
K. E. TIMMERMANN
PH.D.
2000

This is to certify that the

dissertation entitled

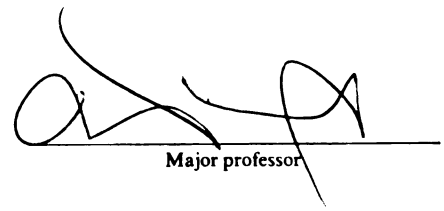
MULTISCALE MODELING AND ESTIMATION OF POISSON PROCESSES WITH APPLICATIONS TO EMISSION COMPUTED TOMOGRAPHY

presented by

Klaus E. Timmermann

has been accepted towards fulfillment
of the requirements for

Ph.D degree in Electrical Eng



Major professor

Date 8/22/00

MULTISCALE MODELING AND ESTIMATION OF POISSON PROCESSES WITH APPLICATIONS TO EMISSION COMPUTED TOMOGRAPHY

By

Klaus Edmond Timmermann

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Electrical and Computer Engineering

2000

ABSTRACT

MULTISCALE MODELING AND ESTIMATION OF POISSON PROCESSES WITH APPLICATIONS TO EMISSION COMPUTED TOMOGRAPHY

By

Klaus Edmond Timmermann

Many important problems in engineering and science are well-modeled by Poisson processes, and in many instances it is of great interest to accurately estimate the intensities underlying the observed Poisson data. This dissertation addresses the problem of general Poisson process estimation, but the work is primarily motivated by the photon-limited imaging problem. First, a Bayesian approach to Poisson intensity estimation based on a multiscale framework is presented. It is shown that the multiscale representation of signals provides a very natural and powerful framework for this problem. Using this framework, a novel multiscale Bayesian prior to model intensity functions is devised. The behavior of the new model is characterized by a study of its correlation properties. The new prior leads to a simple, Bayesian intensity estimation procedure. Practical fast shift-invariant algorithms for the new estimation framework are presented and applied to photon-limited data.

We extend the modeling and estimation approaches for general Poisson processes to the emission computed tomography (ECT) image reconstruction problem. Two multiscale approaches are introduced. The first approach is based on the geometrical

properties of the so-called *natural pixels* of the intensity image. Within this framework, we develop a practical prior model for the sinogram image, which is used to estimate the underlying sinogram intensity from the raw projection data prior reconstruction. The sinogram estimate is then used in conjunction with the standard filtered-backprojection algorithm to produce an improved image reconstruction. The superiority of the new approach over the conventional filtered-backprojection based reconstruction is illustrated with clinical and simulated data.

The second, more sophisticated approach to ECT is based on a new multiscale-based Radon inverse transform. It is shown that the new formulation lends itself to a very natural discretization of the Radon inverse operator which is especially amenable to numerical computation. Within this new reconstruction framework, we develop a prior model for the *cumulative sinogram image* —a new intermediate image representation between a sinogram and the intensity image. The new model is unique in that it exploits the high degree of redundancy of information present in the sinogram.We demonstrate the superiority of the proposed method using real data.

As the multiscale framework to modeling and estimation represents the foundation of the methods presented here, we introduce a new approach to characterizing multiscale models and estimators in order to assess their qualities. This approach is shown to be more general in nature than classical statistical descriptors (*e.g.*, biasness, mean-square error, *etc.*) which are based on limited information. Towards this end, we introduce the information-theoretic definitions of *anomy*, *accuracy precision*, and *resolution power*. Based on criteria developed with these concepts, we explore the advantages of multiscale Bayesian modeling and motivate its application to the estimation problem.

In memory of my Father,

who instilled in me my love for nature,

who taught me the values for order, simplicity and social responsibility,

and who got me started in sciences.

To my Mother,

who nourished the spiritual side in my life,

and who taught me the importance of family and true friendship.

To my beautiful wife,

who has been there for me in times of trial,

and in every other moment to enjoy life together.

To my children, Karl and Derek,

who make it all worth it,

and who remind me day to day that I am not there yet.

We often s

others for the

wish to thank t

ate them all w

First, I wo

time, help and

thank my advi

dedicated to m

and helpfulne

patience as he

and Dr. Hassa

later weigh he

Also, my

supported and

Jim Mills, Jr.,

Willy and Em

Espinoza. An

and Guillermo

I also thank

per workout wi

ACKNOWLEDGMENTS

We often go about our day-to-day business of life not realizing how much we owe others for the many good things that come to us. It is only in the rare occasions we wish to thank them in writing that we come to realize the many they are and that cite them all we cannot. Below is my short list.

First, I would like to thank the members of my guidance committee for their time, help and advice during the various stages of this long endeavor. In particular, I thank my advisor, Dr. Robert Nowak for the many hours of technical discussions he dedicated to me during our years of association; Dr. John Deller for his friendliness and helpfulness throughout these years; Dr. Michael Frazier for his kindness and patience as he introduced me to the fascinating areas of real and multiscale analysis; and Dr. Hassan Khalil for his openness and warmth during an interview which would later weigh heavily in my decision to attend Michigan State.

Also, my gratitude and thanks go to the following very special people who supported and encouraged me during my years of study: Jim and Lois Mills, Jim Mills, Jr., Sue Vollmar, Ken and Jan Dart, and my very dear brother and sister Willy and Emmy, and their spouses, Maria Esther Timmermann and Jesús Alonso Espinoza. And to those whom I owe the most: my wife, Judy and parents, Leonor and Guillermo Timmermann, I thank from the deepest of my heart.

I also thank the National Science Foundation for its very generous financial support without which I would not have been able to pursue this higher academic degree.

TABLE OF CONTENTS

3.1 AM

Pea

3.2 AM

Pea

LIST OF TABLES

LIST OF FIGURES

CHA

Introd

A great numb

Poisson proce

gives rise to th

is encountered

3, and netwo

nsity of a ge

we observe co

and wish to e

one-dimensio

dimension, in

For examp

ing a Poisson

a compact r

ng two-dimen

to λ genera

CHAPTER 1

Introduction

A great number of important phenomena in science and engineering are modeled as Poisson processes. Often, it is of interest to estimate the underlying intensity which gives rise to these phenomena. The intensity estimation problem of Poisson processes is encountered in many fields including medicine [1], astronomy [2], communications [3], and networks [4]. This dissertation considers the problem of estimating the intensity of a general Poisson process from a single observation of the process. That is, we observe counts $\mathbf{c}$ which obey[1]

$$\mathbf{c}|\boldsymbol{\lambda} \sim \text{Poisson}(\boldsymbol{\lambda}), \tag{1.1}$$

and wish to estimate the intensity $\boldsymbol{\lambda}$. The counts $\mathbf{c}$ and intensity $\boldsymbol{\lambda}$ are typically one-dimensional (1-d) signals or 2-d images, but may be of any other higher finite dimension, in general.

For example, the basic photon-limited imaging process is widely regarded as obeying a Poisson law. The problem is described as follows. We observe photon emissions in a compact region of the plane. The photon emissions are the result of an underlying two-dimensional intensity function. We are interested in estimating the intensity

[1] $\mathbf{c}|\boldsymbol{\lambda}$ denotes the random vector $\mathbf{c}$ given the vector $\boldsymbol{\lambda}$.

function fro

plication tha

limitation of

ies. due in p

safety. Becau

post-filtering

intensity [5].

In this dis

mation based

of signals pro

ing this frame

is devised. W

expression for

and shown to

optimal Baye

invariant algor

to photon-limi

As a second

and estimation

ECT is an imp

in functional

associated wit

emission tomo

of radioactive

techniques mak

One import

Compared tomo

function from the counts of photon detections. Nuclear medicine imaging is one application that motivates our study of the photon-limited imaging problem. The major limitation of nuclear medicine imaging is the low-count levels acquired in typical studies, due in part to the limited level of radioactive dosage required to insure patient safety. Because of the variability of low-count images, it is very common to employ a post-filtering or estimation procedure to obtain a "better" estimate of the underlying intensity [5].

In this dissertation we introduce a Bayesian approach to Poisson intensity estimation based on a multiscale framework. It is shown that multiscale representation of signals provides a very natural and powerful framework for this problem. Using this framework, a novel multiscale Bayesian prior to model intensity functions is devised. We look at the nature of the proposed prior by deriving a closed-form expression for its autocorrelation. Some extensions to the basic model are developed and shown to possess very desirable properties. With these new priors, a simple, optimal Bayesian intensity estimation procedure is developed. A practical fast shift-invariant algorithm for the new estimation framework is also presented and applied to photon-limited data.

As a second contribution of this dissertation, we extend the above general modeling and estimation approaches to the emission computed tomography imaging problem. ECT is an important and very active area of research in many fields. For example, in functional neuroimaging, ECT is used to map regions of activity in the brain associated with physical [6] and intellectual [7] tasks; in nuclear waste management, emission tomographic methods may be employed to determine the activity density of radioactive material within cemented barrels [8]; and in electron microscopy, ECT techniques make possible 3-d image reconstruction of chromosomes' structures [9].

One important diagnosis method in nuclear medicine is single-photon emission computed tomography (SPECT). Essentially, the process aims to reconstruct density

maps imag

data collect

well-modele

imaging of n

levels and, th

portance to m

sions of tomog

However, many

other ECT ap

This disser

construction p

so called *natu*

the sinogram i

sinogram ima

the raw projec

conjunction w

an improved i

the convention

We also in

new multis ale

lends itself to

especially ame

to Fourier-bas

inversion opera

to be sampled

approximation

and space pla

maps (images) of radiopharmaceutical distribution within a patient from projection data collected at many angles about the subject. Projection data in SPECT are well-modeled to be the outcome of Poisson processes, and just as in the case of static imaging of nuclear medicine discussed above, they are characterized by low-count levels and, therefore, by low signal-to-noise ratios (SNR). Motivated by its great importance to medicine, and by the challenge posed by its low-count nature, our discussions of tomography focuses throughout on the SPECT problem of nuclear medicine. However, many results obtained here are directly applicable or easily extendable to other ECT applications.

This dissertation presents two new multiscale approaches to the ECT image reconstruction problem. The first approach is based on geometrical properties of the so called *natural pixels* of intensity sinograms as well as on the high structure of the sinogram image. Within this framework, we develop a practical prior model for sinogram images which are used to estimate the underlying sinogram intensity from the raw projection data prior reconstruction. The sinogram estimate is then used in conjunction with the standard filtered-backprojection (FBP) algorithm to produce an improved image reconstruction. We illustrate the superiority of this method over the conventional FBP-based approach using clinical and simulated data.

We also introduce a second and more sophisticated approach to ECT based on a new multiscale-based approach to the Radon inverse transform. This new formulation lends itself to a very natural "discretization" of the Radon inverse operator which is especially amenable to numerical computation. It will be seen that in fact, in contrast to Fourier-based numerical reconstruction methods widely used in practice, the new inversion operator requires no discretization itself, and it is only the data that needs to be sampled. Fourier-based numerical reconstruction methods require conflicting approximations to the Radon operator by simultaneously discretizing the frequency and space planes. In these methods, the intensities must be assumed to be of finite

support in both space and frequency, and such signals do not exist. The method presented here avoids such conundrum and reconstructs intensity images without the associated potential artifacts.

Within the new reconstruction framework, we develop a prior model for the *cumulative sinogram image (intensity)* (CSI). The new model is unique in that it exploits the high degree of redundancy of information present in a sinogram to create a very robust tomographic reconstruction of photon-limited images. We demonstrate the advantage of the proposed method using clinical data.

Multiscale representation and analysis of signals have been used in the past as the basic framework in the modeling and estimation of Poisson, Gaussian, and other types of processes [10, 11, 12, 13, 14, 15, 16]. While in general all these approaches have been successful, except for the case of Gaussian processes, a clear justification for their choice has not been offered. In an attempt to explain the advantage of multiscale modeling and estimation in a most general way, we introduce information-theoretic measures that quantify the degree of goodness of models in various aspects. For this, we introduce the concepts of *anomy*, *accuracy*, *precision*, and *resolution power*. Based on criteria developed with these concepts we gain insight into the advantage of Bayesian estimators within the framework of multiscale representation of processes. The Bayesian approach is shown to give the means for a systematic and maximal use of the available information at each scale of multiscale models. Additionally, we give general guidelines for constructing new multiscale linear transformations which are statistically motivated and applicable to the general Gaussian model. The transformations are potentially better suited than conventional time/frequency multiscale analysis for estimation purposes, and may also be extended to other models as well.

1.1 Organization and Summary of Contributions

The main objective of Chapter 2 is to motivate the multiscale "paradigm" as an approach to modeling for Bayesian estimation. Throughout this dissertation the multiscale modeling of processes and the multiscale representation of signals play very important roles as they provide the underlying framework for the Poisson models and estimators presented in Chapters 3 and 4. Therefore, in Section 2.2 we first present an elementary review of wavelets. Much intuition about multiscale modeling and estimation is gained from its understanding. Also, much of the notation use throughout this dissertation is introduced here and in Section 2.3.

In addition, Chapter 2 offers the following three main contributions. First, in Section 2.3 we present a new, unifying approach to characterize and qualify models and estimators alike. For this, we introduce four new measures that are more general, and we believe more natural, than the conventional statistical measures often used for this purposes. Second, in Section 2.4 the advantages of modeling and estimating within the multiscale framework in general, and within the multiscale Bayesian approach in particular, are established for a class of processes. Third, we give general guidelines to constructing multiscale models which are statistically motivated, and consequently, are better suited for the estimation problem. These guidelines are developed for Gaussian processes, but they can be easily extended to other processes. Much work needs to be done in this respect, but the criteria set forth here opens a wide range of new and exciting possibilities. In Section 2.5 we conclude the chapter with a brief review of some important existing multiscale and estimation approaches.

There are four major contributions in Chapter 3. First, in Section 3.4 we describe a new, multiscale, prior probability model for non-negative intensity functions. This model employs a multiplicative innovations structure in the scale-space domain. Second, based on this new prior, in Section 3.5 we derive a simple and computationally

efficient, Bayesian estimator of the intensity given an observation of counts, under squared error loss. It is shown through examples that the Bayesian estimation procedure significantly outperforms existing wavelet-based methods. Third, we extend in Section 3.6.1 the multiscale intensity prior to a shift-invariant one, and develop a fast shift-invariant estimation procedure. Furthermore, we obtain closed-form expressions for the correlation functions of both priors, and show that the correlation behavior of the shift-invariant prior has $1/f$ spectral characteristics and is more regular than that of the shift-variant prior. Fourth, in Section 3.8 we apply the framework to photon-limited imaging and examine its potential to improve nuclear medicine imaging.

The main contributions in Chapter 4 are three. First, in Section 4.5 we introduce an extension to the prior modeling approach of Chapter 3 which is especially well suited for modeling computed tomography sinograms. The new prior is based on geometrical consideration of the inherent structure of sinograms. The excellent match of the prior to real and synthetic data is seen in the examples. Second, in Section 4.6 we present a new, multiscale-based, Radon-inverse transform algorithm. The transform has three major qualities for ECT applications: it admits a very efficient computational implementation; it allows reconstruction of images at any desired resolution supported by the data with the corresponding computational savings; and it provides a very robust reconstruction of images from photon-limited projection data. This last quality is unique to the new method and is illustrated with an example. Third, in Section 4.7 we introduce a third extension to the intensity modeling approach developed for general Poisson processes and apply it to the cumulative sinogram images of emission-computed tomography. This prior is used in conjunction with the new Radon-inverse transform to reconstruct highly reliable images from photon-limited data. The superiority of this modeling, estimation, and reconstruction method is illustrated with clinical data.

Finally, some comments and conclusions are given in Chapter 5.

CHAPTER 2

The Multiscale Modeling Paradigm

2.1 Preliminaries

In his book "Conceptual Physics" [17], Hewitt identifies a crucial factor that made the evolution of human scientific knowledge possible when he writes: *"Science had its beginnings before recorded history when people first discovered* recurring relationships *around them. Through careful observations of these relationships, they began to know nature and, because of nature's dependability, found they could make predictions that gave them some control over their surroundings."*

Hewitt's recurring relationships refer to the observed patterns of cause and effect that appeared to dictate the course of nature in many instances. It is evident that while our ancestors could not make predictions about the exact shape of a flame in a fire, they could always expect the whole of the flame and smoke to rise. Thus, looked at on a large enough scale, the phenomenon could be satisfactorily predicted.

Probably one of the greatest successes in science before modern times was achieved in predicting the movement of the planets. In their observations of the skies, the Mayans did not have to contend with erratic or chaotic effects, but only had to discover the patterns manifested at very large scales. It is no coincidence that nowadays scientists still seek to discover patterns in whatever is under study in order to advance

their knowledge. After all, if the behavior of artificial neural networks is a hint of how the human brain works, to understand nature means training our brains enough so as to recognize patterns in our environment, for then we may predict its future behavior.

As the subjects studied by people became more and more complex, the patterns to be discovered were less evident and more difficult to perceive. Mathematics then became the primary tool in this quest. After the experimentation phase, models were postulated and tested. At first, the mathematical models were deterministic in nature, but as scientists' interest shifted towards natural phenomena "belonging" to sufficiently small scales, probabilistic models had to be introduced. For example, before the turbulent behavior of the flames in a fire could be understood, statistical thermodynamic models had to be postulated to explain the molecular behavior of gases.

Clearly, the new probabilistic models represented more accurate but less precise[1] models than their deterministic counterparts. They predicted the outcome of an experiment more reliably while providing less detail about such outcome—gas molecules' "typical" behavior could be predicted very successfully; however, very little could be said about any given molecule's state, *e.g.*, its location and momentum. The reason for this was that while small scale phenomena were being modeled, the experiments carried out to undercover their hidden patterns were of much larger scales; consequently, only the patterns displayed at these larger scales provided information about the underlying phenomena. In the case of gas molecules, only their aggregate behavior could be discerned and so, only probabilistic inferences about individual molecules could be made.

[1]The terms accuracy and precision are used here in a general sense, meaning, respectively, the accordance of an assertion with the truth, and the amount of information conveyed by that assertion. Thus, by forecasting rain over the Pacific Ocean this year, one makes a highly accurate assertion but of very low precision for not much information about the time or place of the event is given, for example. In Section 2.3.2 we give precise meaning to these terms.

As advances in technology allowed probing at smaller and smaller scales, the models created became more and more precise. Nevertheless, to maintain accuracy, they necessarily had to be probabilistic in nature, for if we accept that every event in nature has its origins in the smallest of scales, the Heisenberg Principle prevents us from observing the full state (*e.g.*, displacement/momentum, time/frequency, *etc.*) of whatever "lives" at those scales, and so, prevents us from producing infinitely accurate deterministic models—never mind that our known deterministic "laws" of nature do not apply at these extremely small scales.

Although in practice we are often content with deterministic coarse-scale averaged representations of observed phenomena, the above discussion brings to light the intuitive notion that while modeling of phenomena based on coarser scale information alone may be more accurate, it can only be achieved at the expense of precision. Alternatively, coarse-scale models may be constructed more accurately than their fine-scale counterparts, but the coarse-scale models are necessarily less precise.

These two statements are not just reworded versions of one another; however, once the information-theoretic definitions for accuracy and precision have been introduced, we will show them to be equivalent. At that point, we will also be able to gain a greater insight as to their interpretation. Their significance is as follows.

The first assertion establishes that construction of more precise models requires new information, and that such information may only be found in finer scale patterns. This implies the futility of trying to improve the precision of estimators beyond what the observations' scale supports.

The ability to trade precision for accuracy by modeling different scales of a process, as established by the second assertion, is of great significance to the point estimation problem. We will show that under certain conditions the robustness[2] of an estimator

[2]In the literature (see, for example [18]) *robustness* connotes the consistency of an estimator's performance under all possible distributions for the observations, that is, under typical observations

can be enhanced by leveraging the estimation process through this trade. The multiscale Bayesian estimation approach introduced in Chapter 3 will be shown to provide an intrinsic way to achieve this.

Given the crucial roll that multiscale representation of signals plays throughout this dissertation, we next give a brief introduction to the topic. To this end, we make use of wavelet theory as it gives a natural perspective of multiscale analysis. We avoid as much as possible the formalities associated with this topic, however, and emphasize a motivational point of view, as this is the goal of the present chapter.

2.2 Multiresolution through Wavelets

2.2.1 The Fourier Transform and Scales

In the quest to identify the underlying patterns of phenomena and processes, expressing the signals of interest as a linear combination of more elemental functions has often proved invaluable. With this approach, features in the signals which are uniquely characteristic to the event under study have been brought to view and in this manner have helped to identify the patterns.

For most problems encountered in engineering the sets of functions of greatest utility have been those which are complete in $L^2(\mathbb{R})$ (or more generally, in $L^2(\mathbb{R}^N)$)[3] since they can represent any finite energy signal that might arise.[4] Undoubtedly, the Fourier system has had the greatest of influences in this respect since its discovery in 1807. One reason for this is the orthogonality of the set, which leads to its simplicity and wide range of use. However, it is the fact that the elementary functions are

as well as those including outlayers. Here, we use the term to mean a degree of goodness. Later in the chapter we will introduce the concept of *resolution power* and use it for this purpose.

[3]For the sake of simplicity, throughout this dissertation we focus our attention on signals on the real line when this suffices to make the desired point.

[4]Clearly, the fidelity of such a representation is only in the mean-square error sense and so, it is only accurate to within a set of measure zero. We ignore these technicalities for the most part.

eigenfunctions to linear translation invariant operators that has mostly contributed to its great extent of applications.

The Fourier transform decomposes a function into its frequency components; this is particularly easy to visualized in the case of periodic functions—clearly, not L^2 functions. In this case, the components are denumerable and readily identifiable at finite intervals along the frequency axis; thus, admitting a series representation for the signals. For L^2 functions, the situation is radically different as no one component is present in the original signal, for if any one were, it would be so only through its manifestation of energy in some form, no matter how small this is, but we know that no energy is born at any given frequency.

An alternate view of the Fourier representation is that the transform decomposes a signal into a countable number of *scales* corresponding to an arbitrary partitioning of the frequency axis. The system in this case is still orthogonal, but the basic elements, or atoms, are now the functions that the Fourier exponentials integrate to within each of the frequency intervals. The advantage of this system from the temporal viewpoint is that L^2 signals may now be represented as linear combinations of such atoms, with each atom contributing a finite amount of energy—if present at all, that is. From the frequency perspective, we have that the highly frequency-localized analysis of signals is preserved, although not to the infinite frequency resolution of the original Fourier system.

As an example, this scale decomposition may be constructed with a set of Gabor-like functions defined on the frequency line:

$$\hat{h}_{j,k}(\omega) \equiv \frac{1}{\xi_0}\mathbf{1}\left(\frac{\omega - j\xi_0}{\xi_0}\right) e^{i\,k u_0(\omega - j\xi_0)} \quad \text{for all } j,k \in \mathbb{Z}, \tag{2.1}$$

where $\mathbf{1}(\cdot)$ is the indicator function over the $[-1/2, 1/2)$ interval, $i^2 = -1$, and u_0 and ξ_0 are arbitrary parameters. Clearly, the set $\{\hat{h}_{j,k}\}_{j,k}$ partitions the frequency line

into adjacent non-overlapping uniform intervals of width ξ_0, and include modulating factors. Letting $u_0 \xi_0 = 2\pi$, it is easy to show that the frequency spectrum $\hat{f}$ of f may be expressed as (see Appendix A.1)

$$\hat{f}(\omega) = \sum_{j,k} \langle \hat{f}, \hat{h}_{j,k} \rangle \hat{h}_{j,k}(\omega), \tag{2.2}$$

with $\langle \cdot, \cdot \rangle$ denoting inner product. By obtaining the inverse Fourier transform of both sides of this equality and using Parseval formula we obtain,

$$\begin{aligned} f(t) &= \left(\sum_{j,k} \langle \hat{f}, \hat{h}_{j,k} \rangle \hat{h}_{j,k}(\omega) \right)^{\vee} \\ &= \sum_{j,k} \langle \hat{f}, \hat{h}_{j,k} \rangle (\hat{h}_{j,k}(\omega))^{\vee} \\ &= 2\pi \sum_{j,k} \langle f, h_{j,k} \rangle h_{j,k}(t). \end{aligned}$$

The commutation of the summation and integral operators in these steps is assured since every Fourier series is integrable term by term[19]. Here,

$$h_{j,k}(t) = \frac{\sin \frac{\xi_0}{2}(t + ku_0)}{\frac{\xi_0}{2}(t + ku_0)} e^{ij\xi_0 t} \quad \text{for all } j, k \in \mathbb{Z}. \tag{2.3}$$

The partition of the frequency axis could have been chosen to produce a more convenient scale.[5] The octave scale is especially meaningful in acoustics as it matches the sensitivity scale of human hearing more naturally, and is obtained by replacing j by 2^j everywhere in the above expressions.

In many instances, the relevant information in a signal is transitory in nature. In an image, for example, it is the edges and other singularities which often convey the

[5]Throughout, we use the term *scale* with two different but related meanings: one, to signify a single element in the partition of the frequency axis, when we want to stress its place among the rest of the elements in the partition; and two, the partition itself, when we want to stress the specific structure of the partition.

information of interest. This explains why one can frequently discern the content of an image of a real-world scene solely from its high-pass filtered version; the contour of a human face and the silhouette of a house convey much of the total information. Consequently, parsimonious representations of an image with its short-duration features require temporally narrow well-localized atoms. On the other hand, in order to achieve fidelity of representation while maintaining conciseness, long-duration features need also to be represented accurately with only a few atoms. This would insure, for instance, that the presence of texture in an image would not undo the economy or the quality of the linear representation.

A most practical system, then, would consist of both arbitrarily narrow time-localized elements and arbitrarily narrow frequency-localized (temporally long) elements, as well as everything in between. Clearly, neither the Fourier nor the system (2.1)–(2.3) possess these characteristics.

2.2.2 Wavelet Representation of Signals

The theory of wavelets provides a systematic approach to constructing a complete orthogonal set through iterated dilations and translates of an elemental function ψ, called a *mother wavelet*. When this function is chosen to be fairly well localized in time, the wavelet system becomes highly efficient in representing singularities and other local features. This is the case, in particular, with wavelet functions of finite support. However, for this class of wavelets, the scales necessarily overlap to some extent and do not strictly constitute a partition of the frequency axis. Nevertheless, the induced scales are always well defined, if not well localized, due to the inherent nature of wavelets. For this reason, wavelet analysis of signals offers a versatile representation which exposes time-frequency patterns or features not discernible in either the time nor the frequency domain alone.

Multi-resolution Analysis

In general, a wavelet representation of an $L^2(\mathbb{R})$-function f has the form

$$f = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}, \tag{2.4}$$

where $d_{j,k} \equiv \langle f, \psi_{j,k}\rangle$ and $\psi_{j,k}(t) \equiv 2^{-j/2}\psi(2^{-j}t - k)$. Meaning is given to this decomposition through the concept of *multi-resolution analysis*, which is the foundation of every wavelet system representation.

Multi-resolution analysis consists of a family of closed subspace $\{V_j\}_{j\in\mathbb{Z}}$ of $L^2(\mathbb{R})$ having the following properties: [6]

i) $V_{j+1} \subseteq V_j$ for all $j \in \mathbb{Z}$, and $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$.

ii) There exists a function ϕ such that the set $\{\phi(t-k)\}_{k\in\mathbb{Z}}$ is a complete orthonormal set in V_0.

iii) A function $f(t)$ is in V_0 if, and only if, $f(2^{-j}t)$ is in V_j.

iv) $\bigcup_{j\in\mathbb{Z}} V_j$ is dense in $L^2(\mathbb{R})$, *i.e.*, for any $f \in L^2(\mathbb{R})$, there exists a sequence $\{f_n\}_{n\in\mathbb{N}}$ in $\bigcup_{j\in\mathbb{Z}} V_j$ which converges to f in the L^2 sense.

The function ϕ is known as the *scaling function.* Notice that by ii and iii, $\phi(2^{-j}t-k) \in V_j$ for all $j \in \mathbb{Z}$. It is easy to see that since the set $\{\phi(t-k)\}_{k\in\mathbb{Z}}$ is orthonormal, the set $\{\phi(2^{-j}t-k)\}_{k\in\mathbb{Z}}$ is orthogonal, and, with the scaling of $2^{-j/2}$, orthonormal as well. Moreover, the set $\{2^{-j/2}\phi(2^{-j}t-k)\}_{k\in\mathbb{Z}}$ turns out to be complete in the space V_j. Therefore, any element of V_0 can be expressed in terms of this set corresponding

[6] Multi-resolution analysis may be defined for other spaces other than $L^2(\mathbb{R})$, for example, Sobolev spaces $W_p^m \equiv \{f : |\int f^{(k)}(t)|^p\, dt < \infty,\ k = 0, 1, \ldots, m\}$ with $0 < m, p < \infty$; but we only concern ourselves with the former.

to V_{-1} since $V_0 \subseteq V_{-1}$, *i.e.*, there exists a sequence of numbers $\{h_k\}_{k\in\mathbb{Z}}$ such that

$$\phi(t) = \sum_{k\in\mathbb{Z}} h_k 2^{1/2}\phi(2t-k). \tag{2.5}$$

In a similar manner to the definition given earlier for $\psi_{j,k}$, let $\phi_{j,k}(t) \equiv 2^{-j/2}\phi(2^{-j}t-k)$. The conditions for an infinite set of subspaces to be a multi-resolution analysis were first stipulated by Stéphane Mallat [20], who in turn gave the following very important result. Define the sequence $\{g_k\}_{k\in\mathbb{Z}}$ by $g_k \equiv (-1)^{k-1}h^*_{1-k}$ for all $k \in \mathbb{Z}$, where h^*_k designates the complex conjugate of the coefficient h_k in the *scaling equation* (2.5). Define

$$\psi(t) \equiv \sum_{k\in\mathbb{Z}} g_k 2^{1/2}\phi(2t-k), \tag{2.6}$$

Then, $\{2^{-j/2}\psi(2^{-j}t-k)\}_{j,k\in\mathbb{Z}}$ is a complete orthonormal set in $L^2(\mathbb{R})$, *i.e.*, $\{\psi_{j,k}\}_{j,k\in\mathbb{Z}}$ is a wavelet system for $L^2(\mathbb{R})$. This justifies (2.4). Furthermore, the set $\{\psi_{j,k}\}_{k\in\mathbb{Z}}\bigcup\{\phi_{j,k}\}_{k\in\mathbb{Z}}$ is orthonormal, with the consequence that if

$$W_j \equiv \{\sum_{k\in\mathbb{Z}} z_k\psi_{j,k} \,\big|\, (z_k)_{k\in Z} \in l^2(\mathbb{Z})\},$$

then every element of W_j is orthogonal to every element of V_j. The "parent" scaling functions ϕ, and the mother wavelets ψ for the Haar and Daubechies systems are shown in Figure 2.1. Dilation and translations of these atoms give rise to the entire family $\{\psi_{j,k}\}_{k\in\mathbb{Z}}\bigcup\{\phi_{j,k}\}_{k\in\mathbb{Z}}$.

