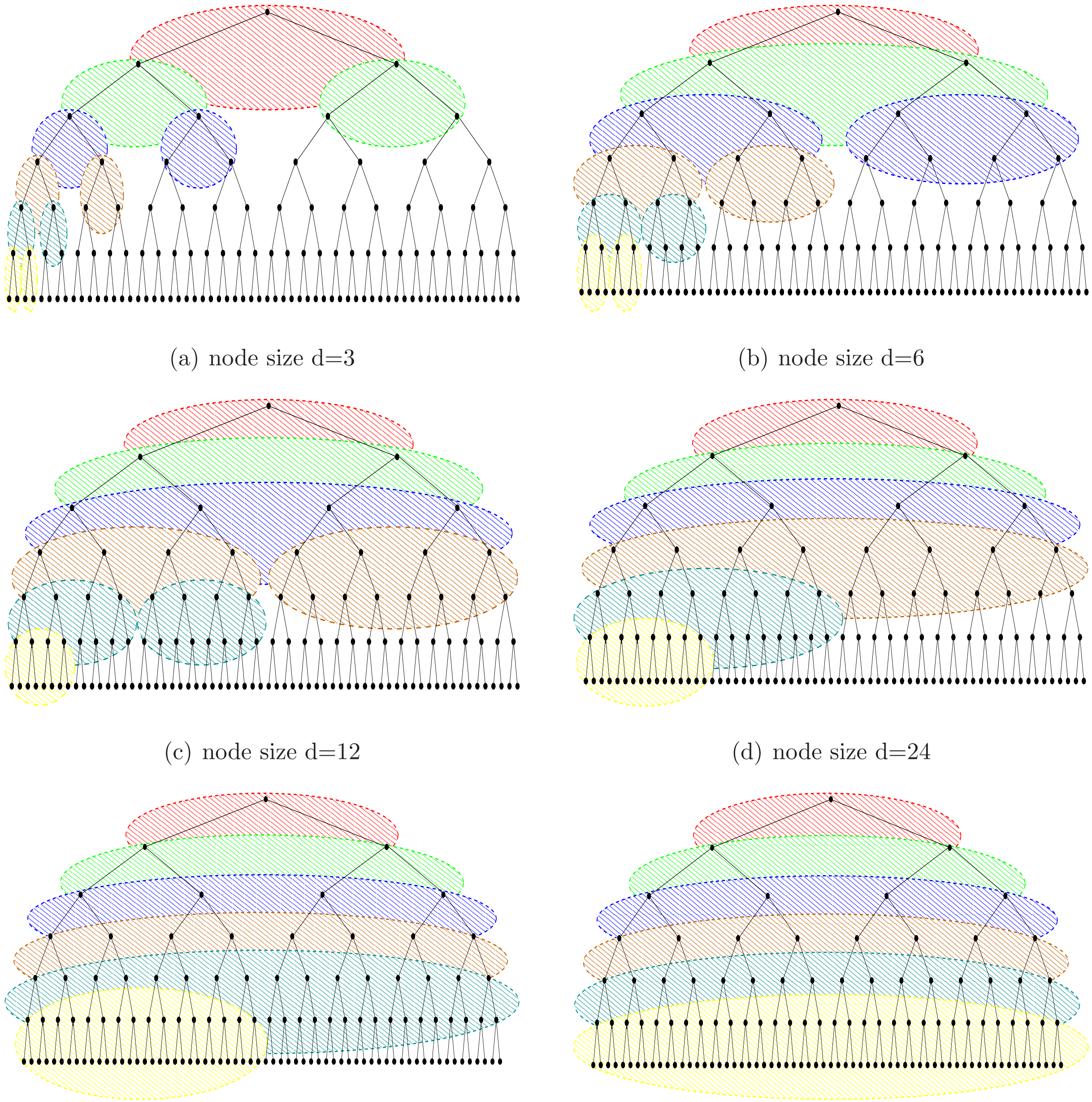
(c) node size d=8

(d) node size d=16

(e) node size d=32

(f) node size d=64

Figure 5.4: Illustration of *first-order* MS modeling with various node sizes applied on a binary tree.



(a) node size d=3

(b) node size d=6

(c) node size d=12

(d) node size d=24

(e) node size d=48

(f) node size d=96

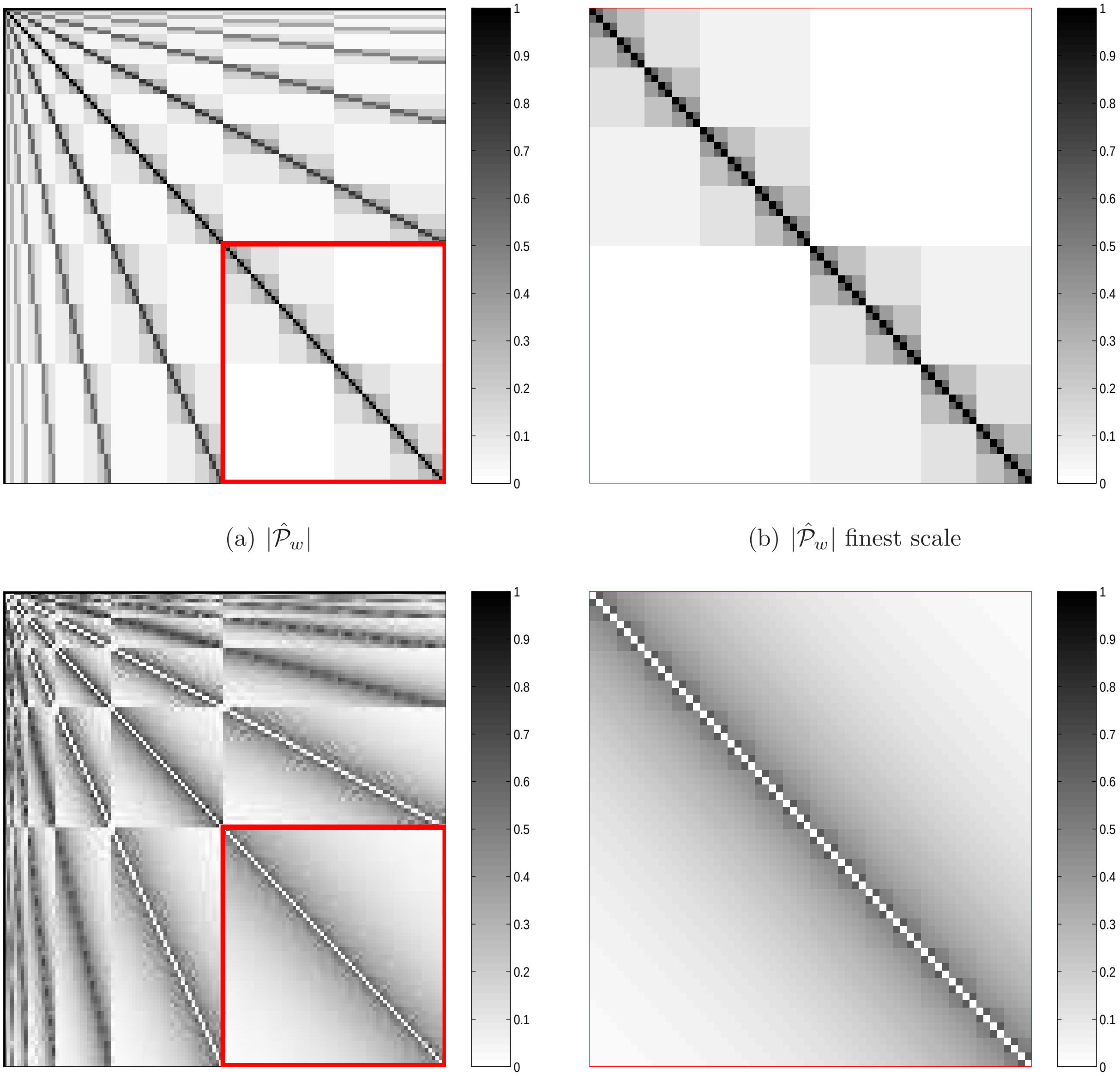
Figure 5.5: Illustration of *second-order* MS modeling with various node sizes applied on a binary tree.

MS Model Order					
1^{st}	2^{nd}	3^{rd}	4^{th}	$\cdots$	$(log_2 n)^{th}$-order
$d = 1(m = n - 1)$	-	-	-	-	-
$d = 2(m = \frac{n}{2})$	$d = 3$	-	-	-	-
$d = 4(m = \frac{n}{4} + 1)$	$d = 6$	$d = 7$	-	-	-
$d = 8(m = \frac{n}{8} + 2)$	$d = 12$	$d = 14$	$d = 21$	-	-
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$d = \frac{n}{2}(m = log_2 n)$	$d = \Sigma_{i=1}^{2} \frac{n}{2^i}$	$d = \Sigma_{i=1}^{3} \frac{n}{2^i}$	$d = \Sigma_{i=1}^{4} \frac{n}{2^i}$	$\cdots$	$d = n - 1(m = 1)$

Complexity of each case is $\mathcal{O}(d^3 m)$, where $m = \#$ of nodes on the tree

Table 5.1: Summary of computational cost of the MS model to approximate a wavelet binary-tree for a 1-D signal of size n. Each number shows the complexity for a combination of MS model order and number of coefficients per node (dimension d).

examine the extended MS framework and test its effectiveness in modeling 2-D wavelet statistics.



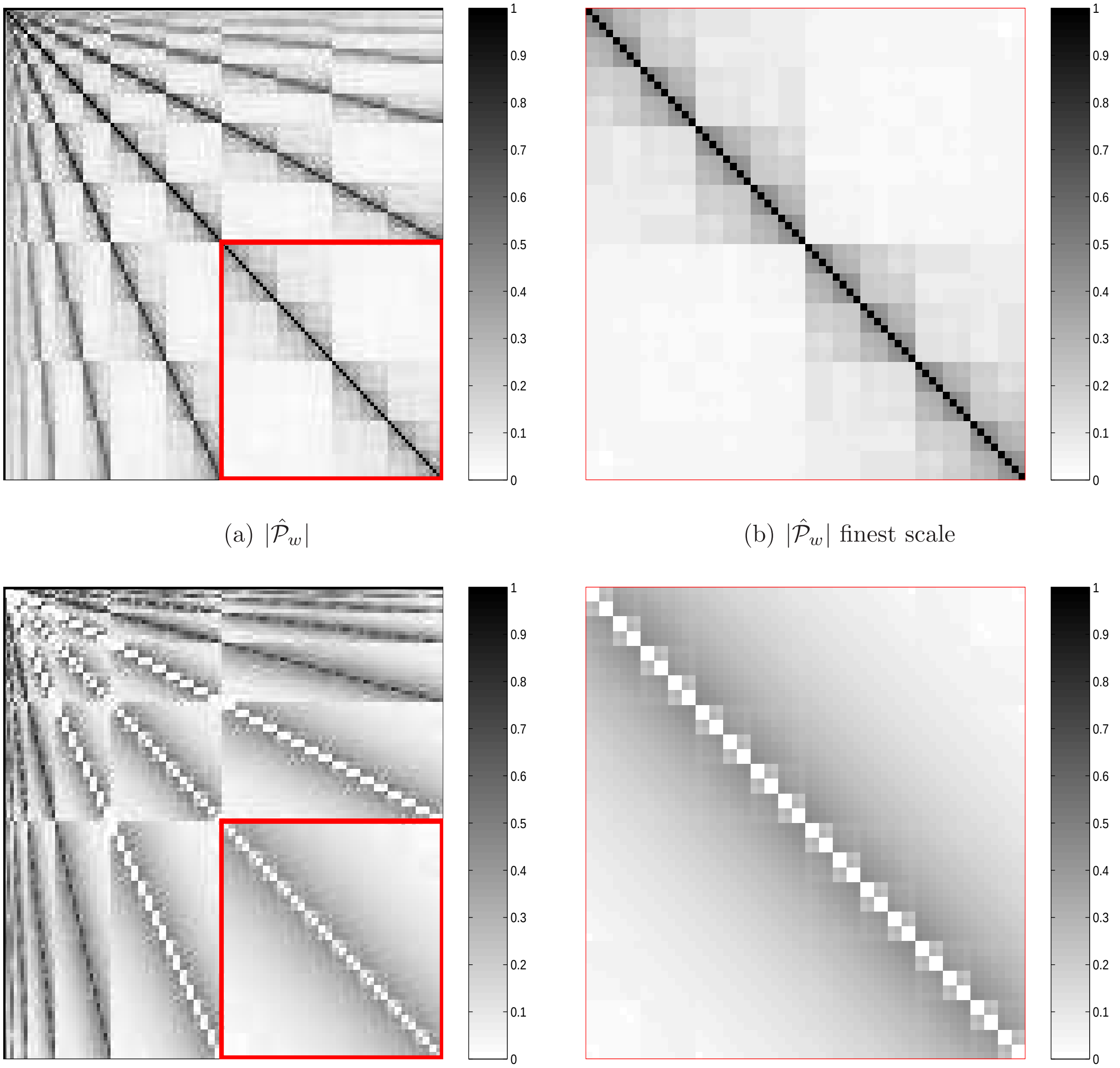
(a) $|\hat{\mathcal{P}}_w|$

(b) $|\hat{\mathcal{P}}_w|$ finest scale

(c) $\left|\mathcal{P}_w - \hat{\mathcal{P}}_w\right|$

(d) $\left|\mathcal{P}_w - \hat{\mathcal{P}}_w\right|$ finest scale

Figure 5.6: Wavelet binary tree first-order MS modeling; node size d=1. (a) shows the approximate $|\hat{\mathcal{P}}_w|$, (b) zooms into the finest scale approximation, and (c-d) show the accuracy of the model through its difference with the original $\mathcal{P}_w$, which was shown in Figure 5.3(a).



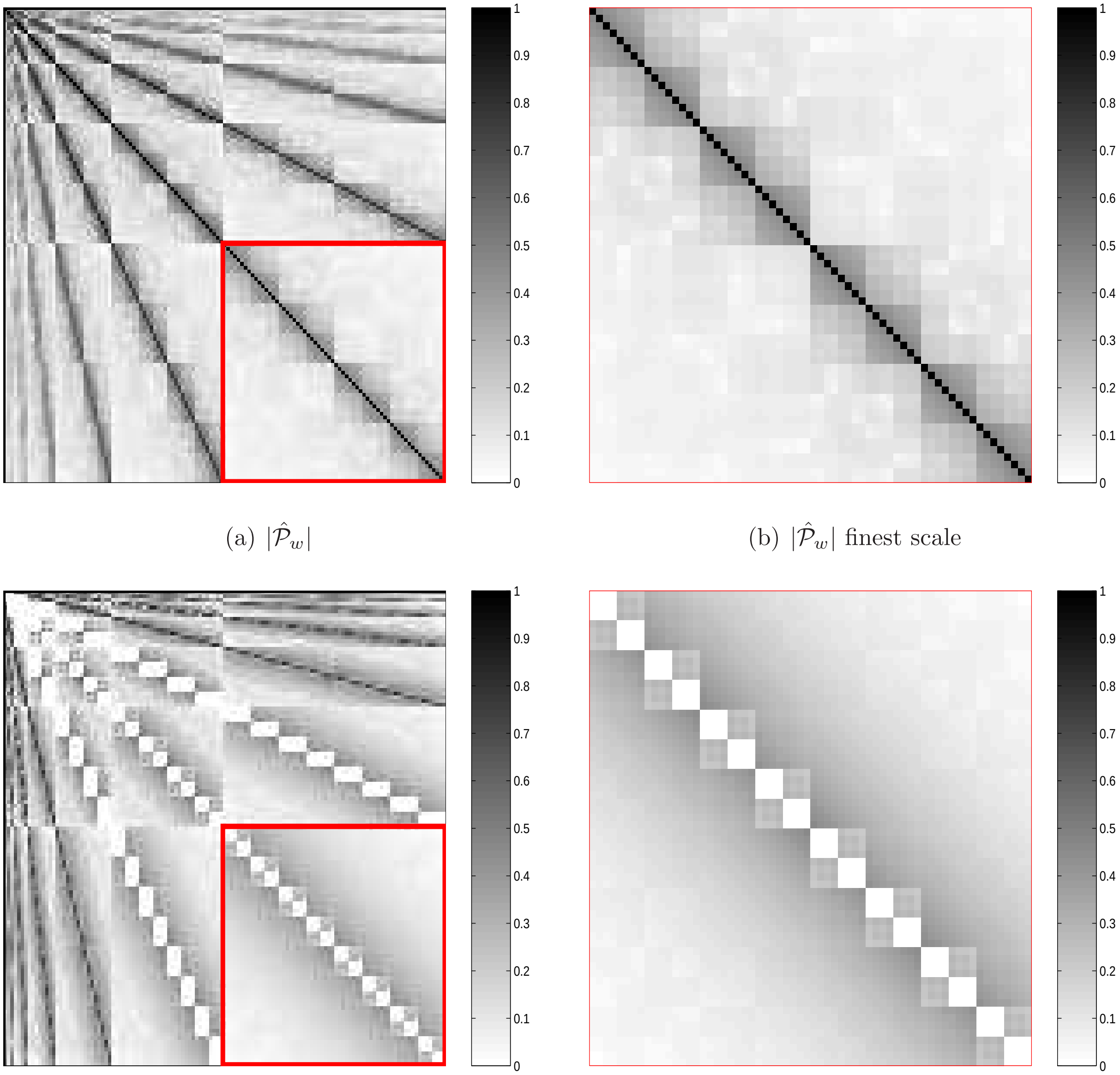
(a) $|\hat{\mathcal{P}}_w|$

(b) $|\hat{\mathcal{P}}_w|$ finest scale

(c) $\left|\mathcal{P}_w - \hat{\mathcal{P}}_w\right|$

(d) $\left|\mathcal{P}_w - \hat{\mathcal{P}}_w\right|$ finest scale

Figure 5.7: Plots parallel to those shown in Figure 5.6, except for node size d=2. The more complex the model, the more accurate the approximation.



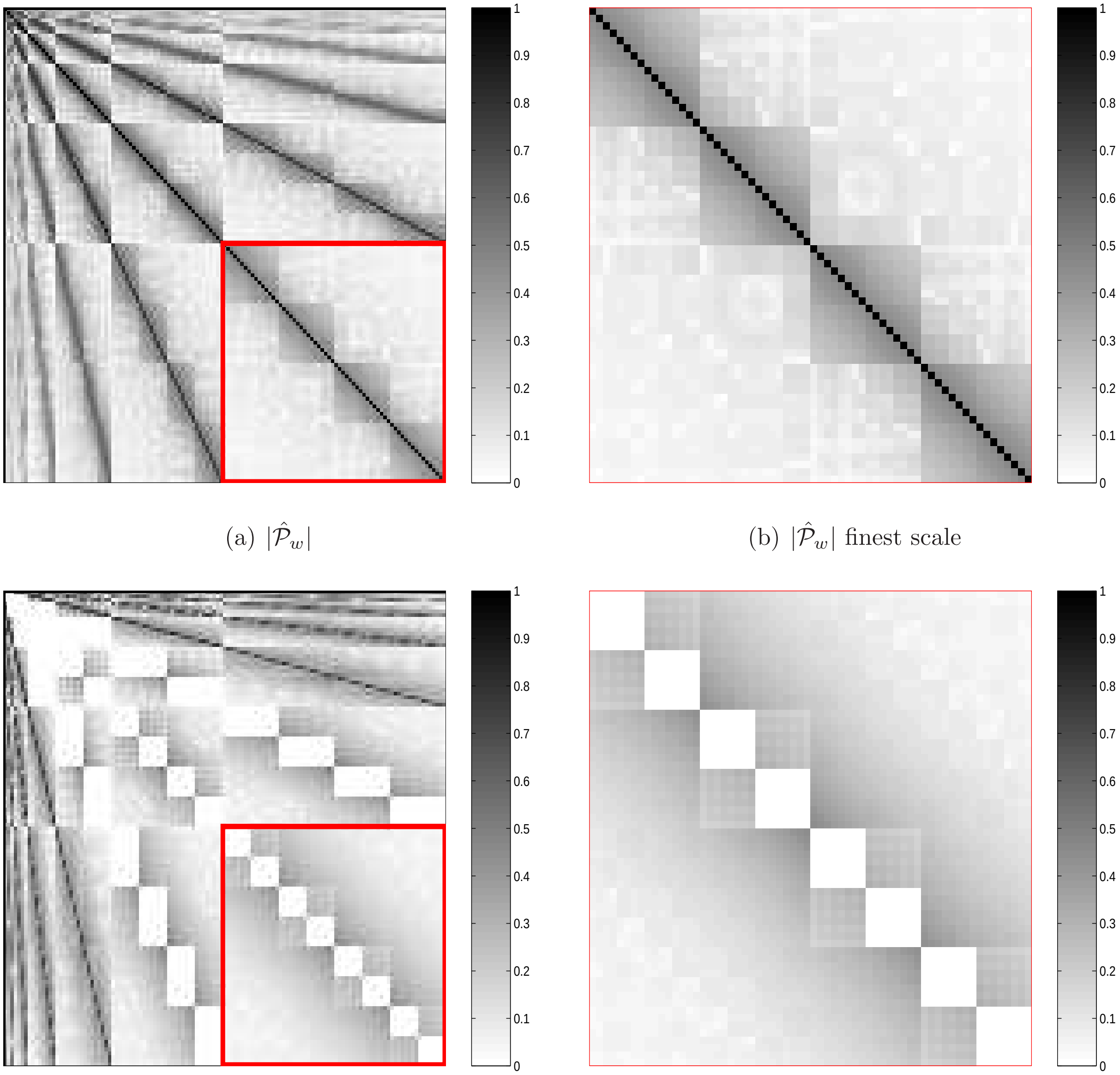
(a) $|\hat{\mathcal{P}}_w|$

(b) $|\hat{\mathcal{P}}_w|$ finest scale

(c) $\left|\mathcal{P}_w - \hat{\mathcal{P}}_w\right|$

(d) $\left|\mathcal{P}_w - \hat{\mathcal{P}}_w\right|$ finest scale

Figure 5.8: Plots parallel to those shown in Figure 5.6, node size d=4. The white color area in panels (c-d) correspond to the node size.



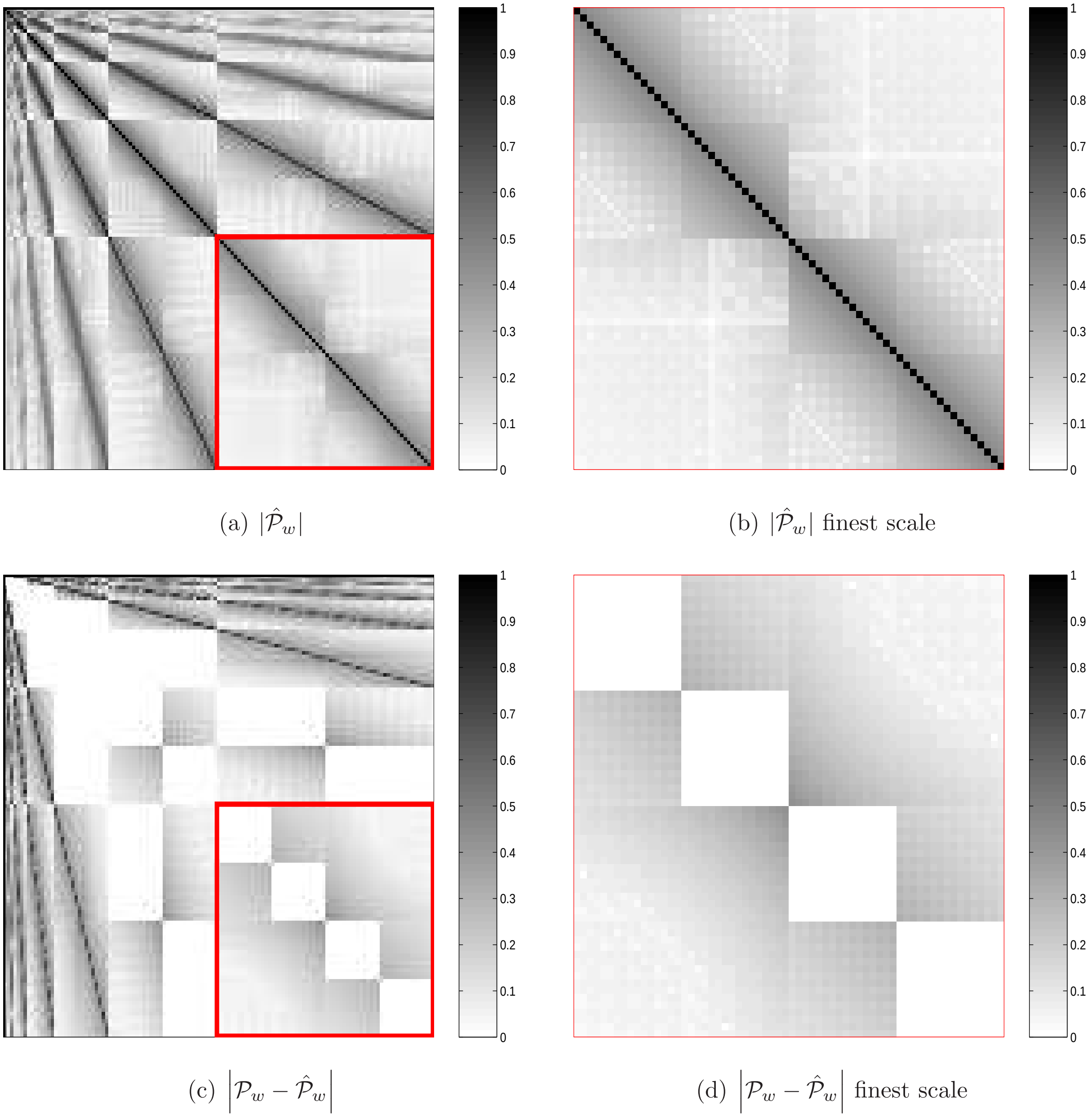
(a) $|\hat{\mathcal{P}}_w|$

(b) $|\hat{\mathcal{P}}_w|$ finest scale

(c) $\left|\mathcal{P}_w - \hat{\mathcal{P}}_w\right|$

(d) $\left|\mathcal{P}_w - \hat{\mathcal{P}}_w\right|$ finest scale

Figure 5.9: Plots parallel to those shown in Figure 5.6, node size d=8.