Mallat also found that $\{W_j, V_j\}$ forms a partition for V_{j-1}, *i.e.*, $W_j, V_j \subseteq V_{j-1}$ and for every element v_{j-1} of V_{j-1}, there exist elements w_j and v_j of W_j and V_j, such that $v_{j-1} = w_j + v_j$. This is expressed concisely as the direct sums $V_{j-1} = W_j \bigoplus V_j$, and are depicted in Figure 2.2.

An immediate consequence of this relationship is the following. Designate the

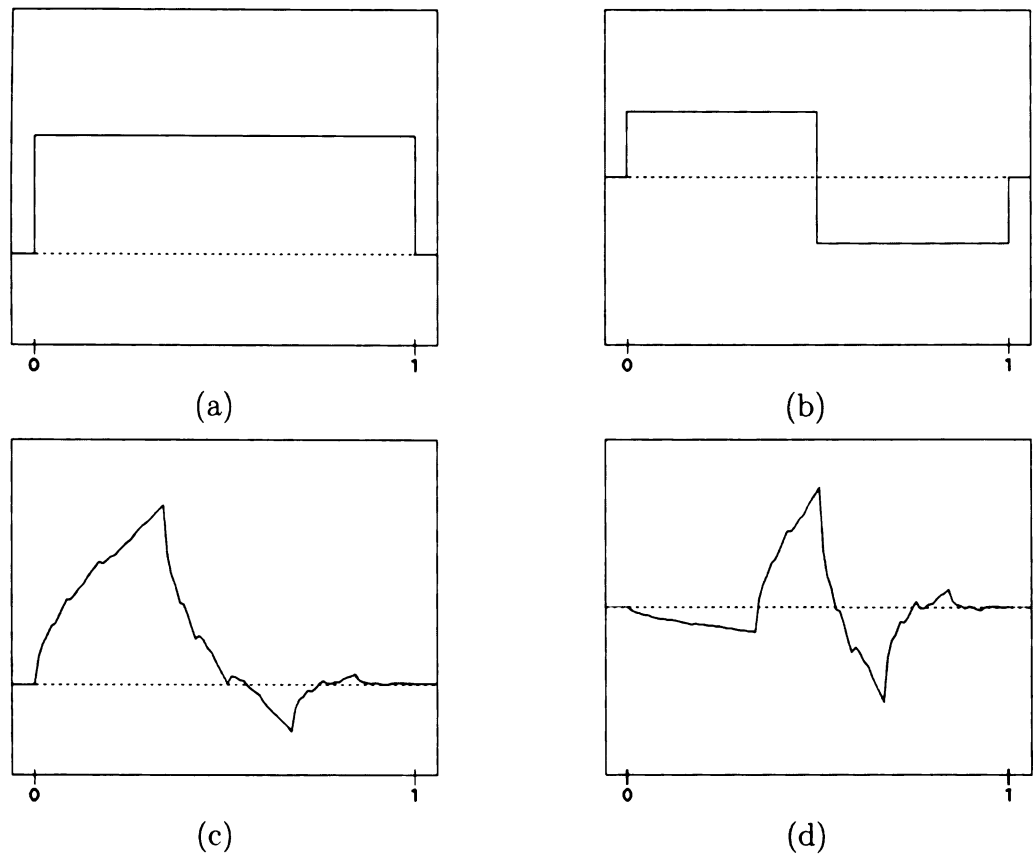


Figure 2.1. *The Haar and Daubechies 4-pt scaling and wavelet functions in the* ***interval.*** *(a) Haar scaling function. (b) Haar mother wavelet. (c) Daubechies scaling* ***function.*** *(d) Daubechies mother wavelet.*

projection of f on V_{j-1} by f_{j-1}, that is,

$$f_{j-1} \equiv \sum_{k \in \mathbb{Z}} c_{j-1,k} \phi_{j-1,k}, \tag{2.7}$$

where $c_{j-1,k} \equiv \langle f, \phi_{j-1,k} \rangle$. We may then also write

$$f_{j-1} = \sum_{k \in \mathbb{Z}} (c_{j,k} \phi_{j,k} + d_{j,k} \psi_{j,k}), \tag{2.8}$$

with $d_{j,k}$ as in (2.4). Equating these two expressions and writing $\phi_{j,k}$ and $\psi_{j,k}$ in terms of $\phi_{j-1,k}$ according to the generalized versions of (2.5) and (2.6)—which are obtained

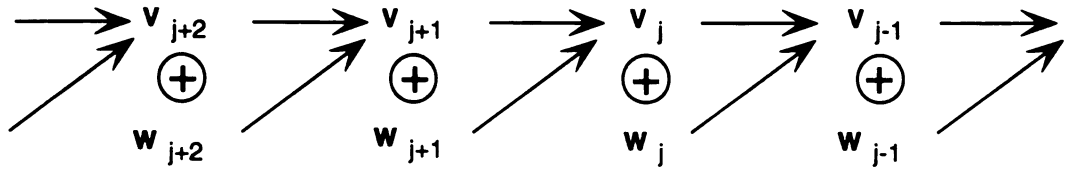


Figure 2.2. *Multiresolution analysis expressed as a sequence of direct sums. Each space V_{j-1} encompasses the functions living in V_j and W_j, and all those functions formed by the sum of any two elements in $V_j \cup W_j$. Since W_j is orthogonal to V_j and, therefore, to every V_l with $l \geq j$, W_j defines the scale of $L^2(\mathbb{R})$ functions with features not representable in any V_l with $l \geq j$.*

by the repeated substitution of the argument t for $2t$, and in each iteration, multiplying the resulting expression by $2^{1/2}$—the *discrete wavelet reconstruction* expression results:

$$c_{j-1,k} = \sum_{l \in \mathbb{Z}} (h_{k-2l} c_{j,l} + g_{k-2l} d_{j,l}), \tag{2.9}$$

for all j, k in $\mathbb{Z}$. The corresponding *discrete wavelet decomposition* relations are

$$c_{j,k} = \sum_{l \in \mathbb{Z}} h_{l-2k} c_{j-1,l} \tag{2.10}$$

and

$$d_{j,k} = \sum_{l \in \mathbb{Z}} g_{l-2k} c_{j-1,l} \tag{2.11}$$

These are obtained by again expressing $\phi_{j,k}$ and $\psi_{j,k}$ in terms of $\phi_{j-1,k}$ according to the generalized versions of (2.5) and (2.6) in $c_{j,k} \equiv \langle f, \phi_{j,k} \rangle$ and $d_{j,k} \equiv \langle f, \psi_{j,k} \rangle$, and writing the resulting integrals in term of $c_{j-1,k}$.

The coefficients h_k and g_k in (2.9), (2.10) and (2.11) are the *wavelet filter coefficients*, often referred to as *quadrature mirror filter* (QMF) coefficients within the

linear-filter-processing community. It is not difficult to show that the sequences (h_k) and (g_k) correspond to low-pass and band-pass filters, respectively.

Signals' Finite Representation

The fact that we can express the coefficients at one scale in terms of those at the scale immediately "above" or "below" without the need of scaling or wavelet functions is key to the usefulness of the wavelet transform, for otherwise, the wavelet transform would have gone the way of the Fourier transform before the FFT (Fast Fourier Transform) was invented. For the transform to be computationally practical, however, it is also necessary that a function may be representable by only a finite number of coefficients. The following shows how under very mild conditions this is possible.

We first rewrite (2.4) as

$$f = \sum_{j \le J'} \sum_{k \in \mathbb{Z}} d_{j,k} \psi_{j,k} + \sum_{J' < j \le J} \sum_{k \in \mathbb{Z}} d_{j,k} \psi_{j,k} + \sum_{J < j} \sum_{k \in \mathbb{Z}} d_{j,k} \psi_{j,k}, \tag{2.12}$$

where $J' < J$, but which are both arbitrary integers otherwise. From Figure 2.2, it is seen that

$$\begin{aligned} f_J &= f_{J+1} + \sum_{k \in \mathbb{Z}} d_{J+1} \psi_{J+1,k} \\ &= f_{J+2} + \sum_{j=J+1}^{J+2} \sum_{k \in \mathbb{Z}} d_{j,k} \psi_{j,k} \\ &= f_{J+3} + \sum_{j=J+1}^{J+3} \sum_{k \in \mathbb{Z}} d_{j,k} \psi_{j,k} \\ &= \cdots = \sum_{J < j} \sum_{k \in \mathbb{Z}} d_{j,k} \psi_{j,k}, \end{aligned}$$

and similarly

$$f_{J'} = \sum_{J'<j} \sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}$$

Therefore, (2.12) can be written as

$$f = (f - f_{J'}) + \sum_{J'<j\leq J} \sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k} + f_J \tag{2.13}$$

It is not hard to show that the projection f_j is the best approximation in V_j to f in the sense that $\|f - f_j\|$ is minimized over all sums $\sum_{k\in\mathbb{Z}} z_k\phi_{j,k}$ for any set $\{z_k\}_{k\in\mathbb{Z}}$ of real numbers. Then, since a multi-resolution analysis is dense in $L^2(\mathbb{R})$ (property iv), the sequence $(f_j)_{j=J'}^{-\infty}$ converges to f, *i.e.*, $\|f - f_j\| \to 0$ as $j \to -\infty$. This implies that if J' is chosen small enough, the $(f - f_{J'})$ term can be ignored maintaining an approximation to f as closely as desired. Doing this in (2.13), and arbitrarily choosing the lowest scale approximation J' to be scale zero[7] we obtain

$$\begin{aligned} f &\simeq \sum_{0<j\leq J} \sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k} + f_J \\ &= \sum_{0<j\leq J} \sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k} + \sum_{k\in\mathbb{Z}} c_{J,k}\phi_{J,k} \end{aligned} \tag{2.14}$$

This expression still involves an infinite number of terms. However, if ψ is of compact support, then the entire wavelet system $\{\psi_{j,k}\}_{j,k}$ is of compact support, and so, if f is also of compact support, there will only exist a finite number of wavelet and scaling coefficients different than zero.

[7]There is no loss of generality here for we may expand or dilate the original function by any amount required so that the approximation to the newly obtained function at scale $J = 0$ represents the desired fit.

The Discrete Wavelet Transform

When a function f admits a finite wavelet representation, it is often convenient to think of the set of coefficients $\{d_{j,k}\}_{k,0<j\leq J} \bigcup \{c_{J,k}\}_k$ in (2.14) as the wavelet representation itself of the function. In this sense, the scaling coefficients $\{c_{0,k}\}_k$ obtained by the reconstruction operation (2.9) is viewed as the (discrete) signal of interest. The operation by which $\{c_{0,k}\}_k$ is transformed into $\{d_{j,k}\}_{k,0<j\leq J} \bigcup \{c_{J,k}\}_k$ is known as the *discrete wavelet transform* (DWT).

Due to the decimation operations indicated in (2.10) and (2.11), the number of scaling (and wavelet) coefficients in a scale is always half the number of scaling coefficients in the prior lower scale. So if the coefficients at scale 0 run from, say, $k = 0$ to $N - 1$, where N is an integer's power of two, it takes $N/2^j$ scaling and wavelet coefficients each at scale j to convey the same information. Therefore, whether the coefficients $\mathbf{c} \equiv \mathbf{c}_0 \equiv (c_{0,0}, \cdots, c_{0,N})^T$ are derived from a function of a continuous variable, or constitute a discrete signal in their own right, the DWT may be expressed as

$$\mathbf{d} \equiv \begin{pmatrix} c_{J,0} \\ \mathbf{d}_J \\ \mathbf{d}_{J-1} \\ \vdots \\ \mathbf{d}_1 \end{pmatrix} = \mathcal{W}\mathbf{c}, \tag{2.15}$$

where, for any j, $\mathbf{d}_j \equiv (d_{j,0}, d_{j,1}, \cdots, d_{j,N/2^j-1})^T$, and $J = \log_2 N$. $(\cdot)^T$ denotes the transpose operation. The elements of the matrix $\mathcal{W}$ are linear combinations of the filter coefficients $\{h_k\}$ and $\{g_k\}$. These are found by expressing (2.10) and (2.11) in matrix form, and applying them iteratively up to the highest possible scale J. In

Figure 2.3. *Wavelet-based Signal Processing. A finite-support function f is first projected to a suitable scale space, e.g., V_0. The scaling coefficients $\mathbf{c} = \mathbf{c}_0$ corresponding to the projection are DWTed. Signal processing algorithms are employed to generate new wavelet coefficients $\tilde{\mathbf{d}}$ from the original. By an IDWT, corresponding scaling coefficients $\tilde{\mathbf{c}} = \tilde{\mathbf{c}}_0$ are obtained from which the desired final function $\tilde{f}$ may be synthesized.*

particular, for $N = 4$, $J = 2$ and

$$\mathcal{W} = \begin{pmatrix} h_0 & h_1 & 0 & 0 \\ g_0 & g_1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} h_0 & h_1 & h_2 & h_3 \\ h_{-2} & h_{-1} & h_0 & h_1 \\ g_0 & g_1 & g_2 & g_3 \\ g_{-2} & g_{-1} & g_0 & g_1 \end{pmatrix}.$$

A significant advantage of the DWT in problems involving finite-support functions of continuous variables is that, once the vector of scaling coefficients $\mathbf{c}_0$ has been obtained, one may operate solely on the coefficients $\mathbf{d}$ resulting from the DWT until the actual desired function is needed; at that point, the new function is synthesized using (2.14) with the newly obtained scaling coefficients following the inverse DWT (IDWT) of the processed coefficients, $\tilde{\mathbf{d}}$. This is the basis for discrete wavelet-based signal processing. A pictorial representation of this idea is shown in Fig. 2.3.

Significance of the DWT coefficients

Each element of the set $\{d_{j,k}\}_{\substack{k=0,\cdots,N-1 \\ j=1,\cdots,J}}$ is associated with a region of the *time-frequency plane* according to the location where the energy of the corresponding wavelet is mostly concentrated. In particular, wavelet coefficient $d_{j,k}$ corresponds to a region

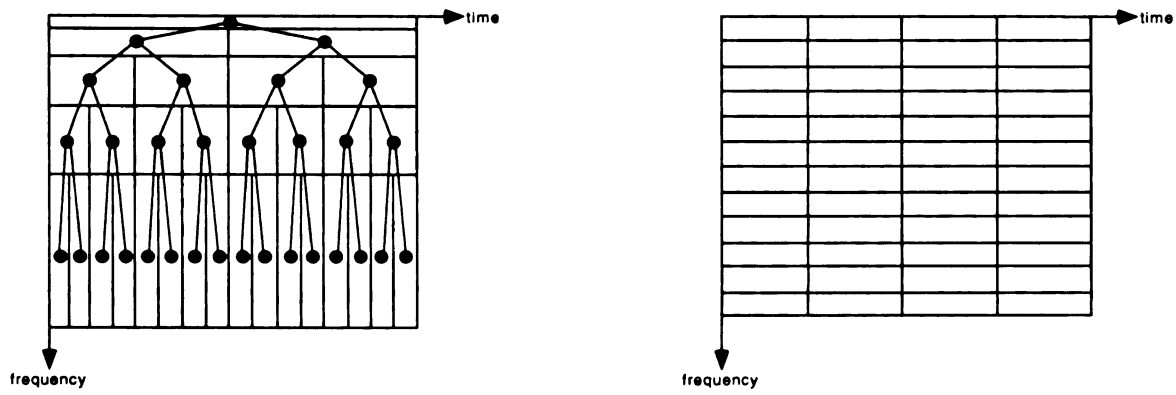


Figure 2.4. *Time-frequency plane. (a) Partition induced by the Haar system, and (b) partition induced by the system (2.3). Nodes and links in (a) display the functional relationships among the various wavelet coefficients. The uppermost node represents wavelet coefficient* $d_{J,0}$, *and those at the bottom, represent coefficients* $d_{0,0}, d_{0,1}, \cdots, d_{0,N-1}$. *See text for an alternate interpretation.*

centered at $(u_{j,k}, \xi_{j,k})$ and of dimensions $\tau_j \times \sigma_j$, where $u_{j,k} \equiv \int t\,|\psi_{j,k}(t)|^2\,dt$, $\xi_{j,k} \equiv \int \omega\,|\hat{\psi}_{j,k}(\omega)|^2\,d\omega$, and for any k, $\tau_j \equiv \int (t-u_{j,k})^2 |\psi_{j,k}(t)|^2\,dt$, and $\sigma_j \equiv \int (\omega - \xi_{j,k})^2 |\hat{\psi}_{j,k}(\omega)|^2\,d\omega$.

The time-frequency partition induced by the Haar system is shown in Figure 2.4(a), and that induced by the system (2.3) is shown in Figure 2.4(b) for reference. The nodes in (a) represent the wavelet coefficients, and the links represent their functional relationships according to (2.9), (2.10) and (2.11).

Another useful interpretation of the tree-like structure superimposed onto the time-frequency plane is that each row of nodes constitutes a projection of the original signal onto a subspace of $L^2(\mathbb{R})$. Specifically, the upper most node always represents $c_{J,0}$, the second row, the $(c_{J-1,0}, c_{J-1,1})$ vector, and so on. The bottom row corresponds to the highest resolution representation available, and as was indicated before, is often taken to be the original signal of interest. In this sense, we speak of the j-scale representation $\mathbf{c}_j \equiv (c_{j,0}, \cdots, c_{j,N/2^j-1})^T$ of $\mathbf{c}_0$, since $\mathbf{c}_j$ are the scaling coefficients of f_j whenever $\mathbf{c}_0$ are of f. Notice that under this alternate interpretation, the signal

represented by the coefficients on any given row, *i.e.*, at any scale, of Figure 2.4(a), may possess energy within the entire frequency band extending from 0 Hz (top of the figure) down to the row of coefficients.

As an illustration of these two interpretations, the values taken by the coefficients in each row of the time-frequency plane are shown in Figure 2.5. At the bottom of the figure, the scaling coefficients corresponding to scale zero constitute the original signal of interest. Right above them, the scaling coefficients corresponding to scales 1, 2, and 3 are shown along (to their left) one translate of the scaling functions used in obtaining them. Each sample on any row corresponds to a node in Figure 2.4 according to the second interpretation. On the other hand, the values taken by each node conforming to the first view are displayed on the right side of the figure. These are the wavelet coefficients obtained by the inner products between the signal at scale zero and the sequence of translates of wavelets shown in the rightmost column. Both views are valuable depending on what is being sought.

Vanishing Moments

There are many wavelet systems in existence and each is better suited for a particular set of applications. The systems may be differentiated by a myriad of properties that they may or may not hold. One very important distinguishing property often stated is the degree of regularity of the family of wavelets. For important families of wavelets this may be measured by the number of *vanishing moments* [21]. A wavelet function ψ has $v \in \mathbb{N}$ vanishing moments if [8]

$$\int t^p \psi(t)\, dt = 0 \quad \text{for } p = 0, \ldots, v-1, \tag{2.16}$$

[8]Throughout this dissertation we avoid writing the limits of integration as much as possible for simplicity of notation, but every integral indicates a definite integration operation. When the limits are omitted, the region of integration is the entire space where the variable of integration is defined.

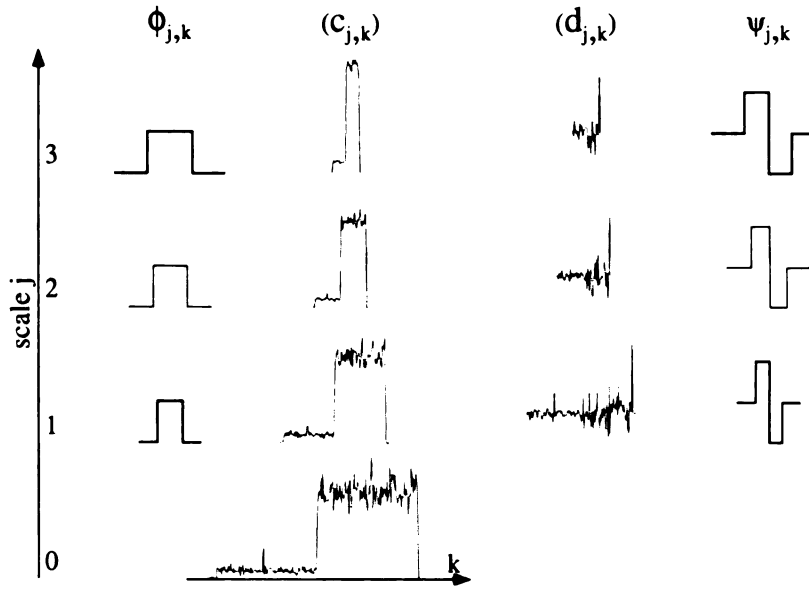


Figure 2.5. *Multiscale Signal Analysis. Scale and Wavelet coefficients (second and third columns, respectively) of the signal at scale 0 (bottom row) are displayed for scales 1, 2, and 3. Columns 1 and 4 give one single translate of a scale function and wavelet used in obtaining the coefficients. These functions correspond to the unnormalized Haar system, which we review more extensively in Chapter 3.*

but not for $p = v$.

By a simple change of variable one can easily show that if ψ has v vanishing moments, then each $\psi_{j,k}$ has v vanishing moments as well. Therefore, any function f which can be closely approximated by a polynomial of order $v - 1$ will have all its wavelet coefficients $d_{j,k}$ be zero or nearly zero. This situation represents a high degree of compression, for only a set of coarse scale scaling coefficients suffices to represent the function.

Beyond parsimonious representations, a sufficiently regular system may be exploited in estimating a signal from within noise. If the true signal is smooth enough, most of the energy in the wavelet coefficients will be from noise, and thus a simple thresholding scheme may remove much of this noise without seriously altering the true signal upon reconstruction. We review some of these wavelet-based estimation

procedures in Section 2.5.

Images

There are various ways to extend wavelet analysis to images or $L^2(\mathbb{R}^2)$ functions. A simple approach often taken consists of constructing a separable 2-d multiresolution analysis V_j^2 as the tensor product of 1-d multiresolution counterparts:

$$V_j^2 \equiv V_j \otimes V_j \quad \text{for all } j \in \mathbb{Z}. \tag{2.17}$$

The resulting family of subspaces $\{V_j^2\}$ of $L^2(\mathbb{R}^2)$ possess 2-d versions of the four properties i–iv given earlier characterizing 1-d multiresolution analyses. Specifically, the scaling property is satisfied by a scaling function defined as

$$\phi_{j,k_1,k_2}(t_1,t_2) \equiv \phi_{j,k_1}(t_1)\phi_{j,k_2}(t_2). \tag{2.18}$$

In two dimensions, the detail space W_j^2 is the space spanned not by one set of wavelets corresponding to scale j, but by three sets, each associated with a given orientation: horizontal, vertical, and diagonal. These wavelets are constructed as follows.

$$\begin{aligned}
\psi^h_{j,k_1,k_2}(t_1,t_2) &\equiv \phi_{j,k_1}(t_1)\psi_{j,k_2}(t_2) \\
\psi^v_{j,k_1,k_2}(t_1,t_2) &\equiv \psi_{j,k_1}(t_1)\phi_{j,k_2}(t_2) \\
\psi^d_{j,k_1,k_2}(t_1,t_2) &\equiv \psi_{j,k_1}(t_1)\psi_{j,k_2}(t_2).
\end{aligned} \tag{2.19}$$

Each set of wavelets defines a subspace of $L^2(\mathbb{R}^2)$ such that the projection of a function f onto it gives the energy (a measure of the amount of detail content) of the function in that orientation. As an illustration, Fig. 2.6 shows the 2-d wavelet decomposition of the standard cameraman picture. On the left is a discretized rendering of the

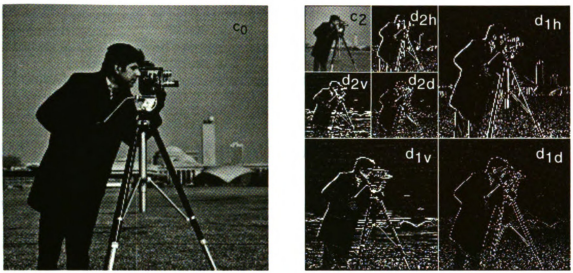


Figure 2.6. *Multiscale representation of images. (Left). The standard cameraman image: highest resolution scaling coefficients, i.e.,* c_0. *(Right). Wavelet representation of the cameraman image:* $\mathbf{c}_2$ *are the scale 2 scaling coefficients,* $\mathbf{d}_1$ *and* $\mathbf{d}_2$ *are the wavelet coefficients at scales 1 and 2 corresponding to the horizontal, vertical, and diagonal orientation according to the h, v, and d designations.*

original scene; hence, we regard it to constitute the set of scaling coefficients at the finest of scales available, and which we arbitrarily denote as scale zero.[9] The wavelet analysis of this image up to scale 2 is shown on the right. The representation is standard: on the upper left, the scaling coefficients corresponding to $j = 2$ form a coarser representation of the original image; the horizontal, vertical, and diagonal wavelet coefficients d^h_{j,k_1,k_2}, d^v_{j,k_1,k_2}, and d^d_{j,k_1,k_2} at the first and second scales provide the details of the original image $\mathbf{c}_0$ not in $\mathbf{c}_2$.

In Chapter 3, we introduce a new two-dimensional multiscale representation of images which is derived from the 1-d Haar system but which is not separable. The

[9] This interpretation in which each pixel in the picture assumes the value of a scaling coefficient (at scale zero, in the present case) is justified if $c_{0,k_1,k_2} \equiv \langle f, \phi_{0,k_1,k_2} \rangle \simeq f(k_1, k_2)$. It may be shown that under proper dilation of the original function f, most wavelet systems, and certainly, all bounded systems of compact support, satisfy this condition whenever f is Lipschitz, *i.e.*, there exist constants $C < \infty$ and $0 < \alpha \leq 1$ such that $|f(t'_1, t'_2) - f(t_1, t_2)| \leq C \left|(t'_1 - t_1)^2 + (t'_2 - t_2)^2\right|^{\alpha/2}$, for all t_1, t_2, t'_1, and t'_2 [22].

new analysis approach is better suited for the modeling of non-negative 2-d functions as the estimates obtained are always non-negative. In general, this does not hold for estimates based on any other 2-d wavelet system.

2.3 Multiscale Modeling and Attributes

2.3.1 The Components of a Model and their Notation

Whether an event is a naturally occurring phenomenon or the result of a man-made process, it ultimately exists only as a manifestation of re-distribution (or conversion) of energy in space. The flow of energy, or its final state, forms a discernible pattern by which the event may be identified. Typically, we record such patterns as signals that we can later manipulate and study. Thus, these signals convey all the information that we may ever have about the process, and might well be regarded as the process itself for modeling purposes. This is reminiscent to the nature of random variables, which encode underlying random events but are of no essence once the random variables are defined.

Because there are no truly spontaneous events, every process is simply the continuation of some prior process, and so, we partition the succession of events by recording signals at intervals of time. Any two consecutive signals constitute the *cause* and *effect* of the process they encompass.[10] Denote these signals by $\lambda \in \Lambda$ and $x \in X$, respectively, where in general, Λ and X are subspaces of $L^2(\mathbb{R}^N)$ or $l^2(\mathbb{Z}^N)$. For simplicity, however, we restrict subsequent discussion to the case where Λ and X are subspaces of $\mathbb{R}^N$, and to stress this, the signals are shown in bold face to remind us

[10]We note that for some processes the cause and effect parameters or signals may seem interchangeable, when in fact, they correspond to two distinctly different processes. For example, in modeling the behavior of ideal gases one choice would be to have the change of temperature of an isolated volume of gas be the result of changing pressure. Another possibility would be to consider the change of temperature to be the cause of the change of pressure.

that they are vectors.

As indicated earlier, a deterministic model for a process may suffice for many applications, but a probabilistic model is always much more general. In fact, it can be argued that the joint distribution $p(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})$ is the right way to convey all known information about the relation between $\boldsymbol{\lambda}$ and $\mathbf{x}$ and, therefore, about the process itself.

Although it is often convenient to distinguish between a random variable and its realizations, for simplicity, we provide no separate symbols for each. Instead, we rely on the context to make the differentiation. Thus, for example, with E representing the expectation operator, the expression $\mathrm{E}[\mathbf{x}|\boldsymbol{\lambda}]$ necessarily implies that $\mathbf{x}$ is a random variable (or a random vector) (*rv*) while $\boldsymbol{\lambda}$ may be a random parameter or a realization of it.

By $p(\mathbf{x})$ we denote the probability density distribution of the *rv* $\mathbf{x}$ if it may take on an uncountable number of possible values. If the range of $\mathbf{x}$ is countable, then $p(\mathbf{x})$ stands for its probability mass distribution. In general, $p(\mathbf{x}|\boldsymbol{\lambda}) \neq p(\mathbf{y}|\boldsymbol{\lambda})$, as we define the densities solely by their arguments. Only in a few instances will it be necessary to be more explicit and write, for example, $p_{\mathbf{x}}(\mathbf{y})$ to mean the density of $\mathbf{x}$ evaluated at $\mathbf{x} = \mathbf{y}$.

Our motivation for modeling events primordially comes from the desire to establish the probable causes of the observed phenomena. That is, given the outcome $\mathbf{x}$ we would like to estimate $\boldsymbol{\lambda}$. This is often an important task for if we know $\boldsymbol{\lambda}$ we may predict or estimate the outcome of some other process dependent on it as well; alternatively, we may be interested in whatever caused the outcome $\boldsymbol{\lambda}$ in the first place. For example, in nuclear medicine, the immediate problem consists of establishing the distribution of radioactive material within the patient given a noisy image of photon counts. This knowledge later aids the physician in diagnosing the state of the patient.

In view of this, we regard only models in which $\boldsymbol{\lambda}$ represents the statistical mean

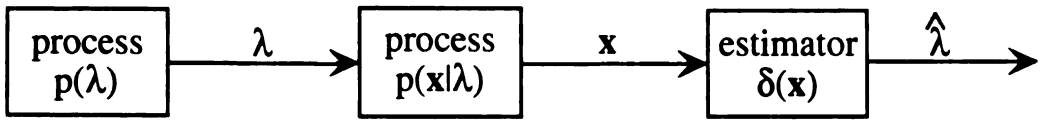


Figure 2.7. *Components of a Model of a Process. The overall process is modeled as a doubly stochastic process in which* $\boldsymbol{\lambda}$ *is the realization of an unknown process, but which is modeled by a prior distribution* $p(\boldsymbol{\lambda})$*. The output* $\mathbf{x}$*, in turn, is modeled as a realization of the process obeying the likelihood* $p(\mathbf{x}|\boldsymbol{\lambda})$*. The purpose of the model is to facilitate the construction of an estimator* δ *for* $\boldsymbol{\lambda}$ *based on the observation* $\mathbf{x}$*, but the estimator is not a component of the model.*

for the *rv* $\mathbf{x}|\boldsymbol{\lambda}$. That is, $\mathbf{x}$ represents a "noisy" realization of $\boldsymbol{\lambda}$ governed by the likelihood $p(\mathbf{x}|\boldsymbol{\lambda})$, which is parameterized by its mean $\boldsymbol{\lambda}$. Furthermore, all derivations in this chapter are made under the assumption that $p(\mathbf{x}|\boldsymbol{\lambda})$ and $p(\boldsymbol{\lambda})$ are absolutely continuous; however, most remarks and conclusions are easily extended to discrete models as well.

A pictorial view of the components of a model are shown in Figure 2.7. In this figure, δ represents an estimator for $\boldsymbol{\lambda}$, the estimate of which is denoted by $\widehat{\boldsymbol{\lambda}}$. The estimator is not an element of the model per se; however, since the objective of the model is to facilitate the construction of the best estimator possible, the chosen model depends on the desired estimator. Bayesian-based estimators can be shown to be optimal over all other types of estimators under most reasonable criteria as long as a suitable prior distribution $p(\boldsymbol{\lambda})$ is available [23]. This is often not the case, however, and constructing one is a difficult problem in general. One very important contribution of this dissertation is, in fact, showing how to formulate practical prior densities for general Poisson processes as well as for photon-limited tomographic processes.

In Section 2.4 we aim to show that under the criteria set forth below, the multiscale framework is the right framework for formulating models of processes suitable for Bayesian inferences.

2.3.2 Anomy, Accuracy, Precision, and Resolution Power

Two important concepts widely used in estimation theory for purpose of evaluating the degree of an estimator's goodness are *unbiasedness* and *minimum variance.* An estimator $\delta(\mathbf{x})$ of $\boldsymbol{\lambda}$ is said to be *unbiased* if $\mathrm{E}[\delta(\mathbf{x})|\boldsymbol{\lambda}] = \boldsymbol{\lambda}$ for all $\boldsymbol{\lambda} \in \Lambda$; and it is said to be *uniform minimum variance unbiased* (UMVU) if $\mathrm{var}\,\delta(\mathbf{x}) \leq \mathrm{var}\,\delta'(\mathbf{x})$ for all $\boldsymbol{\lambda} \in \Lambda$, where $\delta'(\mathbf{x})$ is any other unbiased estimator of $\boldsymbol{\lambda}$, and $\mathrm{var}\,\delta(\mathbf{x}) \equiv \mathrm{E}[(\delta(\mathbf{x}) - \boldsymbol{\lambda})^2|\boldsymbol{\lambda}]$ [18].[11]

Thus, a biased estimator is simply one that on average incurs an error in its estimation, *i.e.*, one which is expected to systematically depart from the truth. On the other hand, a UMVU estimator is one which, in addition to estimating correctly on average, the expected departure from the truth (in a square-error sense) is minimum among all estimators.

Clearly, while biasedness and variance are intuitive and useful concepts, they represent only first and second order measures of the quality of estimators and cannot, therefore, give a complete assessment. Other measures exist for this purpose as well (*e.g.*, equivariance, risk, *etc.*), however, they also give only a partial picture of the merits of estimators. In order to remedy this problem, we introduce four new information-theoretic concepts: *anomy*, *accuracy*, *precision*, and *resolution power.*

Since our interest here is to qualify the degree of goodness of models as well as that of estimators, we take the following unifying and general approach. First, we take the view that models exist independently of any estimator, and that estimators are meaningful only when associated with a model. Second, we only evaluate models per se, but an estimator may be assessed in relation to a model by *augmenting* the model with the estimator and then evaluating the *augmented* model. And third, in

[11] The tests for unbiasedness and UMVU are most often written as $\mathrm{E}[\delta(\mathbf{x})] = \boldsymbol{\lambda}$ and $\mathrm{E}[(\delta(\mathbf{x})-\boldsymbol{\lambda})^2] \leq \mathrm{E}[(\delta'(\mathbf{x}) - \boldsymbol{\lambda})^2]$ in books of statistics [18, 24, 25]; but knowledge of $\boldsymbol{\lambda}$ is implicit in these tests since $\mathbf{x}$ is assumed to be the result of a unique non-random $\boldsymbol{\lambda}$. Here, we prefer to be explicit as $\boldsymbol{\lambda}$ is a random variable itself, and write the conditional dependencies in the tests.

order to unify this view with traditional stands where the various measures apply to estimators rather than models, when necessary we consider the identity estimator, *i.e.*, $\delta(\mathbf{x}) = \mathbf{x}$, in conjunction with a model and apply the various criteria to that estimator. The end result is that whatever is concluded about the estimator applies to the model itself. In this manner, for example, we could talk about the bias of a model by considering the value of $\mathrm{E}[\delta(\mathbf{x})|\boldsymbol{\lambda}] = \mathrm{E}[\mathbf{x}|\boldsymbol{\lambda}]$. For the restricted class of models we are considering, this equals $\boldsymbol{\lambda}$; therefore, this class of models are unbiased. In a similar manner, when referring to the new concepts to be introduced below, they will apply to the model to which we have associated the identity estimator.

The simplicity and convenience of the proposed view of the relationship between models and estimators will become apparent later in the section.

Precision

Let $I(X;\Lambda)$[12] denote the mutual information between the ensembles X and Λ of all possible outcomes $\mathbf{x}$ and $\boldsymbol{\lambda}$, respectively. That is[13]

$$I(X;\Lambda) \equiv \int p(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) \log \frac{p(\mathbf{x}|\boldsymbol{\lambda})}{p(\mathbf{x})}\, d\mathbf{x}\, d\boldsymbol{\lambda}. \tag{2.20}$$

As is well known, $I(X;\Lambda)$ represents the information that, on average, the outcome $\mathbf{x}$ conveys about the *rv* $\boldsymbol{\lambda}$—or that the input $\boldsymbol{\lambda}$ conveys about $\mathbf{x}$—measured in number of bits, nats, or decimal digits depending on whether the log function is base 2, e, or 10, respectively. When the input and output (cause and effect) *rv*'s are scalar, for example, a realization x determines the input λ to within $I(X;\Lambda)$ bits of *precision*, on average. If this quantity were infinite, we would know with all certainty the precise

[12]For convenience, we sometimes simply write $I(\mathbf{x};\boldsymbol{\lambda})$; but we must remember this always represents an averaging process over the entire sample spaces X and Λ.

[13]The mutual information can equivalently be written as $\int p(\boldsymbol{\lambda}|\mathbf{x})p(\mathbf{x}) \log \frac{p(\boldsymbol{\lambda}|\mathbf{x})}{p(\boldsymbol{\lambda})}\, d\mathbf{x}\, d\boldsymbol{\lambda}$, or $\int p(\boldsymbol{\lambda}, \mathbf{x}) \log \frac{p(\boldsymbol{\lambda},\mathbf{x})}{p(\boldsymbol{\lambda})p(\mathbf{x})}\, d\mathbf{x}\, d\boldsymbol{\lambda}$, *etc.*

value of λ that caused x, and we would say that λ is known with infinite precision. Clearly, this occurs when $p(x|\lambda) = \delta_D(x - \lambda)$, *i.e.*, a Dirac delta function. Thus, it appears only natural to define the *precision* $\mathcal{P}$ of a model to be the average mutual information between the cause and effect signals:

$$\mathcal{P} \equiv I(X;\Lambda). \tag{2.21}$$

The chosen terminology is compatible with the common experience. For example, the term precision is formally used among the instrumentation community to indicate the degree of repeatability, over all possible readings, that a measurement can be made [26]. That is, it represents the amount of information in number of *significant* digits that the reading (the *rv* $\mathbf{x}$) conveys about the true state of events (the realization $\boldsymbol{\lambda}$). And this is exactly what $I(X;\Lambda)$ represents: the net information (as an average of positive and negative information) over all possible realization pairs $(\mathbf{x}, \boldsymbol{\lambda})$ [27].

An equivalent and insightful definition for precision is obtained by interpreting it as the entropy of the input $\boldsymbol{\lambda}$ when quantized to the number of bits determined by $I(X;\Lambda)$, averaged over all possible uniform quantization schemes.[14] The proof of the equivalence of the two definitions is given in Appendix A.3, where notation introduced later in this section is used.

The range of $\mathcal{P}$ is clearly the non-negative real numbers. A precision of 0 indicates that $\mathbf{x}$ and $\boldsymbol{\lambda}$ are independent and, therefore, the outcome of one conveys no information about the outcome of the other. For discrete ensembles X and Λ, it is sometimes convenient to talk of the *relative* precision $\mathcal{P}_\mathrm{r}$ of a model which can only

[14]Often we think of quantization of a real variable x, for example, as the mapping $y = \frac{1}{10^n}\lfloor 10^n x \rfloor$, where $\lfloor\cdot\rfloor$ is the integer part function, and where n determines the number of decimal digits of quantization: $n = \infty \Rightarrow$ no quantization, $n = 0 \Rightarrow$ quantization to nearest integer no greater than x, *etc.* It is possible to define, however, a more general quantization operation as follows: for any real $s > 0$ and $r \geq 0$, $y = \frac{1}{s}\lfloor s(x - r) \rfloor + r$.

take values between 0 and 1. A useful definition is given by

$$\mathcal{P}_{\mathrm{r}} \equiv I(X;\Lambda)/H(\Lambda),$$

where $H(\Lambda)$, or simply $H(\boldsymbol{\lambda})$, is the entropy of $\boldsymbol{\lambda}$, *i.e.*, $-\int p(\boldsymbol{\lambda}) \log p(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}$. Clearly, $\mathcal{P}_{\mathrm{r}}$ may be expressed as a percentage as well. For the rest of this chapter, however, only ensembles with absolutely continuous densities will be explicitly considered.

One interesting result stemming from the adopted definition of precision is the fact that the precision $\mathcal{P}$ of any given model is always equal to or greater than the precision $\mathcal{P}_{\boldsymbol{\delta}}$ of any *augmented* model. That is, if the original model is given by the pair $p(\mathbf{x}|\boldsymbol{\lambda})$ and $p(\boldsymbol{\lambda})$, then the augmented model is given by the pair $p(\boldsymbol{\delta}|\boldsymbol{\lambda}) = p(\delta(\mathbf{x})|\boldsymbol{\lambda})$ and $p(\boldsymbol{\lambda})$. This is a consequence of the Data Processing Inequality theorem well known in information theory, and which establishes that $I(X;\Lambda) \geq I(\delta(X);\Lambda)$, for any transformation δ [27]. Thus, we have

$$\mathcal{P}_{\boldsymbol{\delta}} \leq \mathcal{P}. \tag{2.22}$$

As stated earlier, the model $p(\mathbf{x}|\boldsymbol{\lambda}) = \delta_D(\mathbf{x}-\boldsymbol{\lambda})$ is associated with an infinite precision.

Inequality (2.22) is reminiscent to the Cramér-Rao inequality, which states that the mean squared error of any unbiased estimator $\delta(\mathbf{x})$ of the parameter $\boldsymbol{\lambda}$ is lower bounded by the reciprocal of the Fisher information $J(\boldsymbol{\lambda})$ of $\boldsymbol{\lambda}$. Writing this in terms of the reciprocal of the variance, the inequality becomes

$$\frac{1}{\mathrm{var}(\boldsymbol{\delta})} \leq J(\boldsymbol{\lambda}).$$

$\mathrm{var}^{-1}(\boldsymbol{\delta})$ may be regarded as a second-order measure of precision, but it is clear that the inequality (2.22) is stronger for it involves the entire model in all its moments.

Anomy and Accuracy

While precision tells the average amount of information between a model's input and output, it does not give any indication about the utility of the information that either $\mathbf{x}$ conveys about $\boldsymbol{\lambda}$, or that $\boldsymbol{\lambda}$ conveys about $\mathbf{x}$. For instance, it is possible to construct a model in which the output gives much information about its cause (*i.e.*, a very precise model), but which is not all that useful as it is not all that *accurate.* This would occur, for example, in a model characterized by a multimodal density $p(\mathbf{x}|\boldsymbol{\lambda})$ with very well defined ("sharp") modes. Figure 2.8 illustrates one such density when $\boldsymbol{\lambda}$ and $\mathbf{x}$ are scalars.

Assuming a fairly uniform $p(\lambda)$ and that the scenario depicted in Figure 2.8 is representative of the model in general, the integral (2.20) takes on a high value, because the regions on which $\log \frac{p(x|\lambda)}{p(x)}$ becomes large and positive are more likely than those regions on which it assumes negative values. Thus, the model is highly precise.

On average, an outcome x places the value of the input λ to "within" $\mathcal{P}$ bits from its true value. However, since the real line may be partitioned in an infinite number of ways with any given finite number of bits,[15] the model does not necessarily specify a region which includes the input λ, in which case the accuracy of the model is poor. In the example, the $\mathcal{P}$ bits of information conveyed by the realization x places λ to within one of the five regions where $p(x|\lambda)$ is greater than $p(x)$, *i.e.*, within the neighborhoods of the modes, none of which include the input λ. Therefore, the model, while precise, is inaccurate.

This example illustrates a situation where the realizations $\{x_i\}$ are, on average, far removed from their input λ despite the fact that $\frac{1}{n}\sum_1^n x_i \rightarrow \lambda$ as n increases

[15]For example, one bit of information may partition the real line into the positive and negative halves—the sign bit—or may discriminate between numbers with a fractional part in the $[0, .5)$ or $[.5, 1)$ intervals—first bit to the right of the binary point—and so on.

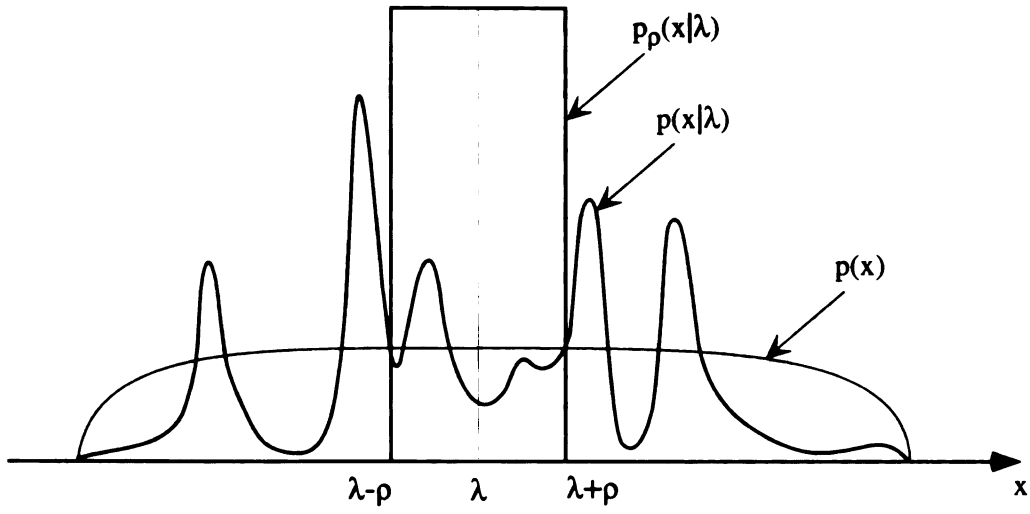


Figure 2.8. *The various probability densities defining* accuracy. $p(\mathbf{x}|\boldsymbol{\lambda})$ *and* $p(\mathbf{x})$ *are pdfs associated with the given model, the precision of which is* $\mathcal{P}$. $p_\rho(\mathbf{x}|\boldsymbol{\lambda})$ *and* $p_\rho(\mathbf{x}) \equiv \int p_\rho(\mathbf{x}|\boldsymbol{\lambda})\, p(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}$ *(not shown) are auxiliary pdfs with the characteristic of representing the same precision* $\mathcal{P}$, *while* $p_\rho(\mathbf{x}|\boldsymbol{\lambda})$ *is the pdf with the smallest support about* $\boldsymbol{\lambda}$.

without bound—Khinchine's Strong Law of Large Numbers [28].

It is important to note that the degree of accuracy of a model is not an indication of its truthfulness. A model, *i.e.*, the joint density $p(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})$, is always assumed correct given the available information. This is not to say that models of varying degrees of accuracy cannot be constructed for the same process; the alternate models may differ in precision as well. Although for particular input-output pairs of realizations a model may be highly inaccurate, on average a (correct) model will possess some degree of accuracy, and its precision bits of information will be useful in that same measure.

If in the example of Figure 2.8, $p(x|\lambda)$ had been unimodal with a very sharp mode at $x = \lambda$, the model would have been highly accurate, and all information given by x would serve to identify the location of λ within its neighborhood. This goes to illustrate that accuracy is some measure of the spread of $p(\mathbf{x}|\boldsymbol{\lambda})$ with respect to the *most accurate* (MA) distribution, which we define to be one that conveys the same

amount of information—one that has the same precision—but whose information most effectively identifies the neighborhood of each $\boldsymbol{\lambda}$ responsible for each outcome $\mathbf{x}$. This requirement brings to mind UMVU models, but as was argued earlier, unbiasedness and minimum variance are only first- and second-order measures of the desired property in the model.

A most general measure of the "spread" of a distribution is given by its entropy. So, we reason that the accuracy $\mathcal{A}$ of a model must be related to the difference $H(X|\Lambda) - H_\rho(X|\Lambda)$ of conditional entropies corresponding to the actual conditional density $p(\mathbf{x}|\boldsymbol{\lambda})$ and to the MA distribution, which we denote by $p_\rho(\mathbf{x}|\boldsymbol{\lambda})$.[16] Here, $H(X|\Lambda) \equiv -\int p(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) \log p(\mathbf{x}|\boldsymbol{\lambda})\, d\boldsymbol{\lambda}\, d\mathbf{x}$, and $H_\rho(X|\Lambda) \equiv -\int p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) \log p_\rho(\mathbf{x}|\boldsymbol{\lambda})\, d\boldsymbol{\lambda}\, d\mathbf{x}$.

The MA distribution $p_\rho(x|\lambda)$ for the particular input λ in the above example is shown in Figure 2.8. In general, $p_\rho(\mathbf{x}|\boldsymbol{\lambda})$ is defined to be a uniform density centered at $\mathbf{x} = \boldsymbol{\lambda}$ and of support the N-dimensional cube $C(\boldsymbol{\lambda}; \rho)$ of sides 2ρ. We denote it by $\mathcal{U}_\rho(\mathbf{x} - \boldsymbol{\lambda})$. This choice is motivated by the fact that of all the probability densities of compact support with a given entropy, the uniform has the smallest of supports, and thus, is most concentrated around its center [29].

To be meaningful, the MA distribution must convey the same precision as is conveyed by the actual model. Denoting the average mutual information between a process' input and output under its MA model by $I_\rho(X;\Lambda)$, *i.e.*,

$$I_\rho(X;\Lambda) \equiv \int p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) \log \frac{p_\rho(\mathbf{x}|\boldsymbol{\lambda})}{p_\rho(\mathbf{x})}\, d\boldsymbol{\lambda}\, d\mathbf{x}, \tag{2.23}$$

where $p_\rho(\mathbf{x}) \equiv \int p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}$, we require that the "radius" ρ of $p_\rho(\mathbf{x}|\boldsymbol{\lambda})$ be such that $I_\rho(X;\Lambda) = I(X;\Lambda)$. Clearly, this requirement makes sense only if $I_\rho(X;\Lambda)$ can

[16] We have departed from the convention of denoting the *rv* as a subscript for its density function. The subscript ρ in $p_\rho(\mathbf{x}|\boldsymbol{\lambda})$ denotes only a parameter.

be shown to be a one-to-one function of ρ, for otherwise, ρ would not be uniquely determined for a given precision of value $I(X;\Lambda)$. In Appendix A.2 we show this to indeed be the case.

Although ρ is not a true radius, thinking of it as such is more descriptive, and helps to better interpret its significance. We shall call it the *equivalent precision radius* or EPR, for short.

Although appealing, measuring the "spread" or "disorder" of the actual model with respect to that of the most accurate model by the difference of their conditional entropies presents a fundamental flaw, which we now illustrate with a simple example. Suppose that $p(\mathbf{x}|\boldsymbol{\lambda})$ also represents a uniform density of support of radius ρ, but which for every $\mathbf{x}$, is centered so that it does not overlap with $p_\rho(\mathbf{x}|\boldsymbol{\lambda})$.[17] This situation results in a difference $H(X|\Lambda) - H_\rho(X|\Lambda) = 0$, leading us to conclude that the actual and MA model have the same accuracy as well as precision. This is clearly not the desired result, as the actual model is completely inaccurate, for its information is misleading for every outcome $\mathbf{x}$.

We now introduce a new measure of the "disorder" of a model which gauges the average degree to which the precision bits fragment the $\mathbb{R}^N$ space in determining the probable input sets for each possible outcome. We call this measure the *anomy* $\mathcal{N}$ of a model, and is defined as

$$\mathcal{N} \equiv D\left(p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) \| p(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})\right), \tag{2.24}$$

where $D(p\|q)$ is the *Kullback Leibler distance* or *relative entropy* between the densities

[17] Strictly, this condition is not realizable if the model is to be correct, however, close approximations may be constructed.

p and q, *i.e.*,

$$D(p(x)\|q(x)) \equiv \int p(x) \log \frac{p(x)}{q(x)}\, dx.$$

More specifically,

$$D\left(p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})\|p(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})\right) = \int p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) \log \frac{p_\rho(\mathbf{x}|\boldsymbol{\lambda})}{p(\mathbf{x}|\boldsymbol{\lambda})}\, d\boldsymbol{\lambda}\, d\mathbf{x}. \tag{2.25}$$

We note that the Kullback Leibler distance is not an ordinary distance in the mathematical sense, for it fails to meet all the conditions of a metric. In particular, it is not the case that $D(p\|q) = D(q\|p)$ for all densities p and q, nor does $D(p\|q) = 0$ imply that $p = q$. Nevertheless, this information-theoretic measure of the "separation" between distributions possess several desirable properties. For example, for all densities p and q, $D(p\|q) \geq 0$ and $D(p\|p) = 0$. Furthermore, because $p_\rho(\mathbf{x}|\boldsymbol{\lambda})$ is uniform conveying the same average information as $p(\mathbf{x}|\boldsymbol{\lambda})$, $D(p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})\|p(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}))$ equals zero only if $p(\mathbf{x}|\boldsymbol{\lambda}) = p_\rho(\mathbf{x}|\boldsymbol{\lambda})$ everywhere except, perhaps, within a set of measure zero. This implies that the anomy of a model can not be zero unless it is the MA model. We prove this assertion in Appendix A.4.

We can establish the similarities and differences between anomy and the difference of entropies discussed earlier by expressing the log function in (2.25) as a difference of log functions: on one hand we have

$$\mathcal{N} = \int p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) \log \frac{1}{p(\mathbf{x}|\boldsymbol{\lambda})}\, d\boldsymbol{\lambda}\, d\mathbf{x} - \int p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) \log \frac{1}{p_\rho(\mathbf{x}|\boldsymbol{\lambda})}\, d\boldsymbol{\lambda}\, d\mathbf{x}, \tag{2.26}$$

while on the other

$$\begin{aligned} &H(\Lambda|X) - H_\rho(\Lambda|X) = \\ &\quad \int p(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) \log \frac{1}{p(\mathbf{x}|\boldsymbol{\lambda})}\, d\boldsymbol{\lambda}\, d\mathbf{x} - \int p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) \log \frac{1}{p_\rho(\mathbf{x}|\boldsymbol{\lambda})}\, d\boldsymbol{\lambda}\, d\mathbf{x}. \end{aligned} \tag{2.27}$$

These two expressions are remarkably similar; however, the averaging process carried out in (2.26) is obtained with respect to the same joint distribution $p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})$, whereas in (2.27), it is not. Thus, the first case represents an average of relative spreads, while the second case represents a difference of average absolute spreads. As a consequence, anomy has the desired property of being a most general measure of separation between actual and the MA model without the shortcomings of the entropy-difference measure.

Accuracy[18] $\mathcal{A}$ of a model may now be defined very naturally as

$$\mathcal{A} \equiv \exp(-\mathcal{N}), \tag{2.28}$$

if the anomy of the model is measured in nats; otherwise, we take $\mathcal{A} \equiv 2^{-\mathcal{N}}$ or $\mathcal{A} \equiv 10^{-\mathcal{N}}$, depending on whether the log base being used is 2 or 10. When the anomy is zero, meaning that the model's average $\mathcal{P}$ bits of information conveyed by $\mathbf{x}$ about $\boldsymbol{\lambda}$ completely determine the input's neighborhood, the accuracy is 1 or, equivalently, 100%, as expected. In the opposite instance, for an infinite anomy—indeed, a complete disorder in the information conveyed—the accuracy is zero.

This extreme case should only occur when the model is erroneous; however, it is easy to construct a model whose anomy is infinite, and yet contains much useful information. In a two-dimensional case, for example, if $p(\mathbf{x}|\boldsymbol{\lambda})$ resembles an upper-half toroid with inner radius greater than $\sqrt{2}\,\rho$ and centered about $\boldsymbol{\lambda}$ so that $p(\mathbf{x}|\boldsymbol{\lambda})$ is zero within the square $C(\boldsymbol{\lambda};\rho)$, $\mathcal{N}$ will be infinity.

To remedy this difficulty inherent in the original definition, we give the following

[18]In classical estimation theory, the inaccuracy of an estimator is measured by the corresponding risk function [18].

more general definition:

$$\mathcal{N} \equiv D(p_\rho(\mathbf{x})||p(\mathbf{x})) \tag{2.29}$$

In Appendix A.5, we demonstrate that this and the definition given in (2.24) are consistent. That is, for "non-pathological" models, the two definitions are completely equivalent. If (2.29) ever becomes infinite, we interpret the model as one of zero accuracy.

Resolution Power

It is natural to assess the "value" of a model by its ability to effectively convey useful information about the process. The amount of information conveyed is measured by its precision, and its utility is measured by the model's accuracy. Therefore, we may measure the quality of the model by the product of precision and accuracy. We call this the *resolution power*, or simply, the *resolution* of the model, and denote it by $\mathcal{R}$:

$$\mathcal{R} \equiv \mathcal{A}\mathcal{P}. \tag{2.30}$$

Essentially, two main mechanisms exist by which one can modify the resolution power of a model, each requiring the introduction of new information into the model. One approach consists of replacing the prior model $p(\boldsymbol{\lambda})$ with one that reflects newly found information about the behavior of $\boldsymbol{\lambda}$. This information may come from a better understanding of the processes that precedes the one of interest, or from simple observations of the signal and a "frequentist" induction of its behavior, for example.

The second mechanism to modifying a model's resolution consists of adding an estimator as shown in Figure 2.7, and incorporating it into the original model. In this case, the estimator function incorporates new knowledge about the relation between

the model's original input and output, knowledge which may be exploited for a better representation of the input.