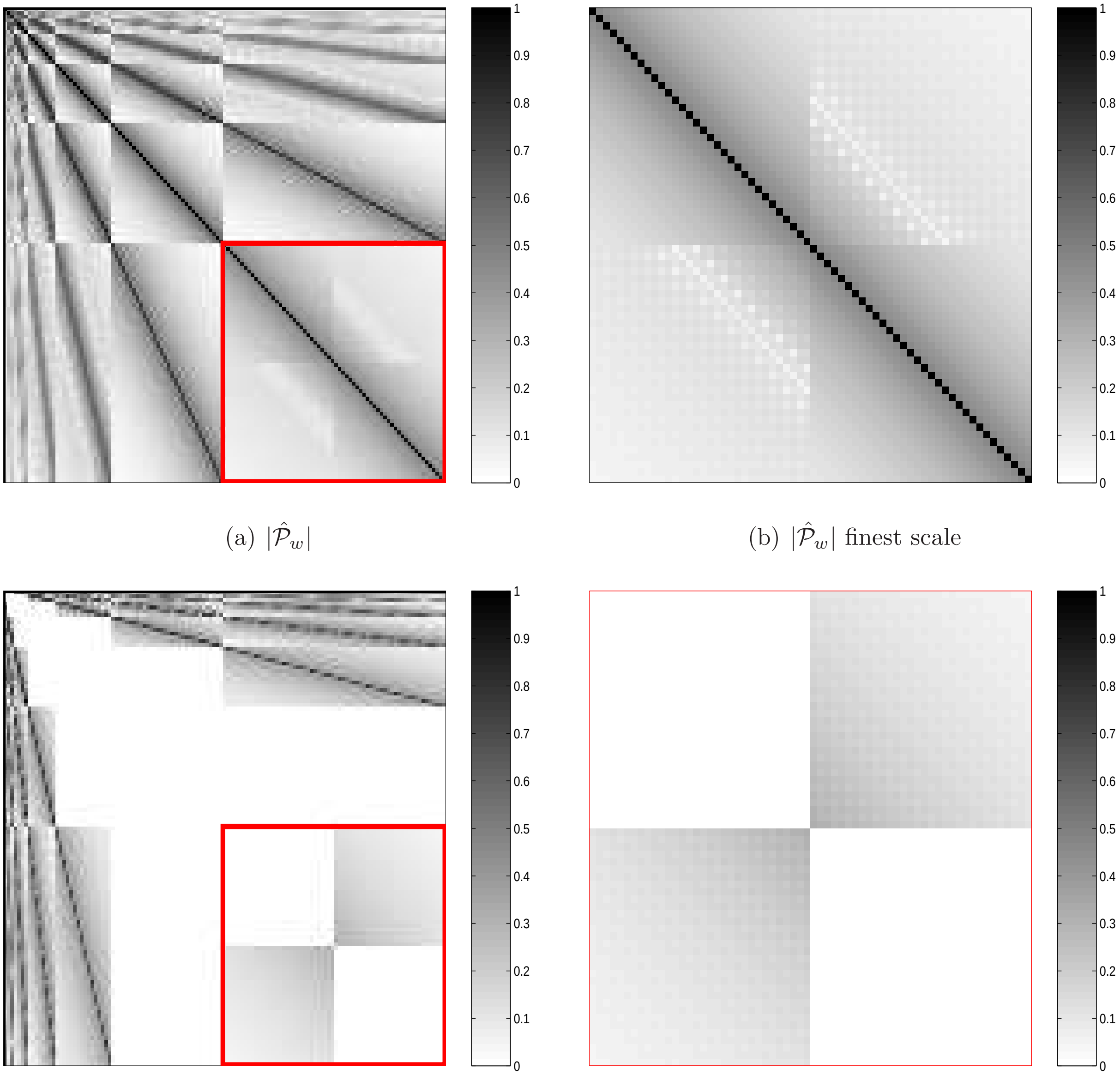
(a) $|\hat{\mathcal{P}}_w|$

(b) $|\hat{\mathcal{P}}_w|$ finest scale

(c) $\left|\mathcal{P}_w - \hat{\mathcal{P}}_w\right|$

(d) $\left|\mathcal{P}_w - \hat{\mathcal{P}}_w\right|$ finest scale

Figure 5.10: Plots parallel to those shown in Figure 5.6, node size d=16.



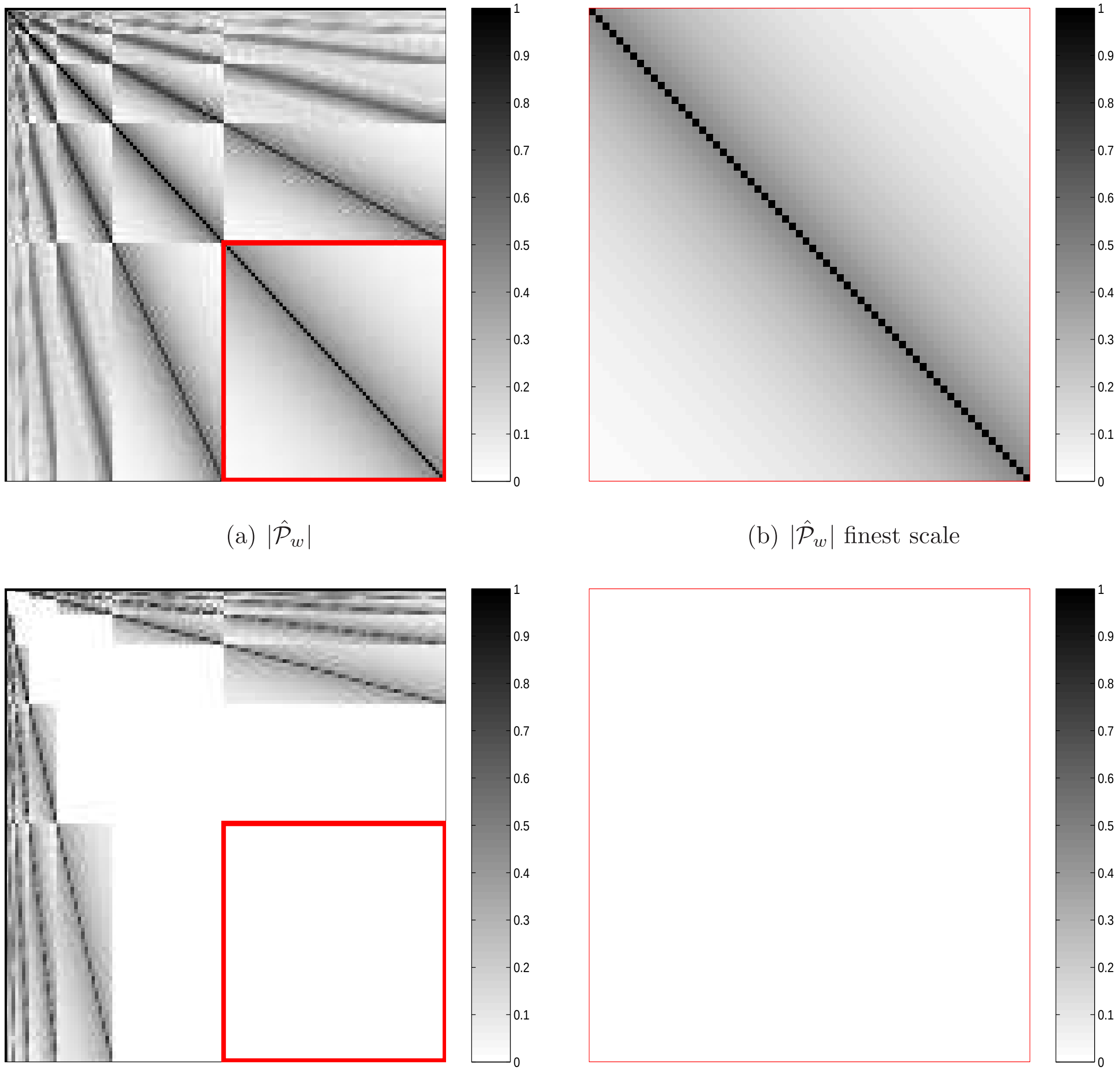
(a) $|\hat{\mathcal{P}}_w|$

(b) $|\hat{\mathcal{P}}_w|$ finest scale

(c) $\left|\mathcal{P}_w - \hat{\mathcal{P}}_w\right|$

(d) $\left|\mathcal{P}_w - \hat{\mathcal{P}}_w\right|$ finest scale

Figure 5.11: Plots parallel to those shown in Figure 5.6, node size d=32.



(a) $|\hat{\mathcal{P}}_w|$

(b) $|\hat{\mathcal{P}}_w|$ finest scale

(c) $\left|\mathcal{P}_w - \hat{\mathcal{P}}_w\right|$

(d) $\left|\mathcal{P}_w - \hat{\mathcal{P}}_w\right|$ finest scale

Figure 5.12: Plots parallel to those shown in Figure 5.6, node size d=64. Note how the accuracy of the MS model can exactly recover the finest scale correlations in this example.

5.1.2 MS Modeling of a Quad Tree

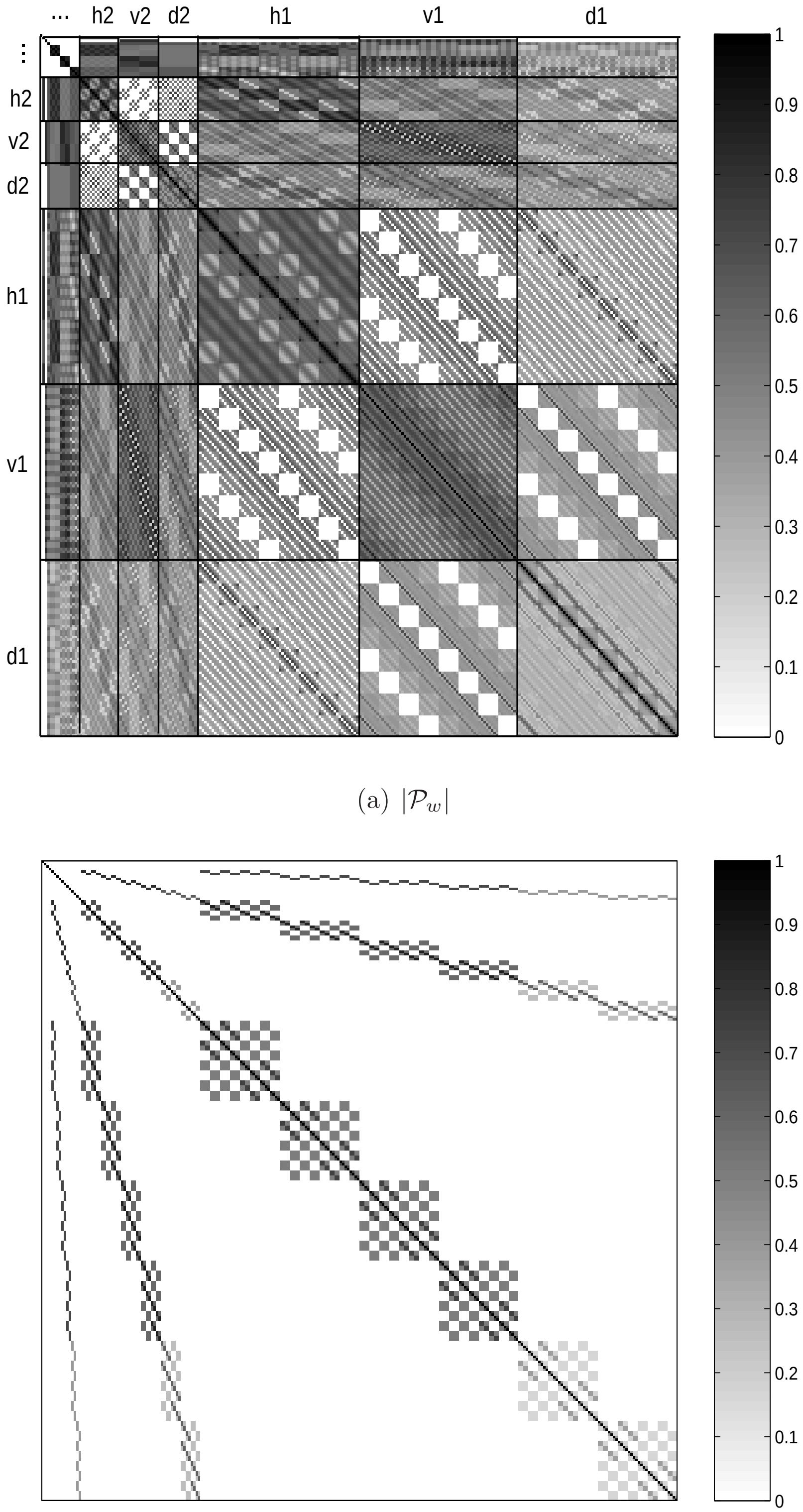
The two-dimensional expansion of the MS model implemented in the previous section is adapted to approximate the quad-tree associated with the 2-D WT. Figure 5.13(b) shows the correlation structure obtained by imposing the first-order scalar MS process (5.2) on the original model of Figure 4.12(h)[3], leading to the following observations:

- The inter-scale correlations, even up to distantly separated scales, are well absorbed by this recursive stochastic model.
- The clear locality of neighborhood dependencies exhibits a within-scale Markovianity.
- The MS modeling is an approach to sparsify the wavelet joint statistics by describing the most significant statistical information between tree parents and children.

Figure 5.14 compares statistical accuracy (RMSE) and complexity (matrix density) of the proposed MS-based wavelet model with the variety of possible wavelet correlation structures examined in § 4.3.1. The MS-based correlation structure is promising and outperforms the estimation based on the decoupling assumption of the WT. The MS-based structure with relatively few coefficients (a sparse structure of the huge covariance matrix) reduces the RMSE.

Regardless of its successful absorption of the across-scale dependencies, this model still demands improvements in describing the within-scale relations. The interrelationship of pixels within a scale (*i.e.*, across and within subbands) is only implicit and very limited (Figure 4.21). Two coefficients may be spatially close, they can be located on distantly separated branches of the wavelet tree. In other words, the correlations between spatially

[3] In order to compare the MS-based approximation with the numerical study of the wavelet covariance structure the same spatial prior as used in § 4.3.1 is adopted here.



(a) $|\mathcal{P}_w|$

(b) $|\hat{\mathcal{P}}_w|$

Figure 5.13: (a) A 2-D wavelet correlation structure and (b) its MS-based approximation.

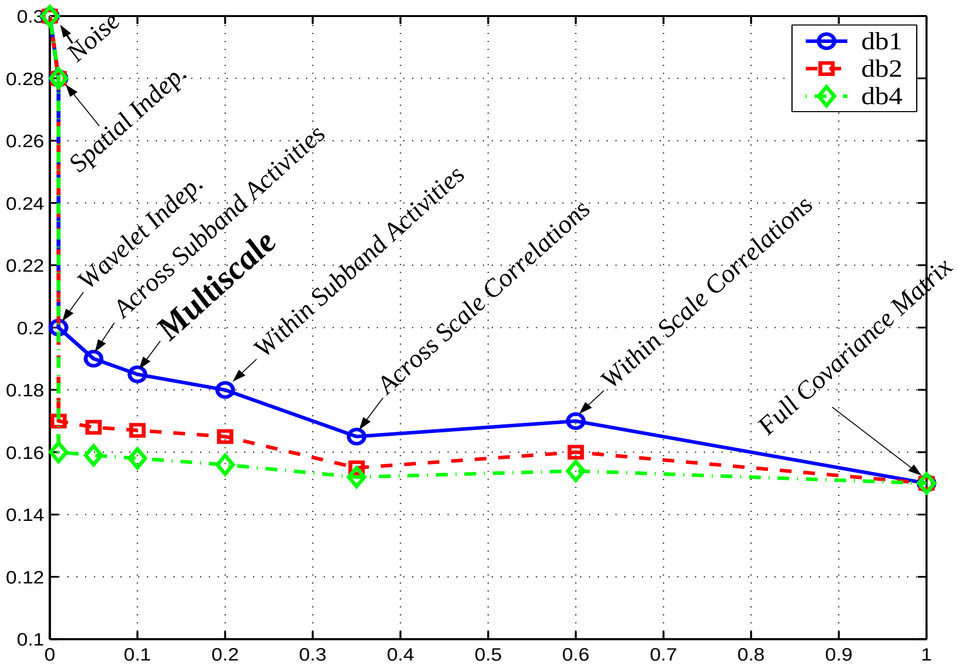


Figure 5.14: RMSE performance as a function of covariance density for an image denoising problem. It is evident how a tiny fraction (but significant) of correlations absorbed by MS model already provides the majority of the improvement.

close siblings are hidden through their parents, because of the MS conditional independence assumption. This performance makes the MS modeling a wavelet orientation independent approach. As is seen in Figure 5.13(a), the diagonal block entities indicate that the direction of correlation between siblings depends explicitly on the subband orientation. However, this important property of the spatial neighborhood is hidden in Figure 5.13(b). Figure 4.21 demonstrates that how the spatial proximity of two coefficients within a subband is different from their location on distantly separated branches of the wavelet tree. Consequently a standard wavelet quad-tree, modeling only parent-child relationships, will only poorly represent spatial interrelationships in those cases where they are found to be significant.

This is a disadvantage of wavelet first-order MS modeling of only parent-child relationships with poor representation of spatial interactions. An alternative is to use higher order MS models with higher node dimensions. This, however, increases the computational cost of the corresponding estimation algorithms [100]. Instead, one can consider a more explicit but appropriate modeling of wavelet statistical dependencies on spatial neighbors. Since correlations are present both within and across scales, there should exist MRF models governing these local dependencies. The remainder of this chapter is focused on investigating random fields modeling of wavelet joint statistics.

5.2 Markov Random Fields Modeling

Assuming that the wavelet coefficients are correlated, a neighborhood structure must first be defined. Consider the wavelet correlation maps discussed in § 4.5 (for convenience Figure 5.15, a copy of Figure 4.24, is included here). These maps demonstrate a correlation structure with obvious dependency at scales and subbands, which is weakly modeled by the MS method. Based on these maps six different symmetric neighborhood structures can be chosen. For a coefficient w_i belonging to the wavelet coefficients set $\underline{w} = \{\underline{w}_h, \underline{w}_v, \underline{w}_d\}$ define

$$p_k(i) = \{p^1(i), \ldots, p^k(i)\}$$
$$c_k(i) = \{c^1(i), \ldots, c^k(i)\}$$

$$s_{ud}(i) : \begin{array}{c} \bullet \\ \times \\ \bullet \end{array} \qquad s_1(i) : \begin{array}{ccc} & \bullet & \\ \bullet & \times & \bullet \\ & \bullet & \end{array} \qquad s_2(i) : \begin{array}{ccc} \bullet & \bullet & \bullet \\ \bullet & \times & \bullet \\ \bullet & \bullet & \bullet \end{array}$$

$$s_{ud}^2(i) : \begin{array}{c} \bullet \\ \bullet \\ \times \\ \bullet \\ \bullet \end{array} \qquad s_{lr}(i) : \bullet \times \bullet \qquad s_{lr}^2(w) : \bullet\ \bullet \times \bullet\ \bullet$$

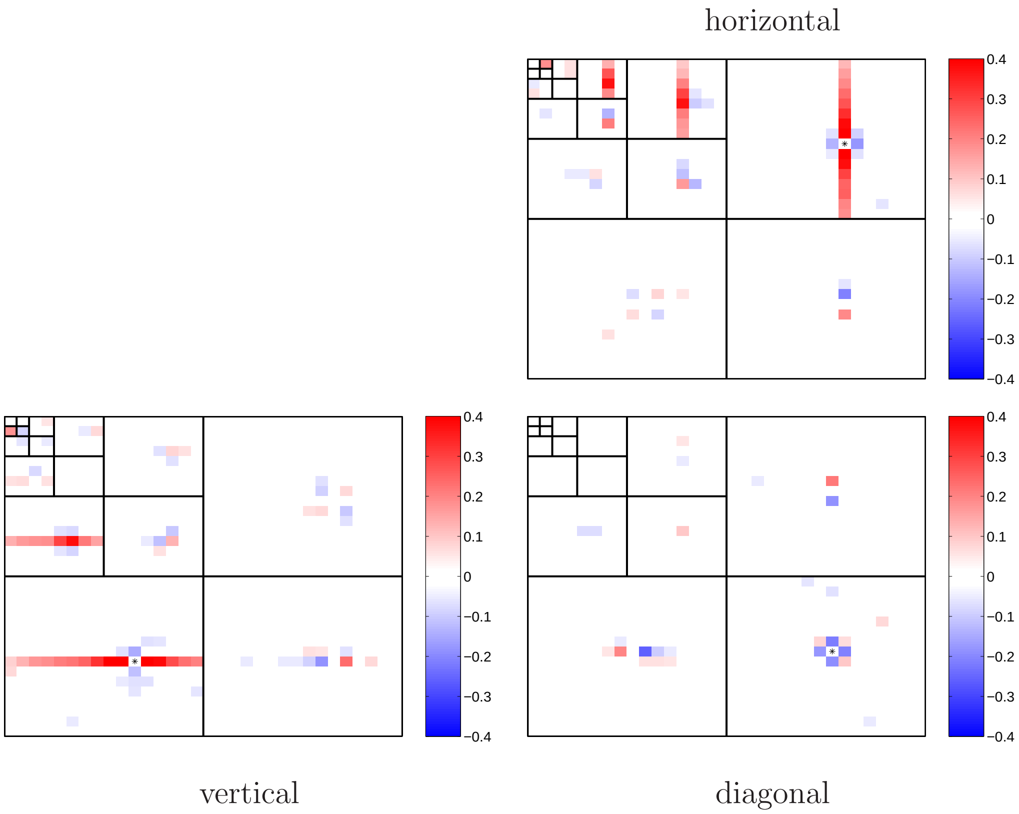


Figure 5.15: Wavelet (db2) correlation structures averaged over a collection of 5000 real images. Each panel contains three plots illustrating the correlations of a given coefficient (•) with its local neighborhoods in the horizontal, vertical, and diagonal subbands. This is a copy of Figure 4.24.

where $p^{\alpha}(i)$ is the ancestor of i of α generations (scales), $c^{\alpha}(i)$ is the set of descendants of i of α generations (scales), and $s_n^{\alpha}(i)$ defines various sibling sets (at the same scale as i).

This allows us to propose six neighborhood structures:

$$\begin{aligned}
\mathcal{N}^1(i) &= \{p_1(i), c_1(i), s_1(i)\} \\
\mathcal{N}^2(i) &= \{p_1(i), c_1(i), s_2(i)\} \\
\mathcal{N}^3(i) &= \{p_2(i), c_2(i), s_1(i)\} \\
\mathcal{N}^4(i) &= \{p_2(i), c_2(i), s_2(i)\}
\end{aligned}$$

$$\mathcal{N}^5(i) = \begin{cases} \{p_2(i), c_2(i), s_2(i), s_{lr}(s_{ud}(v(i))), s_{ud}(d(i))\}, & if \;\; i \in \underline{w}_h \\ \{p_2(i), c_2(i), s_2(i), s_{lr}(s_{ud}(h(i))), s_{lr}(d(i))\}, & if \;\; i \in \underline{w}_v \\ \{p_2(i), c_2(i), s_2(i), s_{lr}(v(i)), s_{ud}(h(i))\}, & if \;\; i \in \underline{w}_d \end{cases} \tag{5.6}$$

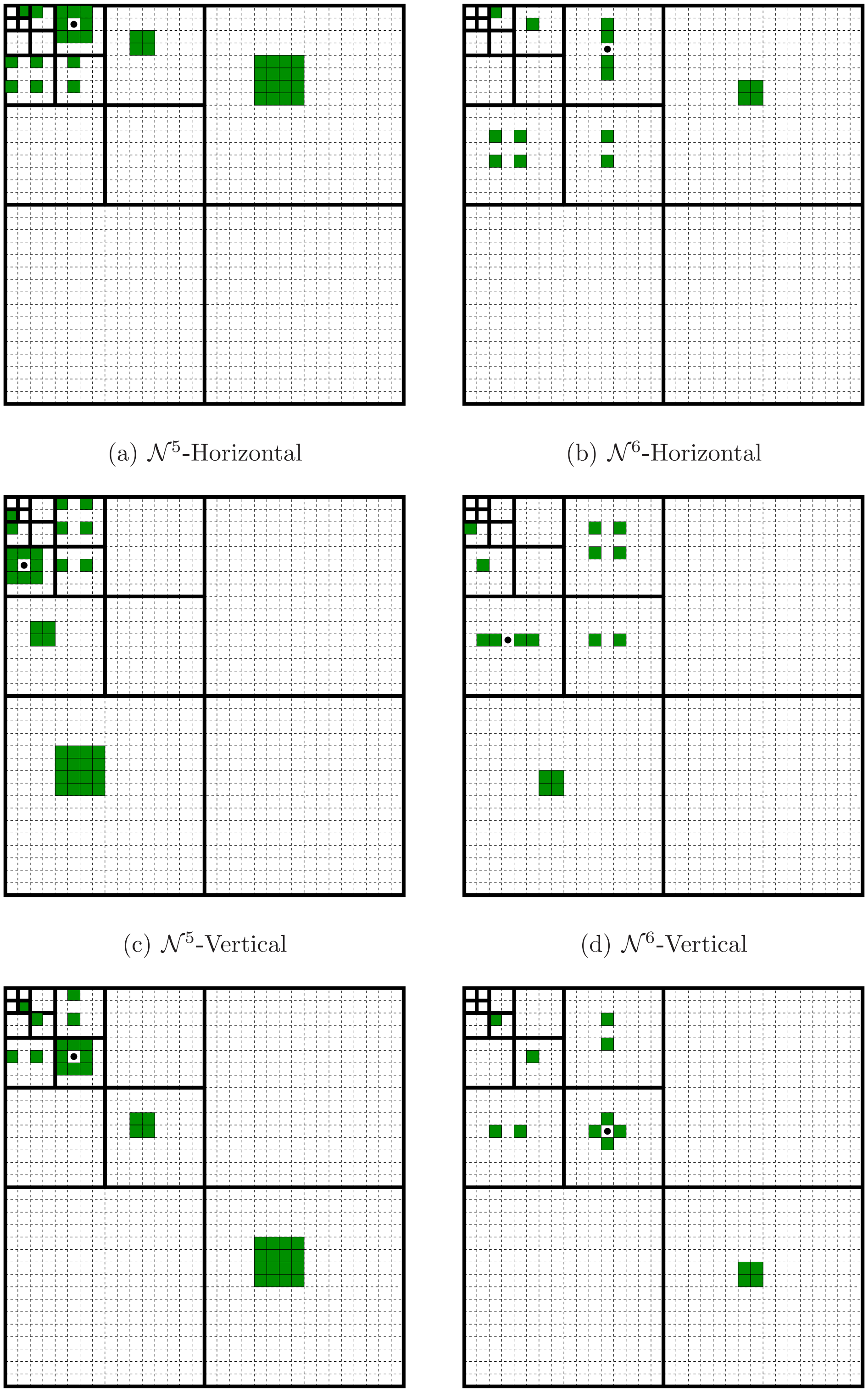
$$\mathcal{N}^6(i) = \begin{cases} \{p_2(i), c_2(i), s_{ud}^2(i), s_{lr}(s_{ud}(v(i))), s_{ud}(d(i))\}, & if \;\; i \in \underline{w}_h \\ \{p_2(i), c_2(i), s_{lr}^2(i), s_{lr}(s_{ud}(h(i))), s_{lr}(d(i))\}, & if \;\; i \in \underline{w}_v \\ \{p_2(i), c_2(i), s_1(i), s_{lr}(v(i)), s_{ud}(h(i))\}, & if \;\; i \in \underline{w}_d \end{cases}$$

where operators $d(i)$, $v(i)$, and $h(i)$ return diagonal, vertical, and horizontal subband counterparts to a given index i. The last two neighborhood systems are visually illustrated by Figure 5.16. With these hypothesized structures in place, the remainder of this section develops and tests two associated models.

5.2.1 Local Estimation

Before any attempt to approximate the wavelet-domain covariance with a sparse structure, the effectiveness of the notion of wavelet locality observed in Figure 5.15 and introduced by (5.6) is examined. I begin with an explicitly local estimator, where only those measurements within the neighborhood are used. Thus, given the noisy measurements

$$\underline{y} = \underline{w} + \underline{\nu}, \qquad \underline{\nu} \sim \mathcal{N}(\underline{0}, R)$$



(a) $\mathcal{N}^5$-Horizontal

(b) $\mathcal{N}^6$-Horizontal

(c) $\mathcal{N}^5$-Vertical

(d) $\mathcal{N}^6$-Vertical

(e) $\mathcal{N}^5$-Diagonal

(f) $\mathcal{N}^6$-Diagonal

Figure 5.16: Structures of two proposed MRF neighborhood systems $\mathcal{N}^5$ and $\mathcal{N}^6$ in (5.6). Note the symmetric structure between the horizontal and the vertical subbands.

and a neighborhood system $\mathcal{N}$, form two neighborhood vectors

$$\begin{aligned} \underline{y}_i &= [y_i, \{y_j; j \in \mathcal{N}_i\}]^T \\ \underline{w}_i &= [w_i, \{w_j; j \in \mathcal{N}_i\}]^T \end{aligned}$$

If $\underline{w}_i$ is assumed jointly Gaussian (as an approximate assumption), the standard estimator follows trivially

$$\begin{aligned} \underline{\hat{w}}_i &= P_{\underline{w}_i,\underline{y}_i} \cdot P_{\underline{y}_i}^{-1} \cdot \underline{y}_i \\ \underline{\hat{w}}_i &= L_i \cdot \underline{y}_i \end{aligned} \qquad (5.7)$$

where

$$\begin{aligned} \hat{w}_i &= \underline{\hat{w}}_i(1) \\ &= E[w_i | \underline{y}_i] \end{aligned} \qquad (5.8)$$

is the only quantity of interest at this stage.

The estimation error covariance for $\underline{\hat{w}}_i = L_i \cdot \underline{y}_i$ is

$$\begin{aligned} cov(\underline{\hat{w}}_i - \underline{w}_i) &= cov(L_i \cdot \underline{y}_i - \underline{w}_i) \\ &= (L_i \cdot \underline{y}_i - \underline{w}_i)(L_i \cdot \underline{y}_i - \underline{w}_i)^T \\ &= L_i(P_{w_i} + R_i)L_i^T - L_i P_{\underline{w}_i} - P_{\underline{w}_i} L_i^T + P_{\underline{w}_i} \\ \tilde{P}_i &= (I - L_i)P_{\underline{w}_i}(I - L_i)^T + L_i R_i L_i^T \end{aligned} \qquad (5.9)$$

For every individual wavelet coefficient the local covariances $P_{\underline{w}_i,\underline{y}_i}$ and $P_{\underline{y}_i}$ are obtained from the original wavelet domain covariance P_w. However, because the true covariance matrix P_w is hard to obtain, then we want to use an approximate estimate, based on the approximated $\bar{P}_{\underline{w}_i,\underline{y}_i}$ and $\bar{P}_{\underline{y}_i}$, which leads (5.7) to

$$\begin{aligned} \hat{\underline{\hat{w}}}_i &= \bar{P}_{\underline{w}_i,\underline{y}_i} \cdot \bar{P}_{\underline{y}_i}^{-1} \cdot \underline{y}_i \\ \hat{\underline{\hat{w}}}_i &= \bar{L}_i \cdot \underline{y}_i \end{aligned} \qquad (5.10)$$

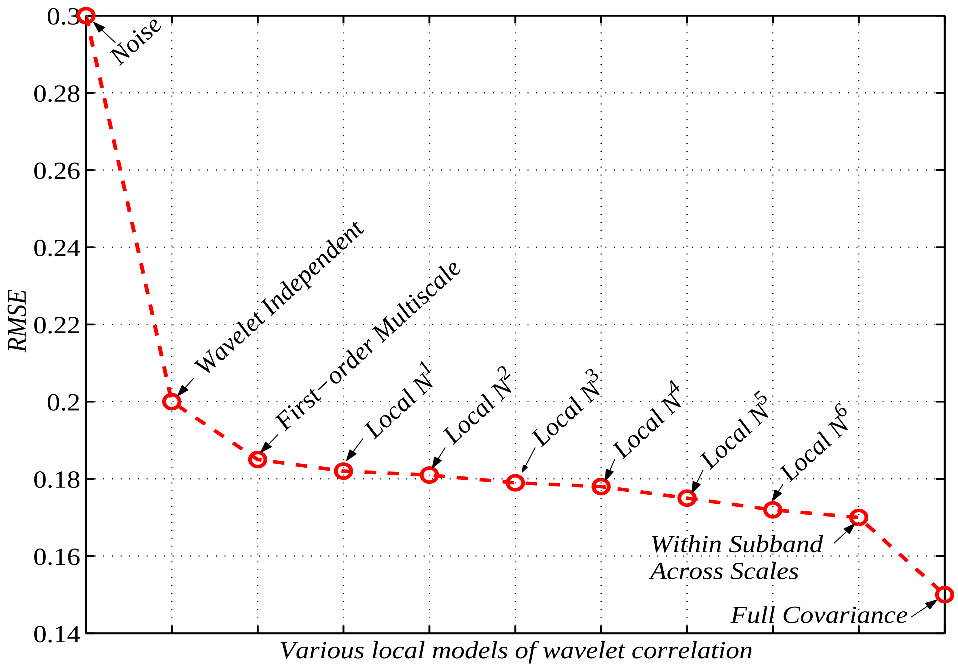


Figure 5.17: RMSE plot as a function of local neighborhood systems of (5.6) used in the explicit local estimation technique of (5.7). The estimation error of even the simplest local system is lower than that of the MS-based estimation.

To calculate the error covariance for $\underline{\hat{\bar{w}}}_i = \bar{L}_i \cdot \underline{y}_i$ assuming $\bar{P}_w$ may be wrong, *i.e.*,

$$\tilde{\bar{P}}_i \neq (I - \bar{L}_i)\bar{P}_{\underline{w}_i}(I - \bar{L}_i)^T + \bar{L}_i R_i \bar{L}_i^T \tag{5.11}$$

because $\bar{P}_w$ is not the true model. Therefore the actual estimation error covariance is

$$\tilde{\bar{P}}_i = (I - \bar{L}_i)P_{\underline{w}_i}(I - \bar{L}_i)^T + \bar{L}_i R_i \bar{L}_i^T \tag{5.12}$$

the only value of interest is $\tilde{\bar{P}}_i(1,1)$ as the true error variance of $\hat{\bar{w}}_i$.

Figure 5.17 plots the estimation error computed by (5.12) as a function of local neighborhood systems of (5.6) used in the above explicit local estimation technique of (5.10). The improvement of the local-based modeling over MS-based modeling is evident. It is of interest to note that a rather simple neighborhood system leads to a lower estimation error in comparison to the MS-based estimation.

The next section explains how the principles of Markov Random Fields (MRFs) can be adopted to approximate the wavelet covariance P_w based on the striking hierarchy of random fields governing the correlation structure of the wavelet coefficients.

5.2.2 MRF-Based Estimation

For the past three decades, MRFs have been used for texture synthesis and analysis with significant improvements over traditional methods [24, 46, 51]. Although, this framework has not been the best choice for image modeling [86], MRFs can be reasonable choices of modeling wavelet domain statistics because the WT can substantially decrease the spatial large neighborhood to more local structures.

An alternative to the explicit use of local structures is an MRF-based modeling approach, where $\bar{P}_w^{-1}$ is sparse. It is known that the sparse values in $\bar{P}_w^{-1}$ are parameters of the Markov model being considered. However, since the true prior P_w is not Markov, then P_w^{-1} won't be sparse. Therefore a Markov model, which approximates P_w, needs to be estimated. The following approximation technique [56] was employed to sparsify P_w^{-1} based on the given neighboring system.

Wavelet Covariance Approximation:

- Choose a neighborhood structure $\mathcal{N}^\alpha$ from (5.6).

- Zero out non-neighbor elements from the true model inverse

$$\Lambda_w^{-1}(i,j) = \begin{cases} P_w^{-1}(i,j) & i,j \in \mathcal{N}^\alpha \\ 0 & i,j \notin \mathcal{N}^\alpha \end{cases} \tag{5.13}$$

forming the simple diagram

$$P_w \rightarrow P_w^{-1} \xrightarrow{\mathcal{N}^\alpha} \Lambda_w^{-1} \rightarrow \Lambda_w$$

- Modify Λ_w to get the same variance values as the original model P_w:

$$\begin{aligned} d_i &= \sqrt{\frac{P_w(i,i)}{\Lambda_w(i,i)}} \\ \bar{P}_w(i,:) &= d_i \cdot \Lambda_w(i,:) \\ \bar{P}_w(:,i) &= d_i \cdot \Lambda_w(:,i) \end{aligned} \tag{5.14}$$

Now $\bar{P}_w$ is a Markov prior. To obtain the actual MRF model coefficients, the following learning process is used.

Parameter Estimation:

- Local interaction for every wavelet coefficient w_i is assumed to be

$$w_i = \sum_{j \in \mathcal{N}_i^\alpha} g_{i,j} w_j + \eta_i \tag{5.15}$$

  with $\eta \sim \Lambda$ as a driven non-white noise process, uncorrelated with the process w_i.

- To find the relationship between w_i and its neighbors $\mathcal{N}_i^\alpha$, let the local parameters $g_{i,j}$ be estimated using the Linear Least Square (LLS) method [51]

$$\underline{w}_{\mathcal{N}_i^\alpha} = [w_j; j \in \mathcal{N}_i^\alpha]^T$$

$$\Theta^{s,o} = \left[\begin{array}{ccccccccc} \underline{w}_{\mathcal{N}_1^\alpha} & | & \cdots & | & \underline{w}_{\mathcal{N}_i^\alpha} & | & \cdots & | & \underline{w}_{\mathcal{N}_k^\alpha} \end{array} \right]$$

  where k denotes the size of the subband, s the scale, and o the orientation, meaning that the data-set Θ is scale and subband dependent. With this neighborhood matrix in hand, the estimated parameters are

$$\underline{\hat{g}}^{s,o} = \left[\Theta^{s,o}\Theta^{s,oT}\right]^{-1} \Theta^{s,o}\ \underline{w}^{s,o} \tag{5.16}$$

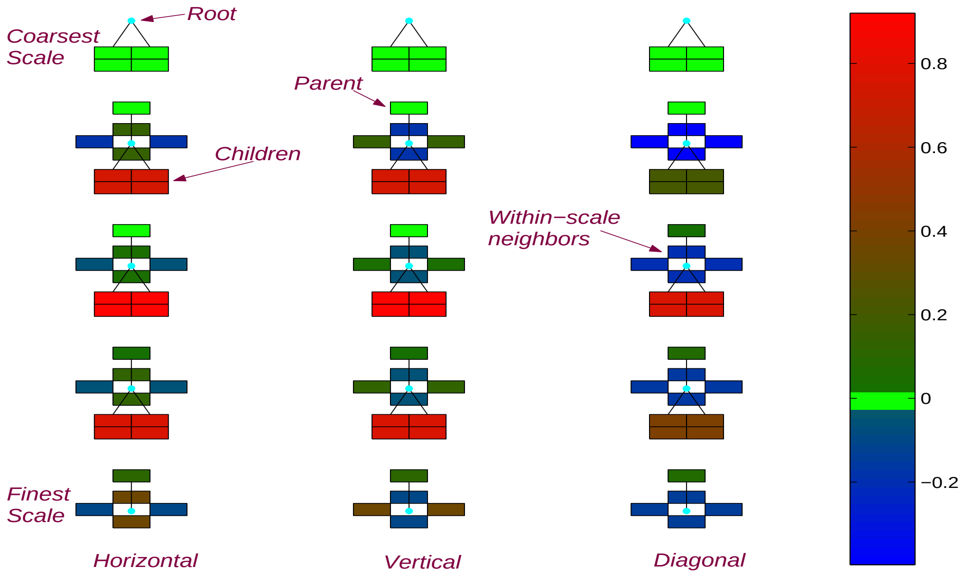


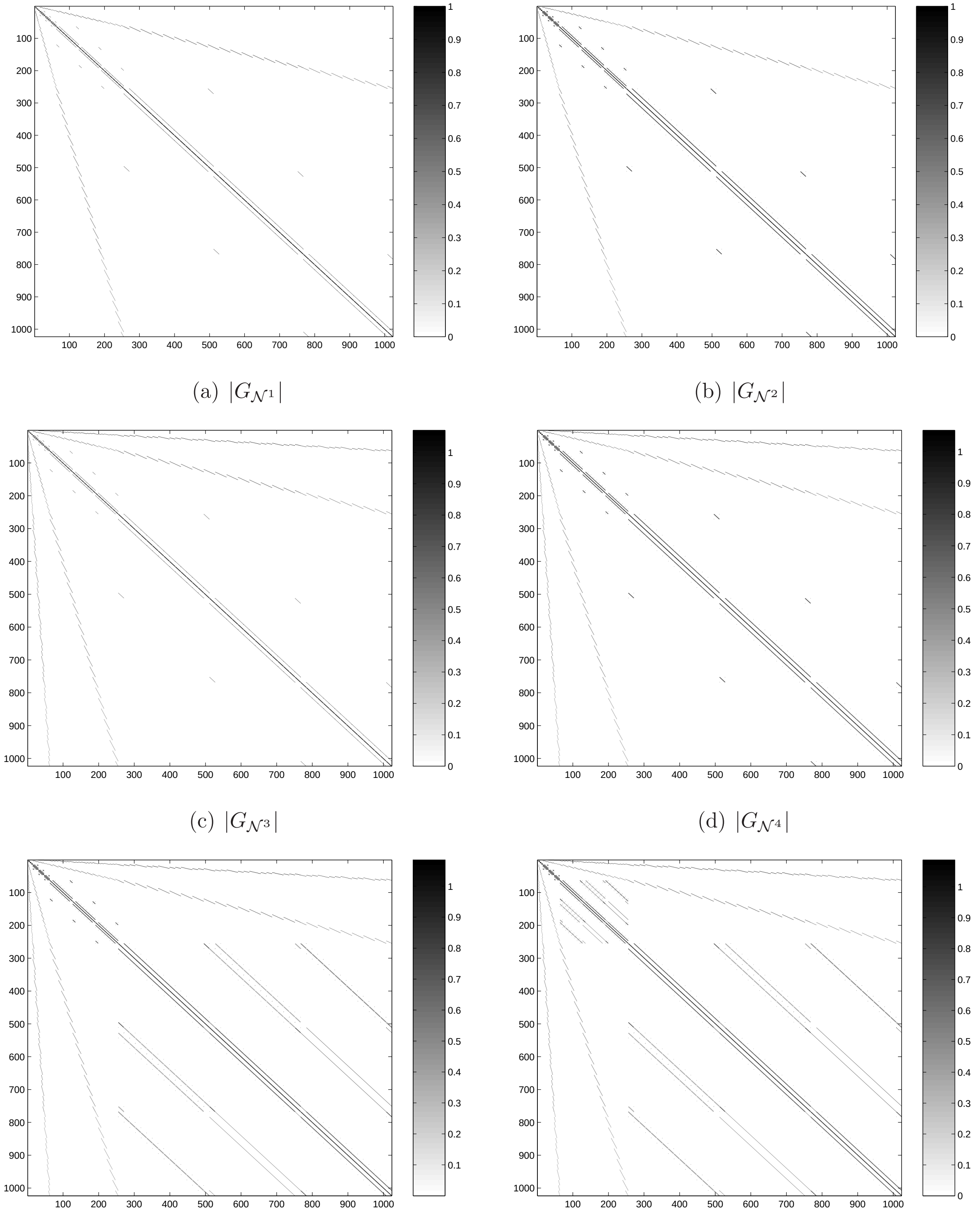
Figure 5.18: The MRF model parameters calculated for a first-order within- and across-scale neighborhood site $\mathcal{N}^1$. The parameters are scale dependent and the hierarchical correlation increases from coarser to finer scales.

> Figure 5.18 displays the model parameters calculated for a simple first-order within- and across-scale neighborhood site $\mathcal{N}^1$ defined in (5.6) for a thin-plate MRF 5-level db2 wavelet transformed. The estimated parameters are scale dependent. They increase the MRF model strength as coefficients dependencies increase from coarse to fine resolutions. The within-scale correlations of horizontal and vertical subbands are symmetrically identical and are stronger than those of the diagonal subband.

To have a clear understanding of the MRF model parameters, the linear system given in (5.15) is re-written in matrix format

$$G\underline{w} = \underline{\eta}$$

where G is a sparse matrix including all the estimated parameters $\underline{g}^{s,o}$ as its off-diagonal



(a) $|G_{\mathcal{N}^1}|$

(b) $|G_{\mathcal{N}^2}|$

(c) $|G_{\mathcal{N}^3}|$

(d) $|G_{\mathcal{N}^4}|$

(e) $|G_{\mathcal{N}^5}|$

(f) $|G_{\mathcal{N}^6}|$

Figure 5.19: All six MRF parameter matrices $G_{\mathcal{N}^\alpha}$ obtained by the LLS estimation algorithm (5.16). The spatial prior was thin-plate MRF and the mother wavelet was db2.

elements and one on its main diagonal. Matrix G shows the structure of wavelet MRF model parameters. Figure 5.19 displays all the G matrices obtained by (5.16) given the six neighborhood systems in (5.6). The parsimony of MRF parameters is obvious, and in all cases the matrix shows a nice sparse structure, which is not necessarily symmetric. Each matrix includes two sets of off-diagonal bands: those parallel to the main diagonal and the finger-like bands. While the former set shows within-scale but across-subband correlations, the latter represents across-scale correlations. Magnitudes of the finger-like bands above the diagonal line are larger than their symmetric counterparts below the diagonal band. Note that, in distinct contrast to the vast majority of planar MRF models in which G is stationary, the structure of the wavelet tree (asymmetry between parent and child, or between siblings) makes G nonstationary and considerably complicates model estimation. This property is theoretically explained in the following discussion.