Strictly speaking, from the assumption that a process is described by its joint density $p(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})$, modifying its model results in a different process than the one intended to be studied. For convenience then, we take the more general view that a model may be modified by updating the prior density $p(\boldsymbol{\lambda})$ and/or by incorporating an estimator $\delta(\mathbf{x})$, and still represent the original model, but augmented with the new information. The key idea here is the invariance of the input and the observable output under consideration; $\mathbf{y}$ is simply a mathematical transformation of $\mathbf{x}$, but not a new observable output per se.

After the model has been modified by either of the above methods, the resulting model may have a lower or higher resolution power than the model from which it originated. To see this, we first consider the prior-modification mechanism, and realize that it is always possible to find a monotone function g such that $F_{\text{new}}(g(\boldsymbol{\lambda})) = F(\boldsymbol{\lambda})$, where F and F_{new} are the cumulative distributions of the original and new priors, respectively [30]. Consequently, the new prior simply constitutes a *rv* transformation $\boldsymbol{\lambda}_{\text{new}} = g(\boldsymbol{\lambda})$ and, by the Data Processing Inequality theorem of information theory, the new model's mutual information can only be less than or equal to the mutual information of the original model [31]. Thus

$$\mathcal{P}_{\text{new}} \leq \mathcal{P}.$$

Updating the prior density of a model, however, can increase, decrease, or leave unaltered the corresponding model's anomy by simply enhancing, reducing, or leaving alone the likelihoods of the sets of $\boldsymbol{\lambda}$ for which the "separation" between $p_\rho(\mathbf{x}|\boldsymbol{\lambda})$ and $p(\mathbf{x}|\boldsymbol{\lambda})$ are greatest.

As for the second mechanism for modifying a model, we saw in page 33 that be-

cause an estimator constitutes a transformation of the output, by the Data Processing Inequality theorem

$$\mathcal{P}_\delta \leq \mathcal{P},$$

where $\mathcal{P}_\delta$ and $\mathcal{P}$ represent the model's precision with and without the estimator. On the other hand, we have that since the estimator in essence gives us a new conditional density $p(\boldsymbol{\delta}|\boldsymbol{\lambda})$, this too can modify the anomy upwards or downwards.

Although these results may appear disconcerting at first, for we may anticipate that new information always increases the resolution power of a model, they are indeed what we should expect. For example, when updating the prior and finding the resolution lowered, it indicates that the leverage given by the model in learning about the input from the output is not as significant as it was before we gained the new information about the input. Likewise, when finding that the inclusion of an estimator only reduces the resolution of the model, it indicates that the newly incorporated information is not constructive, either because it misleads (negative information), or because its gain in accuracy does not offset the loss of precision.

In Section 2.4.2 we will find that resolution power can be employed as a useful figure of merit for comparing models.

2.3.3 Two Illustrative Examples

In this section we present two examples that will help to illustrate the newly introduced concepts of anomy, accuracy, precision, and resolution power. These examples are also intended to give practical results that may be used for further work. In particular, the results of example 1 are used in the next section.

Example 1

In this example we calculate the accuracy, anomy and precision of an N-dimensional Gaussian model. The antecedent or input $\boldsymbol{\lambda}$ is modeled as a vector of jointly Gaussian elements centered around a fixed vector $\boldsymbol{\mu}$ and covariance matrix T. The consequence or output $\mathbf{x}$ is also Gaussian distributed. Its mean is $\boldsymbol{\lambda}$, and its covariance matrix is denoted by K.

The entropy $H(X|\boldsymbol{\lambda})$ is found in [29] to be $\frac{1}{2}\log\{(2\pi e)^N \det K\}$. Since this result is independent of $\boldsymbol{\lambda}$, the conditional entropy $H(X|\Lambda) = E_{\boldsymbol{\lambda}}[H(X|\boldsymbol{\lambda})]$ also takes this form. Likewise, the entropy of the input is $H(\Lambda) = \frac{1}{2}\log\{(2\pi e)^N \det T\}$. Because $\boldsymbol{\lambda}|\mathbf{x}$ is also Gaussian distributed, in this case with covariance matrix $K(K+T)^{-1}T$, $H(\Lambda|X) = \frac{1}{2}\log\{(2\pi e)^N \det K(K+T)^{-1}T\}$. Therefore, from the identity $I(X;\Lambda) = H(X) - H(X|\Lambda)$ we obtain the corresponding model's precision:

$$\mathcal{P} = \frac{1}{2}\log\left\{\frac{\det(K+T)}{\det K}\right\}. \tag{2.31}$$

The units in this case are *nats* because the log function is taken to be base e.

From the definition of anomy,

$$\mathcal{N} = \int p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})\log\frac{1}{p(\mathbf{x}|\boldsymbol{\lambda})}\,d\mathbf{x}\,d\boldsymbol{\lambda} - \int p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})\log\frac{1}{p_\rho(\mathbf{x}|\boldsymbol{\lambda})}\,d\mathbf{x}\,d\boldsymbol{\lambda}.$$

In Appendix A.2 we show that the second integral evaluates to $\log|C|$. Since $p_\rho(\mathbf{x}|\boldsymbol{\lambda}) = \mathcal{U}_\rho(\mathbf{x}-\boldsymbol{\lambda})$ and $p(\mathbf{x}|\boldsymbol{\lambda}) = \frac{1}{\sqrt{(2\pi)^N \det K}}\exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\lambda})^T K^{-1}(\mathbf{x}-\boldsymbol{\lambda}))$, the first

integral can be also expressed as

$$\begin{aligned}
&\frac{-1}{|C|}\int_{\mathbb{R}^N} p(\boldsymbol{\lambda}) \int_{C(\boldsymbol{\lambda};\rho)} \log p(\mathbf{x}|\boldsymbol{\lambda})\, d\mathbf{x}\, d\boldsymbol{\lambda} \\
&\quad = \frac{1}{2}\log\left\{(2\pi e)^N \det K\right\} + \frac{1}{|C|}\int_{\mathbb{R}^N} p(\boldsymbol{\lambda}) \int_{C(\boldsymbol{\lambda};\rho)} \frac{1}{2}\left\{(\mathbf{x}-\boldsymbol{\lambda})^T K^{-1}(\mathbf{x}-\boldsymbol{\lambda})\right\} d\mathbf{x}\, d\boldsymbol{\lambda} \\
&\quad = \frac{1}{2}\log\left\{(2\pi e)^N \det K\right\} + \frac{1}{|C|}\int_{\mathbb{R}^N} p(\boldsymbol{\lambda}) \frac{2^N \rho^{N+2}}{6} \operatorname{tr} K^{-1}\, d\boldsymbol{\lambda} \\
&\quad = \frac{1}{2}\log\left\{(2\pi e)^N \det K\right\} + \frac{\rho^2}{6}\operatorname{tr} K^{-1}
\end{aligned} \tag{2.32}$$

Therefore, the anomy is given by

$$\mathcal{N} = \frac{1}{2}\log\left\{\left(\frac{\pi e}{2\rho^2}\right)^N \det K\right\} + \frac{\rho^2}{6}\operatorname{tr} K^{-1}, \tag{2.33}$$

and consequently, the accuracy is

$$\mathcal{A} = \left(\frac{2\rho^2}{\pi e}\right)^{N/2} \sqrt{\det K^{-1}}\, \exp\left(-\frac{\rho^2}{6}\operatorname{tr} K^{-1}\right). \tag{2.34}$$

In general, it is difficult to find an exact closed form expression for the radius ρ which satisfies the equality $I_\rho(X;\Lambda) = \mathcal{P}$. When the models are not specific enough to apply numerical techniques, we have the option of considering special cases which allow practical simplifications that make the resolution of ρ manageable. The next example is one important case.

Example 2

We consider the same Gaussian model of Example 2, but with the added restriction that the covariance matrices K and T of $p(\mathbf{x}|\boldsymbol{\lambda})$ and $p(\boldsymbol{\lambda})$, respectively, satisfy

$$\max\{k_{i,i} \mid K = \{k_{i,j}\}\} << \min\{t_{i,i} \mid T = \{t_{i,j}\}\}.$$

That is, $p(\boldsymbol{\lambda})$ is a very smooth (slowly varying) function relative to $p(\mathbf{x}|\boldsymbol{\lambda})$.

This model often describes realistic events where $\mathbf{x}$ represents the signal $\boldsymbol{\lambda}$ perturbed by noise whose power is much smaller compared to the signal's own power. In these cases, $p(\mathbf{x}|\boldsymbol{\lambda})$ is the density of the noise, and the diagonal elements $k_{i,i}$ are simply the power in the individual noise elements.

In Appendix A.6 we prove that for sufficiently smooth $p(\boldsymbol{\lambda})$ relative to $p(\mathbf{x}|\boldsymbol{\lambda})$,

$$I_\rho(X;\Lambda) \approx H(\Lambda) - \log|C|, \tag{2.35}$$

which we shall write as an equality. Then, from Example 2,

$$I_\rho(X;\Lambda) = \frac{1}{2}\log\{(2\pi e)^N \det T\} - \log(2\rho)^N.$$

From $I_\rho(X;\Lambda) = \mathcal{P}$, where $\mathcal{P}$ is given in (2.31), the square of the EPR is calculated to be

$$\rho^2 = \frac{\pi e}{2}\left(\frac{\det K \det T}{\det(K+T)}\right)^{1/N}.$$

By applying this expression to (2.33) and (2.34) we obtain the model's anomy and accuracy:

$$\mathcal{N} = \frac{1}{2}\log\left(\frac{\det(K+T)}{\det T}\right) + \frac{\pi e}{12}\left(\frac{\det K \det T}{\det(K+T)}\right)^{1/N} \mathrm{tr}K^{-1}, \tag{2.36}$$

and

$$\mathcal{A} = \sqrt{\frac{\det T}{\det(K+T)}}\exp\left\{-\frac{\pi e}{12}\left(\frac{\det K \det T}{\det(K+T)}\right)^{1/N}\mathrm{tr}K^{-1}\right\}. \tag{2.37}$$

To gain insight into the nature of the accuracy and precision of this model consider the scalar case. With $N = 1$, let $\sigma^2 = \det K$ and $\gamma^2 = \det T$. From the original

assumption we have that $\sigma^2 << \gamma^2$, and (2.31) and (2.37) reduce to[19]

$$\mathcal{P} = \frac{1}{2}\log\left(\frac{\sigma^2+\gamma^2}{\sigma^2}\right) \approx \log\left(\frac{\gamma}{\sigma}\right)$$

and

$$\mathcal{A} = \sqrt{\frac{\gamma^2}{\sigma^2+\gamma^2}}\,\exp\left(-\frac{\pi e}{12\sigma^2}\,\frac{\sigma^2\gamma^2}{\sigma^2+\gamma^2}\right) \approx e^{-\frac{\pi e}{12}}.$$

Thus, for the scalar doubly random Gaussian model where the "noise" power is small compared to the power of the signal, the precision of the model increases logarithmically to the inverse decay of the noise power. This relationship certainly appeals to intuition, especially when written in terms of the signal-to-noise ratio and expressed in bits: $\mathcal{P} \approx .166\,\mathrm{SNR}_{\mathrm{dB}}$ (bits). As an example, we have that in order to achieve a precision of 8 bits, an SNR of 48 dB is required.

Meanwhile, we observe the peculiar behavior of the model's accuracy of being approximately constant for all values of signal and noise powers (within the allowed ranges.) Therefore, the resolution power behaves like the precision within a factor of close to one-half, *i.e.*,

$$\mathcal{R} = \mathcal{A}\mathcal{P} \approx .49\log\left(\frac{\gamma}{\sigma}\right).$$

For example, using this model with an SNR of 48 dB, the outcomes x convey, on average, 4 bits of information about the true value of their corresponding antecedents λ. The other 4 of the 8 bits of precision also discern the inputs, but to within sets too "dispersed" to be constructive. Only with a detailed study of the forms of $p(x|\lambda)$ and $p(\lambda)$ could one possibly exploit the extra information for a particular outcome. In fact, for the models where $p(x|\lambda)$ are not unimodal, on average the extra bits may be outright misleading.

[19]The square of the EPR becomes $\rho^2 = \frac{\pi e}{2}\left(\frac{\sigma^2\gamma^2}{\sigma^2+\gamma^2}\right) \approx \frac{\pi e}{2}\sigma^2$, and $\rho \approx 2\sigma$.

2.4 The Multiscale Modeling and Estimation Advantage

Early in the chapter we asserted that coarse-scale models may be constructed more accurately than their fine-scale counterparts, but that the coarse-scale models are necessarily less precise. Our argument was in no way rigorous, and was based only on an intuitive notion of the concepts of precision and accuracy. Now, we are in a position to formalize this statement using the definitions introduced in Section 2.3.2, and give some insight into the conditions under which it holds.

We shall refer to the ability of trading accuracy for precision (and vice versa) as we move through scales as the *Accuracy/Precision* (A/P) *property* of multiscale models. Likewise, we shall say that a model having this property is an A/P model and satisfies the A/P conditions.

Perhaps the major consequence of the A/P property is in estimation. For models possessing this property, the following estimation approach is possible:
Under the multiscale framework, the estimation process may be started at a coarse scale, where the model is typically highly accurate. Then, one may proceed with the next finer scale, leveraging the new estimate with the accurate estimate from the previous coarser scale. By proceeding in this fashion, moving in each step to the next finer scale, a sequence of increasingly more precise and highly accurate estimates can be obtained.

In Chapters 3 and 4, we employ this estimation approach with great success in the "recovery" of the underlying intensity signals which give rise to Poisson processes.

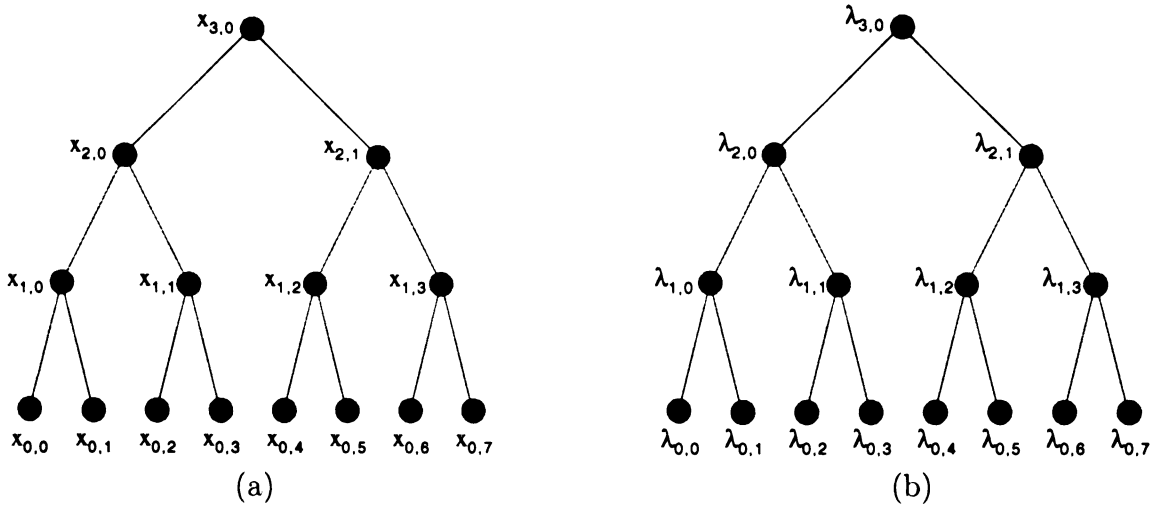


Figure 2.9. *Multiscale representation of a 1-d model of size $N = 8$. (a) Observations $\mathbf{x}_j = (x_{j,0}, \ldots, x_{j,\frac{N}{2^j}-1})$ at the various scales j, and (b) their corresponding intensities $\boldsymbol{\lambda}_j = (\lambda_{j,0}, \ldots, \lambda_{j,\frac{N}{2^j}-1})$. In this example, a two-to-one relation between elements of adjacent scales is illustrated; however, more generally the relations are m-to-one, where $2 \leq m \leq N$.*

2.4.1 A/P Models and their Properties

A model for the highest resolution representations $\boldsymbol{\lambda}_0$ and $\mathbf{x}_0$ of the input and output of a process is given by the joint probability density $p(\mathbf{x}_0, \boldsymbol{\lambda}_0) = p(\mathbf{x}_0|\boldsymbol{\lambda}_0)p(\boldsymbol{\lambda}_0)$. Likewise, the model corresponding to the j-scale representation of the same process is given by $p(\mathbf{x}_j, \boldsymbol{\lambda}_j) = p(\mathbf{x}_j|\boldsymbol{\lambda}_j)p(\boldsymbol{\lambda}_j)$. Recall that the subindex j denotes the scale, with higher numbers corresponding to coarser scales (see Figure 2.9.) We denote the anomy, accuracy, and precision associated with this model by $\mathcal{N}_j$, $\mathcal{A}_j$, $\mathcal{P}_j$, respectively.

The estimation process of progressing from coarser to finer scales so that more precise and highly accurate estimates are obtained at each step, derives from the Accuracy/Precision conditions assumably satisfied by some multiscale models. These conditions are summarized below.

The Accuracy/Precision Conditions

j-scale Model		$(j+1)$-scale Model
$p(\mathbf{x}_j\|\boldsymbol{\lambda}_j)p(\boldsymbol{\lambda}_j)$		$p(\mathbf{x}_{j+1}\|\boldsymbol{\lambda}_{j+1})p(\boldsymbol{\lambda}_{j+1})$
$\mathcal{P}_j$	$\mathcal{P}_j > \mathcal{P}_{j+1}$	$\mathcal{P}_{j+1}$
$\mathcal{A}_j$	$\mathcal{A}_j < \mathcal{A}_{j+1}$	$\mathcal{A}_{j+1}$

Throughout we have confined all discussions to finite length signals and their models, but we believe extensions can easily be made to models of signals with infinite lengths, which appear to represent the simpler case; we do not pursue them here, however.

Whether the A/P conditions are met at any given scale j depends on the form of the model $p(\mathbf{x}_j|\boldsymbol{\lambda}_j)p(\boldsymbol{\lambda}_j)$ and on the particular transformation used to generate the next scale model. Thus, a model $p(\mathbf{x}_0|\boldsymbol{\lambda}_0)p(\boldsymbol{\lambda}_0)$ is an *A/P model* if a set of orthonormal linear transformations $\{\mathcal{W}_j\}$[20] exists such that each of the models $p(\mathbf{x}_j|\boldsymbol{\lambda}_j)p(\boldsymbol{\lambda}_j)$ satisfy the A/P conditions.

The models of interest here all satisfy (**1**) $\mathrm{E}[\mathbf{x}_j|\boldsymbol{\lambda}_j] = \boldsymbol{\lambda}_j$, and (**2**) $p(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_{j+1}) = p(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_j)$ (*i.e.*, $\boldsymbol{\lambda}_{j+1}$ is a sufficient statistics for $\mathbf{x}_{j+1}$). If in addition to these conditions, we impose the mild restriction that for all scales (**3**) $H(\boldsymbol{\lambda}_j|\mathbf{x}_{j+1}) > H(\boldsymbol{\lambda}_j|\mathbf{x}_j)$ (*i.e.*, that the model be *non-trivial*), the inequality $\mathcal{P}_j > \mathcal{P}_{j+1}$ of the A/P conditions

[20]It is understood from our previous discussions that the set of transformations must be such that

$$\begin{pmatrix} \boldsymbol{\lambda}_{j+1} \\ \boldsymbol{\theta}_{j+1} \end{pmatrix} = \mathcal{W}_j \boldsymbol{\lambda}_j$$

where $\boldsymbol{\lambda}_{j+1}$ is a "coarser scale representation" and with half the length of $\boldsymbol{\lambda}_j$. When the transformation is a wavelet transformation, $\boldsymbol{\theta}_{j+1}$ is the vector of wavelet coefficients at scale $j+1$.

is trivially satisfied:

$$\begin{aligned}\mathcal{P}_j = I(X_j;\Lambda_j) &= H(\Lambda_j) - H(\Lambda_j|X_j) \\ &> H(\Lambda_j) - H(\Lambda_j|X_{j+1}) = I(X_{j+1};\Lambda_j) \\ &\geq I(X_{j+1},\Lambda_{j+1}) = \mathcal{P}_{j+1}.\end{aligned}$$

The second inequality in this expression is due to the Data Processing Inequality theorem, because the mapping $\boldsymbol{\lambda}_j \mapsto \boldsymbol{\lambda}_{j+1}$ is non-invertible (see Section 2.2.2).

The idea of non-triviality of a multiscale model is perhaps better understood by considering the idea of a *trivial* model. We define a trivial model to be one that for any valid scale j, the output $\mathbf{x}_j$ conveys no more information about the input $\boldsymbol{\lambda}_j$ than $\mathbf{x}_{j+1}$ does.

This concept is analog to that of finite bandwidth signals, where we regard an oversampled signal s_j to be a trivial representation of its decimated counterpart s_{j+1}. In this instance, the oversampled signal conveys no more information than s_{j+1} does, and so, it is also the case that $H(t_j|s_j) = H(t_j|s_{j+1})$ for any *rv* t_j.

Now we can add to the requirements that a model must meet and show that under the augmented set of conditions the inequality $\mathcal{A}_j < \mathcal{A}_{j+1}$ is also satisfied. Our effort in this direction has only produced conditions which are much less intuitive than the required and sufficient condition $\mathcal{N}_j > \mathcal{N}_{j+1}$. This is because, in addition to the already intuitive nature of anomy as a "distance", this measure is much more encompassing than more traditional ones involving only a finite number of statistical moments. Consequently, it becomes cumbersome trying to characterize general A/P models using conventional statistics.

In Appendix A.7, we develop a sufficient condition for the accuracy inequality. We use it in the following example as a starting point to delineate a general approach to finding a set of transformations $\{\mathcal{W}_j\}$ for a Gaussian model such that the A/P

conditions hold. We use the notation of Example 2 of Section 2.3.3, and those of Appendix A.7. Please refer to those sections as needed, especially Figure A.1. Also, recall that $|C(\boldsymbol{\lambda}_j; \rho_j)| = (2\rho_j)^{N_j}$ denotes the volume of the cube $C(\boldsymbol{\lambda}_j; \rho_j)$ centered at $\boldsymbol{\lambda}_j$, of sides twice the EPR ρ_j, and of dimension $N_j = N_0/2^j$.

Example

Condition (A.12) may be simplified by noting that $\mathrm{A}_{j+1}(\alpha) \geq 0$. So, a new (and more stringent) condition for the accuracy condition to hold is

$$\log\left(\frac{\rho_j}{\rho_{j+1}}\right)^{N_j} < \int_{\mathbb{R}^{N_j}} p(\boldsymbol{\lambda}_j)\,\{\mathrm{A'}_j(p) - \mathrm{A}_j(p)\}\; d\boldsymbol{\lambda}_j. \tag{2.38}$$

Resolution of the integral $-\int p(\boldsymbol{\lambda}_j)\, A_j(p)\, d\boldsymbol{\lambda}_j$ was already obtained in (2.32):

$$-\int_{\mathbb{R}^{N_j}} p(\boldsymbol{\lambda}_j)\, A_j(p)\, d\boldsymbol{\lambda}_j = \frac{1}{2}\log\left\{(2\pi e)^{N_j} \det K_j\right\} + \frac{\rho_j^2}{6}\,\mathrm{tr}K_j^{-1}.$$

And the integral $\int p(\boldsymbol{\lambda}_j)\, A'_j(p)\, d\boldsymbol{\lambda}_j$ can similarly be solved for:

$$\begin{aligned}\int_{\mathbb{R}^{N_j}} p(\boldsymbol{\lambda}_j)\, A'_j(p)\, d\boldsymbol{\lambda}_j &= -\frac{1}{2}\log\left\{(2\pi e)^{N_j} \det \mathcal{W}_j K_j \mathcal{W}_j^T\right\} - \frac{\rho_{j+1}^2}{6}\,\mathrm{tr}\mathcal{W}_j K_j^{-1}\mathcal{W}_j^T \\ &= -\frac{1}{2}\log\left\{(2\pi e)^{N_j} \det K_j\right\} - \frac{\rho_{j+1}^2}{6}\,\mathrm{tr}K_j^{-1}.\end{aligned}$$

Recall that K_j is the covariance matrix of the N_j-dimensional Gaussian pdf $p(\mathbf{x}_j|\boldsymbol{\lambda}_j)$. Substituting these two results in (2.38) and letting $\alpha_j \equiv \frac{\mathrm{tr}K_j^{-1}}{6N_j}$ the condition is

$$\log\left(\frac{\rho_j}{\rho_{j+1}}\right) < \alpha_j\,(\rho_j^2 - \rho_{j+1}^2).$$

Let $f(\rho_{j+1}) \equiv \log\left(\frac{\rho_{j+1}}{\rho_j}\right) - \alpha_j\,(\rho_{j+1}^2 - \rho_j^2)$, then we have as a condition

$$f(\rho_{j+1}) > 0. \tag{2.39}$$

$f(\rho_{j+1})$ is a concave function with one maximum at $\rho_{j+1} = 1/\sqrt{2\alpha_j}$. The set of $\rho_{j+1} > 0$ for which (2.39) is satisfied is always an interval, except when $\rho_j = 1/\sqrt{2\alpha_j}$, where there is no solution at all. If $\rho_j > 1/\sqrt{2\alpha_j}$, the interval of solutions lies immediately to the left of ρ_j; and if $\rho_j < 1/\sqrt{2\alpha_j}$, the interval of solutions lies immediately to the right of ρ_j. Because ρ_j and α_j are already known quantities at the time of search for the transform $\mathcal{W}_j$ that satisfies condition (2.39), we can readily determine whether we should seek a transform that induces a new EPR that is greater than or less than ρ_j.

Let $\zeta_1 \geq \cdots \geq \zeta_{N_j}$ be the eigenvalues of K_j, and $\mathbf{m}_1, \ldots, \mathbf{m}_{N_j}$ the corresponding eigenvectors normalized to 1, *i.e.*, $\|\mathbf{m}_i\| = 1$. Since the axes of the N_j-dimensional ellipsoid $Q_j(\mathbf{z}) \equiv \mathbf{z}K_j^{-1}\mathbf{z}^T = 1$ are $\sqrt{\zeta_1}, \ldots, \sqrt{\zeta_{N_j}}$ [32], the point $1/\sqrt{2\alpha_j}$ is the square root of the harmonic average of the square of the ellipsoid's axes, scaled by $\sqrt{3}$.

Ignoring this last factor, this averaging operation heavily favors the smaller axes, and so, if the number of large axes is not disproportionately greater than the number of small ones, it is convenient to begin the search for the desired transform by choosing $\mathcal{W}_j$ to be such that ρ_{j+1} takes the smallest value possible among the values of the ellipsoid's axes. By starting with the smallest possible value and incrementing it at each step of the algorithm, we are assured not to miss the interval of solutions. So we shall proceed in this fashion of favoring simplicity over optimality of search.

Let us first assume that $\rho_j > 1/\sqrt{2\alpha_j}$. Then, our initial choice is $\mathcal{W}_j^0 = M_j^0 \equiv [\mathbf{m}_1, \ldots, \mathbf{m}_{N_j}]$, which produces the smallest EPR ρ_{j+1}^0 possible because $p(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_{j+1})$ is obtained by integrating $p(\mathbf{x}_j|\boldsymbol{\lambda}_j)$ over the space spanned by the first $N_j/2$ eigenvectors $\mathbf{m}_1, \ldots, \mathbf{m}_{N_j/2}$. Again, these correspond to the direction of slowest change of $p(\mathbf{x}_j|\boldsymbol{\lambda}_j)$ [32]. Consequently, as $p(\mathbf{x}_j|\boldsymbol{\lambda}_j)$ is projected onto the $\mathbf{x}_{j+1}$-$(\mathbf{w}_{j+1} \times \mathbf{x}_{j+1})$ hyper-plane it becomes the density $p(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_{j+1})$ with the smallest possible profile. See Figure 2.10 for a depiction of this at one step ahead in the iteration process.

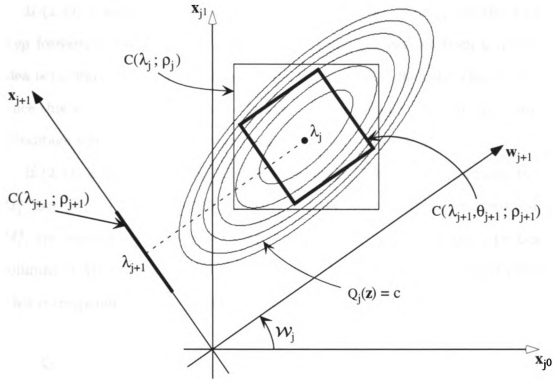


Figure 2.10. *Method for obtaining consecutively increasing EPRs* $\rho_{j+1}^0, \rho_{j+1}^1, \ldots$ *for a Gaussian model. The transform* $\mathcal{W}_j$ *gets updated so that the projections of* $p(\mathbf{x}_j|\boldsymbol{\lambda}_j)$ *onto the* $\mathbf{x}_{j+1}$-$(\mathbf{w}_{j+1} \times \mathbf{x}_{j+1})$ *hyper-plane progress from its narrow side to its broadside. The process is depicted at an intermediate step near the beginning of the algorithm.*

If $\rho_{j+1} = \rho_{j+1}^0$ does not satisfy condition (2.39), we next rotate the axes $\mathbf{w}_{j+1}$-$\boldsymbol{\lambda}_{j+1}$ a small angle in a direction which tends to broaden the projection of the density and make the corresponding EPR bigger. The possibilities of rotation paths increase very rapidly with N_j; in fact, when $N_j > 2$ the number is clearly uncountable. One simple approach, however, is to take $\mathcal{W}_j^1 = M_j^1$, the *first rearrangement* of M_j^0 corresponding to a reordering of the eigenvalues obtained by interchanging the middle two:

$$\zeta_1 \quad \cdots \quad \zeta_{N_j/2-1} \quad \zeta_{N_j/2+1} \quad \zeta_{N_j/2} \quad \zeta_{N_j/2+2} \quad \cdots \quad \zeta_{N_j}.$$

Thus, $M_j^1 \equiv [\mathbf{m}_1, \ \ldots \ , \ \mathbf{m}_{N_j/2-1}, \ \mathbf{m}_{N_j/2+1}, \ \mathbf{m}_{N_j/2}, \ \mathbf{m}_{N_j/2+2}, \ \ldots \ , \ \mathbf{m}_{N_j}]$.