In (5.15) η_i was assumed to be a non-white noise process, uncorrelated with the process w_i. Assuming $j \in \mathcal{N}_i^\alpha$ and that w_j and w_i belong to different subbands, then

$$\begin{aligned}
E\left[\eta_i \cdot \eta_j\right] &= E\left[\eta_i \cdot \left\{w_j - \sum g_{j,k} w_k\right\}\right] \\
&= -g_{j,i} E\left[\eta_i \cdot w_i\right] \\
&= -g_{j,i} E\left[\eta_i \cdot \left\{\sum g_{i,j} w_j + \eta_i\right\}\right] \\
&= -g_{j,i} E\left[\eta_i \cdot \eta_i\right] \\
&= -g_{j,i} \lambda_i^2 \\
E\left[\eta_j \cdot \eta_i\right] &= -g_{i,j} \lambda_j^2 \\
\frac{g_{i,j}}{g_{j,i}} &= \frac{\lambda_i^2}{\lambda_j^2} \qquad (5.17)
\end{aligned}$$

where the above ratio does not have to equal one, because λ_i^2 and λ_j^2 are the variances of two nodes belonging to two different regions. In other words, one can say that parent-child dependency in the wavelet domain is direction dependent, clearly because the wavelet subbands have different variance values with larger values at coarser scales. Thus, the model parameter $g_{i,j}$ which describes the status of a parent w_i (greater variance) by its child w_j (smaller variance) is not necessarily equal to $g_{j,i}$ describing a child w_j by its parent w_i.

This section started with the strong belief that the structure of wavelet correlation was not limited to the parent-child dependencies, meaning that the MS-based modeling assumed only a subset of wavelet correlations. The empirical investigations in this section indicate that although any spatial Markovianity is lost with WT of an image [51], the MRF-based modeling still results in an approximated wavelet covariance with a very sparse structure which preserves the significant correlations within and across scales and orientations. Before leaving this chapter, a quantitative evaluation of the proposed random field models, in terms of their accuracy and complexity will be presented.

5.3 Quantitative Evaluations

The six different wavelet neighborhood structures of (5.6) are examined here. Clearly each choice of neighborhood will differ in its statistical accuracy. The six local and MRF-based results are compared with the null estimator

$$\hat{w}_i = y_i$$

and the pointwise estimator

$$\hat{w}_i = \frac{\sigma^2_{w_i}}{\sigma^2_{y_i}} y_i$$

with the RMSE of all cases plotted in Figure 5.20. It is clear that significant benefit is obtained from relatively few coefficients in the locality of the center coefficient. Empirically, the presence of within-scale (and across-orientation) correlation in these simulations (from

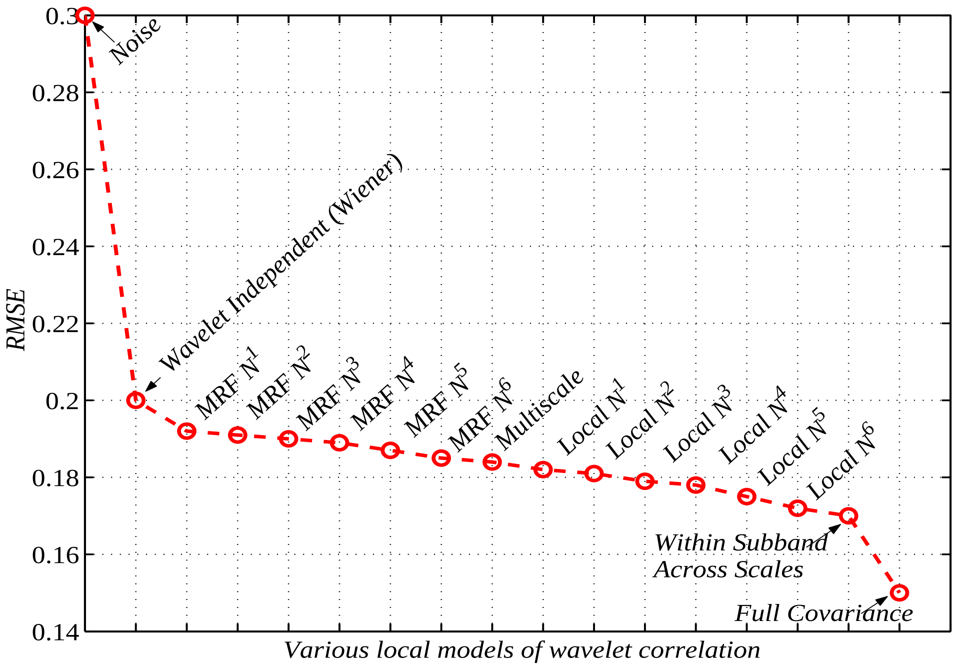


Figure 5.20: RMSE plot as a function of local neighborhood systems of (5.6) used in the MRF-based approximation techniques. Evidently, the estimation error the MRF-based models for each local structure is lower than that of the MS-based estimation. This is a comprehensive plot comparing to the one displayed in Figure 5.17.

$\mathcal{N}^1$ to $\mathcal{N}^6$) reduces the estimation error. Evidently, the local estimation (5.10) outperforms the MRF-based method (5.14). The local estimate is an explicit use of the actual local covariances P_w rather than the approximated values $\bar{P}_w$ of (5.14), but with no assumption of correlations beyond a per-defined locality. The MRF-based estimate uses its approximate model rather than the true model, but it asserts conditional independence, in which the long range correlations are implicitly considered.

The second aspect of comparison is computational complexity. In increasing order of complexity is a) Pointwise, b) Local, c) Multiscale, d) MRF, e) Full model.

Clearly the pointwise method is a linear approach with its complexity growing linearly as the number of wavelet coefficients n increases. On the other side, the complexity of the MS-based estimator is $\mathcal{O}(d^3 n)$, where d shows the tree-node's dimensionality (in the simplest case $d = 1$); see Table 5.1.

Re-write (5.7) to investigate the complexity of the local models

$$\underline{\hat{w}}_i = P_{\underline{w}_i,\underline{y}_i} \cdot P_{\underline{y}_i}^{-1} \cdot \underline{y}_i = L_i \cdot \underline{y}_i \qquad 1 \leq i \leq n \tag{5.18}$$

with matrix L_i of size $m \times m$, where $m \ll n$ denotes the neighborhood size. The complexity of calculating L_i is of order $\mathcal{O}(m^3)$. The cost of the estimator, however, depends on the prior, which can be stationary or not. Both cases are considered at this point.

- Stationary wavelet prior model:

  In this case the complexity of the model estimation process L is fixed to $\mathcal{O}(m^3)$, because $L_i = L_j$, if $j \neq i$. Thus total complexity of the estimation process $\underline{w} = L\underline{y}$ is $\mathcal{O}(m^3 + n \cdot m^2)$.

- Non-stationary wavelet prior model:

  It is known that a stationary prior projected into the wavelet domain changes to nonstationary because of the multiscale nature of the wavelet domain. In this case the complexity of the model estimation process L is $\mathcal{O}(n \cdot m^3)$, because $L_i \neq L_j$, if $j \neq i$. Thus, the overall complexity of the local estimator $\underline{\hat{w}} = L\underline{y}$ is $\mathcal{O}(n \cdot m^3)$.

The computational cost for the MRF-based estimator is more complicated. In this work, only the simple linear case, *i.e.*, a Gaussian prior is assumed. Consider the MRF prior

$$G\underline{w} = \underline{\eta}, \qquad \underline{\eta} \sim \mathcal{N}(\underline{0}, \Lambda)$$

and the measurement

$$\underline{y} = \underline{w} + \underline{\nu}, \quad \underline{\nu} \sim \mathcal{N}(\underline{0}, R)$$

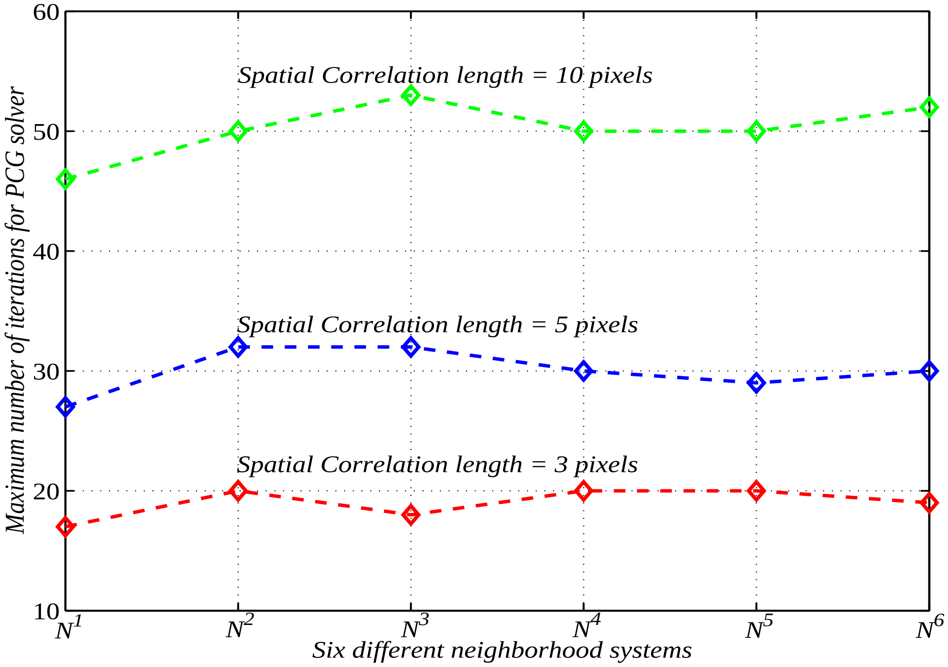


Figure 5.21: Maximum number of iterations for the PCG to solve (5.19) as a function of wavelet domain MRF neighborhood systems (5.6). These simulations were run for thin-plate prior GMRF with three different correlation lengths. The mother wavelet was db2.

Define a linear estimator to find $\underline{w}$ which minimizes

$$\begin{aligned} &\arg_{\underline{w}} \min \left|\left|\underline{y} - \underline{\hat{w}}\right|\right|_{R^{-1}} + ||G\underline{\hat{w}}||_{\Lambda^{-1}} \\ &\Rightarrow \underline{\hat{w}} = \left(R^{-1} + G^T\Lambda^{-1}G\right)^{-1} R^{-1}\underline{y} \end{aligned} \tag{5.19}$$

which is a linear system of equations to be solved. Several iterative solvers such as Gauss-Sidel [84] and Preconditioned Conjugate Gradient (PCG) [84] were tested. The empirical results for the PCG algorithm is discussed here. The computational complexity of the PCG is $\mathcal{O}(itrn \cdot m \cdot n)$, where m is the neighborhood size and $itrn$ shows number of iterations for a solver to converge to a predefined tolerance value. The experimental results indicate a surprisingly fast convergence speed. The PCG algorithm was run for

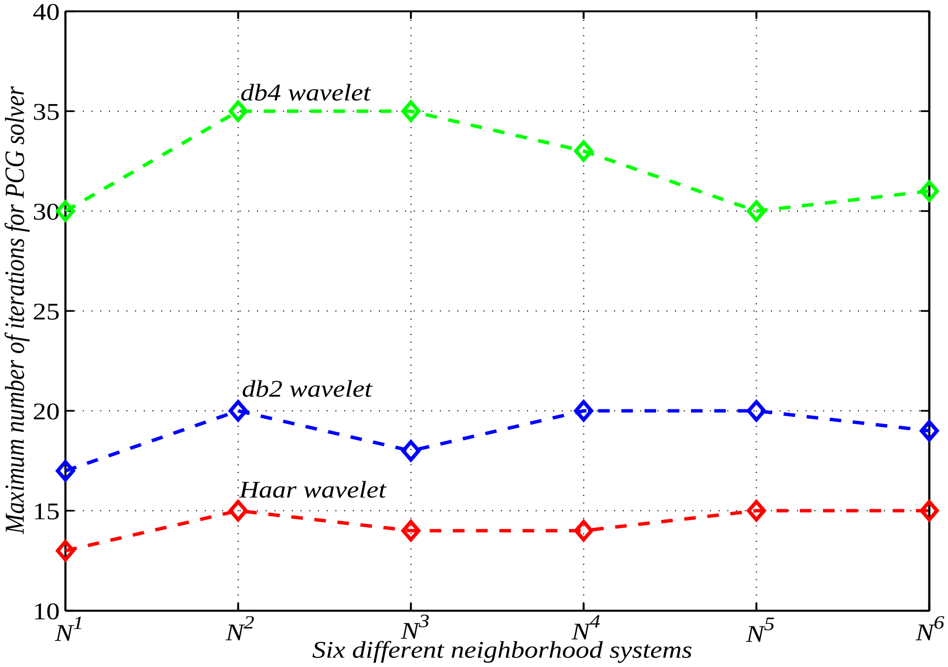


Figure 5.22: Maximum number of iterations for the PCG adopted in solving (5.19) as a function of wavelet domain MRF neighborhood systems (5.6). The simulations were run for thin-plate GMRF with fixed spatial correlation length. The mother wavelets db1-4 were considered.

six different MRF-based neighborhood systems. In these experiments a thin-plate prior model with three different correlation lengths was considered. For the purpose of simplicity textures of size 32 × 32 projected by Daubechies wavelets are examined.

Figure 5.21 illustrates the *itrn* number for the PCG to solve (5.19) for all six neighborhood systems, where the mother wavelet was fixed to be db2. A thin-plate prior of three different correlation lengths was used. The results indicate that the convergence speed is very fast for the PCG. For a fixed correlation length, *itrn* number remains almost unchanged for different neighborhood sizes. However, the increment of correlation length, *i.e.*, the larger extent of pixels connectivity (smoothness), results in estimators with higher

Methodology	**Assumption**	**Complexity**
Pointwise	independence	$\mathcal{O}(n)$
Local	stationarity	$\mathcal{O}(m^3 + m^2 \cdot n)$
Local	non-stationarity	$\mathcal{O}(m^3 \cdot n)$
Multiscale	first-order	$\mathcal{O}(n)$
Multiscale	$log_4 n^{th}$-order	$\mathcal{O}(3(\frac{n}{4})^3)$
MRF	Guass-Sidel	$\mathcal{O}(itrn \cdot m \cdot n)$
MRF	PCG	$\mathcal{O}(itrn \cdot m \cdot n)$

Table 5.2: List of the model-based wavelet estimators' complexities. m: neighborhood size, n: data size, $itrn$: no. of iterations.

computational cost.

To examine the sensitivity of the proposed MRF-based technique with change of basis, various Daubechies wavelets were investigated. Figure 5.22 shows the $itrn$ number of the PCG solver for all $\mathcal{N}^\alpha, \alpha = 1, \cdots, 6$ systems, where a thin-plate GMRF with a fixed correlation length was used. The term $itrn$ remains small for all cases and grows slowly where more regular wavelets is considered. In all experiments $itrn$ is a relatively small number which represents the low complexity for the wavelet linear MRF-based estimator.

Table 5.2 lists the complexities associated with the wavelet estimators based on the proposed models. The relatively fast convergence rate of the PCG method makes the MRFs reasonable choices of wavelet local models.

5.4 Chapter Summary

A probabilistic study of wavelet joint statistics was presented in this chapter. An examination of the coefficient correlations, within or across scales, revealed that there exist local stochastic models (explicit or MRF) governing these local dependencies. The proposed hierarchical random fields models exhibit a sparse neighborhood structure which absorbs correlation of any coefficient with the rest of the wavelet tree. The accuracy of the proposed models was evaluated in RMSE sense and in terms of their computability and complexity.

This chapter was only one step in the broader goal (Figure 1.1) of building more capable joint models for image representation. The principal motivation of this work is to devise an estimation or correlated shrinkage algorithm which takes into account the proposed wavelet joint model and results in an optimum estimation error and low computational cost. Chapter 6 will focus on the development of a non-linear correlated empirical Bayesian shrinkage algorithm, with illustrations and evaluations of its estimation results.

Chapter 6

Correlated Wavelet Shrinkage

This chapter proposes a novel correlated shrinkage method based on wavelet joint statistics. The objective is to demonstrate the effectiveness of the wavelet correlation models [4, 9] of Ch. 5 in estimating the original signal from a noisy observation.

The structures of the existing wavelet correlations were studied

1. In § 4.3: Finding the wavelet sample covariance over a large collection of real images and adopting the standard diagram of 2-D WT to display the correlation map for every individual coefficient.

2. In § 4.5: Empirically observing that the wavelet spatial statistics are strongly orientation dependent, structures which are surprisingly not considered by state-of-the-art wavelet modeling techniques.

3. In Ch. 5: Studying probabilistic models including the multiscale as well as Markov random fields to describe the exhibited wavelet neighborhood structure.

Having accomplished the above empirical steps, this chapter is focused on the development of a non-linear correlated empirical Bayesian shrinkage algorithm. Simulation results will demonstrate the advantages of the new correlated shrinkage function. In comparison with popular nonlinear shrinkage algorithms [13, 26], it improves the denoising results. The goal is to obtain a shrinkage method which outperforms the current joint-model based shrinkage algorithms such as HMMs [26] and GSM [79] based techniques.

6.1 Independent Wavelet Shrinkage

Suppose a random field $\underline{x}$ is projected into the wavelet domain with a resulting coefficient vector $\underline{w}$. The objective is to estimate $\underline{\hat{w}}$, given the noisy observation $\underline{y}$:

$$\begin{aligned} \underline{y} &= \underline{w} + \underline{\nu} \qquad \underline{\nu} \sim \mathcal{N}(\underline{0}, \Sigma_\nu) \\ y_i &= w_i + \nu_i \qquad \nu_i \sim \mathcal{N}(0, \sigma_\nu^2) \end{aligned}$$

where $\underline{\nu}$ is assumed additive *i.i.d.* random noise. In general, if the coefficients are assumed *independent* and normally distributed, then the linear Bayesian estimate [38] is optimum in the mean squared error sense

$$\hat{w}_i = E[w_i|y_i] = \frac{\sigma_w^2}{\sigma_w^2 + \sigma_\nu^2} y_i \tag{6.1}$$

However, since the wavelet marginal prior is well-known to be non-Gaussian (generalized Gaussian [14] or generalized Laplacian [86]), then $E[\underline{w}|\underline{y}]$ is non-linear. One of the superior non-linear shrinkage methods, known as BayesShrink [14], determines a threshold $T_{Bayes} = \frac{\sigma_\nu^2}{\sigma_w}$ for each subband assuming a Generalized Gaussian Distribution (GGD) for the coefficients. Chang *et al.* [14] observed that the threshold value T_{Bayes} is very close to the optimum threshold. BayesShrink performs soft thresholding, with its data-driven, subband dependent threshold. The results obtained by BayesShrink visually look more appealing than those obtained using VISUShrink and SUREShrink.[1]. This makes BayesShrink a reasonable choice of comparison with the experimental results obtained in this chapter.

[1] Section 3.3 has already covered a comprehensive review of wavelet shrinkage.

All of these shrinkage algorithms treat the non-Gaussian coefficients as independent, however, based on our observations of the wavelet joint statistics we propose a correlated shrinkage method whose non-linearity is approximated through an empirical Bayesian approach. As opposed to the state-of-the-art wavelet-domain HMMs [26] where the coefficients non-linear relationships are considered through their conditional independence assumption, the proposed algorithm is based on the wavelet linear joint statistics followed by a non-linear Bayesian estimator. The future improvements of this shrinkage algorithm is to be compared with the shrinkage results of the HMMs.

6.2 Empirical Bayesian Estimation

The generalized Gaussian [14] or generalized Laplacian [86] priors for wavelets are, at best, heuristics or approximations. Different classes of images will necessarily have different wavelet priors. It is, therefore, very difficult to talk about or even formulate the optimum Bayesian estimates, making an empirical approach attractive.

6.2.1 Marginal Bayesian Estimate

Given a vast number of $\{w_i, y_i\}$ pairs, the optimum Bayesian expectation can be formulated as a sample mean

$$\hat{w}_i = E[w_i|y_i] \simeq average\{w_j|y_j \simeq y_i\} \tag{6.2}$$

where the $\{w_i\}$ are marginally considered. This is non-linear shrinkage, as the conditional mean will normally not be a linear function of y_i, implemented through the steps in Algorithm 1, where the parameter *winsize* specifies the one-sided length of a window containing those coefficients whose magnitude are close.

Algorithm 1 Empirical Bayesian Estimate

```
1: sort $\{w_i\}$ and $\{y_i\}$ based on $\underline{w}$
2: select a windowing size $winsize$
3: $\hat{w}_i = \sum_{|j-i|<winsize} \alpha_{j-1} w_j, \quad \alpha_0 = 0$
```

Clearly the choice of *winsize* is very important in this method. The larger the averaging window size, the lower the estimation error but the smoother the estimated result. For the purpose of clarification, the independent Bayesian estimate (6.2) was implemented for the “Lena” image and was compared to the universal soft and hard thresholding as well as BayesShrink. Figure 6.1 plots MSE of the estimation method as a function of the additive noise standard deviation. Although, in this experiment the empirical Bayesian was applied universally, *i.e.*, it did not benefit from the scale and subband dependency of BayesShrink, it still outperforms the other shrinkage methods at some noise regions, even for different choices of *winsize*.

This simple demonstration showed that the empirical Bayesian approach is a good estimator to approximate the original data. With this confirmation in hand, this rather independent empirical approach will be expanded to an estimation which depends on the locality of wavelet coefficients.

6.2.2 Joint Bayesian Estimate

To define the joint Bayesian estimate, it must be noticed that

$$E[w_i|\underline{y}] \neq E[w_i|y_i]$$

because the $\{y_i\}$ are not assumed independent because of the correlation in the $\{w_i\}$ [4, 8].

To solve the joint estimate one normally limits the attention to some neighborhood $\mathcal{N}$

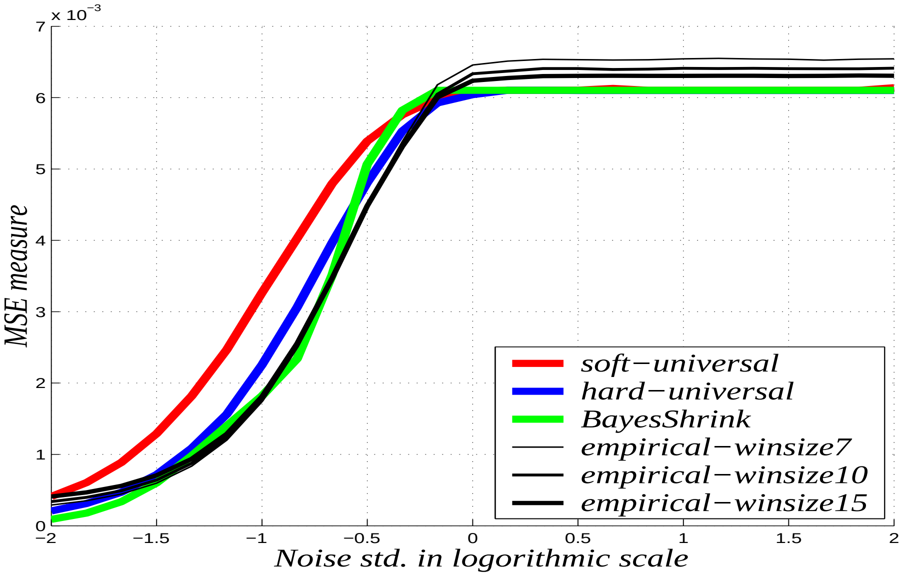


Figure 6.1: Comparison of MSE measures of uniformed-window empirical wavelet shrinkage with universal shrinkage (VISUShrink) and BayesShrink. "Lena" image was 4-level decomposed into the db4 wavelet basis.

of the coefficients

$$E[w_i|\underline{y}] \simeq E\left[w_i|\{y_j; j \in \mathcal{N}_i\}\right] \tag{6.3}$$

This is a high dimensional problem, where it is hard to find the sample average. In principle (6.3) can be solved as before, using empirical Bayes Algorithm 1, but where we now take a sample mean over similar neighborhoods

$$E[w_i|\underline{y}] \simeq average\left\{w_k|y_l \simeq y_j; \quad l = \mathcal{N}_{k,m} \quad j = \mathcal{N}_{i,m}\right\} \tag{6.4}$$

where $\mathcal{N}_{i,m}$ is the m^{th} element index in the neighborhood of i. However, the required data grows exponentially with neighborhood size and is impractical for all but the smallest neighborhoods.

An alternative is

$$E[w_i|\underline{y}] \simeq E\left[w_i|f(\{y_j; j \in \mathcal{N}_i\})\right] \tag{6.5}$$

where $f(\cdot)$ can be any simple or complicated function.

Instead, imagine combining (6.2) and (6.5), using a linear function $f(\cdot)$ in (6.5) to take into account the joint relationships (*e.g.*, the standard linear prediction $\hat{w}_i = \sum_{j\in\mathcal{N}_i} g_{i,j} y_j$, where $j \in \mathcal{N}_i$, if y_j is related to w_i). However, since this linear function does not account for the non-linearity of the shrinkage algorithm the empirical Bayes (6.2) is used to infer any needed non-linearity to find a good estimate. The development of such an approach follows.

6.3 Correlated Shrinkage

This section is meant to develop a probabilistic shrinkage algorithm which considers models of wavelet localities. It was observed that the wavelet correlations are scale and subband dependent, meaning that the correlated shrinkage must be subband dependent.

Based on the correlation map of Figure 5.15 and neighborhood symbols (5.6), §5.2 defined various different neighborhoods, of which only two structures are repeated here. For a coefficient w_i belonging to the wavelet coefficients set $\underline{w} = \{\underline{w}_h, \underline{w}_v, \underline{w}_d\}$ define two *asymmetric* neighborhood structures:

$$\mathcal{N}^{asym1}(i) = \begin{cases} \{s_{ud}(i), p_1(i)\}, & if \;\; i \in \underline{w}_h \\ \{s_{lr}(i), p_1(i)\}, & if \;\; i \in \underline{w}_v \\ \{s_{lr}(i)), s_{ud}(i), p_1(i)\}, & if \;\; i \in \underline{w}_d \end{cases}$$

$$\mathcal{N}^{asym2}(i) = \begin{cases} \{s_{ud}(i), s_{lr}(i), s_{lr}(s_{ud}(v(i))), s_{ud}(d(i)), p_1(i)\}, & if \;\; i \in \underline{w}_h \\ \{s_{ud}(i), s_{lr}(i), s_{lr}(s_{ud}(h(i))), s_{lr}(d(i)), p_1(i)\}, & if \;\; i \in \underline{w}_v \\ \{s_{ud}(i), s_{lr}(i), s_{lr}(v(i)), s_{ud}(h(i)), p_1(i)\}, & if \;\; i \in \underline{w}_d \end{cases} \tag{6.6}$$

where operators d, v, and h return diagonal, vertical, and horizontal subband counterparts.

With these local structures defined, this section proposes correlated wavelet shrinkage:

1. Neighborhood Selection: The given random field $\underline{x}$ is projected into the wavelet domain with the resulting coefficient vector $\underline{w}$ and the noisy observation:

$$y_i = w_i + \nu_i$$

A neighborhood system $\mathcal{N}$ is chosen and two neighborhood vectors are formed:

$$\begin{aligned} \underline{y}_i &= [y_i, \{y_j; j \in \mathcal{N}(i)\}]^T \\ \underline{w}_i &= [w_i, \{w_j; j \in \mathcal{N}(i)\}]^T \end{aligned}$$

2. Linear Estimate: If $\underline{w}_i$ is assumed correlated (albeit a highly non-Gaussian joint correlation) with the calculated local covariances, then the best linear relaxing operation on the noisy coefficients is

$$\underline{z}_i = P_{\underline{w}_i} \cdot P_{\underline{y}_i}^{-1} \cdot \underline{y}_i \tag{6.7}$$

where we are only interested in

$$z_i = \underline{z}_i(1) = E[w_i|\underline{y}_i]$$

For every individual wavelet coefficient w_i the quantities $P_{\underline{w}_i}$ and $P_{\underline{y}_i}$ are obtained numerically (by sampling).

3. Non-linear Estimate: The estimate $\hat{w}_i$ is found via the non-linear empirical Bayes (6.2)

$$\hat{w}_i = E[w_i|z_i] \simeq average\{w_j|z_j \simeq z_i\}$$

4. The computation of $average\{\cdot\}$ can be done in different ways, such as uniform or triangular windowing.

This is only a starting and simple approach to a correlated shrinkage, while a complete schema of the Correlated empirical Bayesian Shrinkage (CBS) algorithm is

$$\underline{x} \xrightarrow{\mathcal{WT}} \underline{w} \xrightarrow{corrupted} \underline{y} \xrightarrow{local\ map} \underline{z} \xrightarrow{empirical\ Bayes} \underline{\hat{w}} \xrightarrow{\mathcal{WT}^{-1}} \underline{\hat{x}} \tag{6.8}$$

whose effectiveness in estimating from a noisy measurement is discussed in the following section.

6.4 Experimental Results

To test the performance of the proposed CBS algorithm, it was applied to a class of Gauss Markov random fields, as well as a collection of real images. The simulation results were compared with BayesShrink and independent empirical Bayesian estimate (6.2). The objective is to compare this shrinkage whit the state of the art shrinkage algorithms such as GSM and HMMs ones.

6.4.1 CBS and Gauss Markov Random Fields

Sample statistics were found over a class of GMRF, including five textures (grass, pigskin, tree-bark, calf leather, and thin-plate) shown in Figure 4.10. The averaged sample covariance over all five fields with association of the asymmetric neighborhood structures defined by (6.6) was used in (6.8) to linearly calculate $\underline{z}$. Then, non-linear uniform averaging was adopted in (6.8) to estimate $\underline{\hat{w}}$. The simulation results of triangular averaging technique are similar to those of the uniform widowing method.

Figure 6.2 plots RMSE for BayesShrink as well as independent and correlated empirical Bayesian shrinkage with various averaging window sizes, applied on a GMRF corrupted with a large range of noise variances. The results were obtained by using "db2" mother wavelet and the neighborhood system $\mathcal{N}^2$. All panels show that shrinkage with the assumption of correlated coefficients always outperforms the shrinkage with independent assumption. The correlated shrinkage is mostly better than the BayesShrink and as the window size gets bigger, this superiority becomes apparent.

The sensitivity of the CBS algorithm with the non-linear part of the averaging window size was also studied. The plots in Figure 6.3 show RMSE of the above methods measured as a function of window size for four different noise levels. BayeShrink is windowing size independent; it is displayed for the purpose of comparison only. The results are consistent across all four plots indicating that regardless of the noise level, the window size is an almost fixed number (about 15 for this particular simulation).

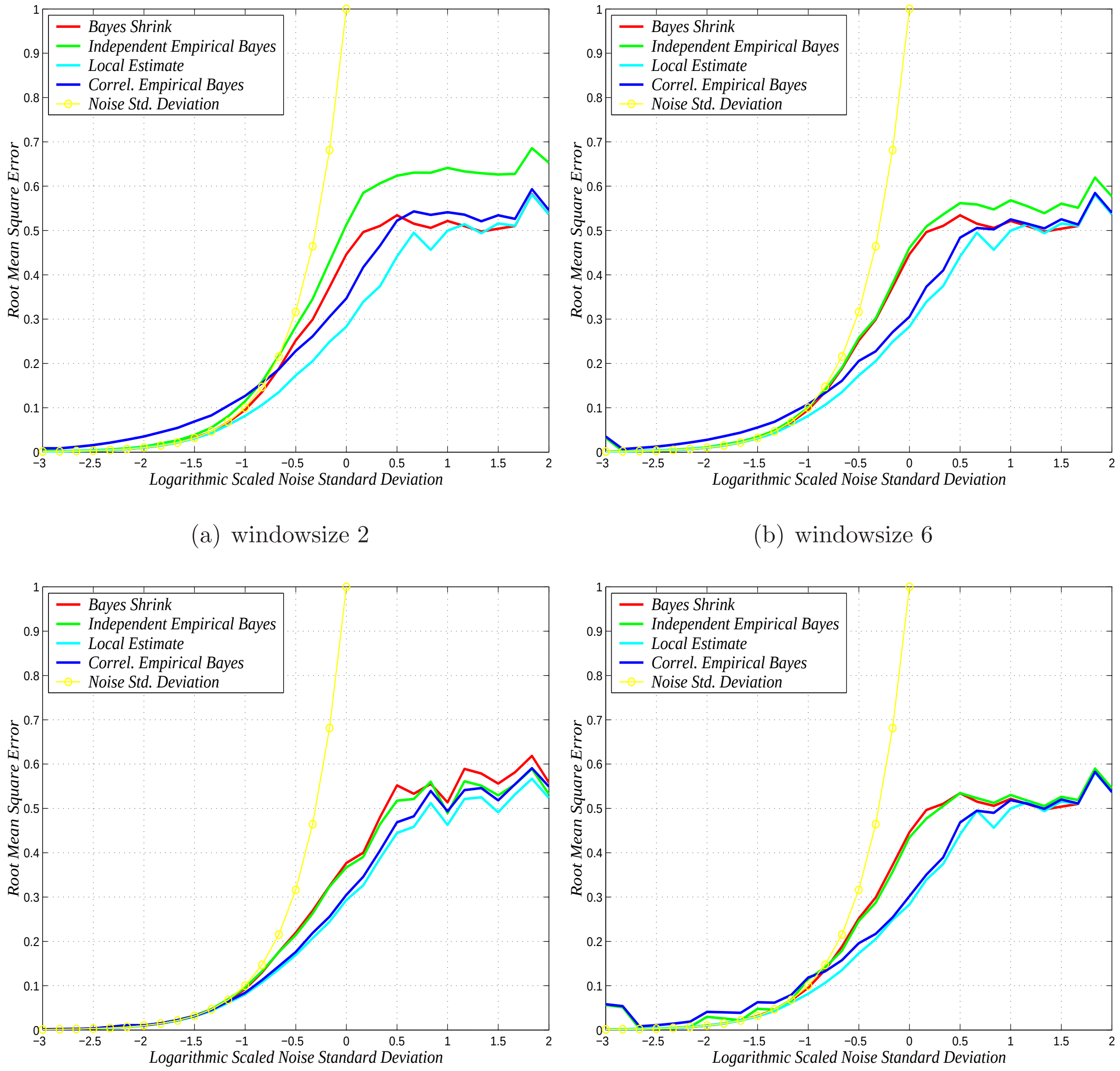
The above experiment with GMRFs was repeated with "db1" and "db4", leading to consistent observations and similar conclusions. The GMRFs played just as a necessary medium to test accuracy of our proposed model. The complementary step is to examine this framework with real world images, for which stationary is not assumed.

6.4.2 CBS and Real Images

The above framework was also applied on several standard real images. Results for the two images shown in Figure 6.4 are discussed here. The bottleneck in the implementation of the CBS method includes computing the local covariances and finding the optimum averaging window size:

- Wavelet Local Covariances:

  For each test image, the local covariances, *i.e.*, $P_{\underline{w}_i}$ in (6.7), were calculated. The structures of these local statistics depend on the regularity of the chosen mother wavelet and connectivity of the image pixels. Figure 6.5 through 6.7 display the



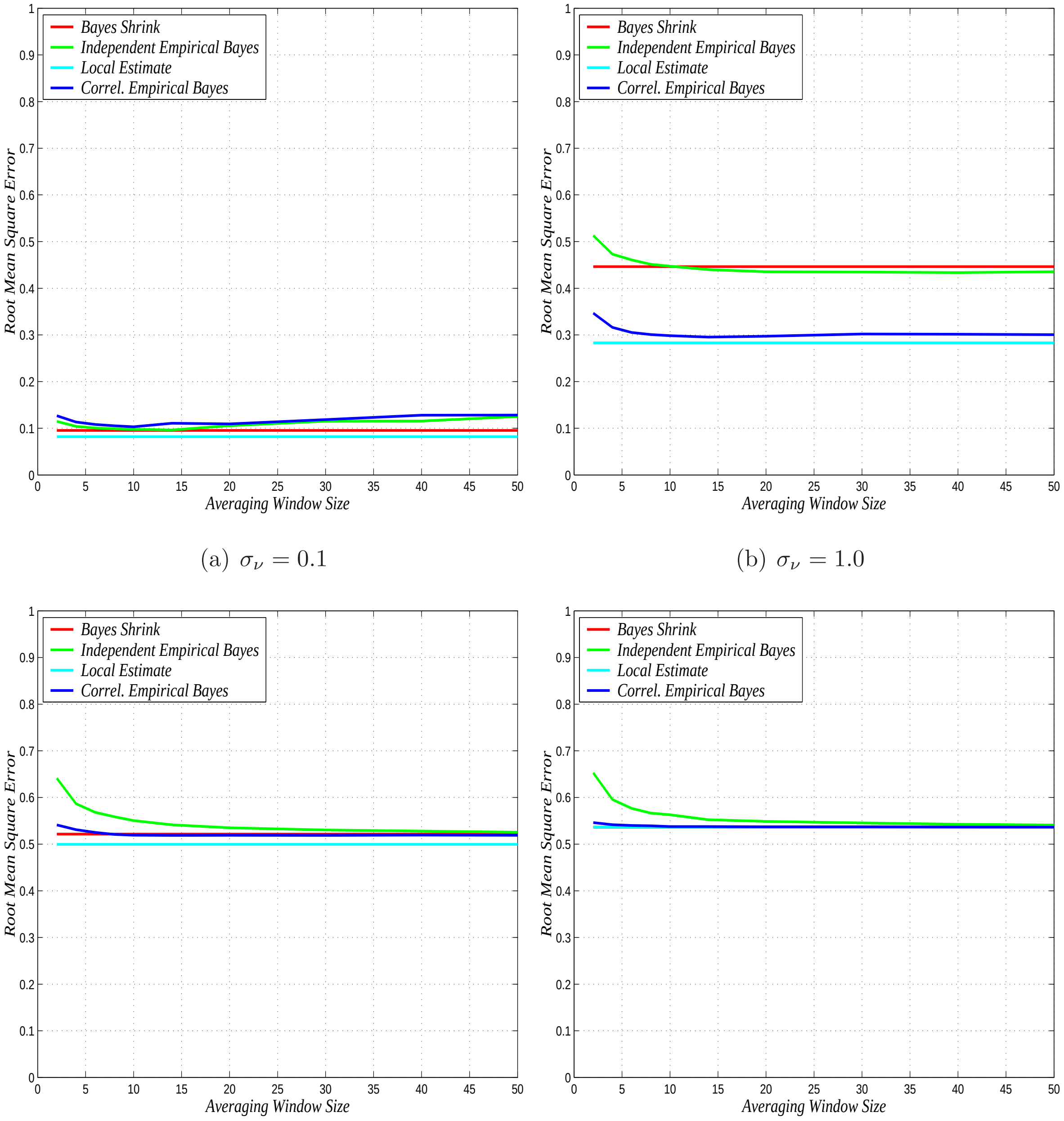
(a) windowsize 2

(b) windowsize 6

(c) windowsize 10

(d) windowsize 20

Figure 6.2: Plots of RMSE measurement for db2 BayesShrink as well as independent and correlated empirical Bayesian shrinkage, with thin-plate as the prior and uniform averaging window sizes of 2, 6, 10 and 20.



(a) $\sigma_\nu = 0.1$

(b) $\sigma_\nu = 1.0$

(c) $\sigma_\nu = 10.0$

(d) $\sigma_\nu = 100.0$

Figure 6.3: Plots of RMSE measurement for db2 BayesShrink as well as independent and wavelet local correlation based empirical Bayesian shrinkage, with Thin-plate as the prior and given noise level.

(a) Gold Hill (b) Lena

Figure 6.4: Two real images were tested with the CBS algorithm.

orientation-dependent sample covariances obtained for four-level "db1" to "db4" wavelet transform of the "Goldhill" image. In this experiment $\mathcal{N}^{asym2}$ was used, *i.e.*, the order of the entities in each sample covariance is associated with:

$$\underline{w}_i = \left\{ w_i, w_j | j \in \mathcal{N}^{asym2} \right\}$$

where for local structure $\mathcal{N}^{asym2}$ and for scale $j = 1, \cdots, J-1$

$$\begin{array}{r|l} \left|\underline{w}_i^h\right| = 12 & P_i^{j,h} : 12 \times 12 \\ \hline \left|\underline{w}_i^v\right| = 12 & P_i^{j,v} : 12 \times 12 \\ \hline \left|\underline{w}_i^d\right| = 10 & P_i^{j,d} : 10 \times 10 \end{array} \tag{6.9}$$

where $|\underline{w}|$ denotes size of the vector $\underline{w}$. In these plots, dark red squares show maximum correlation and the dark blue ones maximum anti-correlation. Clearly, the horizontal-band coefficients are vertically correlated and vertical-band coefficients are horizontally related. Both the horizontal- and vertical-band coefficients are strongly correlated with their parents (a phenomenon which is less significant for the diagonal-band coefficients). Similar simulations were performed for the "Lena" image with the associated local covariances displayed by Figure 6.8 through Figure 6.10.

The striking consistency between these figures and the plots in Figure 5.15 is very interesting. Apparently the wavelet local maps are more orientation dependent than scale dependent, *e.g.*, the color distributions in $P_i^{j,h}$ are almost identical across scales. These local maps were substituted in (6.7), completing the local estimation part.

- Averaging Window Size:

  The next subtlety was the notion of local averaging size, which was studied in this simulation. As is illustrated by Figure 6.11 and Figure 6.12, the averaging window size depends on the resolution as well as the additive noise standard deviation. The coarser the resolution (*i.e.*, the less information), the smaller is the averaging window size. It is observed that window sizes at adjacent scales are one to four proportional. To some points the optimum window size also depends on the original image attributes.

Having the local covariances and optimum averaging window sizes in hand, now the CBS algorithm is ready to estimate the above real images from their noisy observations. Each panel in Figure 6.13 and Figure 6.14 compares BayesShrink and our CBS algorithm applied on a real image in RMSE sense with different wavelet decomposition level J. It is evident that regardless of decomposition level, the CBS works better.

The final demonstration, but the most important one, is the qualitative comparison. The proposed CBS algorithm was tested on above two images for a variety of wavelet bases consistent results visualized by Figure 6.15 through Figure 6.18. The improvement

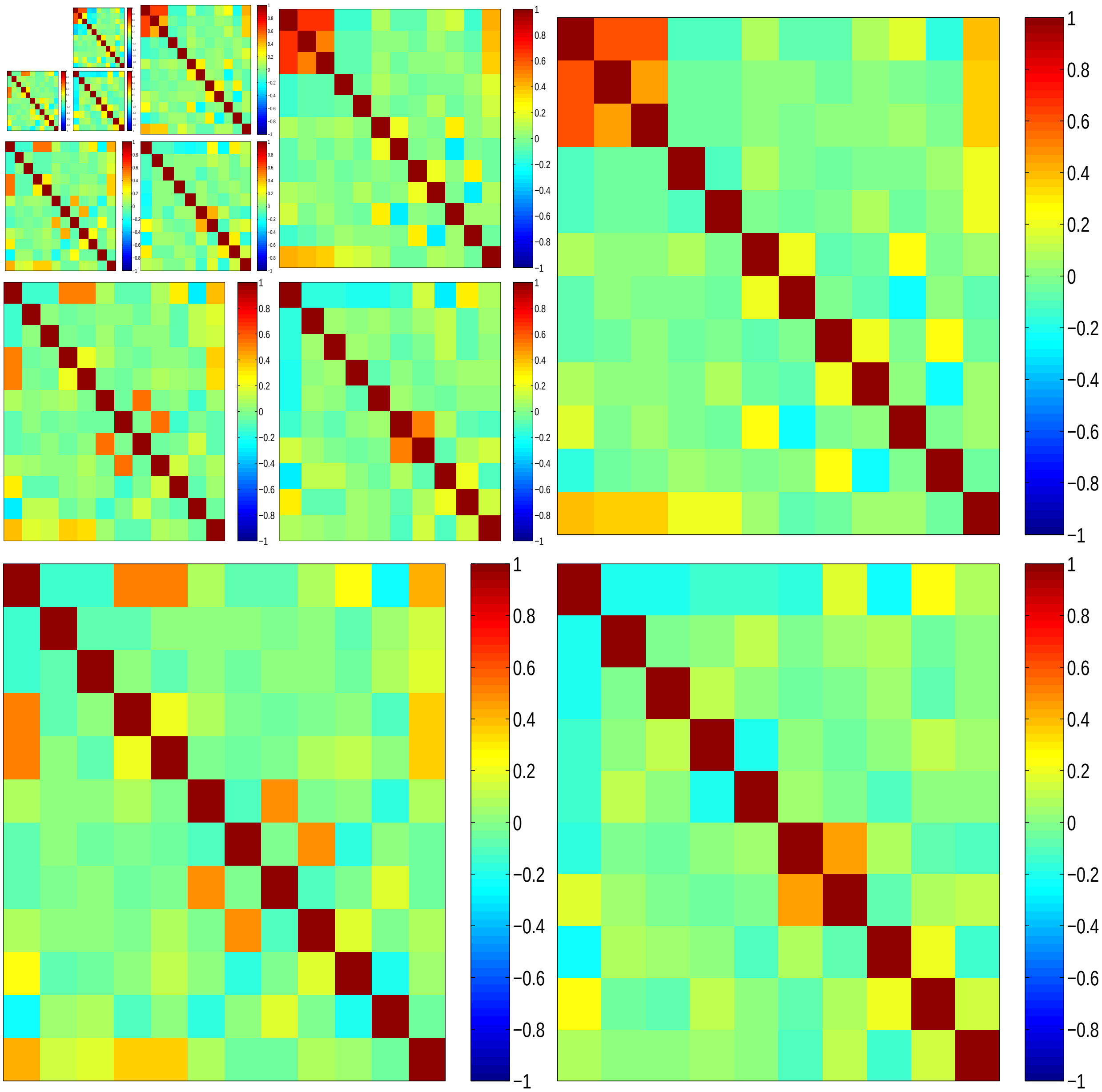


Figure 6.5: The scale-dependent sample covariances we obtained for four-level db1 wavelet transform of the “Goldhill” image. The order of the entities in each sample covariance is associated with that of the elements in $\mathcal{N}^{asym2}$ given by (6.6).

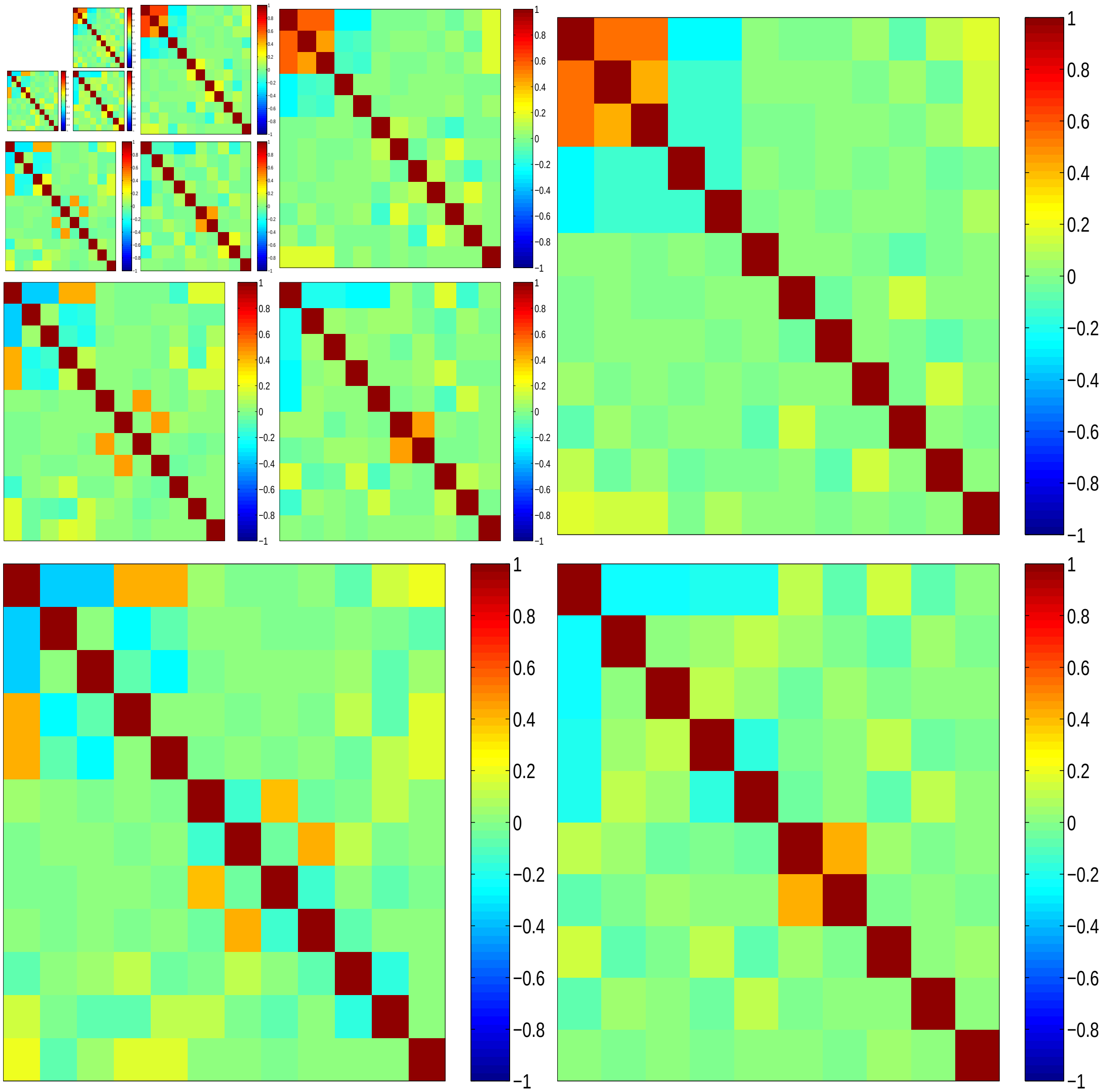


Figure 6.6: As in Figure 6.5, for Goldhill, using a db2 wavelet transform.