If (2.39) is satisfied with the presently computed $\rho_{j+1} = \rho_{j+1}^0$, we still move one step forward in the algorithm just as if the condition had not been satisfied. The idea is to proceed until (2.39) is satisfied with an EPR sufficiently close to $1/\sqrt{2\alpha_j}$, since this will insure a large spread between $\mathcal{N}_j$ and $\mathcal{N}_{j+1}$,[21] and therefore, a greater advantage when estimating within the multiscale framework.

If (2.39) is once again satisfied with the new EPR $\rho_{j+1} = \rho_{j+1}^1$ induced by $\mathcal{W}_j^1 = M_j^1$, but $|\rho_{j+1}^0 - 1/\sqrt{2\alpha_j}| < |\rho_{j+1}^1 - 1/\sqrt{2\alpha_j}|$, we stop; otherwise, we proceed using M_j^2, the *second rearrangement* of M_j. In this case, we interchange the next two columns of M_j^1 nearest to the center but which have not been exchanged previously. This corresponds to the eigenvalue reordering

$$\zeta_1 \quad \cdots \quad \zeta_{N_j/2-2} \quad \zeta_{N_j/2+2} \quad \zeta_{N_j/2+1} \quad \zeta_{N_j/2} \quad \zeta_{N_j/2-1} \quad \zeta_{N_j/2+3} \quad \cdots \quad \zeta_{N_j}.$$

We proceed in this manner until (2.39) is satisfied and we can no longer reduce the distance between the computed EPR and $1/\sqrt{2\alpha_j}$.

If at the outset $\rho_j < 1/\sqrt{2\alpha_j}$, the approach is the same as before, but in this case we do not test (2.39) while $\rho_{j+1} < \rho_j$, for we know the solution, if one exists, must lie to the right of ρ_j. Once a suitable transform $\mathcal{W}_j$ is found, the search for $\mathcal{W}_{j+1}$ may begin in a similar fashion.

Comments

Several points are worth noting here about the searching approach just delineated. First is the fact that some intermediate transforms may be required between two consecutive trial transforms M_j^i and M_j^{i+1} if ρ_{j+1}^{i+1} "overshoots" the target value. It is easy to envision methods for constructing the sequence of transforms M_j^i which

[21] Since we are attempting to satisfy only a sufficient condition, it is clear that the optimal EPR, one for which the difference $|\mathcal{N}_j - \mathcal{N}_{j+1}|$ is maximum, may be found to be other than $1/\sqrt{2\alpha_j}$.

represent finer rotation changes at each step than those provided here. One possible way would be, for example, to consider all possible permutations of the eigenvalues of K_j, and construct the transforms M_j^i of normalized eigenvectors with the corresponding ordering which generates the monotonic rotational increasing sequence from M_j^0 to $M_j^{N_j!-1} = [\mathbf{m}_{N_j}, \mathbf{m}_{N_j-1}, \ldots, \mathbf{m}_2, \mathbf{m}_1]$. We shall not dwell on this issue as our intention is solely to illustrate the general idea as to how one may find the set of transforms $\{\mathcal{W}_j\}$.

Another important point is that the sequence of transformations obtained in this manner, and which guarantee that $\mathcal{N}_j > \mathcal{N}_{j+1}$ at all valid scales, may be viewed as a new multiscale transform which is statistically motivated. This is in contrast with wavelet based and other existing multiscale transforms that operate in the time or space domain. Clearly, the possibility exists that the two types of multiscale filtering coincide for a class of models, which we would call *scale ergodic* models. For this class of models the A/P conditions may be satisfied through the following mechanism.

From the discussions of Section 2.2, we know that the coarser scale representation $\mathbf{x}_{j+1}$ of $\mathbf{x}_j$ is obtained by a linear transformation that produces elements $\{x_{j+1,i}\}_i$ that are more highly correlated than the elements $\{x_{j,i}\}_i$ of $\mathbf{x}_j$ are. Consequently, for the models of interest here, for which $\mathrm{E}[\mathbf{x}_j|\boldsymbol{\lambda}_j] = \boldsymbol{\lambda}_j$, the density function of $\mathbf{x}_{j+1}|\boldsymbol{\lambda}_{j+1}$ tends to be sharper than the one for $\mathbf{x}_j|\boldsymbol{\lambda}_j$, concentrating a greater probability mass about its mean. This, in turn, tends to reduce the size of EPRs, leading to a better match between the densities $p(\mathbf{x}_j|\boldsymbol{\lambda}_j)$ and $p_{\rho_j}(\mathbf{x}_j|\boldsymbol{\lambda}_j)$, and so, increasing the accuracy of the model with scale.

On a more general point, determining at the outset whether a model admits transformations $\mathcal{W}_j$ such that (A.12) is satisfied at each scale is, in general, a difficult task, and we believe that each case, or at least, each family of models, needs to be looked at separately. We defer this effort for future work. However, under the assumption that a model does admit such a sequence of transformations, we can use condition (A.12) to

aid us in their search in a manner very similar to what we have done here—although, this may entail an intensive iterative process. With knowledge of the sequence of transformations, we can generate the sequence of multiscale representations $\{\mathbf{x}_j\}$ of the data $\mathbf{x}_0$, which can then be used to exploit the estimation advantage referred to earlier in the section.

Coarse-Scale-Data Limited Models

Earlier in the chapter we stated that "while modeling of phenomena based on coarser scale information alone may be more accurate, it can only be achieved at the expense of precision." We shall say that a model that possesses this property satisfies the *coarse-scale-data limited model conditions*, or CSDL model conditions, for short, and refer to such models as CSDL models. At that time we also asserted the equivalence of this phenomenon to the fact that coarse-scale models may be constructed more accurately than their fine-scale counterparts, but always at the expense of precision, which is the defining property of A/P models. We relegate the proof of this equivalence to Appendix A.8.

2.4.2 Bayesian Multiscale Models and Estimation

In the past section we have seen that an advantage exists in estimating an underlying intensity within a multiscale framework whenever the model is an A/P model. However, the advantage is only potential, because whether it is realized or not depends not only on the model itself, but also on the particular estimator used at each scale. To illustrate this we take an extreme example.

Suppose the model at hand is indeed an A/P model, and suppose that within the top-down multiscale leveraging estimation framework we obtain at each scale an estimate $\widehat{\boldsymbol{\lambda}}_j$ of $\boldsymbol{\lambda}_j$ by randomly choosing a signal among all those that satisfy $\widehat{\boldsymbol{\lambda}}_{j+1} = \mathcal{W}_j\widehat{\boldsymbol{\lambda}}_j$. Here, $\mathcal{W}_j \in \mathbb{R}^{\frac{N_j}{2}\times N_j}$, and $\widehat{\boldsymbol{\lambda}}_{j+1}$ denotes the previously obtained estimate for

$\boldsymbol{\lambda}_{j+1}$, which at the coarsest scale J considered we let $\widehat{\boldsymbol{\lambda}}_J = \mathbf{x}_J$. Clearly, the potential advantage in this case does not, on average, lead to a better estimate of $\boldsymbol{\lambda}_0$ than one would obtain, for example, by simply letting $\widehat{\boldsymbol{\lambda}}_0 = \mathbf{x}_0$. It is not hard to see that under the first estimation scheme one has $\mathcal{P}_0 = 0$ and $\mathcal{N}_0 = 1$, and thus $\mathcal{R}_0 = 0$; while for the second method, $\mathcal{P}_0$ and $\mathcal{N}_0$ are the original values corresponding to the model $p(\mathbf{x}_0|\boldsymbol{\lambda}_0)\,p(\boldsymbol{\lambda}_0)$, which for practical models give $\mathcal{R}_0 > 0$.

This example highlights the nontrivial aspect of how to exploit available information in a multiscale estimation scheme. So, we ask what is the optimal estimator $\delta(\mathbf{x}_j)$ for $\boldsymbol{\lambda}_j$ that most efficiently and systematically exploits knowledge of the estimates $\widehat{\boldsymbol{\lambda}}_{j+1}, \ldots, \widehat{\boldsymbol{\lambda}}_J$? Clearly, the answer must satisfy two conditions: the estimator must have the means to extract all relevant information (patterns) from coarser scales, and it must optimally reduce the estimation error.

It is well known that Bayesian estimators are optimal in the mean-square error sense, and in fact, under various other reasonable criteria as well, whenever a suitable prior density of the antecedent is available [23]. Furthermore, since a Bayesian-based estimator can make explicit use of the posterior distribution $p(\boldsymbol{\lambda}_j|\widehat{\boldsymbol{\lambda}}_{j+1}, \ldots, \widehat{\boldsymbol{\lambda}}_J)$—the most comprehensive statement of $\boldsymbol{\lambda}_j$'s dependency on the prior information—they are the most natural and optimal multiscale estimators of choice. A Bayesian multiscale estimator $\boldsymbol{\delta} \equiv \{\delta(\mathbf{x}_j)\}$[22] can be formulated as follows.

For the desired estimate $\widehat{\boldsymbol{\lambda}}_0 = \delta(\mathbf{x}_0)$ to be optimal, each indiviual estimate[23] $\widehat{\boldsymbol{\lambda}}_j = \delta(\mathbf{x}_j)$ must also be optimal. The optimal estimate at scale j is the posterior

[22]Note that each estimator δ is scale dependent, but we let the argument indicate this. That is, $\delta(\mathbf{x}_j) \neq \delta(\mathbf{x}_{j+l})$ for $l \neq 0$.

[23]When referring to the estimate $\widehat{\boldsymbol{\lambda}}_j = \delta(\mathbf{x}_j)$, it is clear that $\mathbf{x}_j$ is taken to be a realization of the corresponding random variable, which we also denote by $\mathbf{x}_j$—a clear abuse of notation. Here, however, the terms estimate and estimator are used interchangeably in order to reduce some of the repetitive vocabulary.

mean

$$\widehat{\boldsymbol{\lambda}}_j \equiv \mathrm{E}\left[\boldsymbol{\lambda}_j \,\middle|\, \mathbf{x}_j, \widehat{\boldsymbol{\lambda}}_{j+1}, \ldots, \widehat{\boldsymbol{\lambda}}_J\right] = \int \boldsymbol{\lambda}_j \, p(\boldsymbol{\lambda}_j | \mathbf{x}_j, \widehat{\boldsymbol{\lambda}}_{j+1}, \ldots, \widehat{\boldsymbol{\lambda}}_J) \, d\boldsymbol{\lambda}_j. \tag{2.40}$$

A Bayesian approach facilitates the solution of (2.40) by expressing it in terms of the known model distribution of $\mathbf{x}_j$ given $\boldsymbol{\lambda}_j$. Applying Bayes' theorem to (2.40) and using the equalities $p(\mathbf{x}_j|\boldsymbol{\lambda}_j, \widehat{\boldsymbol{\lambda}}_{j+1}, \ldots, \widehat{\boldsymbol{\lambda}}_J) = p(\mathbf{x}_j|\boldsymbol{\lambda}_j)$ and $p(\boldsymbol{\lambda}_j|\widehat{\boldsymbol{\lambda}}_{j+1}, \ldots, \widehat{\boldsymbol{\lambda}}_J) = p(\boldsymbol{\lambda}_j|\widehat{\boldsymbol{\lambda}}_{j+1})$[24] we obtain the desired form

$$\widehat{\boldsymbol{\lambda}}_j = \frac{\int \boldsymbol{\lambda}_j \, p(\mathbf{x}_j|\boldsymbol{\lambda}_j) p(\boldsymbol{\lambda}_j|\widehat{\boldsymbol{\lambda}}_{j+1}) \, d\boldsymbol{\lambda}_j}{\int p(\mathbf{x}_j|\boldsymbol{\lambda}_j) p(\boldsymbol{\lambda}_j|\widehat{\boldsymbol{\lambda}}_{j+1}) \, d\boldsymbol{\lambda}_j}. \tag{2.41}$$

The prior $p(\boldsymbol{\lambda}_j|\widehat{\boldsymbol{\lambda}}_{j+1})$ can be obtained from the distribution $p(\boldsymbol{\lambda}_j|\boldsymbol{\lambda}_{j+1})$ allowing for uncertainty in the deviation of $\widehat{\boldsymbol{\lambda}}_{j+1}$ with respect to $\boldsymbol{\lambda}_{j+1}$. If the data $\mathbf{x}_j$, and therefore, $\mathbf{x}_{j+1}, \ldots, \mathbf{x}_J$, are determined to be reliable (accurate), then $p(\boldsymbol{\lambda}_j|\widehat{\boldsymbol{\lambda}}_{j+1})$ can be taken to be well approximated by $p_{\boldsymbol{\lambda}_j|\boldsymbol{\lambda}_{j+1}}(\boldsymbol{\lambda}_j|\widehat{\boldsymbol{\lambda}}_{j+1})$. Since for A/P models, the accuracy of the data increases with scale, the approximation is increasingly better with scale as well.

This represents an important advantage of the multiscale estimation framework: While Bayesian estimation facilitates the realization of the multiscale leveraging estimation advantage by fully exploiting all available information in the scales, a multiscale model enhances the reliability of the Bayesian estimator by virtue of the more accurate data at those higher scales—the A/P property.

Another benefit of the multiscale description of a Bayesian estimator is that the formulation of realistic priors $\{p(\boldsymbol{\lambda}_j|\boldsymbol{\lambda}_{j+1})\}$ is often much simpler than devising the

[24] Clearly, $p(\boldsymbol{\lambda}_j|\boldsymbol{\lambda}_{j+1}, \widehat{\boldsymbol{\lambda}}_{j+1}, \ldots, \widehat{\boldsymbol{\lambda}}_J) = p(\boldsymbol{\lambda}_j|\boldsymbol{\lambda}_{j+1})$, and assuming the optimality of the estimate $\widehat{\boldsymbol{\lambda}}_{j+1}$, it is also reasonable to assume $p(\boldsymbol{\lambda}_{j+1}|\widehat{\boldsymbol{\lambda}}_{j+1}, \ldots, \widehat{\boldsymbol{\lambda}}_J) = p(\boldsymbol{\lambda}_{j+1}|\widehat{\boldsymbol{\lambda}}_{j+1})$. Therefore, $p(\boldsymbol{\lambda}_j|\widehat{\boldsymbol{\lambda}}_{j+1}, \ldots, \widehat{\boldsymbol{\lambda}}_J) = \int p(\boldsymbol{\lambda}_j|\boldsymbol{\lambda}_{j+1}, \widehat{\boldsymbol{\lambda}}_{j+1}, \ldots, \widehat{\boldsymbol{\lambda}}_J) \, p(\boldsymbol{\lambda}_{j+1}|\widehat{\boldsymbol{\lambda}}_{j+1}, \ldots, \widehat{\boldsymbol{\lambda}}_J) \, d\boldsymbol{\lambda}_{j+1} = \int p(\boldsymbol{\lambda}_j|\boldsymbol{\lambda}_{j+1}) \, p(\boldsymbol{\lambda}_{j+1}|\widehat{\boldsymbol{\lambda}}_{j+1}) \, d\boldsymbol{\lambda}_{j+1} = p(\boldsymbol{\lambda}_j|\widehat{\boldsymbol{\lambda}}_{j+1})$.

prior $p(\boldsymbol{\lambda}_0)$ alone. In the next chapter, we elaborate this point further, and show how to construct a practical prior especially suited for general Poisson processes.

Also, within the multiscale Bayesian approach, it is sometimes possible to formulate the problem such that the inferences can be computed very efficiently. This is particularly true if the prior is chosen to be conjugate to the conditional density of the model [23]. The estimator developed in the next chapter is an excellent example of this.

Resolution Power of Multiscale Bayesian Models

We now consider the precisions and accuracies of a model augmented by a Bayesian estimator outside and within the multiscale framework. We shall refer to the resulting models as the *standard* and *multiscale* augmented models, respectively. With these formulations at hand, we can gain some insight into the mechanics of the multiscale Bayesian approach that lead to its resolution power advantage. For simplicity, we shall assume the estimators to be invertible, which is certainly the case for the Poisson processes addressed in Chapter 3. One immediate consequence of this assumption is that the precision of either augmented model is unchanged by the estimators, and that these precisions corresponding to the standard and multiscale augmented models are the same.

Since our interest ultimately lies in whatever occurs at the finest of scales, we restrict the following derivations and discussions to scale $j = 0$, but all the results apply equally to every other scale. This is important because the benefits of the multiscale approach can only be guaranteed if every intermediate scale enjoys similar benefits. We first consider the standard augmented model.

Let $\boldsymbol{\zeta}_0$ be a Bayesian estimator for $\boldsymbol{\lambda}_0$ without the benefit of any conditioning

from higher scales:

$$\boldsymbol{\zeta}_0 \equiv \mathrm{E}[\boldsymbol{\lambda}_0|\mathbf{x}_0] = \int \boldsymbol{\lambda}_0\, p(\boldsymbol{\lambda}_0) \left(\frac{p(\mathbf{x}_0|\boldsymbol{\lambda}_0)}{p(\mathbf{x}_0)} \right) d\boldsymbol{\lambda}_0. \tag{2.42}$$

We recognize the factor $\frac{p(\mathbf{x}_0|\boldsymbol{\lambda}_0)}{p(\mathbf{x}_0)}$ as the ratio in the mutual information $\log \frac{p(\mathbf{x}_0|\boldsymbol{\lambda}_0)}{p(\mathbf{x}_0)}$ of $\mathbf{x}_0$ and $\boldsymbol{\lambda}_0$ at the particular realizations at which the ratio is evaluated. In general, as we know, this ratio is a measure of the dependence between the two random variables: when it takes values near one, it indicates highly independent entities; when it takes large values, it indicates that the variables are highly dependent; and when the ratio is very small, it indicates near mutual exclusiveness. In our applications we do not encounter this third case as we assume $\mathbf{x}$ to be positively correlated to $\boldsymbol{\lambda}$.

Inspecting (2.42), we see that the estimator produces an average based almost exclusively on the information provided by the prior when it deems that the *rv* $\mathbf{x}_0$ conveys very little information regarding $\boldsymbol{\lambda}_0$. On the contrary, when the dependence between the variables is high, $\frac{p(\mathbf{x}_0|\boldsymbol{\lambda}_0)}{p(\mathbf{x}_0)}$ will de-emphasize those values of $\boldsymbol{\lambda}_0$ which are far from the regions that make the ratio large so that the computed average corresponds to the average of those regions that favor the particular outcome $\mathbf{x}_0$.

This is the general mechanics of a Bayesian estimator, and gives insight into its powerful approach to producing an estimate. Now, the mechanics that makes a multiscale-based Bayesian estimator superior to its standard version can be seen by computing the difference of the anomie corresponding to their augmented models:

$$\begin{aligned} \mathcal{N}_{\zeta_0} - \mathcal{N}_{\delta_0} = & \int p_{\rho_0}(\boldsymbol{\zeta}_0|\boldsymbol{\lambda}_0)\, p(\boldsymbol{\lambda}_0) \log \frac{p_{\rho_0}(\boldsymbol{\zeta}_0|\boldsymbol{\lambda}_0)}{p(\boldsymbol{\zeta}_0|\boldsymbol{\lambda}_0)}\, d\boldsymbol{\zeta}_0\, d\boldsymbol{\lambda}_0 \\ & - \int p_{\rho_0}(\boldsymbol{\delta}_0|\boldsymbol{\lambda}_0)\, p(\boldsymbol{\lambda}_0) \log \frac{p_{\rho_0}(\boldsymbol{\delta}_0|\boldsymbol{\lambda}_0)}{p(\boldsymbol{\delta}_0|\boldsymbol{\lambda}_0)}\, d\boldsymbol{\delta}_0\, d\boldsymbol{\lambda}_0, \end{aligned}$$

where $\boldsymbol{\delta}_0$ denotes the estimator (2.41) for $j = 0$, and $\boldsymbol{\zeta}_0$ is the estimator (2.42). Note that the EPRs associated with both of these estimators is the same. This is

a consequence of the invariance of the precision under the standard and multiscale model augmentation scheme by an invertible estimator. The above difference reduces to

$$\mathcal{N}_{\zeta_0} - \mathcal{N}_{\delta_0} = \int p_{\rho_0}(\boldsymbol{\sigma}_0|\boldsymbol{\lambda}_0)\, p(\boldsymbol{\lambda}_0) \log \frac{p_{\boldsymbol{\delta}_0}(\boldsymbol{\sigma}_0|\boldsymbol{\lambda}_0)}{p_{\boldsymbol{\zeta}_0}(\boldsymbol{\sigma}_0|\boldsymbol{\lambda}_0)}\, d\boldsymbol{\sigma}_0\, d\boldsymbol{\lambda}_0,$$

where $\boldsymbol{\sigma}_0$ is a dummy variable. The ratio of densities may be interpreted in accordance with the definitions of the estimators $\boldsymbol{\delta}_0$ and $\boldsymbol{\zeta}_0$:

$$\frac{p_{\boldsymbol{\delta}_0}(\boldsymbol{\sigma}_0|\boldsymbol{\lambda}_0)}{p_{\boldsymbol{\zeta}_0}(\boldsymbol{\sigma}_0|\boldsymbol{\lambda}_0)} = \frac{p(\mathrm{E}[\boldsymbol{\lambda}_0|\mathbf{x}_0, \boldsymbol{\lambda}_1]|\boldsymbol{\lambda}_0)}{p(\mathrm{E}[\boldsymbol{\lambda}_0|\mathbf{x}_0]|\boldsymbol{\lambda}_0)}.$$

It is intuitive that this ratio is non-increasing as a function of the distance $|\boldsymbol{\sigma}_0 - \boldsymbol{\lambda}_0|$. That is, while the probability of finding the estimate $\mathrm{E}[\boldsymbol{\lambda}_0|\mathbf{x}_0, \boldsymbol{\lambda}_1]$ near the true value $\boldsymbol{\lambda}_0$ is greater than the probability of finding $\mathrm{E}[\boldsymbol{\lambda}_0|\mathbf{x}_0]$ in the same region, the probability of $\mathrm{E}[\boldsymbol{\lambda}_0|\mathbf{x}_0, \boldsymbol{\lambda}_1]$ decreases more rapidly than that of $\mathrm{E}[\boldsymbol{\lambda}_0|\mathbf{x}_0]$ as their difference to $\boldsymbol{\lambda}_0$ increases, ie, $p(\mathrm{E}[\boldsymbol{\lambda}_0|\mathbf{x}_0, \boldsymbol{\lambda}_1]|\boldsymbol{\lambda}_0)$ is more "concentrated" about $\boldsymbol{\lambda}_0$ than $p(\mathrm{E}[\boldsymbol{\lambda}_0|\mathbf{x}_0]|\boldsymbol{\lambda}_0)$ is. Assuming this behavior yields

$$\mathcal{N}_{\zeta_0} - \mathcal{N}_{\delta_0} > \int p(\boldsymbol{\lambda}_0) \lim_{t\to\infty} \frac{1}{(2t)^{N_0}} \int_{C(\boldsymbol{\lambda}_0;t)} \log \frac{p_{\boldsymbol{\delta}_0}(\boldsymbol{\sigma}_0|\boldsymbol{\lambda}_0)}{p_{\boldsymbol{\zeta}_0}(\boldsymbol{\sigma}_0|\boldsymbol{\lambda}_0)}\, d\boldsymbol{\sigma}_0\, d\boldsymbol{\lambda}_0,$$

which in terms of the density of $\mathbf{x}_0|\boldsymbol{\lambda}_0$ may be written as

$$\mathcal{N}_{\zeta_0} - \mathcal{N}_{\delta_0} > \int p(\boldsymbol{\lambda}_0) \lim_{t\to\infty} \frac{1}{(2t)^{N_0}} \int_{C(\boldsymbol{\lambda}_0;t)} \log \frac{p(\mathbf{x}_0|\boldsymbol{\lambda}_0)/|J_{\boldsymbol{\delta}_0}(\mathbf{x}_0)|}{p(\mathbf{x}_0|\boldsymbol{\lambda}_0)/|J_{\boldsymbol{\zeta}_0}(\mathbf{x}_0)|}\, d\mathbf{x}_0\, d\boldsymbol{\lambda}_0,$$

where $J_{\boldsymbol{\delta}_0}(\mathbf{x}_0) \equiv \frac{\partial \boldsymbol{\delta}_0}{\partial \mathbf{x}_0^T}$ and $J_{\boldsymbol{\zeta}_0}(\mathbf{x}_0) \equiv \frac{\partial \boldsymbol{\zeta}_0}{\partial \mathbf{x}_0^T}$ are the Jacobians of $\boldsymbol{\delta}_0$ and $\boldsymbol{\zeta}_0$ with respect to $\mathbf{x}_0$. Thus,

$$\mathcal{N}_{\zeta_0} - \mathcal{N}_{\delta_0} > \lim_{t\to\infty} \frac{1}{(2t)^{N_0}} \int_{C(\mathbf{0};t)} \log \frac{|J_{\boldsymbol{\zeta}_0}(\mathbf{x}_0)|}{|J_{\boldsymbol{\delta}_0}(\mathbf{x}_0)|}\, d\mathbf{x}_0, \tag{2.43}$$

For simplicity, we consider the 2-d case. We let $(\lambda_1, \theta_1)^T = \mathcal{W}\boldsymbol{\lambda}_0$ and $(x_1, w_1)^T = \mathcal{W}\mathbf{x}_0$, where $\mathcal{W}$ is a linear orthonormal transformation. Then, due to the linearity of the expectation operator

$$\frac{|J_{\boldsymbol{\zeta}_0}(\mathbf{x}_0)|}{|J_{\boldsymbol{\delta}_0}(\mathbf{x}_0)|} = \frac{\left|\frac{\partial \mathrm{E}[(\lambda_1,\theta_1)^T|x_1,w_1]}{\partial(x_1,w_1)}\right|}{\left|\frac{\partial \mathrm{E}[(\lambda_1,\theta_1)^T|x_1,w_1,\lambda_1]}{\partial(x_1,w_1)}\right|} = \frac{|\frac{\partial\hat{\lambda}_1(x_1,w_1)}{\partial x_1}\frac{\partial\hat{\theta}_1(x_1,w_1)}{\partial w_1} - \frac{\partial\hat{\lambda}_1(x_1,w_1)}{\partial w_1}\frac{\partial\hat{\theta}_1(x_1,w_1)}{\partial x_1}|}{|\frac{\partial\hat{\lambda}_1(x_1,w_1,\lambda_1)}{\partial x_1}\frac{\partial\hat{\theta}_1(x_1,w_1,\lambda_1)}{\partial w_1} - \frac{\partial\hat{\lambda}_1(x_1,w_1,\lambda_1)}{\partial w_1}\frac{\partial\hat{\theta}_1(x_1,w_1,\lambda_1)}{\partial w_1}|}.$$

Typically, the dependency of the estimate $\widehat{\lambda}_1$ on the crossterm w_1, and the dependency of the estimate $\widehat{\theta}_1$ on the crossterm x_1 are weak compared to the dependencies on x_1 and w_1, respectively. Therefore, the behavior of (2.43) may be inferred from the approximation

$$\frac{|J_{\boldsymbol{\zeta}_0}(\mathbf{x}_0)|}{|J_{\boldsymbol{\delta}_0}(\mathbf{x}_0)|} \approx \frac{|\frac{\partial\hat{\lambda}_1(x_1,w_1)}{\partial x_1}\frac{\partial\hat{\theta}_1(x_1,w_1)}{\partial w_1}|}{|\frac{\partial\hat{\lambda}_1(x_1,w_1,\lambda_1)}{\partial x_1}\frac{\partial\hat{\theta}_1(x_1,w_1,\lambda_1)}{\partial w_1}|}.$$

This ratio is, at least on average, greater than one, for the functional dependency of $\widehat{\lambda}_1$ on λ_1, and that of $\widehat{\theta}_1$ on θ_1, makes the denominator less sensitive to changes of x_1 and w_1. A more rigorous argument can be formulated, but we leave this effort for future work. From (2.43), we conclude that $\mathcal{N}_{\delta_0} < \mathcal{N}_{\zeta_0}$, and so, $\mathcal{A}_{\delta_0} > \mathcal{A}_{\zeta_0}$. Since the precision associated with the standard and multiscale formulation of the Bayesian mode is the same, we conclude that the multiscale-based Bayesian estimator achieves a greater resolution power than the traditional Bayesian estimator: $\mathcal{R}_{\delta_0} > \mathcal{R}_{\zeta_0}$.

Although some of the derivations in this section have not been carried out with all the desired rigour, they lead to plausible results which cast light on the advantage of the multiscale-based Bayesian estimation framework over the traditional Bayesian approach. The main objective of this chapter up to this point has been to promote this approach by establishing, at least in its beginning form, a foundation under which models and estimators alike can be studied under a common all-comprehensive set of criteria that applies equally to multiscale and traditional frameworks. These criteria should not only become useful in comparing various estimation approaches,

but should also become the basis for designing guidelines for new estimators.

The new characterization of models and estimators also reflects our attempt to put forward a clearer view of the interplay between scale (space/frequency) resolution and information resolution (resolution power). Their relation depends on the specific transformations used in creating the various scale models from the original high-resolution model, and is reflected in the accuracies and precisions attained at those scales.

2.5 Other Multiscale Modeling and Estimation Approaches

In this section we briefly review some other important multiscale modeling and estimation approaches. These are the threshold smoothing methods: *Hard and Soft thresholding, Universal, and the SureShrink Method; Cross-Validation Method; and False Discovery Rate.* This list is not all exhaustive, but in conjunction with the multiscale Bayesian approach, they represent the most important classical methods. Their importance derives not only from their wide use in practice, but also from the fact that they represent estimation paradigms from which many other methods have later derived.

2.5.1 Threshold Smoothing Methods

The standard model used to represent the input-output phenomena of a very large class of processes[25] is

$$\mathbf{x} = \boldsymbol{\lambda} + \boldsymbol{\eta}, \tag{2.44}$$

[25]Clearly, this model is not as general as a Bayesian-based model.

where $\mathbf{x}$ is the "noisy" data, $\boldsymbol{\lambda}$ the underlying intensity, and where $\boldsymbol{\eta}$ represents additive noise [33]. Within the thresholding methods, estimation of the intensity from the data occurs in the frequency-space (wavelet) domain, where the intensity of typical real-world signals can be projected onto a relatively small subspace compared to that occupied by the noise's projection. The segregation of the signals' energies is what makes their separation from noise possible. The general approach consists of the following steps:

Wavelet transform the data, reduce the smaller wavelet coefficients to zero according to some thresholding rule and, inverse transform the coefficients to recover an estimate of the intensity.

The samples of signals of interest (the elements of $\boldsymbol{\lambda}$) are typically highly correlated. This induces a high structure among the wavelet coefficients containing significant signal energy. Specifically, they exhibit the properties of *clustering* and *persistence across scales* [34]. Clustering is the property of coefficients tending to take values in the order of those of its neighbors; and persistence across scales means that large/small values of wavelet coefficients tend to propagate across scales. These phenomena will be illustrated in Chapter 3. A consequence of this high structure is the sparseness in the wavelet domain representation of signals, which is to say that only a relatively few coefficients convey the signal's features.

Meanwhile, noise samples (the elements of $\boldsymbol{\eta}$) are often well-modeled as being independent, and therefore, the noise displays a "flat" distribution across the frequency spectrum where the intensity lives. This causes the noise energy to become distributed among the wavelet coefficients of a much larger subspace than that occupied by the signal. Consequently, large wavelet coefficients are associated with signal energy, and small coefficients with noise contribution. Thus, reducing the smaller coefficients to zero effectively removes noise from the intensity, with only a minor smoothing effect on the reconstructed intensity.

The literature on thresholding schemes often assumes the noise to be independent of the intensity, and much of that work also assumes it to be Gaussian distributed [12, 33]. These two points are to be found in high contrast with the methods developed later on in this dissertation. These assumptions greatly simplify the estimation problem because the wavelet transform, being a linear orthogonal operator, sustains the two assumptions over the transformation. These common assumptions are made in the following descriptions.

Hard and Soft Thresholding

Let $\mathcal{W}$ denote the DWT. Applying it to (2.44)

$$\mathbf{w} = \boldsymbol{\theta} + \mathcal{W}\boldsymbol{\eta} \tag{2.45}$$

results. Here, $\mathbf{w} \equiv \mathcal{W}\mathbf{x}$ and $\boldsymbol{\theta} \equiv \mathcal{W}\boldsymbol{\lambda}$, the wavelet coefficients of the data and of the signal (see (2.15)). For any given threshold $\tau > 0$ and an element w_k of $\mathbf{w}$, there are two standard ways of modifying the coefficient. These are the *hard* and *soft* thresholding. Hard thresholding produces a new wavelet coefficient according to the rule

$$\widehat{w}_k = \begin{cases} w_k & \text{if } |w_k| \geq \tau \\ 0 & \text{otherwise.} \end{cases}$$

Soft thresholding uses a rule that modifies every coefficient, even those with strong signal to noise ratios. The thresholding condition is

$$\widehat{w}_k = \begin{cases} w_k - \tau & \text{if } w_k > \tau \\ 0 & \text{if } |w_k| \leq \tau \\ w_k + \tau & \text{if } w_k < -\tau \end{cases}$$

When applied to the denoising of images, soft thresholding is claimed to give more pleasing estimates than the hard thresholding [35]; however, this can be due to over-smoothing effects, that often has this visual appeal but which in fact may represent a degraded estimate under most conventional error measures. The smoothing phenomenon will be illustrated in Chapter 4 within the context of tomographic image reconstruction. For a more in-depth treatment of thresholding techniques, see [10, 12, 35, 36].

Universal Method

This method can take the form of a hard or soft thresholding rule. In the latter case, Donoho & Johnson [35], the developers of the method, call it *VisuShrink*. This is because the method usually oversmooths the noisy signal, which as we noted before, tends to produce visually appealing estimates. A more accurate method is also given in [35], which is a *minimax thresholding* method that is optimal in terms of L^2 risk.

Both universal and minimax based methods are global thresholding approaches because the chosen threshold is applied to all the wavelet coefficients. In practice, however, it is typical to threshold only coefficients at the finest resolution (*i.e.*, scale $j = 0$), since coarser-scale features of the data tend to belong to the original intensity signal.

The distinguishing characteristic of the universal method is the basis on which the threshold is computed. Assuming the noise or error term for each element of $\mathbf{w}$ is i.i.d. normally distributed like $N(0, \sigma^2)$, the elements of $\mathcal{W}\boldsymbol{\eta}$ are also distributed according to $N(0, \sigma^2)$. In this case, the simplest of the universal thresholds is calculated to be $\tau_N = \sqrt{2\sigma^2 \log N}$.

This threshold insures that as N increases, the probability that all noise coefficients get "rooted out" tends to one. Clearly, because the threshold τ_N is derived from asymptotic arguments, the universal method does not perform well for signals

of short lengths [33].

SureShrink Method

In contrast to global thresholding approaches, the SureShrink method is a data-dependant selection procedure. This very popular method, also introduced by Donoho & Johnstone [12], finds thresholdings τ_j at each scale j so that the L^2 risk of the estimator $\widehat{\boldsymbol{\lambda}}$ for $\boldsymbol{\lambda}$ will be small. Using the equality $\mathrm{E}[\boldsymbol{\lambda}^T\boldsymbol{\lambda}] = \mathrm{E}[\boldsymbol{\theta}^T\boldsymbol{\theta}]$, the risk can be expressed in the wavelet domain:

$$R(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) \equiv \mathrm{E}\left[\frac{1}{N}\sum_{k=0}^{N-1}(\widehat{\lambda}_k - \lambda_k)^2\right] \propto \mathrm{E}\left[\sum_{j,k}(\widehat{\theta}_{j,k} - \theta_{j,k})^2\right]. \tag{2.46}$$

The wavelet coefficient estimates $\{\widehat{\theta}_{j,k}\}_k$ are obtained from the data wavelets $\{w_{j,k}\}_k$ by a soft thresholding rule with threshold τ_j.

Using Stein's Unbiased Risk Estimator (SURE) defined as

$$\mathrm{SURE}(\tau_j; \mathbf{w}_j) \equiv N + (\tau_j^2 - 2)\sum_{k=0}^{N-1}\mathrm{I}(w_{j,k} \leq \tau_j) + \sum_{k=0}^{N-1} w_{j,k}^2\,\mathrm{I}(w_{j,k} > \tau_j),$$

where $\mathrm{I}(\cdot)$ stands for the indicator function and $\mathbf{w}_j \equiv (w_{j,k})_k$, an unbiased estimate of the risk of the wavelet coefficient estimates $\{\widehat{\theta}_{j,k}\}_k$ can be obtained. Then, with the set of thresholds $\{\tau_j\}$ such that for each j, $\tau_j = \arg\min_{\tau>0}\mathrm{SURE}(\tau; \mathbf{w}_j)$, we can expect that the risk $R(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\lambda})$ be asymptotically minimized, because due to the Law of Large Numbers, the SURE criterion is asymptotically close to the true risk [33].

Cross-Validation Method

The general method of cross-validation (CV) was initially adapted to wavelet regression by Weyrich & Warhola [37] and Nason [38]. Other important related work have followed since then (see for example [39, 40].) Similarly to the SureShrink method,

the multiscale CV approach aims to minimize the risk (2.46). Here, however, a global threshold τ is chosen such that the average risk corresponding to two different estimates for $\boldsymbol{\lambda}$ is minimized. The first of these estimates, $\widehat{\boldsymbol{\lambda}}^{\text{odd}}$, is obtained by soft thresholding the data wavelets corresponding to the odd samples of $\mathbf{x}$; and the second estimate, $\widehat{\boldsymbol{\lambda}}^{\text{even}}$, is similarly computed from the wavelets associated with the even elements. The threshold used is one which minimizes

$$M(\tau) \equiv \frac{1}{4N}\left\{\sum_i (\widehat{\lambda}_i^{\text{odd}} - x_i^{\text{even}})^2 + \sum_i (\widehat{\lambda}_i^{\text{even}} - x_i^{\text{odd}})^2\right\},$$

where x_i^{even} and x_i^{odd} are interpolated samples of $\mathbf{x}$ required to align the "data" samples to the even and odd estimates.

Essentially, the CV approach seeks to obtain the threshold value that reduces the squared error incurred when predicting half of the data samples with estimates that are based solely on the other half of the data elements. Clearly, due to the required interpolation of the data, the separation between the portion of the data used in estimating and that used in testing—that is, used in computing the associated risk—is not complete in the case of the multiscale approach.

False Discovery Rate

The False Discovery Rate (FDR) approach to computing the required threshold is due to Abramovich & Benjamini [41]. They structure the problem in terms of a multiple hypothesis test. There are $N-1$ null hypotheses $H_0 : \theta_{j,k} = 0$, which are regarded to be highly probable on average. The aim of the approach is to reduce the erroneous possibility that a coefficient satisfying the null hypothesis is included in the reconstruction of the signal estimate, or that a coefficient not satisfying H_0 is included with the wrong sign. Thus, if R is the number of coefficients that are included in the reconstruction (erroneously or not), and Q the number of coefficients incorrectly

included, then Abramovich & Benjamini attempt to include as many coefficients as possible maintaining the expected value of Q/R below a user-specified value.

Comparison between this and the *VisuShrink* method shows that the FDR approach performs better for signals that include some abrupt changes, while the *VisuShrink* method is superior when the function space only includes smooth intensities [41].

CHAPTER 3

Multiscale Modeling and Estimation of Poisson Processes

In this chapter, we present and analyze a new multiscale Bayesian framework for the modeling of Poisson processes. In the previous chapter, we motivated this approach for arbitrary processes by showing that multiscale representation of signals makes possible the utilization of all available information in the signal. These results, however, did not specifically show the way to model any particular process in order to gain the multiscale advantage. There are always many ways of achieving this even within the multiscale framework, but not all lead to simple and practical models. The approach introduced here will be shown to provide a very powerful and natural framework to studying a wide variety of Poisson processes. The new framework makes full use of the Poisson probability model and enables the incorporation of realistic prior information into the estimation process. We will also show how it can be applied to photon-limited imaging.

3.1 Preliminaries

The problem of estimating the intensity $\boldsymbol{\lambda}$ of a general Poisson process from a single observation[1] $\mathbf{c}$ of the process has been studied in great depth. For example, many earlier approaches to Poisson intensity estimation were based on the idea of modeling the variability of the process by Gaussian fluctuations with non-stationary characteristics, *e.g.*, [42, 43]. Recently, simple wavelet-based approaches to this problem make use of the square-root of the counts (a variance stabilizing transformation that makes the data approximately Gaussian) and then apply standard wavelet thresholding techniques for Gaussian noise removal [12]. More sophisticated wavelet-based estimation procedures attempt to deal with the Poisson statistics directly. Kolaczyk has developed a wavelet-based thresholding scheme for the estimation of a special class of Poisson processes termed "burst-like" processes [14]. The burst-like Poisson process is characterized by a homogeneous, low intensity background with spatially isolated bursts of high intensity, and is motivated by problems in astronomical imaging. Nowak and Baraniuk propose a wavelet-based method for the estimation of more general Poisson intensities in [44] using the cross-validation estimator developed in [40]. This method is applied to nuclear medicine image estimation in [45]. Both methods [14, 44] can provide satisfactory results in certain situations. However, neither method adopts a Bayesian perspective, and hence they do not explicitly make use of prior information that may be available. As we noted at the end of last chapter, several wavelet-based Bayesian estimation procedures have been proposed for Gaussian data, *e.g.*, [46, 34, 11, 15, 47], however, such methods are not applicable to the Poisson problem considered here.

[1]For completeness, in Section 3.5.4 we address the case of multiple observations of the same and related processes, but our main thrust throughout will be the single-observation case.

3.2 Notation

To simplify the presentation we work with one-dimensional intensity functions in the interval $[0, 1]$. In a later section, we extend the modeling and estimation approaches to two-dimensional problems. Furthermore, we assume that the intensity function is discretized so that $\boldsymbol{\lambda}$ is represented as a vector of length N with elements $(\lambda_k)_{k=0}^{N-1}$. The counts c_k are the elements of the vector $\mathbf{c}$, also of length N.

We follow the wavelet representation and notation introduced in Section 2.2.2 throughout. Specifically, let $\mathbf{c}_0$ be the data sequence of counts $\mathbf{c}$ of length N, where N is assumed to be a positive integer power of two, and let $c_{0,k}$ be its k^{th} element. As before, the subscript 0 denotes the finest scale (resolution) of analysis. Similarly, let $\boldsymbol{\lambda}_0$ be the finest resolution representation of the intensity sequence $\boldsymbol{\lambda}$, *i.e.*, $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}$, also of length N. Then,

$$\mathbf{c}_0 | \boldsymbol{\lambda}_0 \sim \text{Poisson}(\boldsymbol{\lambda}_0), \tag{3.1}$$

and the objective is to estimate $\boldsymbol{\lambda}_0$ from the observation $\mathbf{c}_0$.

From Figures 2.1-(a) and -(b), and expressions (2.10) and (2.11) the wavelet filter coefficients corresponding to the Haar system are $(h_0, h_1) = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $(g_0, g_1) = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$. For reasons soon to be addressed, it is convenient to work with the *unnormalized* Haar system. The filter coefficients corresponding to this system are simply $(h_0, h_1) = (1, 1)$ and $(g_0, g_1) = (1, -1)$. Therefore, a multiscale analysis of $\mathbf{c}_0$ can be obtained by iterating

$$c_{j,k} = c_{j-1,2k} + c_{j-1,2k+1}, \tag{3.2}$$

$$d_{j,k} = c_{j-1,2k} - c_{j-1,2k+1}, \tag{3.3}$$

for $j = 1, \ldots, J$ and $k = 0, \ldots, N/2^j - 1$, and $J = \log_2(N)$. As before, J denotes the

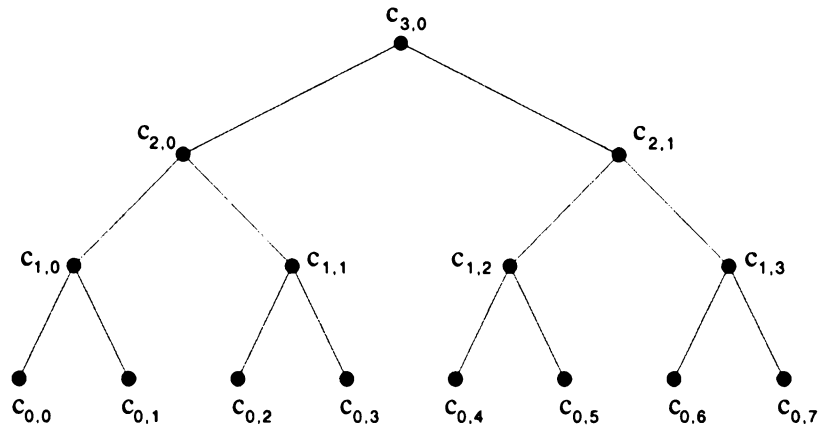


Figure 3.1. *Multiscale scaling coefficients* $\{c_{j,k}\}$. *At the top, we have scaling coefficients at the coarsest resolution. At the bottom, we have the finest resolution, expressed by the data themselves. The connecting segments illustrate the functional dependencies among the various scaling coefficients* $c_{j,k}$ *according to expression* (3.2).

coarsest scale of analysis, and $c_{j,k}$ and $d_{j,k}$ denote the scaling and wavelet coefficients of the data, respectively, at scale j and position (shift) k. The scaling coefficients $\mathbf{c}_j = (c_{j,k})_{k=0}^{N/2^j-1}$ represent a lower resolution representation of the data $\mathbf{c}_{j-1}$. The "detail" information in $\mathbf{c}_{j-1}$, which is absent in $\mathbf{c}_j$, is conveyed by the sequence of wavelet coefficients $\mathbf{d}_j = (d_{j,k})_{k=0}^{N/2^j-1}$. Using (2.9), $\mathbf{c}_{j-1}$ can be perfectly reconstructed from $\mathbf{c}_j$ and $\mathbf{d}_j$. Figure 3.1 is an annotated version of tree-structured representation of Figure 2.4-(a), which shows the functional dependencies among the various scaling coefficients of a sequence $\mathbf{c}$ of length $N = 8$.

Similarly, as for $\mathbf{c}_0$, we define the scaling coefficients $\lambda_{j,k}$ and the wavelet coefficients $\theta_{j,k}$ of the intensity function $\boldsymbol{\lambda}_0$:

$$\lambda_{j,k} = \lambda_{j-1,2k} + \lambda_{j-1,2k+1}, \tag{3.4}$$

$$\theta_{j,k} = \lambda_{j-1,2k} - \lambda_{j-1,2k+1}. \tag{3.5}$$

When the intensity of interest is of a discrete nature, $\boldsymbol{\lambda}_0$ is simply the sequence $\boldsymbol{\lambda}$

itself. If, on the contrary, the intensity signal is a function of a continuous variable $t \in [0,1]$, say $\lambda(t)$, then $\boldsymbol{\lambda}_0$ corresponds to the sequence of scaling coefficients at scale 0. That is, in accordance with the definition of the unnormalized Haar wavelet transform on the interval, $\lambda_{0,k} = \langle \lambda, \phi_{0,k} \rangle = \int_{k/N}^{(k+1)/N} \lambda(t)\, dt$, and more generally,

$$\lambda_{j,k} = \langle \lambda, 2^{j/2} \phi_{j,k} \rangle = \int_{2^j k/N}^{2^j (k+1)/N} \lambda(t)\, dt \tag{3.6}$$

3.3 Why the Unnormalized Haar Transform?

Multiscale analysis based on the unnormalized version of the Haar transform has the unique property that every scaling coefficient is the sum of two finer-scale scaling coefficients, and consequently, due to the reproducing property of the Poisson distribution,[2] every scaling coefficient is Poisson distributed. Furthermore, it is well known that given two Poisson variates, $c_1|\lambda_1 \sim \text{Poisson}(\lambda_1)$ and $c_2|\lambda_2 \sim \text{Poisson}(\lambda_2)$, the conditional distribution of c_1 given λ_1, λ_2, and the sum $c_1 + c_2$ is binomial [48]. This reveals a very simple "parent-child" relationship between the scaling coefficients across scales. In Section 3.5, these facts are crucial in the development of the proposed intensity estimator. Similar attributes (reproducibility and simple parent-child relationship) do not hold for more general multiscale analyses of Poisson processes based on other wavelet systems. This is in marked contrast to the Gaussian case, in which such attributes hold for a wide variety of wavelet analyses (including all orthogonal wavelet systems). In short, multiscale transforms of Poisson processes other than the unnormalized Haar transform are much more difficult to analyze and process. The natural match between the unnormalized Haar transform and the Poisson process is the primary motivation for choosing it.

The use of the unnormalized Haar wavelet transform to carry out the multiscale

[2] $c_i|\lambda_i \sim \text{Poisson}(\lambda_i)$, $c_i|\lambda_i$ independent $\Rightarrow \sum c_i | \sum \lambda_i \sim \text{Poisson}(\sum \lambda_i)$.

analysis of the data has additional benefits springing from the following points. Poisson processes result from counting independent events occurring in disjoint regions of time or space of equal size. In such cases, the unnormalized Haar scaling coefficients correspond exactly to these counts occurring at intervals of sizes varying according to scale. Thus, the scaling and wavelet coefficient have a very natural interpretation according to (3.6) and (3.3). The Haar basis functions also have the property of being completely localized in space. By this we mean that at each scale, scaling functions, as well as wavelets, do not overlap. Therefore, at each scale, scaling coefficients are conditionally independent, that is,

$$p(\mathbf{c}_j|\boldsymbol{\lambda}_j) = \prod_k p(c_{j,k}|\lambda_{j,k}). \tag{3.7}$$

Also, Poisson processes are inherently nonnegative; therefore, a good estimator should always produce intensity estimates that are either positive or zero. An estimator based on the Haar system may be designed with this quality.

3.4 A New Probability Model for Intensity Images

3.4.1 Multiscale Signal Model Framework

To formulate a Bayesian estimator for this problem, we must first propose a prior probability model for the unknown intensity $\boldsymbol{\lambda}$. The observed data $\mathbf{c}$ is regarded as the realization of a Poisson process spawned by the unknown realization $\boldsymbol{\lambda}$ of some random sequence with prior density $p(\boldsymbol{\lambda})$. In last chapter's terms, $\boldsymbol{\lambda}$ and $\mathbf{c}$ are respectively the cause and effect of the process completely determined by $p(\mathbf{c}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})$, where $p(\mathbf{c}|\boldsymbol{\lambda})$ is the Poisson probability mass distribution.

Just as we did in the last chapter for general processes, we can formulate the present model within a multiscale framework, and similarly arrive at the optimal

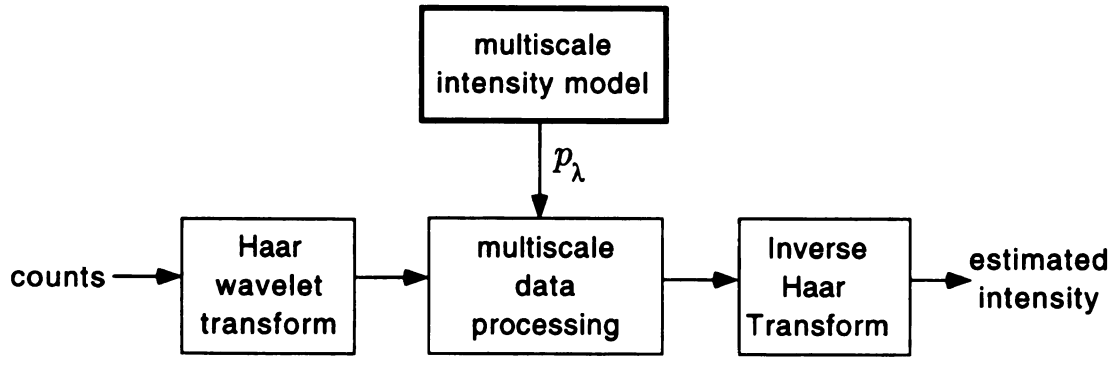


Figure 3.2. *Structure of a Haar-based intensity estimator.*

estimator (see (2.41))

$$\widehat{\boldsymbol{\lambda}}_j = \frac{\int \boldsymbol{\lambda}_j \, p(\mathbf{c}_j|\boldsymbol{\lambda}_j) p(\boldsymbol{\lambda}_j|\widehat{\boldsymbol{\lambda}}_{j+1}) \, d\boldsymbol{\lambda}_j}{\int p(\mathbf{c}_j|\boldsymbol{\lambda}_j) p(\boldsymbol{\lambda}_j|\widehat{\boldsymbol{\lambda}}_{j+1}) \, d\boldsymbol{\lambda}_j}. \tag{3.8}$$

This Bayes estimator poses two interrelated problems. First, the specification of a meaningful and useful prior $p(\boldsymbol{\lambda}_j|\boldsymbol{\lambda}_{j+1})$, which we deem to be an excellent approximate for $p(\boldsymbol{\lambda}_j|\widehat{\boldsymbol{\lambda}}_{j+1})$ (see Section 2.4.2). Second, the numerical computation of the estimator. The role the above quantities play in the estimation process is illustrated in Figure 3.2. The remainder of this section describes a new prior probability model for the Haar scaling and wavelet coefficients of a non-negative intensity that leads to a very simple specification of $p(\boldsymbol{\lambda}_j|\boldsymbol{\lambda}_{j+1})$. In Section 3.5, we derive an efficient algorithm for computing the optimal estimator (3.8).

There are two important reasons for adopting a multiscale approach to this problem:

- Prior models that are mathematically tractable, computationally practical, and empirically supported can be specified very naturally.

- Poisson data is much more reliable (accurate) at coarse scales than at fine reso-

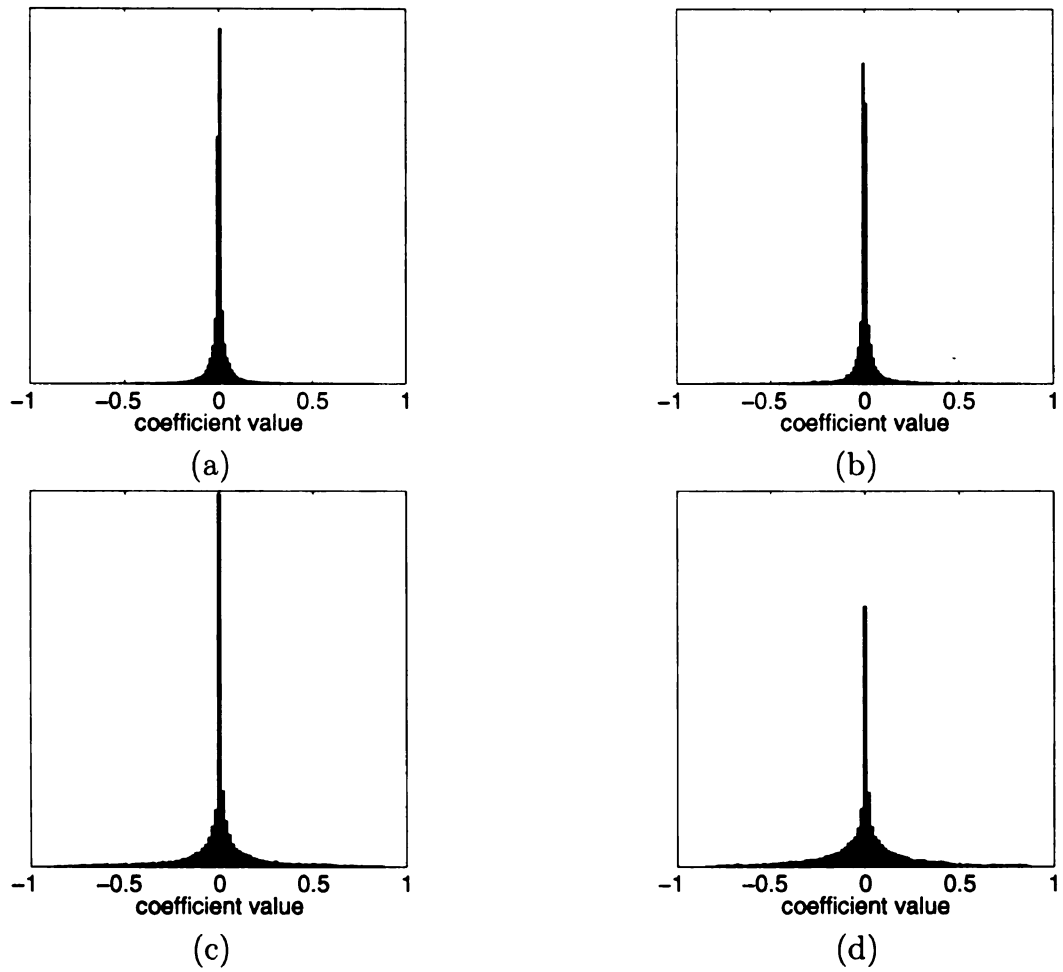


Figure 3.3. *Histogram of perturbation variates ($\delta = \theta/\lambda$) for scale (a) $j = 1$, (b) $j = 2$, (c) $j = 3$, and (d) $j = 4$ of the cameraman image of Figure 3.10(a). The general invariance of the distributions' structure across scales illustrates a self-similar property of real-world image statistics.*

lutions (higher counts $\Rightarrow$ higher signal-to-noise ratio). Therefore, more reliable coarse-scale estimates can be leveraged to improve high resolution estimators.[3]

The first point is due partly to the fact that multiscale decompositions of real-world intensities are often statistically *self-similar*. By this we mean the property that the various scale representations preserve the major features and characteristics

[3]The good match between the Poisson and Gaussian distributions that occurs at high counts suggests a similar positive relation between the Poisson models' anomies and corresponding signal-to-noise ratios as seen to exists in the Gaussian case of Example 3 of Section 2.3.3.

of the original object, except for the usual gradual loss of resolution. In particular, it has been widely recognized that the distribution of the wavelet coefficients of real-world signals tend to be similar at all scales of analysis, and are usually concentrated around the origin and unimodal [34]. The self-similarity captured by the Haar multiscale analysis is illustrated by considering the distributions of the wavelet coefficients at various scales. Figure 3.3 illustrates this phenomena. The histograms in this figure correspond to the wavelet coefficients [4] at scales 1, 2, 3, and 4 for the cameraman image of Figure 2.6. The similarity between these distributions facilitates the specification of Bayesian prior for the intensity in a very natural way.