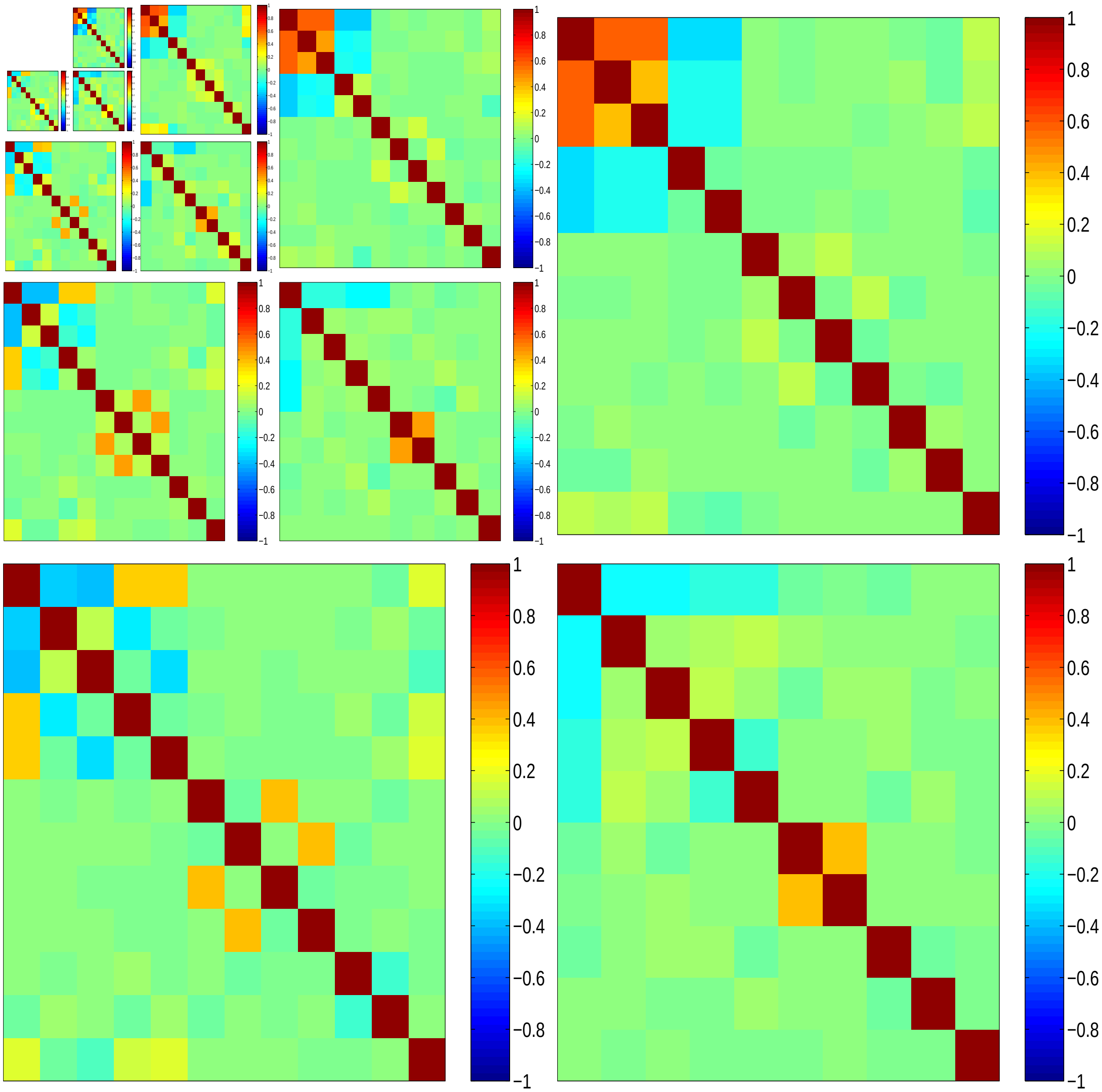


Figure 6.7: As in Figure 6.5, for Goldhill, using a db4 wavelet transform.

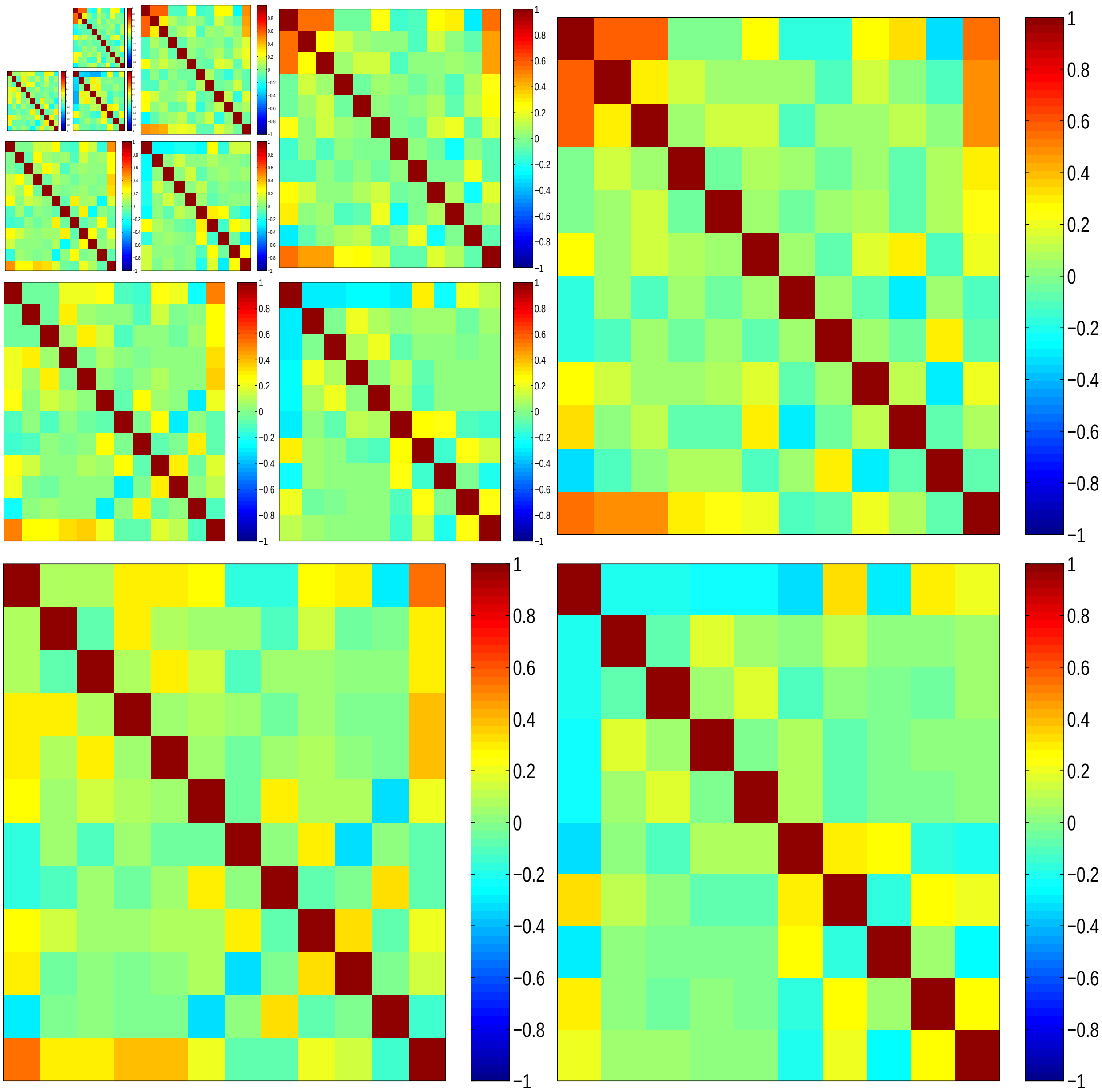


Figure 6.8: Parallel to Figure 6.5, The scale-dependent sample covariances obtained for db1 wavelet transform of the "Lena" image. The order of the entities in each sample covariance is associated with that of the elements in $\mathcal{N}^{asym2}$ given by (6.6).

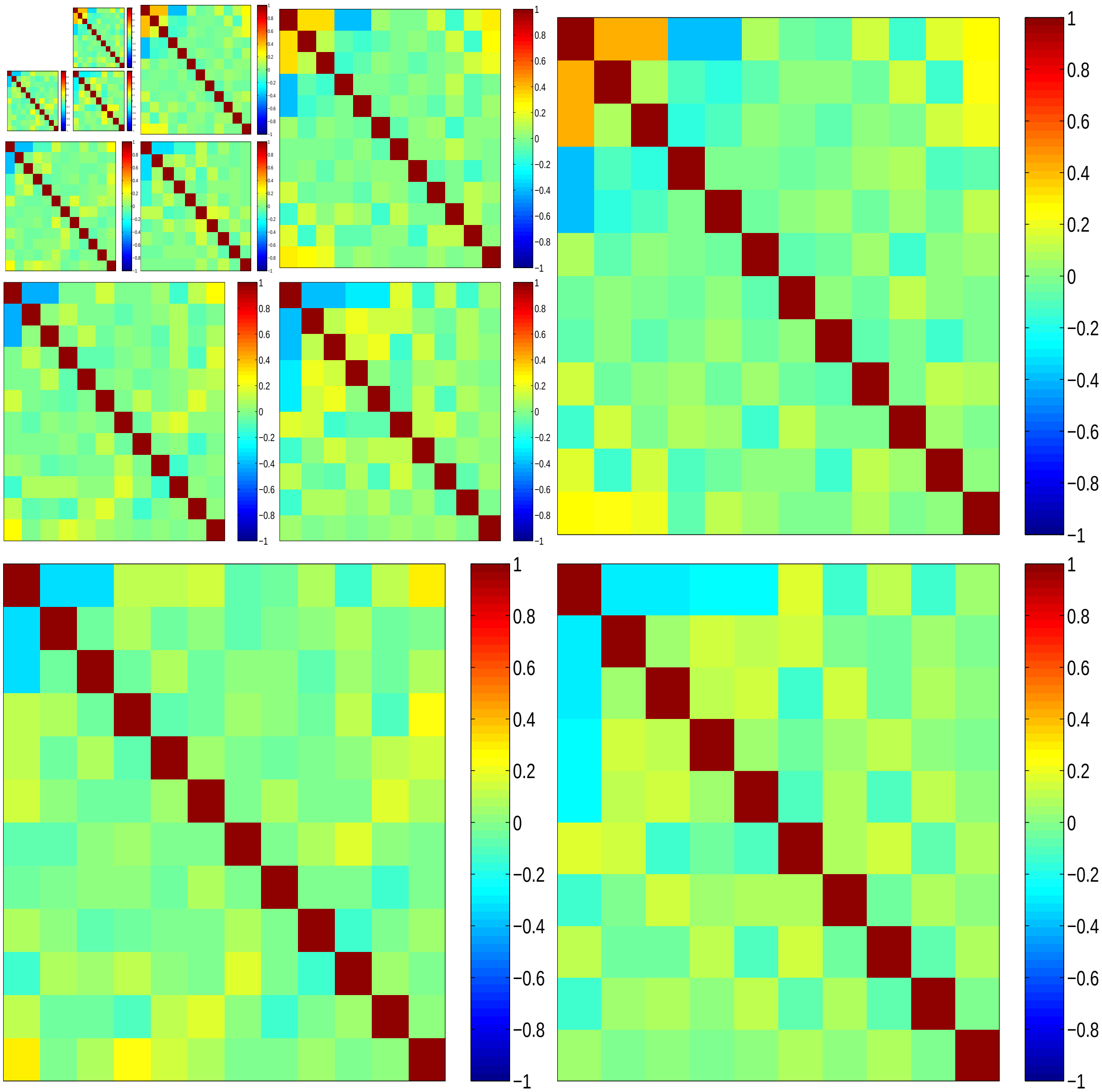


Figure 6.9: As in Figure 6.8, for Lena, using a db2 wavelet transform.

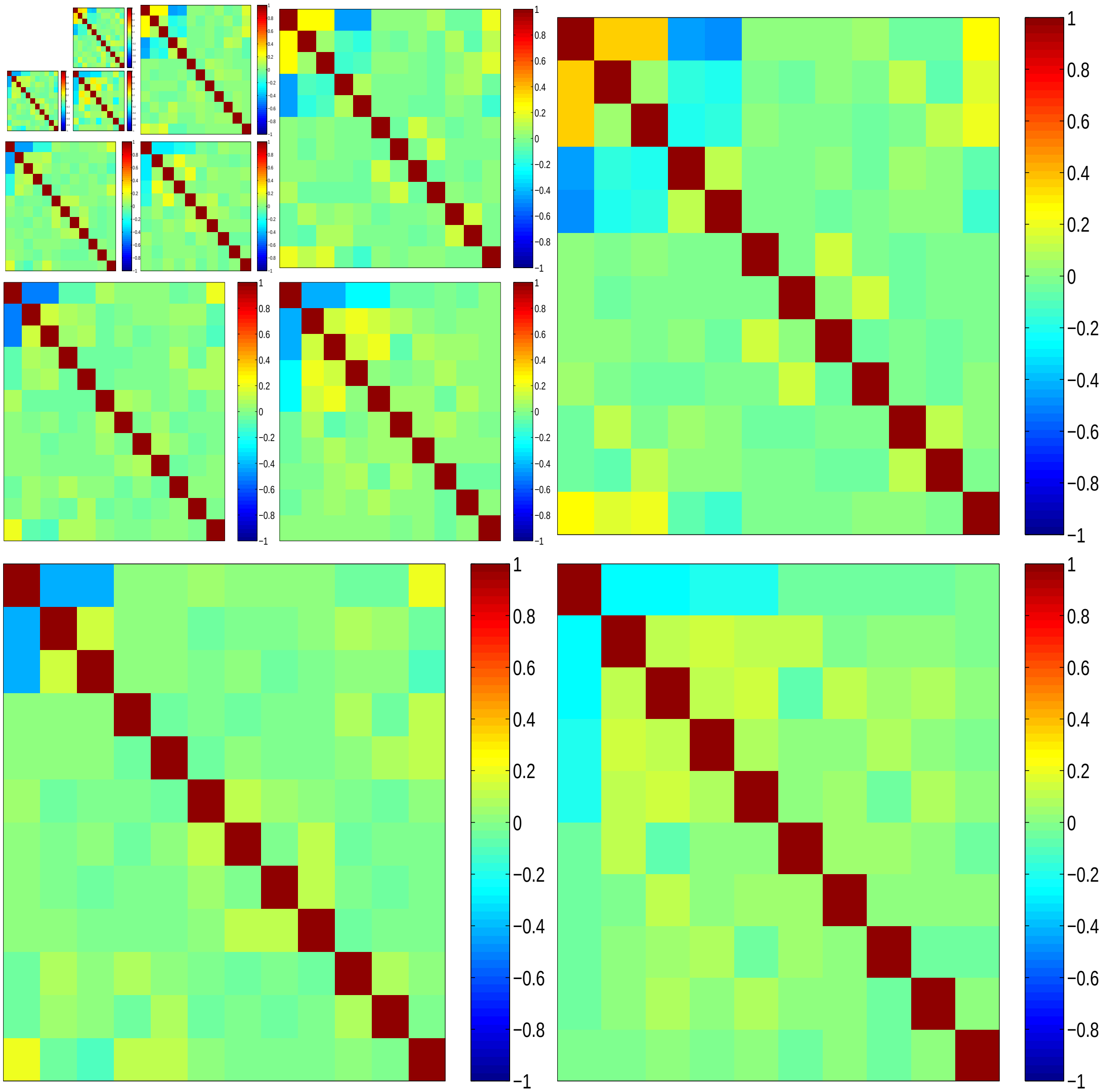
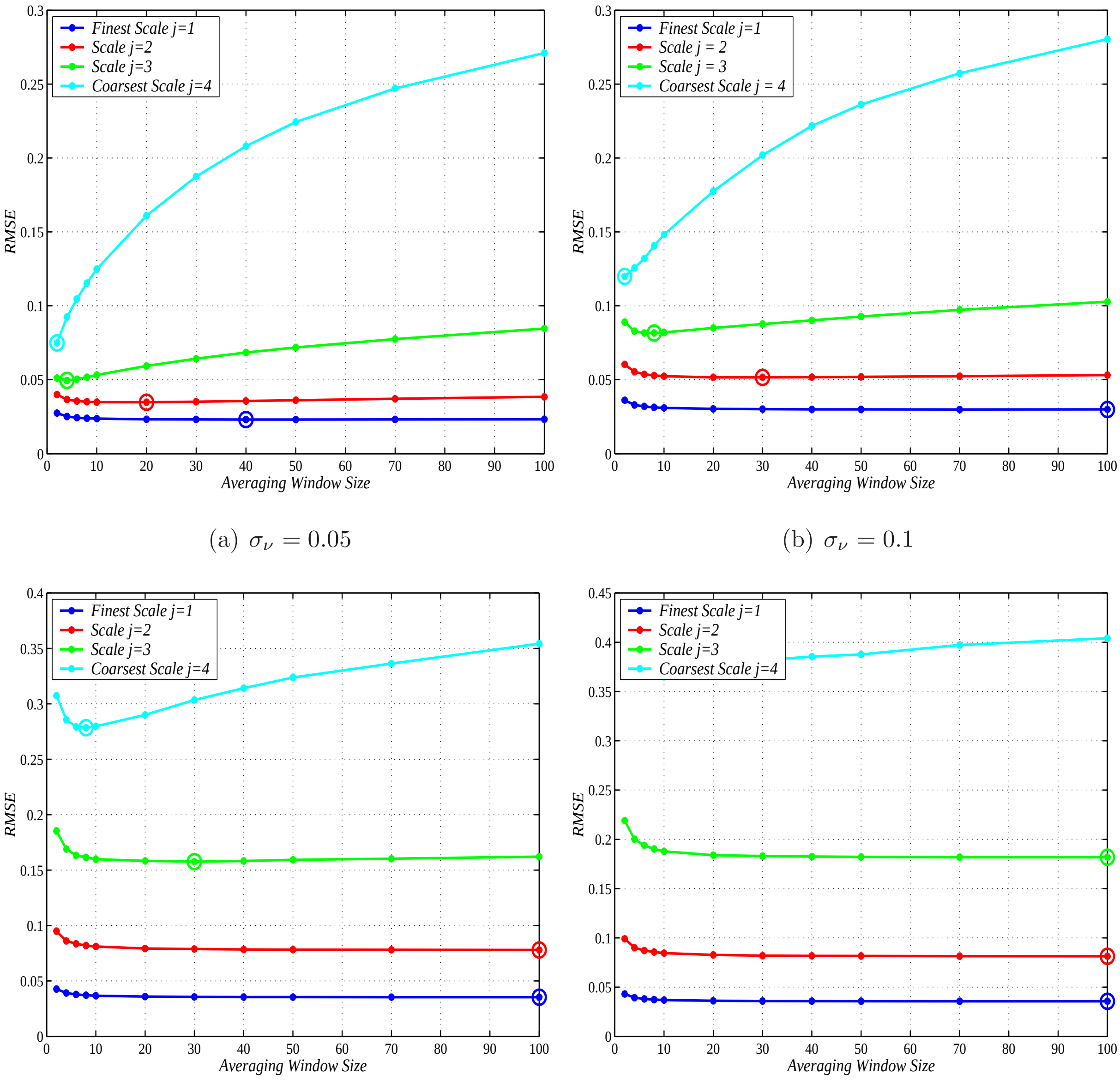


Figure 6.10: As in Figure 6.8, for Lena, using a db4 wavelet transform.