The second point above motivates an estimation process that evolves from coarse to fine scales (See Chapter 2). It is easily verified that the signal-to-noise ratio (SNR) in a Poisson process increases linearly with the underlying intensity (signal). Thus, according to (3.2), $c_{J,0} = \sum_{k=0}^{N-1} c_{0,k}$, and so the signal-to-noise ratio at scale J is 2^J times as large as that for the average data point $c_{0,k}$. For example, for a 128 by 128 pixel image, this represents a SNR improvement of 42 dB.

3.4.2 Multiscale Multiplicative Innovations Model

We now describe a new Haar-based probability model for the intensity. Let $\lambda_{j,k}$ and $\theta_{j,k}$ denote the random variables corresponding to the j,k-th scaling and wavelet coefficient of the intensity, respectively. At the coarsest scale $j = J$, the single scaling coefficient $\lambda_{J,0}$ has a density with support on $\mathbb{R}^+$. In this work, we choose the gamma density, since it is especially easy to use in conjunction with the Poisson mass function, and because it provides a reasonable mechanism for incorporating prior knowledge of the intensity range. However, as noted earlier, the SNR in the count $c_{J,0}$ is typically

[4] More precisely, here we are plotting the histogram of the ratio of the wavelet coefficient relative to the corresponding scaling coefficient. That is, each wavelet coefficient is divided by the corresponding scaling coefficient at the same scale and position. If the scaling coefficient is zero, the operation maps to zero. The motivation for this ratio will become apparent in the following sections.

very high, and therefore any reasonable prior with support on $\mathbb{R}^+$ will not significantly influence the estimation of $\lambda_{J,0}$.[5]

Next, introduce statistically independent perturbation variables $\{\delta_{j,k}\}$ and model the wavelet coefficients by

$$\theta_{j,k} = \lambda_{j,k}\,\delta_{j,k}. \tag{3.9}$$

Each wavelet coefficient is modeled as an independent perturbation of its corresponding scaling coefficient. Furthermore, the perturbations at all scales and positions are assumed to be mutually independent. Applying recursions (3.4) and (3.5) to these coefficients, we find that $\lambda_{j-1,k} = \frac{1}{2}(\lambda_{j,\lfloor k/2\rfloor} + (-1)^k \theta_{j,\lfloor k/2\rfloor})$, where $\lfloor\cdot\rfloor$ stands for the integer part of the argument.

To gain some insight into this model, consider the random variable $y_{j,k}$ defined by

$$y_{j,k} \equiv \frac{1}{2}\,(1+\delta_{j,k})\,. \tag{3.10}$$

The variable $y_{j,k}$ can be viewed as the *canonical* multiscale parameter for Poisson processes because of the following parent-child relationship. It is well known that given two Poisson variates c_1 and c_2 such that $c_1|\lambda_1 \sim \text{Poisson}(\lambda_1)$ and $c_2|\lambda_2 \sim \text{Poisson}(\lambda_2)$, the conditional distribution of c_1 given the sum $c_1 + c_2$ is binomial with parameter $y = \frac{\lambda_1}{\lambda_1+\lambda_2}$ [48]. In the context of our multiscale analysis, this special property implies a very simple parent-child relationship. Specifically, the conditional distribution of child $c_{j-1,2k}$ given the parent $c_{j,k} = c_{j-1,2k} + c_{j-1,2k+1}$ is binomial with parameter $y_{j,k}$. This relationship demonstrates the fundamental role of $y_{j,k}$ in the multiscale analysis of Poisson processes.

[5]In fact, in practice we often use the estimate $\widehat{\lambda}_{J,0} = c_{J,0}$.

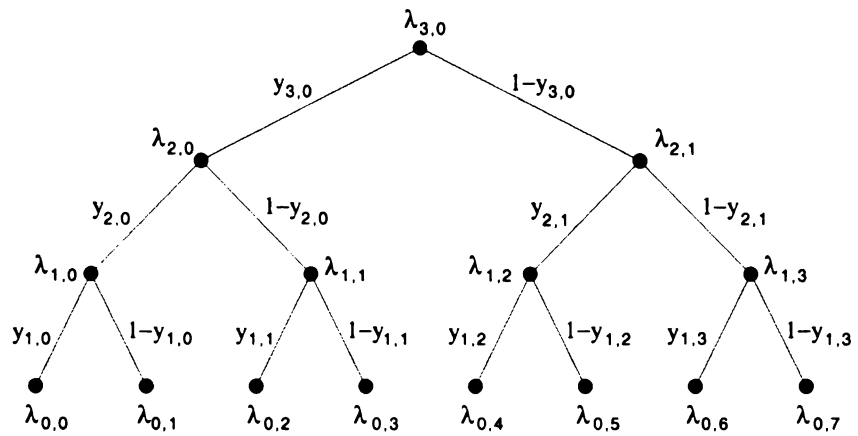


Figure 3.4. *MMI model interpreted as a probabilistic tree. The MMI model can be viewed as a tree-structured probability model in which the intensity $\lambda_{j,k}$ at coarse scale j is refined (split) via the multiplicative innovation $y_{j,k}$ to obtain two new intensities $\lambda_{j-1,2k}$ and $\lambda_{j-1,2k+1}$ at the next finer scale of analysis $j-1$. The innovations variates $\{y_{j,k}\}$ are mutually independent at all scales j and positions k.*

Using $y_{j,k}$ in conjunction with (3.9) we have

$$\lambda_{j-1,2k} = \lambda_{j,k}\, y_{j,k}, \tag{3.11}$$

$$\lambda_{j-1,2k+1} = \lambda_{j,k}\,(1 - y_{j,k}). \tag{3.12}$$

We can interpret these refinements as a multiscale innovations structure, with the innovations $y_{j,k}$ and $1-y_{j,k}$ entering in a multiplicative fashion, in contrast to the more standard additive innovations structure encountered in Gaussian estimation problems [49]. We call this model a *multiscale multiplicative innovations* (MMI) model. The model is graphically depicted in Figure 3.4. The following are some key properties of the MMI model.

So long as the distributions for the perturbations $\{\delta_{j,k}\}$ are chosen to be similar across scales, the MMI model gives rise to self-similar intensity representations, which as discussed in Section 3.4.1, are typical of real-world intensities. Also, here we

only consider temporally homogeneous processes; it would be undesirable if the prior depended on the observation time interval in a complicated manner. Due to the multiplicative innovations structure, only the coarsest scale of the prior is dependent on the observation time interval. Thus, the model is essentially invariant to the length of the observation time. Moreover, in Section 3.5 the MMI model is shown to provide a mathematically tractable match to the Poisson nature of the data which leads to a very simple estimator formulation.

The MMI model is closely related to other models studied in physics and statistics. The MMI model belongs to the class of *cascade models*, which are used in statistical physics for modeling a variety of natural phenomena including turbulence modeling [50] and rainfall distributions [51]. In fact, because MMI model is a type of cascade model, it can be shown that the MMI model is a random multifractal [52]. It is also interesting to note that the MMI model is a special case of a *Polya tree* [53, 54]. In the statistics community, Polya trees are used to model probability distributions, a problem analogous to modeling a non-negative intensity function.

3.4.3 Prior Distribution for Innovations

A prior distribution is determined by the nature of the ensemble of objects to be modeled, and so, the better defined the ensemble is, the more informative the prior, and consequently, the better the model. Towards this end, we emphasize photon-limited images, particularly, nuclear medicine images; however, as noted earlier, a very large class of other real-world images are well characterized by the same features of the prior $p(\delta)$ for the perturbations $\delta_{j,k}$: statistical self-similarity across scale, symmetry about the origin, unimodality, concentration around zero (see Section 3.4.1), and support on the $[-1, 1]$ interval.

This last property is due to the fact that the range of $\theta_{j,k}$ is $[-\lambda_{j,k}, \lambda_{j,k}]$. The rest of the properties are based on the characteristics of observed wavelet coefficients' dis-

tributions resulting from natural signals [34], and which have been exploited in other areas including wavelet-based compression [55]. These properties are also illustrated in the histograms of Figure 3.3. All the properties, however, may equally be inferred *a priori* from a few observations.

The images of interest have the common characteristic of being mostly smooth with a few discontinuities. Typically, these discontinuities form edges that are long enough to span several scales. On average, neither the edges nor the overall intensity variations have any preferred orientation, as dictated by nature. It is the slowly varying intensity typical of images which is responsible for the high concentration of wavelet coefficients around the origin, and it is its non-oriented nature that causes their symmetry. The few outlayers coefficients are due to the discontinuities, which are found nearly in the same numbers in opposite orientations—think about the silhouette of a phase, for example. The statistical self-similarity is due to the span of images feature across scales. For example, large regions of smooth intensities and discontinuities alike are observed unaltered at many scales.

One very general class of probability density functions that possesses the desired characteristics and which reflect the above image characteristics are beta-mixture densities of the form

$$p(\delta) = \sum_{i=1}^{M} p_i \frac{(1-\delta^2)^{s_i-1}}{B(s_i, s_i)\, 2^{2s_i-1}}, \tag{3.13}$$

for $-1 \leq \delta \leq 1$, where B is the Euler beta function, $0 \leq p_i \leq 1$ is the weight of the i-th beta density $\frac{(1-\delta^2)^{s_i-1}}{B(s_i,s_i)\,2^{2s_i-1}}$ with parameter $s_i \geq 1$, and $\sum_{i=1}^{M} p_i = 1$. Figure 3.5 depicts a mixture of three beta densities. A similar method was also recently proposed using a prior based on a mixture of a Dirac impulse and a single beta distribution [56].

Other classes of density functions may also provide the desired characteristics, but

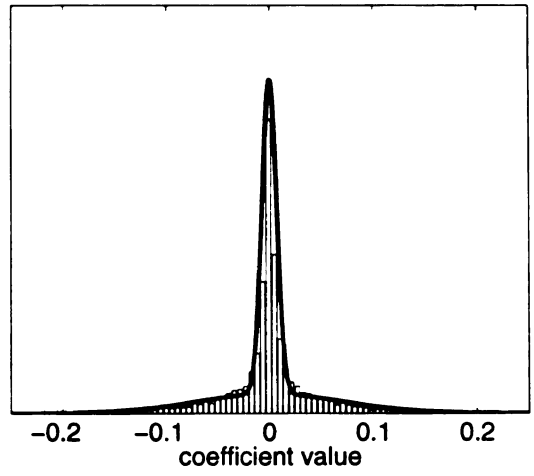


Figure 3.5. *Three component Beta-mixture distribution (solid line) superimposed on the histogram of the perturbation variates ($\delta = \theta/\lambda$) of Figure 2.6 of Section 2.2.2. The beta mixture parameters here are* $s_1 = 1$, $s_2 = 100$, $s_3 = 10000$, $p_1 = 0.001$, $p_2 = 0.400$, *and* $p_3 = 0.599$.

the beta family has a significant computational advantage. As pointed out above, $y_{j,k}$ parameterizes the conditional distribution of the child $c_{j-1,2k}$ given the parent $c_{j,k}$. This conditional distribution is binomial. Hence, from a practical perspective, the use of a prior that is conjugate to the binomial will greatly facilitate computations [23]. It is well known that the beta family is conjugate to the binomial. For this reason, the beta mixture prior described above leads to a very simple, closed-form estimator which is discussed in the next section. However, for the sake of brevity, in our derivation of the optimal estimator, given in Section B.2, we directly compute the posterior means based on a beta mixture prior, without explicitly noting the use of conjugacy.

3.5 Estimation

3.5.1 Bayesian Multiscale Intensity Estimator

We shall focus on the posterior mean estimator, although other estimators (e.g., MAP) may also be considered within our framework. The posterior mean is the optimal Bayes estimate under a quadratic loss function. The posterior mean estimate of an intensity $\lambda_{j,k}$, given all the information available in the data $\mathbf{c}_0$ and the MMI prior model $p_{\boldsymbol{\lambda}}$, is the conditional mean $\widehat{\lambda}_{j,k} \equiv \mathrm{E}[\lambda_{j,k}|\mathbf{c}_0]$. In this section we derive simple closed-form expressions for the posterior mean.

First, based on the analysis in Section B.1 of Appendix B,

$$\widehat{\lambda}_{j,k} = \mathrm{E}[\lambda_{j,k}|\mathbf{c}_j].$$

This implies that a simple coarse-to-fine procedure can be employed in the estimation process.

At the coarsest scale $j = J$, the intensity is represented by a single scaling coefficient $\lambda_{J,0}$. Let us begin by considering the estimation of $\lambda_{J,0}$. We have $\widehat{\lambda}_{J,0} \equiv E[\lambda_{J,0}|\mathbf{c}_0] = E[\lambda_{J,0}|c_{J,0}]$. As argued in Section 3.4, the corresponding count $c_{J,0}$ is itself usually a very good estimate for $\lambda_{J,0}$ provided the total number of counts is sufficiently large. This choice has the added advantage of insuring the preservation of total number of counts, *i.e.*, $\sum_k \widehat{\lambda}_{0,k} = \sum_k c_{0,k}$.

The posterior mean estimate for the wavelet coefficient $\theta_{j,k}$ is given by

$$\begin{aligned} \widehat{\theta}_{j,k} &\equiv \mathrm{E}[\theta_{j,k}|\mathbf{c}_0] \\ &= \mathrm{E}[\lambda_{j,k}|\mathbf{c}_0]\ \mathrm{E}[\delta_{j,k}|\mathbf{c}_0], \end{aligned}$$

where we have used (3.9), and have exploited the independence between $\lambda_{j,k}$ and $\delta_{j,k}$.

Now, we may simply write

$$\widehat{\theta}_{j,k} = \widehat{\lambda}_{j,k}\,\widehat{\delta}_{j,k}, \tag{3.14}$$

with the obvious definition $\widehat{\delta}_{j,k} \equiv \mathrm{E}\left[\delta_{j,k}|\mathbf{c}_0\right]$.

The desired form for $\widehat{\lambda}_{j,k}$ may be obtained using (3.4) and (3.5), and the linearity property of the expectation operator:

$$\begin{aligned}\widehat{\lambda}_{j,k} &= \mathrm{E}\left[\frac{1}{2}\left(\lambda_{j+1,[k/2]} + (-1)^k\,\theta_{j+1,[k/2]}\right)\middle|\,\mathbf{c}_0\right] \\ &= \frac{1}{2}\left(\widehat{\lambda}_{j+1,[k/2]} + (-1)^k\,\widehat{\theta}_{j+1,[k/2]}\right).\end{aligned} \tag{3.15}$$

Here is where we exploit the multiscale framework for estimating the desired intensity. The estimate for $\widehat{\lambda}_{j,k}$ is leveraged by the more robust coarser-scale scaling coefficient's estimate $\widehat{\lambda}_{j+1,[k/2]}$.

In Section B.2 we show that

$$\widehat{\delta}_{j,k} = d_{j,k}\,\frac{\sum_i p_i \frac{B(s_i+c_{j-1,2k},\,s_i+c_{j-1,2k+1})}{B(s_i,s_i)\,(2s_i+c_{j,k})}}{\sum_i p_i \frac{B(s_i+c_{j-1,2k},\,s_i+c_{j-1,2k+1})}{B(s_i,s_i)}}. \tag{3.16}$$

The parameters $\{p_i, s_i\}$ are the defining parameters for the beta mixture model in (3.13). Note that $\widehat{\delta}_{j,k}$ guarantees nonnegativity of the resulting intensity estimates. That is, $\widehat{\lambda}_{j,k} \geq 0$ for $j = 0, \ldots, J$ and all k. To verify this, simply rewrite (3.16) with the factor $d_{j,k}$ inside the upper summand and consider the fact that $\left|\frac{d_{j,k}}{2s_i+c_{j,k}}\right| \leq 1$ for all i.

The overall algorithm is described below.

Bayesian Multiscale Intensity Estimation

1. Estimate coarsest scale coefficient

$$\widehat{\lambda}_{J,0} = c_{J,0}$$

2. For $j = J$ down to 1 and $k = 0$ to $N/2^j - 1$

Compute:

$\widehat{\delta}_{j,k}$ according to (3.16)

$$\widehat{\theta}_{j,k} = \widehat{\lambda}_{j,k}\,\widehat{\delta}_{j,k}$$

Refine:

$$\widehat{\lambda}_{j-1,2k} = \tfrac{1}{2}\left(\widehat{\lambda}_{j,k} + \widehat{\theta}_{j,k}\right)$$
$$\widehat{\lambda}_{j-1,2k+1} = \tfrac{1}{2}\left(\widehat{\lambda}_{j,k} - \widehat{\theta}_{j,k}\right)$$

These simple procedural steps produce posterior mean estimates for finer and finer representations of the underlying intensity, and terminate with the desired, finest-scale estimate $\widehat{\boldsymbol{\lambda}}_0$. The complexity of the proposed estimator is $O(N)$, the same order as the fast wavelet transform itself.

For large data sets it is possible that the full $J = \log_2(N)$ iterations over the scales in Step 2 above are not necessary. For instance, for long data sequences the estimator may be initiated at a scale $J^* < \log_2(N)$, for which the estimate $\widehat{\boldsymbol{\lambda}}_{J^*} = \mathbf{c}_{J^*}$ is already very reliable. In practice, even for low-count data, we have found that $J^* = 5$ provides excellent results. Using J^* other than J is equivalent to dividing the original data sequence into equal subsequences, estimating their underlying intensities separately, and concatenating the individual results to obtain the overall intensity estimate. However, in Section 3.6 we introduce a shift-invariant version of the estimator above, for which the equivalence just described does not hold.

3.5.2 Selection and Analysis of the Beta-Mixture Prior

Our experiments and analysis with real-world intensity functions have led to several conclusions.

1. The perturbation densities of many real-world intensities are very well characterized by a weighted combination of three beta densities with shape parameters $s_1 = 1$, $s_2 = 100$, and $s_3 = 10000$.

2. The $s_1 = 1$ component is the uniform density, and we have found that fixing the corresponding weight to a small positive constant (e.g., $p_1 = 0.001$), to insure that the prior density p_δ is bounded from below over the entire $[-1, 1]$ interval, is appropriate in most situations we have studied.

3. The key parameter that distinguishes the characteristics of different intensity functions is the trade-off between the $s_2 = 100$ component and the lower variance $s_3 = 10000$ component. The trade-off is parameterized by the probability $p_2 \in [0, 1 - p_1]$. (Note $p_3 = 1 - p_1 - p_2$.)

To gain some insight into the functioning of the MMI model estimator, consider Figure 3.6 which plots $\widehat{\delta}_{j,k}$ versus the ratio $d_{j,k}/c_{j,k}$ for three cases: low ($c = 10$), medium ($c = 30$), and high counts ($c = 1000$). These are, respectively, the dashed, solid, and dot-dashed curves. The ratio $d_{j,k}/c_{j,k}$ may be regarded as an empirical counterpart to the perturbation variate $\delta_{j,k}$. Figure 3.6(a) and (b) correspond to two δ-priors with $p_2 = 0.1$, and $p_2 = 0.9$, respectively. The rest of the parameters take the values given in points 1–3 above.

From these figures we can observe the following. At high counts, when the SNR is high, the estimator regards $d_{j,k}/c_{j,k}$ to be a good estimate of $\delta_{j,k}$ for almost every value of $d_{j,k}/c_{j,k}$; thus, the resemblance to the unit-slope linear function. In contrast, for low counts, the SNR is much lower, and consequently the estimator severely attenuates

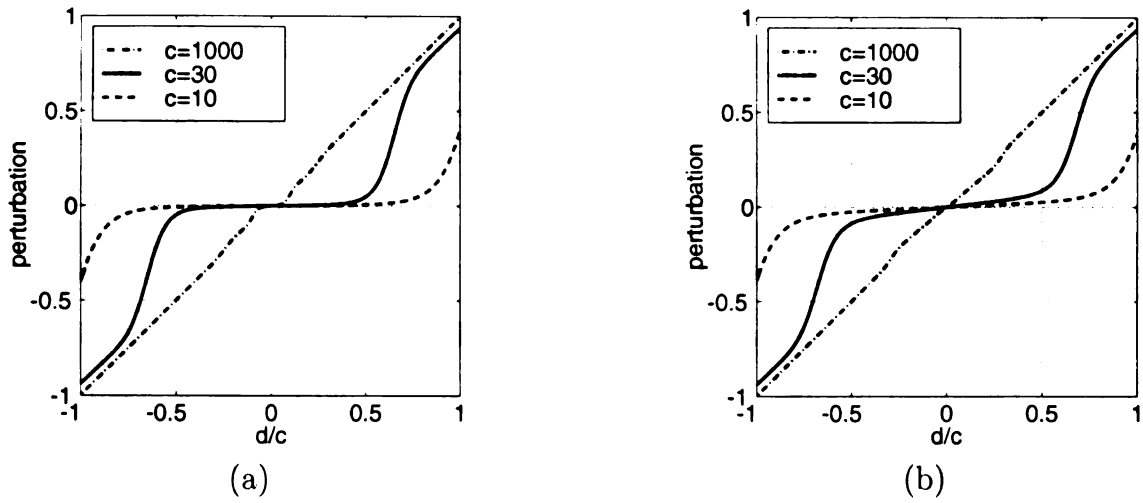


Figure 3.6. *Perturbation estimate* $\widehat{\delta}$ *as a function of* d/c *for* $c = 10$ *(dash),* $c = 30$ *(solid), and* $c = 1000$ *(dot-dash). The estimator's defining parameters are* $s_1 = 1$, $s_2 = 100$, $s_3 = 10000$, $p_1 = 0.01$, *and (a)* $p_2 = 1 - p_1 - p_3 = 0.100$, *and (b)* $p_2 = 1 - p_1 - p_3 = 0.900$.

$d_{j,k}/c_{j,k}$. This phenomenon is reminiscent of the behavior of wavelet-domain threshold estimators designed for additive Gaussian white noise (AGWN) [12], except that the threshold is adaptive to the local intensity.

The MMI model estimator minimizes the expected squared error with respect to the MMI model prior. Of course, the error is minimized by balancing the trade-off between fidelity to the data and fitting to the prior model. If the data are very 'reliable,' then the estimator favors the data. If the data are unreliable, then the estimator favors the prior. As noted, at low counts the ratios $d_{j,k}/c_{j,k}$ are not accurate estimates for δ, and so, the Bayesian estimator attenuates them to minimize the expected squared error in accordance with the prior. The nonlinearities in Figure 3.6(a) correspond to a lower-variance prior (smaller p_2) than that giving rise to the nonlinearities of Figure 3.6(b). As a result, the nonlinearities in Figure 3.6(a) display a 'dead-band' characteristic, since the prior requires that a greater number of $d_{j,k}/c_{j,k}$ samples be mapped towards zero, in contrast to the higher-variance prior which displays a less

harsh attenuation of small $d_{j,k}/c_{j,k}$ in Figure 3.6(b). This behavior is similar to that observed in other wavelet-based Bayesian estimators designed for AGWN [46, 11, 15].

However, the functioning of the MMI model estimator is in contrast to that of estimators for AGWN processes. In the latter cases, all wavelet coefficients are typically attenuated independently and in disregard to the values of the corresponding scaling coefficients [46, 11, 15]. The MMI model estimator, on the other hand, adapts not just to the statistics of a particular scale, but also naturally incorporates information from coarser scales. This is crucial in the Poisson problem since the coarse-scale intensities (scaling coefficients) are indicative of the statistical reliability of the wavelet coefficient.

3.5.3 Estimation of Prior Parameters

The Haar wavelet coefficient distributions of real-world intensities often fit a profile which resembles that of Figure 3.5 as previously discussed. However, while many distributions follow this general characteristic, one expects that subtle variations will exist from application to application. Therefore, it is of interest to adapt the prior to the problem at hand. Here we give a very simple approach to fitting the prior based on a moment-matching method. This adaptation can be viewed as an empirical Bayesian extension of framework described above.

Recall, we assume that for each scale j, the set $\{\delta_{j,k}\}$ is independent identically distributed (i.i.d.) with an M-component beta-mixture density distribution with parameters $\{p_{j,i}, s_i\}_{i=1}^{M}$. We let the mixing probabilities $\{p_{j,i}\}$ depend on scale to enable variations of the density across scale. Also note that here the index i does not refer to shift (as does k in $y_{j,k}$ for example), but rather it refers to the i-th component of the mixture density. Then, by (3.10), at each scale j, $\{y_{j,k}\}_{k=0}^{N/2^j-1}$ is also i.i.d. with an M-component standard-beta-mixture density distribution with parameters

$\{p_{j,i}, s_i\}_{i=1}^{M}$:

$$p_j(y) = \sum_{i=1}^{M} p_{j,i} \frac{y^{s_i-1}(1-y)^{s_i-1}}{B(s_i, s_i)}, \quad 0 \le y \le 1. \tag{3.17}$$

Since $y_{j,k}$, is independent of $\lambda_{j,k}$, $\mathrm{E}[\lambda_{j-1,2k}^n] = \mathrm{E}[\lambda_{j,k}^n]\,\mathrm{E}[y_{j,k}^n]$ and

$$\mathrm{E}\left[y_{j,k}^n\right] = \frac{\mathrm{E}[\lambda_{j-1,2k}^n]}{\mathrm{E}[\lambda_{j,k}^n]}. \tag{3.18}$$

The moments $\mathrm{E}[y_{j,k}]$ need not be computed using this expression since the prior model (3.17) gives the mean $\frac{1}{2}$ for any choice of parameters. For $n \ge 2$, the moments $\mathrm{E}[\lambda_{j-1,2k}^n]$ and $\mathrm{E}[\lambda_{j,k}^n]$ are easily estimated from the data. Since $\mathrm{E}[c_{j,k}|\lambda_{j,k}] = \lambda_{j,k}$,

$$\mathrm{E}[\lambda_{j,k}] = \mathrm{E}[\mathrm{E}[c_{j,k}|\lambda_{j,k}]] = \mathrm{E}[c_{j,k}].$$

And in general, it can be shown that for all n there exists a degree n polynomial $p_n(\cdot)$ such that

$$\mathrm{E}[\lambda_{j,k}^n] = \mathrm{E}[p_n(c_{j,k})].$$

For example,

$$\mathrm{E}[\lambda_{j,k}^2] = \mathrm{E}[c_{j,k}^2 - c_{j,k}] \approx \frac{1}{N/2^j} \sum_{k} (c_{j,k}^2 - c_{j,k}).$$

Substituting these estimates for various n into (3.18) we can obtain empirical estimates for the various moments of $y_{j,k}$. Equating these to the moments of the beta-mixture model (3.17) produces a set of equations that can be solved for the parameters $\{p_{j,i}, s_i\}_{i=1}^{M}$.

As mentioned above, we have found that the choice of a three component prior

beta-mixture model suffices for many real-world intensities. At all scales j, we suggest the shape parameters $s_1 = 1$, $s_2 = 100$, and $s_3 = 10000$, with weights $p_{j,1} = .001$, $p_{j,3} = 1\text{-}p_{j,1}\text{-}p_{j,2}$, and $p_{j,2}$ adapted at each scale using the second moment estimate for $y_{j,k}$:

$$\mathrm{E}[y_{j,k}^2] = \int_0^1 y^2 p_j(y)\, dy = \sum_{i=1}^{3} p_{j,i} \frac{s_i + 1}{4s_i + 2} \approx \frac{\sum_m (c_{j-1,2m}^2 - c_{j-1,2m})}{\sum_m (c_{j,m}^2 - c_{j,m})}. \qquad (3.19)$$

3.5.4 Optimal Estimation from Large Ensembles

Until now, we have strived to construct an estimator which optimally estimates the intensities of Poisson processes from a single observation of the process. There are situations, however, when several observations of the same process are available, or, when a large set of realizations corresponding to distinct elements of the ensemble to which the object of interest belongs is at hand or may be created.[6] These two conditions require distinct approaches in order to exploit all information available and optimally estimate the underlying intensity.

In the first situation, one arrives at a suitable approach by viewing the set of observations (of the one process) as sub-intervals of a longer integration time for the process. Then, the counts corresponding to this larger integration interval are simply the sum of all the counts available on a sample (pixel) basis, and represent sufficient statistics for the intensity pixels [18]. This well known result is due to the reproducing property of the Poisson distribution,[7] and may be stated as $\widehat{\boldsymbol{\lambda}} \equiv \mathrm{E}[\boldsymbol{\lambda}|\mathbf{c}_1, \ldots, \mathbf{c}_m] = \mathrm{E}[\boldsymbol{\lambda}|\mathbf{c}_1 + \cdots + \mathbf{c}_m]$.

[6] Under certain conditions, estimates for the object's distribution may be obtained from simulated large scale ensembles, for example, with the use of the bootstrap method [57, 58].

[7] See footnote 2 in page 74.

In the second scenario, an optimal estimator would be desirable if, for example, a large data bank of nuclear medicine images of, say, knees are available from past clinical studies. These data then could be used to leverage the estimate of the nuclear images of a new patient's knees. Situations like this admit two possible solutions. First, the moment-matching method described in the last section could be employed with the larger observation data set to obtain very robust estimates of the prior parameters. A robust prior clearly would lead to a robust estimate of the intensity of interest.