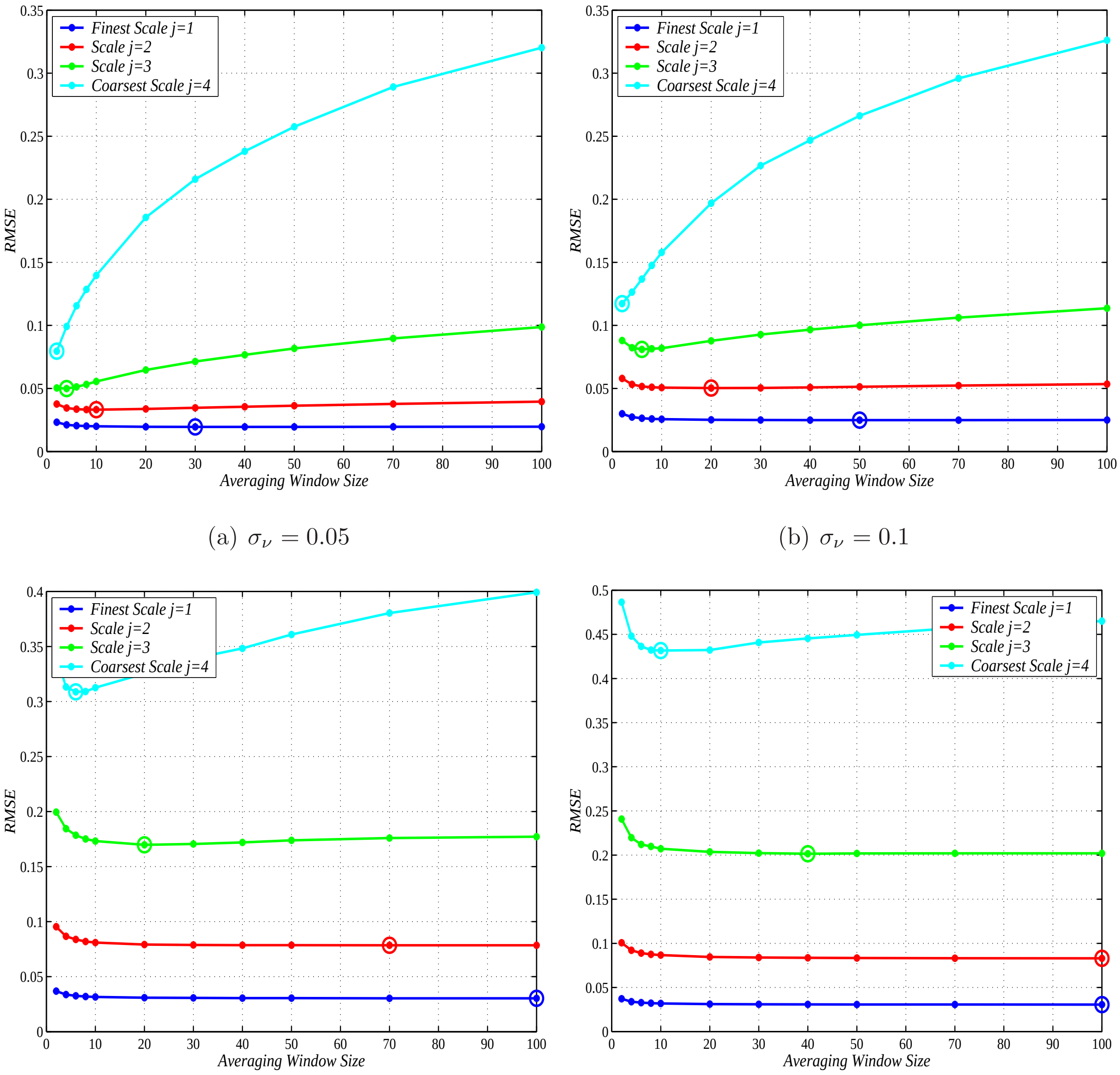
(a) $\sigma_\nu = 0.05$

(b) $\sigma_\nu = 0.1$

(c) $\sigma_\nu = 0.5$

(d) $\sigma_\nu = 1.0$

Figure 6.11: RMSE of the CBS method calculated at several scales as a function of the averaging window sizes. The optimum window size at each scale ($\circ$) depends on the resolution as well as the additive noise level σ_ν. Test image: "Goldhill"



(a) $\sigma_\nu = 0.05$

(b) $\sigma_\nu = 0.1$

(c) $\sigma_\nu = 0.5$

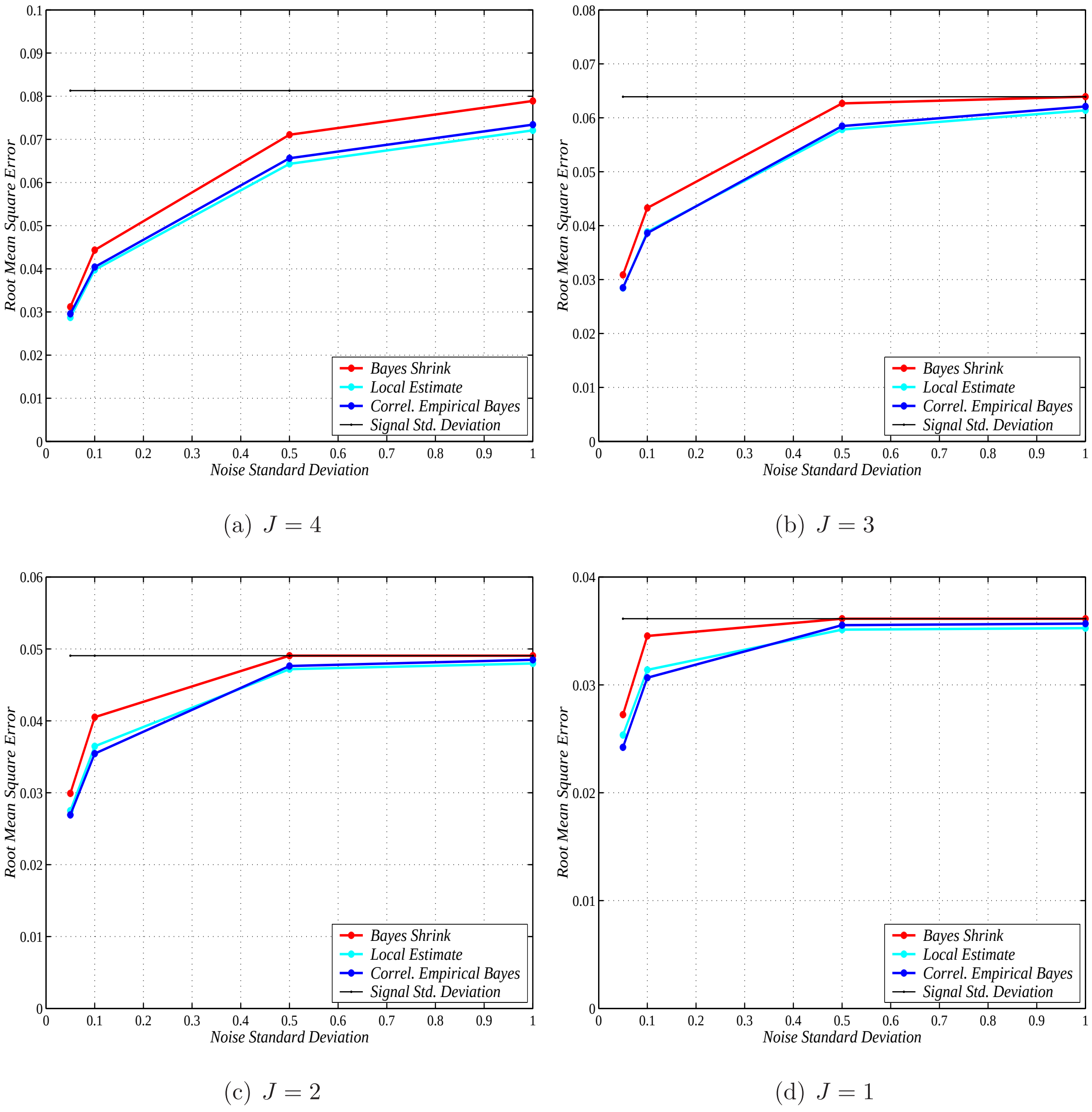
(d) $\sigma_\nu = 1.0$

Figure 6.12: As in Figure 6.11, RMSE of the CBS method calculated at several scales as a function of the averaging window sizes. Test image: “Lena”

of the CBS algorithm in depicting less blocky edges and in removing the artifacts appear in BayesShrink results is quite clear in these figures.

6.5 Chapter Summary

This chapter proposed a new correlation-based shrinkage scheme with considerable improvement over the performance of the well-known shrinkage methods with their assumption of generalized Gaussian or generalized Laplacian as the coefficients prior. The proposed CBS algorithm adopts joint statistics of the underlying image and results in a smaller estimation error and better visualization.



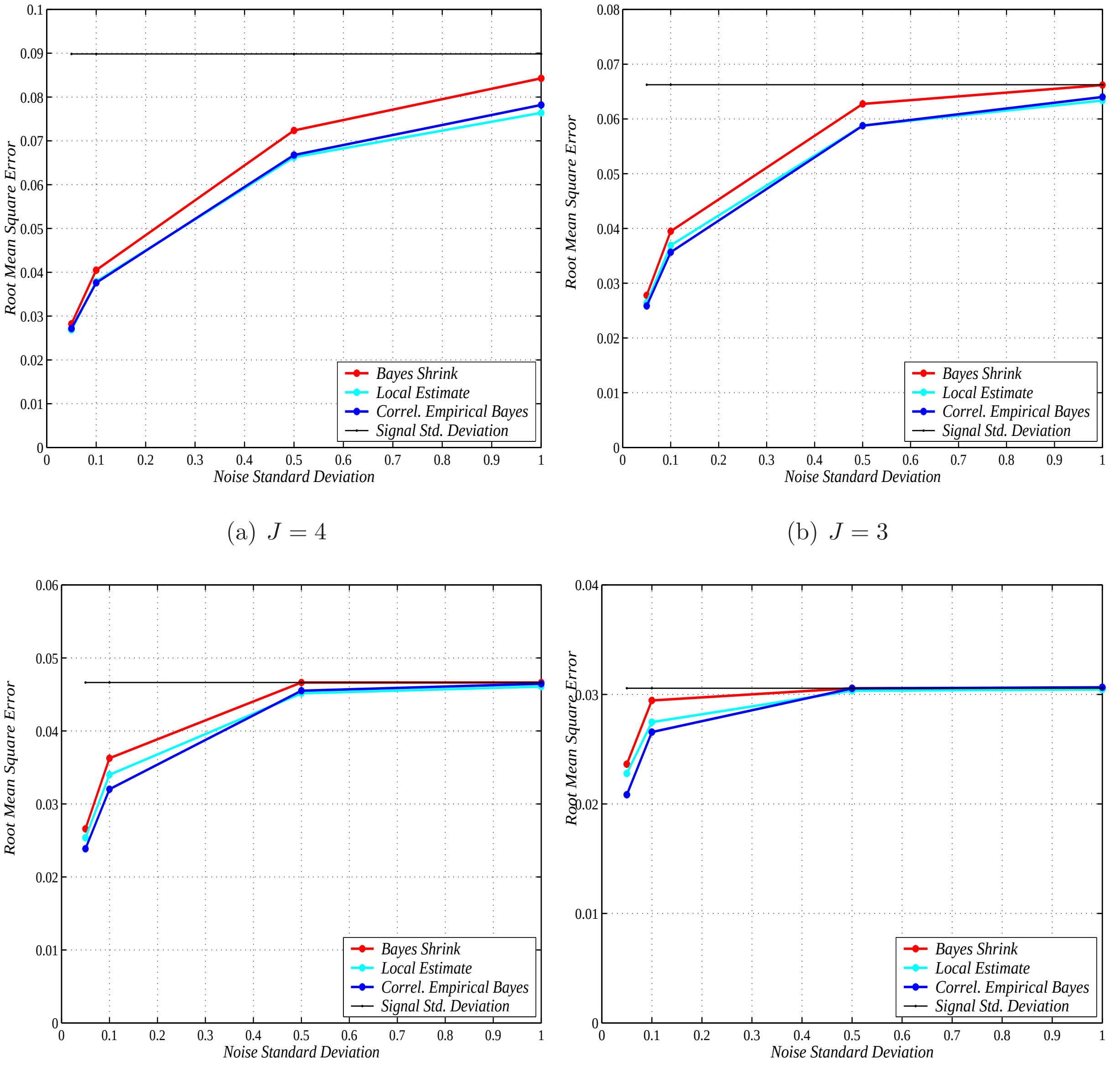
(a) $J = 4$

(b) $J = 3$

(c) $J = 2$

(d) $J = 1$

Figure 6.13: RMSE comparison of BayesShrink and CBS algorithm applied on Goldhill image as a function of noise level and wavelet decomposition level J. The proposed CBS always results in lower estimation error.



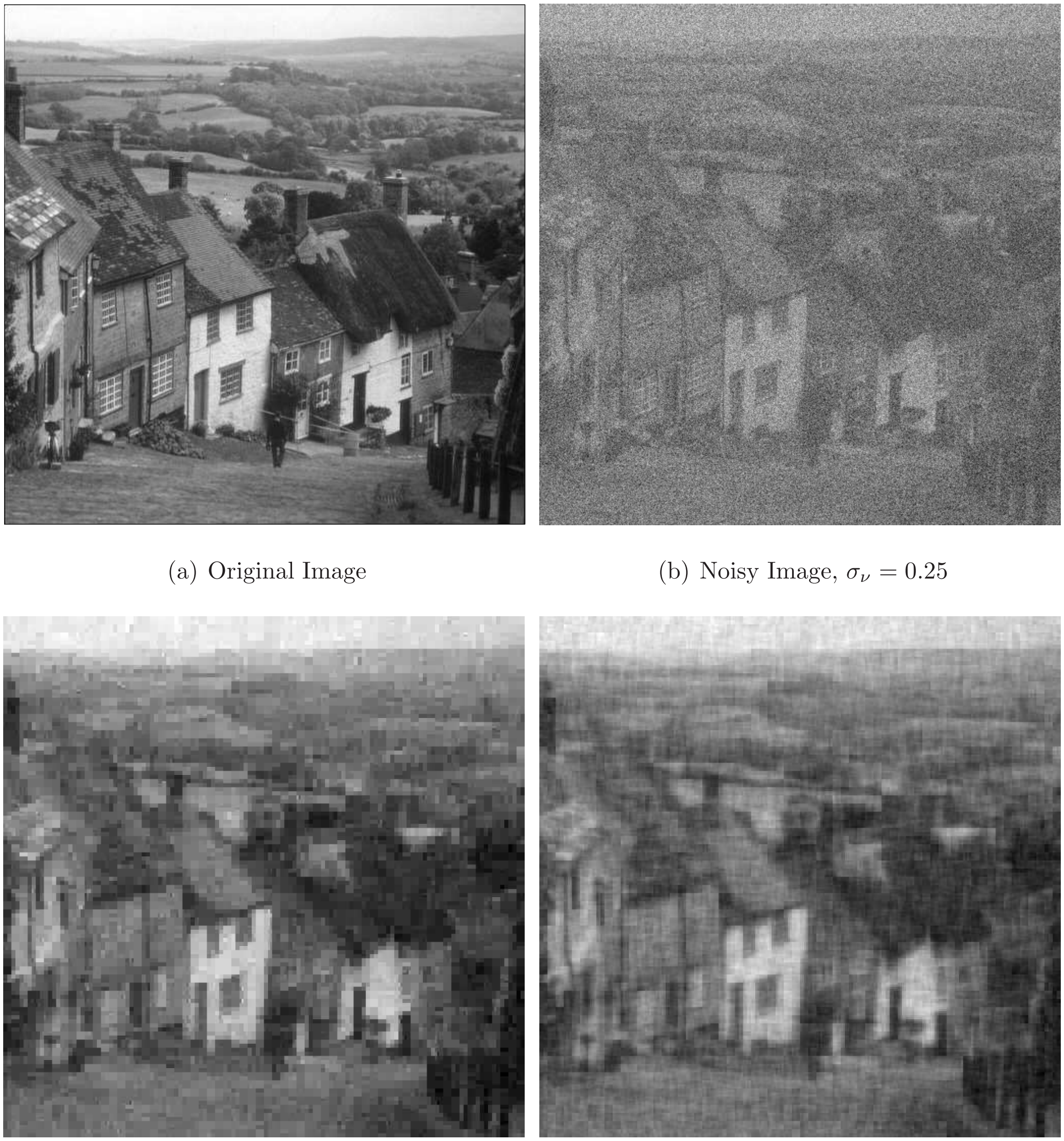
(a) $J = 4$

(b) $J = 3$

(c) $J = 2$

(d) $J = 1$

Figure 6.14: Similar to Figure 6.13, for Lena image.



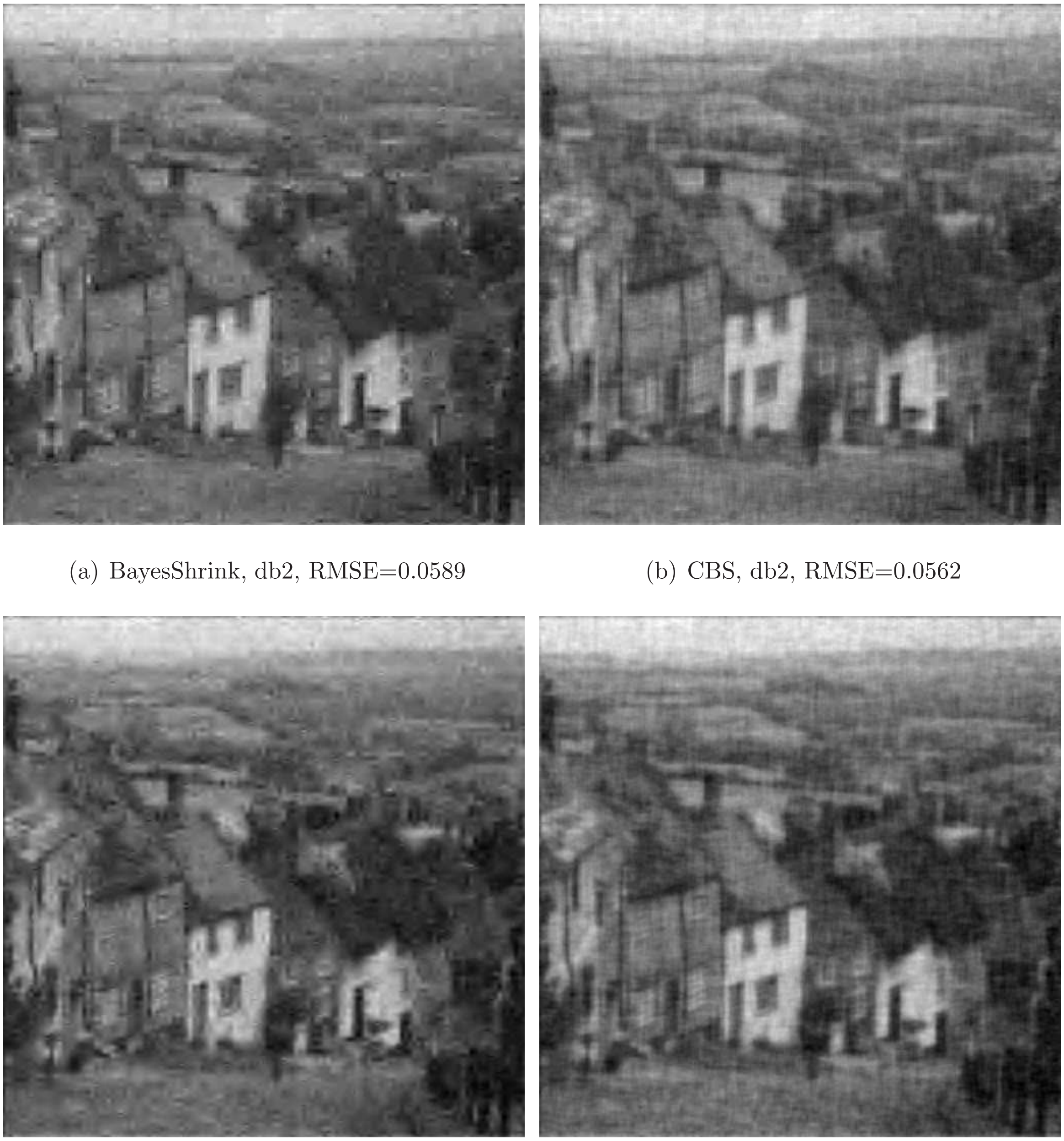
(a) Original Image

(b) Noisy Image, $\sigma_\nu = 0.25$

(c) BayesShrink, db1, RMSE=0.0603

(d) CBS, db1, RMSE=0.0558

Figure 6.15: The proposed CBS algorithm was applied on the Goldhill image with db1 basis function. It was successful to reduce the block artifacts appear in BayesShrink results and to depict more clear edges.



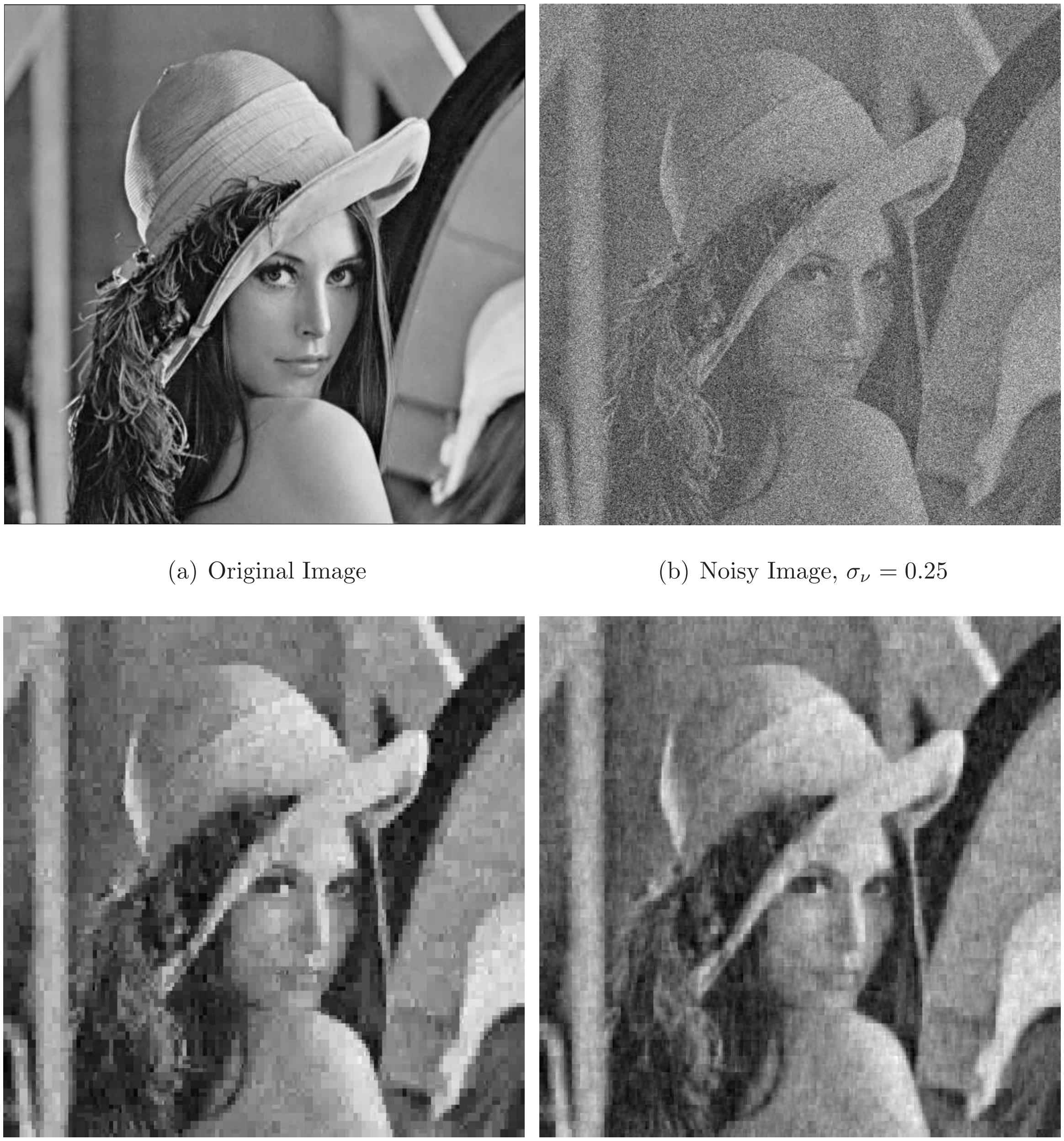
(a) BayesShrink, db2, RMSE=0.0589

(b) CBS, db2, RMSE=0.0562

(c) BayesShrink, db4, RMSE=0.0568

(d) CBS, db4, RMSE=0.0540

Figure 6.16: The proposed CBS algorithm was applied on the Goldhill image with db2 and db4 basis functions. The CBS results in reduced block artifacts and clear edges.



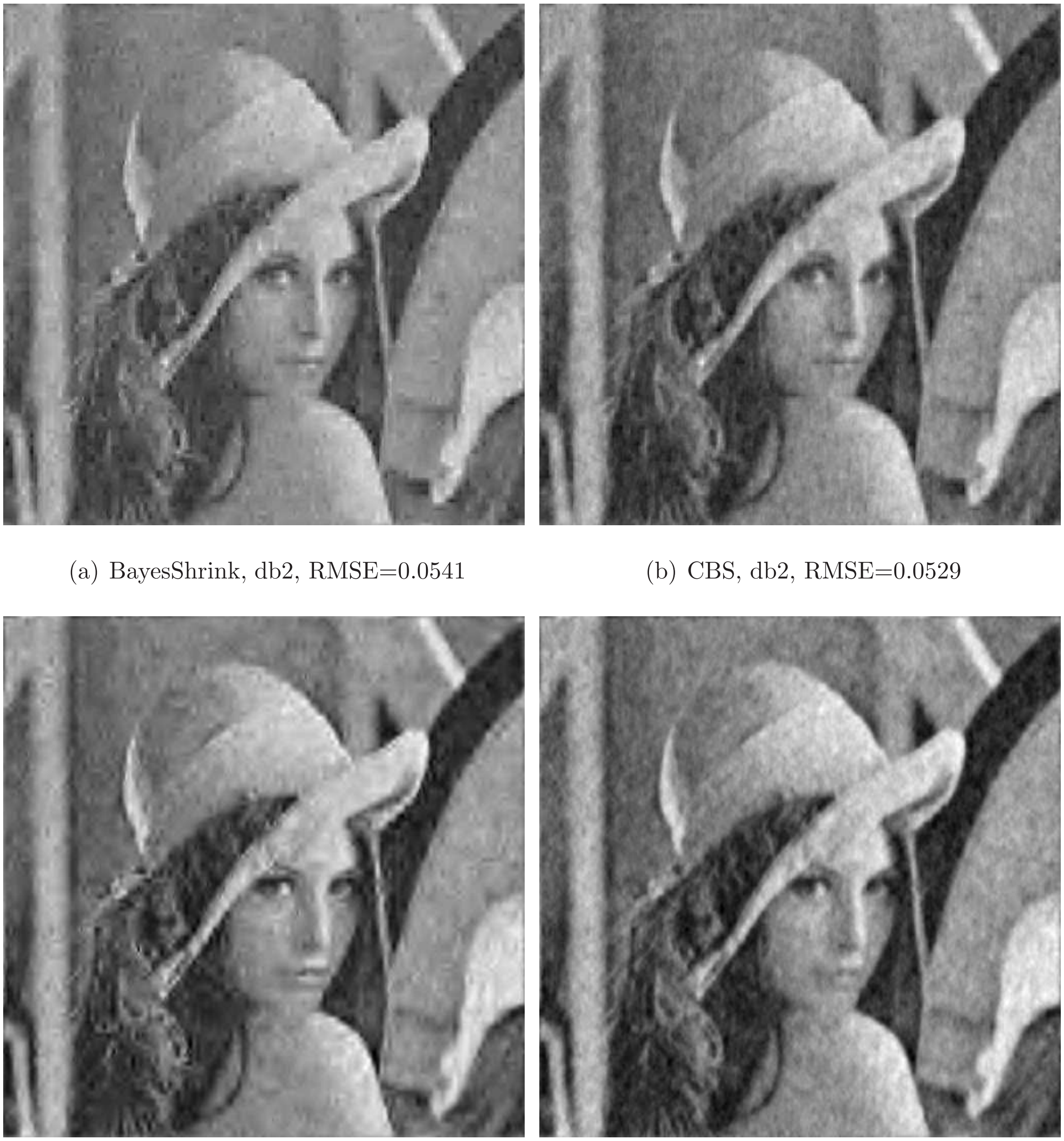
(a) Original Image

(b) Noisy Image, $\sigma_\nu = 0.25$

(c) BayesShrink, db1, RMSE=0.0583

(d) CBS, db1, RMSE=0.0554

Figure 6.17: The proposed CBS algorithm applied on Lena image with db1 basis function. It reduced the block artifacts appear in BayesShrink results and depicted more crisp edges.



(a) BayesShrink, db2, RMSE=0.0541

(b) CBS, db2, RMSE=0.0529

(c) BayesShrink, db4, RMSE=0.0515

(d) CBS, db4, RMSE=0.0504

Figure 6.18: The proposed CBS algorithm applied on Lena image with db2 and db4 basis functions. The CBS results in reduced block artifacts and clear edges.

Chapter 7

Conclusions and Future Perspectives

This thesis addressed the problem of developing probabilistic descriptions for image elements when they are projected into the orthogonal wavelet domain and shed further light on understanding wavelet joint statistics. This chapter draws some conclusions by pointing out the original contributions and explaining how the work in this thesis raises directions for future research.

7.1 Research Contributions

A probabilistic study of image models for wavelet domain statistics was the primary goal of this research. The goal has been met with the following contributions:

- **Structures of Wavelet Statistics:** A thorough study of empirical 2-D wavelet correlations was performed. The significant achievement at this stage was a novel invention of an approach to localize the coefficients joint structures. Representing the wavelet domain correlation maps with the 2-D WT diagram is a useful tool which not only shows the exact patterns of wavelet statistics, without any guess or approximation, but also displays the significance of those correlations. This was found to be the most crucial step towards defining the wavelet priors which govern those sparse statistics.

- **Models of Wavelet Statistics:** According to the achievements of statistical dependencies between the wavelet coefficients we proposed to model the wavelet coefficients not as independent, but as governed by a Markov random field. Since correlations are present both within and across scales, a random field model for the wavelet coefficients with itself needs to be hierarchical. The development of Markov random field methods on hierarchies has some past literature, but is still relatively new [43]. The contribution of this model is its capability in absorbing all direct correlations among the wavelet coefficients without any further consideration such as hidden states.

  A technique which approximates wavelet covariance based on an MRF neighborhood assumption was devised. The model parameters indicate a non-symmetric structure in the wavelet covariance which is an important consideration in the future model-based wavelet algorithms.

- **Applicability of the Wavelet Joint Models**: The effectiveness of the proposed model was examined by applying it to the wavelet shrinkage problem. By coupling the wavelet coefficients, the shrinkage problem is complicated considerably, in that the processing of the wavelet coefficients now depends on all others, in precisely the same way that inverting a banded matrix is much harder than a diagonal one. The very first step was to account for those correlations within a linear framework (least square estimate) which still outperforms the state-of-the-art BayesShrink algorithm.

7.2 Future Research Directions

The work presented in this thesis is only one single step in the broad evolution of the model-based wavelet estimation. This research provided answers to some open problems and perhaps raises new questions. At this point I outline ideas for future research related to this work:

- **Non-linear shrinkage**: The initial direction involves a further improvement to the correlated Bayesian shrinkage (CBS) algorithm [10]. Significant challenges still remain to be addressed. There are several directions and challenges associated with this kind of undertaking, with the first being the notion of non-Gaussianity for real images. The proposed CBS will be studied for a large class of real images with highly non-Gaussian joint statistics, where the joint linear estimator will not be an optimum method, thus the effect of the non-linear estimator will become significant.

- **Improvements to HMMs**: The wavelet HMMs assign a state to each coefficient and consider dependencies among those discrete states. Although, the conditional independence assumption connects all states (and consequently the coefficients) implicitly, the locality assumed for hidden states can be modified based on the proposed correlation maps. For instance, the conditional independence of a state associated with a horizontal-band coefficient can be considered when a state set including the parent state, states of the within-subband vertical siblings and across-subband cousins are given. This dependency map is anticipated to outperform the HMT-2 and HMT-3S methods where the state neighborhoods are assumed intuitively and based on heuristics.

- **Extensions to other transforms**: The limitation caused by shift-variance associated with wavelet transforms has led to a variety of alternatives such as the overcomplete steerable pyramids [87], and the complex WT [53]. The image singularity detection shortcoming of the wavelet transform has also caused a new generation of transforms including the wedgelet [31], ridgelet [12] and curvelet [89] transforms, which combine ideas of multiscale analysis and geometry. In these frameworks a large number of orientations results in many subbands, *i.e.*, there exists a significant degree of across-subband correlations. I believed that the statistics of these newly proposed transformations can be modeled by extensions of our modeling framework in conjunction with some explicit geometrical constraints.

- **Change of basis**: There exists past literature on the use of wavelets for the preconditioning of linear systems problems [102]. Such preconditioning is mathematically very similar to the wavelet change of basis in wavelet shrinkage, and may have insights to offer.

- **Applications**: Because of the great impact of the wavelet transform on the broad spectrum of information processing, the applicability of the proposed wavelet model is to be explored in many different disciplines, including biophysics, medicine, remote sensing, earth science. For example, in the area of medical imaging one can investigate possible improvements in model-based and adaptive image enhancement, motion tracking, and template matching by incorporating the wavelet joint models.

Bibliography

[1] F. Abramovich, T. Besbeas, and T. Sapatinas. Empirical Bayes approach to block wavelet function estimation. *Comput. Statist. Data Anal.*, 39:435–451, 2002.[1] [53]

[2] F. Abramovich, T. Sapatinas, and B. W. Silverman. Wavelet thresholding via a Bayesian approach. *J. R. Statis. Soc. B*, 60:725–749, 1998. [5, 41, 43, 48, 49, 50]

[3] D. Andrews and C. Mallows. Scale mixtures of normal distributions. *J. R. Statist. Soc*, 36:99, 1974. [53]

[4] Z. Azimifar, P. Fieguth, and E. Jernigan. Wavelet-domain joint statistics. *submitted to IEEE Trans. on Image Processing.* [149, 152]

[5] Z. Azimifar, P. Fieguth, and E. Jernigan. Wavelet shrinkage with correlated wavelet coefficients. *Proceedings of the 8th ICIP*, 2001. [70]

[6] Z. Azimifar, P. Fieguth, and E. Jernigan. Hierarchical multiscale modeling of wavelet-based correlations. *Proceedings of the 9th SSPR*, 2002. [50, 110]

[7] Z. Azimifar, P. Fieguth, and E. Jernigan. Towards random field modeling of wavelet statistics. *Proceedings of the 9th ICIP*, 2002. [50, 77, 106, 110]

[1] *The numbers at the end of each reference show the page number where that reference was cited.*

[8] Z. Azimifar, P. Fieguth, and E. Jernigan. Hierarchical Markov models for wavelet-domain statistics. *Proceedings of the 12th IEEE Statistical Signal Processing Workshop*, 2003. [50, 110, 152]

[9] Z. Azimifar, P. Fieguth, and E. Jernigan. Textures and wavelet-domain joint statistics. *Springer-Verlag: Lecture Notes in Computer Science: Image Analysis and Recognition*, 3212:331–339, 2004. [149]

[10] Z. Azimifar, P. Fieguth, and E. Jernigan. Correlated wavelet shrinkage: Models of local random fields across multiple resolutions. *Proceedings of the 12th ICIP*, 2005. [179]

[11] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *J. Royal Society, Series E*, 36:192–236, 1974. [17]

[12] E. Candes and D. Donoho. Ridglets: the key to higher-dimensional intermittency? *Phil. Trans. Roy. Soc. London*, 357:2495–2509, 1999. [37, 180]

[13] S. Chang, B. Yu, and M. Vetterli. Spatially adaptive wavelet thresholding with context modeling for image denoising. *IEEE Trans. on Image Processing*, 9:1522–1531, 2000. [43, 45, 46, 50, 53, 150]

[14] S. Chang, B. Yu, and M. Vetterli. Spatially adaptive wavelet thresholding with context modeling for image denoising. *IEEE Trans. on Image Processing*, 9:1522–1531, 2000. [46, 50, 52, 150, 151]

[15] H. Chipman, E. Kolaczyk, and R. McCulloch. Adaptive Bayesian wavelet shrinkage. *J. Amer. Statis. Assoc.*, pages 92–99, 1997. [5, 41, 48, 49, 50, 55, 56, 57]

[16] H. Chipman and L. Wolfson. Prior elicitation in the wavelet domain. *Wavelets and Statistics*, pages 83–94, 1999. [50]

[17] D. Cho and T. Bui. Multivariate statistical modeling for image denoising using wavelet transforms. *J. Signal Processing: Image Communication*, 20:78–89, 2005. [6]

[18] H. Choi and R. Baraniuk. Image segmentation using wavelet-domain hidden Markov models. *Proceedings of SPIE*, 3816, 1999. [55]

[19] H. Choi and R. Baraniuk. Multiscale image segmentation using wavelet-domain hidden Markov models. *IEEE Trans. on Image Processing*, 10:1309–1321, 2001. [50, 102, 112]

[20] K. Chou, A. Willsky, and A. Benveniste. Multiscale recursive estimation, data fusion, and regularization. *IEEE Trans. on Automatic Control*, 39:468–478, 1994. [21, 23]

[21] C. Christopoulos, A. Skodras, and T. Ebrahimi. The JPEG2000 still image coding system: An overview. *IEEE Trans. on Consumer Electronics*, 46:1103–1127, 2000. [3]

[22] I. Cohen, S. Raz, and D. Malah. Orthonormal shift-invariant wavelet packet decomposition and representation. *Signal Processing*, 57:251–270, 1997. [37]

[23] R. Coifman and D. Donoho. Translation-invariant de-noising. *Technical report, Stanford University*, 1997. [37]

[24] G. Cross and A. Jain. Markov random field texture models. *IEEE Trans. on PAMI*, pages 25–39, 1983. [136]

[25] M. Crouse and R. Baraniuk. Contexual hidden Markov models for wavelet-domain signal processing. *Proceedings of 31st Asilomar Conference in Signals, Systems and Computers*, 1997. [50, 52, 57, 59, 60]

[26] M. Crouse, R. Nowak, and R. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. on Signal Processing*, 46:886–902, 1998. [4, 39, 49, 50, 52, 54, 55, 56, 57, 58, 63, 150, 151]

[27] M. Daniel and A. Willsky. A multiresolution methodology for signal-level fusion and data assimilation with applications to remote sensing. *Proceedings of the IEEE*, pages 164–180, 1997. [23]

[28] I. Daubechies. *Ten Lectures on Wavelets.* PA: SIAM, Philadelphia, 1992. [3, 4, 25, 26, 28]

[29] I. Daubechies, S. Mallat, and A. Willsky. Special issue on wavelet transforms and multiresolution signal analysis. *IEEE Trans. on Information Theory*, 38:529–860, 1992. [4]

[30] D. Donoho. De-noising by soft thresholding. *IEEE Trans. on Information Theory*, 41(3):613–627, 1995. [43]

[31] D. Donoho. Wedgelets: Nearly minimax estimation of edges. *Ann. Statist.*, page 859897, 1999. [180]

[32] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994. [41, 43, 44]

[33] D. Donoho and I. Johnstone. adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Stat. Assoc.*, 90:1200–1224, 1995. [5, 41, 42, 44, 50, 51, 85]

[34] R. Dufour and E. Miller. Statistical signal restoration with $1/f$ wavelet domain prior models. *Signal Processing*, 78:209–307, 1998. [39]

[35] G. Fan and X. Xia. Image denoising using a local contextual hidden Markov model in the wavelet domain. *IEEE Signal Processing Letters*, 8:125–128, 2001. [50, 52, 57, 60]

[36] G. Fan and X. Xia. Improved hidden Markov models in the wavelet-domain. *IEEE Trans. on Signal Processing*, 49:115–120, 2001. [50, 57, 58]

[37] G. Fan and X. Xia. Wavelet-based texture analysis and synthesis using hidden Markov models. *IEEE Trans. on Circuits and Systems*, 50:106–120, 2003. [50, 52, 55, 57, 60, 63, 102]

[38] P. Fieguth. *Multidimensional Signal Modeling and Estimation.* Lecture Notes, University of Waterloo, 2004. [12, 13, 16, 64, 65, 73, 89, 150]

[39] P. Flandrin. Wavelet analysis and synthesis of fractional Brownian motion. *IEEE Trans. on Information Theory*, 38:910–916, 1992. [39]

[40] I. Fodor and C. Kamath. Denoising through wavelet shrinkage: An empirical study. *Journal of Electronic Imaging*, 12(1):151–160, 2003. [5, 41, 44]

[41] D. Geman. Random fields and inverse problems in imaging. *Lecture Notes in Mathematics, Springer-Verlag*, pages 113–193, 1990. [3, 14, 99]

[42] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans on PAMI*, 6:721–741, 1984. [14]

[43] S. Golden. *Identifying Multiscale Statistical Models Using The Wavelet Transform.* Ma. Sc. Thesis, Massachusetts Institute of Technology, 1991. [114, 178]

[44] C. Graffigne, F. Heitz, and P. Perez. Hierarchical Markov random field models applied to image analysis: a review. *SPIE*, 1995. [6]

[45] A. Haar. Zur theorie der orthogonalen funktionensysteme. *PhD Dissertation*, 1909. [26, 32]

[46] M. Hassner and J. Sklansky. The use of Markov random fields as models of texture. *Computer Graphics and Image Processing*, 12:357–370, 1980. [14, 136]

[47] J. Huang and D. Mumford. Statistics of natural images and models. *Proceedings of CVPR*, 1999. [45, 49, 50]

[48] B. Hubbard. *The world according to wavelets*. A K Petters, Ltd., 1996. [1]

[49] B. Jawerth and W. Sweldens. An overview of wavelet based multiresolution analysis. *SIAM Review*, 36:377–412, 1994. [28, 32, 33, 74]

[50] E. Jernigan. Digital image processing. *Lecture Notes, University of Waterloo*, 2003. [73]

[51] R. Kashyap, R. Chellappa, and A. Khotanzad. Texture classification using features derived from random field models. *Pattern Recognition Letters*, pages 43–50, 1982. [136, 137, 141]

[52] R. Kinderman and J. Snell. Markov random fields and their applications. *American Mathematical Society, Providence, Rhode Island*, pages 253–258, 1980. [14]

[53] N. Kingsbury. Image processing with complex wavelets. *Phil. Trans. Royal Society London A*, 357:2543–2560, 1999. [37, 180]

[54] E. Kretzmer. Statistics of television signals. *Bell Syst. Tech. J.*, 31:751–763, 1952. [2]

[55] H. Krim, W. Willinger, A. Juditski, and D. Tse. Special issue on multiscale statistical signal analysis and its applications. *IEEE Trans. on Information Theory*, 45:825–1062, 1999. [4, 50]

[56] S. Lakshmanan and H. Derin. Gaussian Markov random fields at multiple resolutions. *Chellappa and Jain editors: Markov Random Fields: Theory and Applications*, pages 131–157, 1993. [4, 6, 14, 136]

[57] J. Lee. Digital image enhancement and noise filtering by use of local statistics. *IEEE Trans. on PAMI*, PAMI-2:165–168, 1980. [53]

[58] S. Li. *Markov Random Field Modeling in Computer Vision.* Springer-Verlag, New York, 1995. [3, 13, 14, 16, 18, 20]

[59] J. Liu and P. Moulin. Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients. *IEEE Trans. on Image Processing*, 10:1647–1658, 2001. [50]

[60] M. Luettgen, W. Karl, A. Willsky, and R. Tenney. Multiscale representation of Markov random fields. *IEEE Trans. on Signal Processing*, 41(12):3377–3396, 1993. [6]

[61] D. Mackenzie. Wavelets seeing the forest and the trees. *National Academy of Sciences*, 2002. [34]

[62] M. Malfait and D. Roose. Wavelet-based image denoising using a Markov random field a priori model. *IEEE Trans. on Image Processing*, 6:549–565, 1997. [50, 52]

[63] S. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. on PAMI*, 11:674–693, 1989. [4, 26, 29, 30, 31, 32, 33]

[64] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, second edition, 1999. [3, 4, 28, 35, 36, 37]

[65] S. Mallat and W. Hwang. Singularity detection and processing with wavelets. *IEEE Trans. on Information Theory*, 38:617–643, 1992. [39]

[66] S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Trans on PAMI*, 14:710–732, 1992. [39]

[67] Y. Meyer. *Orthonormal wavelets*. Inverse Probl. Theoret. Imaging. Springer, Berlin, 1989. [26, 74]

[68] Y. Meyer. Wavelets: their past and their future. *Progress in wavelet analysis and applications*, pages 9–18, 1993. [4]

[69] M. Mihcak, I. Kozintsev, and K. Ramchandran. Low-complxity image denoising based on statistical modeling of wavelet coefficients. *IEEE Signal Processing Letters*, 6:300–303, 1999. [50, 52, 53, 60]

[70] P. Moulin and J. Liu. Analysis of multiresolution image denoising schemes using a generalized Gaussian and complexity priors. *IEEE Trans. on Information Theory*, 45:909–919, 1999. [45, 50, 51, 52]

[71] P. Muller and Eds B. Vidakovic. *Bayesian Inference in Wavelet-Based Models*. Springer, New York, 1999. [49]

[72] P. Muller and B. Vidakovic. An introduction to wavelets. *Bayesian inference in Wavelet based Model*, pages 1–18, 1999. [27, 28]

[73] R. Nowak. Multiscale hidden Markov models for Bayesian image analysis. *Technical Report, Michigan State University*, 1998. [50, 55, 57]

[74] R. Nowak. Multiscale hidden Markov models for Bayesian image analysis. *Wavelet Based Models, B. Vidakovic and P. Muller, Eds*, Springer-Verlag, 1999. [50, 52, 54, 55]

[75] M. Orchard and K. Ramchandran. An investigation of wavelet-based image coding using an entropy-constrained quantization framework. *Proceedings of Data Compression Conference (Snowbird, Amer. Utah)*, pages 341–350, 1994. [39]

[76] T. O'Rourke and R. Stevenson. Human visual system based wavelet decomposition for image compression. *J. Vis. Commun. Image Represent.*, 6:109–121, 1995. [3]

[77] A. Pizurica, W. Philips, I. Lemahieu, and M. Acheroy. A joint inter- and intrascale statistical model for Bayesian wavelet based image denoising. *IEEE Trans. on Image Processing*, 11:545–557, 2002. [49, 50, 52]

[78] J. Portilla and E. Simoncelli. Image denoising via adjustment of wavelet coefficient magnitude correlation. *Proceedings of the 7th ICIP*, 2000. [4, 50, 51]

[79] J. Portilla, V. Strela, M. J. Wainwright, and E.P. Simoncelli. Image denoising using Gaussian scale mixtures in the wavelet domain. *IEEE Trans. on Image Processing*, 12:1338–1351, 2003. [6, 50, 52, 53, 150]

[80] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of IEEE*, pages 257–286, 1989. [54]

[81] B. Ripley. *Statistical inference for spatial processes.* Cambridge University Press, 1988. [3, 14]

[82] J. Romberg, H. Choi, and R. Baraniuk. Bayesian tree-structured imaged modeling using wavelet-domain hidden Markov models. *IEEE Trans. on Image processing*, 10:1056–1068, 2001. [39, 50, 52, 54, 55, 58]

[83] A. Rosenfeld. *Image Modeling*. Academic Press, 1982. [11]

[84] Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, Second Edition, 2003. [144]

[85] J. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. on Signal Processing*, 41:3445–3462, 1993. [52]

[86] E. Simoncelli. Modeling the joint statistics of images in the wavelet domain. *Proceedings of the SPIE 44th Annual Meeting*, 1999. [49, 50, 52, 54, 136, 150, 151]

[87] E. Simoncelli, W. Freeman, E. Adelson, and D. Heeger. Shiftable multiscale transforms. *IEEE Trans. on Information Theory*, 38:587–607, 1992. [37, 180]

[88] A. Srivastava. Stochastic models for capturing image variability. *IEEE Signal Processing Magazine*, 19(5):63–76, 2002. [51]

[89] J. Starck, E. Candes, and D. Donoho. The curvelet transform for image denoising. *IEEE Trans. on Image Processing*, 11(6):670–684, 2002. [37, 180]

[90] G. Strang. Wavelets and dilation equations: a brief introduction. *SIAM Rev.*, 31(4):614–627, 1989. [3, 4]

[91] G. Strang. Wavelet transforms versus Fourier transforms. *Bull. Amer. Math. Soc. (N.S.)*, 28:288–305, 1993. [26, 27]

[92] V. Strela, J. Potrilla, and E. Simoncelli. Image denoising using a local Gaussian scale mixture model in the wavelet domain. *Proceedings of the SPIE*, 2000. [50, 54]

[93] A. Tewfik and M. Kim. Correlation structure of the discrete wavelet coefficients of fractional brownian motion. *IEEE Trans. on Information Theory*, 38(2):904–909, 1992. [39]

[94] M. Vannucci and F. Corradi. Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective. *J. R. Statis. Soc. B*, 61:971–986, 1999. [52, 53]

[95] M. Vetterli and J. Kovacevic. *Wavelets and Subband Coding.* Prentice-Hall, Englewood Cliffs, 1995. [4, 28]

[96] B. Vidakovic. Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *J. Amer. Statis. Assoc.*, pages 173–179, 1998. [5, 41, 43, 50]

[97] B. Vidakovich and F. Ruggeri. BAMS Methods: Theory and simulations. *Indian Journal of Statistics, Special Issue on Wavelet Methods*, 63, 2001. [43, 48]

[98] M. Wainwright, E. Simoncelli, and A. Willsky. Random cascade on wavelet trees and their use in modeling natural images. *Appl. Comput. Harmon. Anal*, 11:89–123, 2001. [50, 51, 52, 53]

[99] G. Wang, J. Zhang, and G. Pan. Solution of inverse problems in image processing by wavelet expansion. *IEEE Trans. on Image Processing*, 4:579–593, 1995. [4]

[100] A. Willsky. Multiresolution Markov models for signals and image processing. *Proceedings of the IEEE*, 90(8):1396–1458, 2002. [4, 13, 20, 22, 110, 113, 130]

[101] Y. Xu, J. Weaver, D. Healy, and J. Lu. Wavelet transform domain filters: a spatially selective noise filtration technique. *IEEE Trans. on Image Processing*, 3(6):747–758, 1994. [50, 52, 54]

[102] M. Yaou and W. Chang. Fast surface interpolation using multiresolution wavelet transform. *IEEE Trans. on PAMI*, 16:673–688, 1994. [180]

[103] Y. Yoo, A. Ortega, and B. Yu. Image subband using coding context–based classification and adaptive quantization. *IEEE Trans. on Image Processing*, 8:1702–1715, 1999. [50]

[104] X. Zhang and M. Desai. Adaptive denoising based on SURE risk. *IEEE Signal Processing Letters*, 5(10):265–267, 1998. [43, 44]