If the data set is large enough that the ratios

$$\frac{p(c_{j-1,2k} = c_1 + 1 | c_{j,k} = c + 1)}{p(c_{j-1,2k} = c_1 | c_{j,k} = c)}$$

can be computed sufficiently accurately, a prior-independent optimal estimator for the innovations $\{y_{j,k}\}$ can be devised as follows. Abbreviating the indexing notation, we have

$$\begin{aligned} \widehat{y} &= \int y\, p(y|c_1, c)\, dy \\ &= \int y\, p(c_1|c, y) \frac{p(y|c)}{p(c_1|c)}\, dy. \end{aligned}$$

Since $p(c_1|c, y) = \binom{c}{c_1} y^{c_1} (1-y)^{c-c_1}$ (see [48]), and since the innovations are independent of any *rv* at their own and coarser scales,

$$\begin{aligned} \widehat{y} &= \frac{\int y \binom{c}{c_1} y^{c_1} (1-y)^{c-c_1} p(y)\, dy}{p(c_1|c)} \\ &= \frac{c_1 + 1}{c + 1} \frac{\int \binom{c+1}{c_1+1} y^{c_1+1} (1-y)^{c-c_1} p(y)\, dy}{p(c_1|c)}. \end{aligned}$$

Thus, we obtain

$$\widehat{y}_{j,k} = \frac{c_1 + 1}{c + 1} \frac{p(c_{j-1,2k} = c_1 + 1 | c_{j,k} = c + 1)}{p(c_{j-1,2k} = c_1 | c_{j,k} = c)}. \tag{3.20}$$

Clearly, to compute the ratio of probabilities in this expression we only need, for example, a table for $p(c_1|c)$. For instance, for a realization $c_{j,k} = 0$, necessarily $c_{j-1,2k} = 0$, and the innovations estimate reduces to $\widehat{y}_{j,k} = p(c_{j-1,2k} = 1 | c_{j,k} = 1)$, which typically takes on values close to $\frac{1}{2}$. For very large counts of $c_{j,k}$, one would expect that $p(c_1|c) \approx p(c_1 + 1|c + 1)$. Therefore, for very large counts $\widehat{y}_{j,k} \approx \frac{c_1}{c}$, reflecting the fact that at high counts the SNR is high and the data is, therefore, reliable.

3.5.5 Example of Estimation from a Single Observation

For our first illustration of the performance of the Bayes' estimator we give the following simple example. We offer more comparative examples in Section 3.7 once a more sophisticated model and estimator is introduced in the next section. In Section 3.8, two-dimensional examples are given.

Consider the test intensity function depicted in Figure 3.7 (a). Figure 3.7 (b) depicts a realization of counts from this intensity. Note that the first "bump" in the low intensity region of (a) is statistically reliable—it is easily distinguished in the count data (b). The second bump in the high intensity region is equal in size to the first bump, but it is not statistically reliable as is seen in the count data in (b). Therefore, we expect that a good estimator should recover the first bump from the data and not attempt to recover the second. Figure 3.7 (c) depicts the estimate resulting from the PRESS-optimal estimator[8] proposed in [59]. It is an improvement

[8]This estimator is a wavelet threshold type operation that is derived using the statistical method of cross-validation [40].

over the raw data, but still exhibits quite a bit of variability. Figure 3.7 (d) depicts the intensity estimated by the Bayesian approach introduced in Section 3.5.1.

Due to the very simple—and unnatural—structure of the intensity where the corresponding wavelet coefficients may be found in one of only two states, zero and "large", a simple two-component beta-mixture density model sufficed for the innovations $y_{j,k}$. The parameters of the beta densities were $s_1 = 1$ and $s_2 = 10000$. The mixing parameter $p_1 = 1 - p_2$ was adapted at each scale using the data-based moment matching approach proposed in Section 3.5.3. Note that the Bayes' estimate does an excellent job at correctly estimating the reliable features (edges and bump in low intensity region). The experiment shown in Figure 3.7 was repeated in 1000 independent trials. The normalized MSE of the raw count estimate was 1.000, the MSE of the PRESS-optimal estimator was 0.375, and the MSE of the new Bayes' estimator was 0.106.

3.6 Stationary Intensity Models and Estimators

One potential limitation of the MMI model is that it is not stationary due to the fact that the Haar wavelet transform is shift-dependent. That is, the analysis and estimation depends on the alignment between the Haar basis functions and the data. Moreover, the coarse scale approximation by Haar wavelets is piece-wise constant, an unrealistic intensity model. To circumvent such problems, *shift-invariant wavelet transforms* have been proposed in literature [60, 61, 62, 63, 36, 64]. In this section, we provide a unified Bayesian framework for shift-invariant analysis and estimation based on the MMI model. We formulate a fast shift-invariant estimator from this analysis, and illustrate it with an example. Also, we characterize the autocorrelation functions of the MMI model (non-stationary) and a shift-invariant MMI model (stationary), and show that the shift-invariant MMI model has a more regular correlation behavior

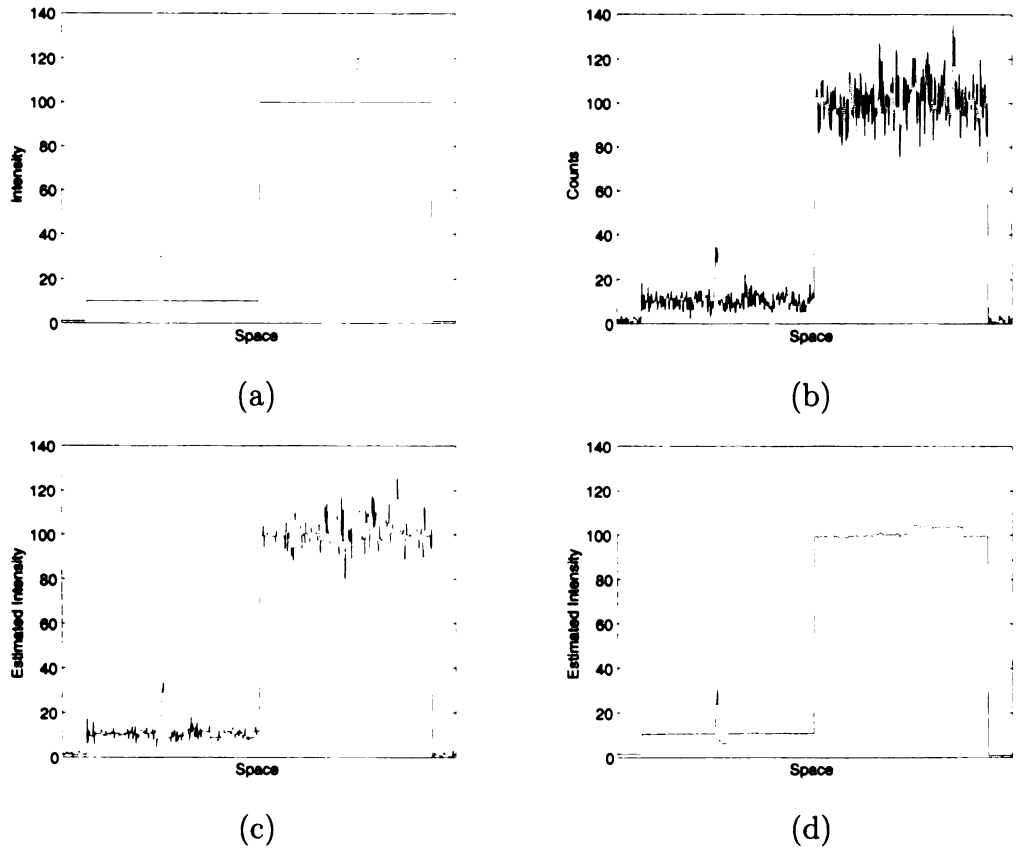


Figure 3.7. *Poisson intensity estimation. (a) Test intensity function. (b) Realization of Poisson process with intensity in (a). (c) PRESS-optimal estimate using algorithm in [59]. (d) Bayes' estimate using algorithm described in Section 3.5.1.*

which may be better suited for modeling real-world intensities.

3.6.1 A Shift-Invariant MMI Model

Shift-invariant MMI intensity models can be easily constructed within a Bayesian framework. Specifically, the shift of the intensity function with respect to the Haar wavelet system can be viewed as an additional degree of freedom in the model, and a probability measure on the shift parameter can be introduced. It is assumed that all shifts are circular. If we regard the original model as "unshifted" (shift = 0), then the standard MMI model introduced in Section 3.4 is denoted $p(\boldsymbol{\lambda}|\text{shift} = 0)$. Now

let $p(\text{shift} = m)$ denote a probability mass function for the shift, and consider the averaged MMI model

$$p(\boldsymbol{\lambda}) = \sum_m p(\boldsymbol{\lambda}|\text{shift} = m)p(\text{shift} = m). \tag{3.21}$$

If $p(\text{shift} = m)$ is the uniform distribution, a non-informative prior expressing no preference for any particular shift, then the averaged MMI model provides a shift-invariant (stationary) intensity prior and we call it the shift-invariant (SI) MMI model. It is shown in Section 3.6.3 that the SI-MMI model is more regular than the basic MMI model.

Estimation using the SI-MMI model is easily carried out by computing the optimal Bayes shift-dependent estimator given in Section 3.5.1 for each shift, and then computing an average of the results. The complexity of the shift-invariant estimator is $O(N^2)$ operations. However, note that if we employ a J^*-scale SI-MMI model, with $J^* < \log_2(N)$, then the estimator is invariant to shifts modulo 2^{J^*}. This is due to the fact that the scaling functions at the coarsest scale have support over 2^{J^*} samples. In general the J^*-scale SI-MMI model only requires a uniform shift prior over a 2^{J^*} sample region of support, rather than over the entire range of circular shifts. Thus, if $J^* < \log_2(N)$, then the complexity of shift-invariant estimation is only $O(N2^{J^*})$. As pointed out earlier (see Section 3.5.1), in many applications it suffices to take J^* smaller than $\log_2(N)$.

Note that the fast shift-invariant methods described in [61, 62] are not applicable in this case due to the dependence between wavelet coefficients across scale. This dependence stems from the multiplicative relationship between scaling and wavelet coefficients.[9] However, fast shift-invariant estimation algorithms are feasible if not "homogeneous" (see below). We next present one such approach.

[9]The fast shift-invariant methods treat all wavelet coefficients independently.

3.6.2 A Fast Shift-Invariant MMI Estimator

Referring to the tree representation of the intensity model of Figure 3.4 in Section 3.4.2, we know that the mapping $\{\lambda_{0,k}\}_k \mapsto \{\lambda_{J,0}\} \cup \{y_{j,l}\}_{j,l}$ is invertible. Therefore, $p(\boldsymbol{\lambda}) \propto p(\lambda_{J,0}, y_{J,0}, \ldots, y_{1,N/2-1})$, and so, the shift-invariant model (3.21) may also be expressed as

$$p(\boldsymbol{\lambda}) \propto p(\lambda_{J,0}) \sum_m \prod_{j,k} p(y_{j,k}^m)\, p(\text{shift} = m), \tag{3.22}$$

where we have exploited the invariance of $\lambda_{J,0}$ with respect to shift, and the independence of the set $\{y_{j,k}\}_{j,k}$. The notation $p(y_{j,k}^m)$ is simply a more economical way of expressing $p(y_{j,k}|\text{shift} = m)$. Note that with this notation, $\lambda_{0,0}^m \equiv \lambda_{J,0}\, y_{J,0}^m\, y_{J-1,0}^m \cdots y_{1,0}^m$, for example, is a component of $\lambda_{0,m}$ and not of $\lambda_{0,0}$ in the averaging process.

The shift-invariant model of the previous section is characterized by the assumption that any two distributions $p(y_{j,k}^{m_1})$ and $p(y_{j,k}^{m_2})$ are equal for all shifts m_1 and m_2. If, however, we choose probabilities such that $p(y_{j,k}^m) = 0$ for all $k \neq 0$ the above model becomes

$$p(\boldsymbol{\lambda}) \propto p(\lambda_{J,0}) \sum_m \prod_{j} p(y_{j,0}^m)\, p(\text{shift} = m). \tag{3.23}$$

Clearly, this new model is also shift invariant for it is obtained by averaging over all possible shifts, which we take to be equally likely. However, we say the model is *not homogeneous* because a different model results if we choose a new set of distributions to be the non-zero distributions in (3.22), say, for example, $p(y_{j,k}^m) = 0$ for all j and k, except whenever $k \neq 1$.

Despite the lack of homogenity, this intensity model has proved to be valuable as it leads to an efficient estimator of complexity $\mathcal{O}(N)$. Also, note that more sophisticated fast algorithms can be created from this general prescription by simply choosing not

one, but a combination of "legs" of the tree to represent the elements of the intensity sequence as the tree shifts. We have experimented with these notions and have formulated the corresponding estimators. Although the resulting models from the added complexity in the construction process are theoretically more appealing for they "close" the gap between the homogeneous and non-homogeneous models, we have found that they add very little as to the quality of the intensity estimates. For this reason, we have not pursued them here any further.

The procedure for the fast estimator derived from the simpler intensity model (3.23) is the following.

Fast Shift-Invariant MMI Intensity Estimation

1. Estimate coarsest scale coefficient

$$\widehat{\lambda}_{J,0} = c_{J,0}$$

2. For $m = 0$ to $N-1$ and $j = J$ down to 1

Compute:

$$\widehat{\delta}_{j,0}^{m} \text{ according to (3.16)}$$
$$\widehat{y}_{j,0}^{m} = \tfrac{1}{2}\left(1 + \widehat{\delta}_{j,0}^{m}\right)$$

Refine:

$$\widehat{\lambda}_{j-1,0}^{m} = \widehat{\lambda}_{j,0}^{m}\, \widehat{y}_{j,0}^{m}$$

The desired intensity estimate is $\widehat{\boldsymbol{\lambda}} = (\widehat{\lambda}_{0,0}^{0}, \widehat{\lambda}_{0,0}^{1}, \ldots, \widehat{\lambda}_{0,0}^{N-1})$. We note that these procedural steps call for $N/2$ redundant shifts in order to compute the N elements of the intensity sequence. Although it is not hard to describe the above estimation algorithm void of these redundancies, we feel that the description is clearer in the form presented.

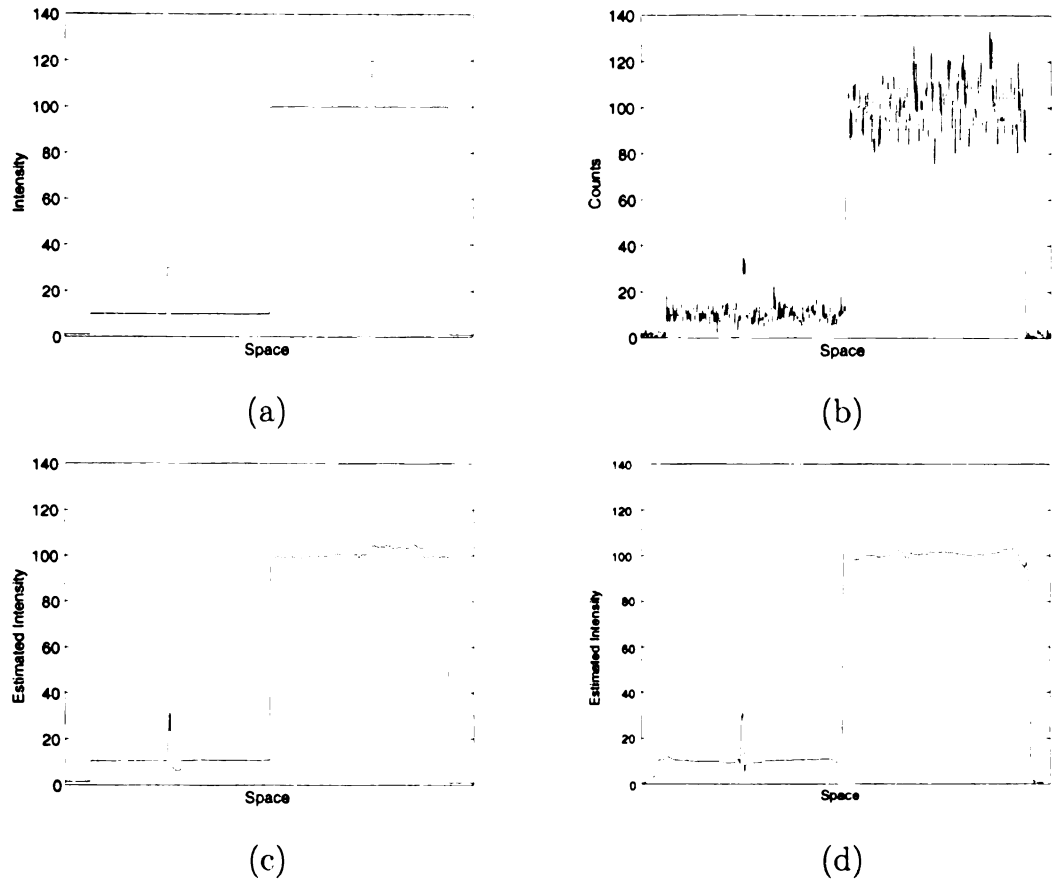


Figure 3.8. *Fast Poisson intensity estimation. (a) Test intensity function. (b) Realization of Poisson process with intensity in (a). (c) Bayes' estimate using the algorithm described in Section 3.5.1. (d) Bayes' estimate using the fast shift-invariant algorithm described here.*

Example

In order to illustrate the quality of the estimates obtained by the fast shift-invariant estimator we repeated the example of Section 3.5.5. We created a sequence of Poisson counts from the intensity in Figure 3.8 (a), and used them to estimate the intensity using the fast algorithm. The result is depicted in Figure 3.8 (d). For comparison, we reproduced Figure 3.7 (d) in Figure 3.8 (c), which shows the estimate based on the MMI model of Section 3.5.1.

By repeating this experiment with a large number of different count realizations,

we found that the quality of the estimate evident in Figure 3.8 (d) was representative of the fast shift-invariant estimator.

3.6.3 Autocorrelation Functions of MMI and SI-MMI Models

The underlying beta mixture densities capture the key heavy-tailed, non-Gaussian nature of wavelet coefficient distributions. However, the shift-dependent nature of the Haar wavelet transform generates a non-stationary correlation structure as illustrated next. For the sake of illustration, we focus on the 1-d case. Extensions to higher dimensions are straightforward. Also, to keep things simple, we assume that the maximum number of scales $J = \log_2(N)$ are computed in the analysis, where N is the length of the intensity vector. The results easily extend to other choices for $J^* \neq J$.

First, consider that basic MMI model (shift= 0) and let us introduce the following notation. Let $\mu_J^2 \equiv \mathrm{E}[\lambda_{J,0}^2]$, the second moment of the scaling coefficient at the coarsest scale, and let $\rho_j^2 \equiv \mathrm{E}\left[y_{j,k}^2\right] = \mathrm{E}\left[(1-y_{j,k})^2\right]$, the second moment of the innovations variates. Recall that we assume the distribution of $y_{j,k}$ does not depend on position k. The variables $y_{j,k}$ and $1-y_{j,k}$ have a common second moment due to the symmetry of the distribution of $y_{j,k}$ about $\frac{1}{2}$.

For illustration consider the correlation between the intensities $\lambda_{0,0}$ and $\lambda_{0,2}$ in the MMI model depicted in Figure 3.4 in Section 3.4.2. Note that the finest scale for which $\lambda_{0,0}$ and $\lambda_{0,2}$ have a common predecessor above in the tree is $j = 2$ and the predecessor is $\lambda_{2,0}$. Therefore we can write

$$\begin{aligned} \lambda_{0,0} &= y_{1,0}\, y_{2,0}\, \lambda_{2,0}, \\ \lambda_{0,2} &= y_{1,1}\, (1-y_{2,0})\, \lambda_{2,0}. \end{aligned}$$

The correlation between the two intensities is computed

$$\mathrm{E}\left[\lambda_{0,0}\lambda_{0,2}\right] = \mathrm{E}\left[y_{1,0}\; y_{1,1}\; y_{2,0}\; (1 - y_{2,0})\; \lambda_{2,0}^2\right].$$

Exploiting the independence of the innovations variates, we have

$$\mathrm{E}\left[\lambda_{0,0}\lambda_{0,2}\right] = \mathrm{E}\left[y_{1,0}\right]\; \mathrm{E}\left[y_{1,1}\right]\; \mathrm{E}\left[y_{2,0}\; (1 - y_{2,0})\right]\; \mathrm{E}\left[\lambda_{2,0}^2\right].$$

Examining the individual product terms and making use of the moments defined above,

$$\begin{aligned} \mathrm{E}\left[y_{1,0}\right] = \mathrm{E}\left[y_{1,1}\right] &= 2^{-1}, \\ \mathrm{E}\left[y_{2,0}\; (1 - y_{2,0})\right] &= 1/2 - \rho_2^2, \\ \mathrm{E}\left[\lambda_{2,0}^2\right] &= \mu_3^2\rho_3^2. \end{aligned}$$

Hence,

$$\mathrm{E}\left[\lambda_{0,0}\lambda_{0,2}\right] = 2^{-2}\; (1/2 - \rho_2^2)\; \mu_3^2\rho_3^2.$$

Now consider a general case in which we are interested in the correlation between two intensities, say λ_{0,k_1} and λ_{0,k_2}. Let j^* be the finest scale for which λ_{0,k_1} and λ_{0,k_2} have a common predecessor (above) $\lambda_{j^*,k}$ in the MMI model tree. The scale j^* can be explicitly calculated using the binary representations of k_1 and k_2, and depends on the exact positions of k_1 and k_2 with respect to the alignment of the Haar basis

functions. Assuming that $k_1 < k_2$, we have

$$\lambda_{0,k_1} = \prod_{i=1}^{j^*-1} y_{i,k}\; y_{j^*,k}\; \lambda_{j^*,k}, \tag{3.24}$$

$$\lambda_{0,k_2} = \prod_{i=1}^{j^*-1} y'_{i,k}\; (1 - y_{j^*,k})\; \lambda_{j^*,k}. \tag{3.25}$$

In the expressions for λ_{0,k_1} and λ_{0,k_2} above, we use a generic spatial index k, since the distributions of independent innovations variates do not depend on position. We do, however, use $y_{j,k}$ and $y'_{j,k}$ to distinguish the independent innovations variates corresponding to λ_{0,k_1} and λ_{0,k_2}, respectively. The correlation is given by

$$\mathrm{E}\left[\lambda_{0,k_1}\lambda_{0,k_2}\right] = \mathrm{E}\left[\prod_{i=1}^{j^*-1} y_{i,k} \prod_{i=1}^{j^*-1} y'_{i,k}\; y_{j^*,k}\; (1 - y_{j^*,k})\; \lambda_{j^*,k}^2\right]. \tag{3.26}$$

Again exploiting the independence of the innovation variates and making use of the moments defined above, we have $\mathrm{E}\left[\lambda_{j^*,k}^2\right] = \mu_J^2 \prod_{i=j^*+1}^{J} \rho_i^2$ and

$$\begin{aligned} r(k_1, k_2) &\equiv \mathrm{E}\left[\lambda_{0,k_1}\lambda_{0,k_2}\right] \\ &= 2^{-2(j^*-1)}\; (1/2 - \rho_{j^*}^2)\; \mu_J^2 \prod_{i=j^*+1}^{J} \rho_i^2. \end{aligned} \tag{3.27}$$

Note that because j^* depends on the alignment between the intensity function and the Haar basis, $r(k_1, k_2)$ is not a function of the difference $k_1 - k_2$ alone. This shows that the intensity distribution represented by the MMI model is non-stationary. Two columns (fixed k_1) of the autocorrelation function of a 256 length MMI model intensity prior are shown in Figure 3.9. Note also that the autocorrelation function is highly irregular (piecewise-constant), which is an undesirable model for real-world intensities.

Now consider the autocorrelation function of the SI-MMI model (3.21). Given a displacement of n between two intensities, say $\lambda_{0,0}$ and $\lambda_{0,n}$, we can compute the probability of the finest scale j^* of a common predecessor, with respect to the uniform

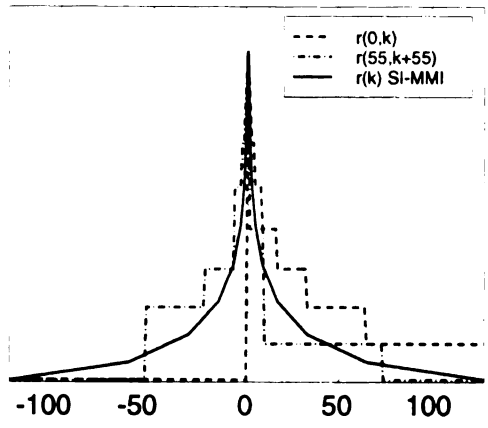


Figure 3.9. *Correlation functions for MMI and SI-MMI 256-point ($J = 8$) intensity priors. MMI model autocorrelation $r(0,k)$ (dash-dash) and $r(55,k+55)$ (dash-dot) plotted as a function of k. Stationary SI-MMI model autocorrelation $r(k)$ (solid). For both the MMI and SI-MMI models in this example, $\mu_j^2 = 1$ and $\rho_j^2 = 0.26$, $j = 1,\ldots,J$.*

distribution over the shift parameter, as follows. We need to count the number of shifts that give rise to each possible value of j^*, or more precisely the probability of the set of shifts that give rise to each possible j^*. For example, suppose $n = 1$ and consider the tree in Fig. 3.4. In this case, four shifts (out of eight) result in $j^* = 1$, another two shifts result in $j^* = 2$, and two shifts (one wrapping around due to circularity) result in $j^* = 3$. Hence, we compute $p(j^* = 1|n = 1) = 1/2$, $p(j^* = 2|n = 1) = 1/4$, and $p(j^* = 3|n = 1) = 1/4$. Similarly, if $n = 2$, then $p(j^* = 1|n = 2) = 0$, $p(j^* = 2|n = 2) = 1/2$, and $p(j^* = 3|n = 2) = 1/2$, where again we must be careful to account for the wrap-around effect of the circular shifting.

Given a displacement of $n = k$ between two scaling coefficients in a J-scale MMI model, the probability of the scale j^* is determined by inspecting the associated binary tree. We need only consider $|k| \leq 2^{j-1}$ due to the periodicity of the MMI. First note that we have $p(j^* \leq J \,|\, n = k) = 1$. Next, notice that $(2^m - k)_+$ is the number of shifts of a length-k sequence that fit within a length-2^m sequence, where

$(x)_+ = \max(x, 0)$. In other words, $(2^m - k)_+$ is the total number of scaling coefficient pairs spaced k apart at the bottom of a m-level binary tree. Also note that the number of m-level subtrees, $m < J$, at the bottom of a larger J-level binary tree is precisely 2^{J-m}. Hence, for $m < J$

$$\begin{aligned} p(j^* \leq m \,|\, n = k) &= (2^m - |k|)_+ \, 2^{J-m} 2^{-J} \\ &= (2^m - |k|)_+ \, 2^{-m}. \end{aligned} \tag{3.28}$$

It follows that

$$\begin{aligned} p(m|k) &\equiv p(j^* = m | n = k) \\ &= \begin{cases} 0, & m < \lceil \log_2(|k|+1) \rceil \\ 1 - 2^{-m}|k|, & m = \lceil \log_2(|k|+1) \rceil \\ 2^{-m}|k|, & \lceil \log_2(|k|+1) \rceil < m < J \\ |k| 2^{-J+1}, & m = J, \end{cases} \end{aligned}$$

where $\lceil \log_2(|k|+1) \rceil$ denotes the smallest integer greater than or equal to $\log_2(|k|+1)$.

With these probabilities defined, the autocorrelation function is given by

$$r(k) \equiv \mathrm{E}\left[\lambda_{0,l} \lambda_{0,l+k}\right] = \sum_{m=1}^{J} \nu_m \, p(m|k), \tag{3.29}$$

where

$$\nu_m \equiv 2^{-2(m-1)} \, (1/2 - \rho_m^2) \, \mu_J^2 \prod_{i=m+1}^{J} \rho_i^2 \tag{3.30}$$

is simply the autocorrelation between two intensities at the bottom of an MMI binary tree model for which $j^* = m$, as in (3.27). Note that the SI-MMI autocorrelation is stationary. The autocorrelation function for a 256 length SI-MMI intensity prior

is shown in Figure 3.9. Unlike the autocorrelation of the MMI model, the SI-MMI model's autocorrelation is piece-wise linear (compared to piece-wise constant) and hence is more regular and potentially better suited for the analysis of natural intensities. These results are similar in spirit to the analysis in [61], where it is observed that the shift-invariant Haar wavelet transform is line-preserving.[10] Moreover, the results here suggest possible schemes for choosing the parameters of the SI-MMI model. For example, the decay of the autocorrelation function may be tailored by appropriate choices of the ρ_i^2, $i = 1, \ldots, J$. Larger values for the ρ_i^2 cause $r(k)$ to decay faster. We also note that similar correlation analysis may be carried out for related wavelet-domain signal models developed for Gaussian data [46, 34, 11, 15, 47]. We also note that the issue of combining or mixing trees (which is essentially what is being done in the SI-MMI model) has been studied in the more general setting of the Polya tree. Some very interesting theoretical results concerning the continuity of densities generated by mixtures of Polya trees (SI-MMI models are a special case) are given in [54]. These results may provide further insights into the SI-MMI model and may suggest other extensions of our framework, but we have not pursued this here.

3.6.4 SI-MMI Model and $1/f$ Processes

Remarkably, the SI-MMI correlation function has a fractal $1/f$-like character. Fractal $1/f$ random process models are commonly used in image modeling, and it has been observed that natural signals and images often display a correlation structure similar to that of the SI-MMI model [65, 66]. To see that the SI-MMI correlation is $1/f$, consider the simple case in which the second moments of the innovations are constant

[10] These conclusions extend to the SI-MMI model as well.

independent of scale, *i.e.*, $\rho_j = \rho,\ j = 1, \ldots, J$. Then

$$\nu_m = 2^{-2(m-1)}\,(1/2 - \rho^2)\,\mu_J^2 \rho^{2(J-m)} \tag{3.31}$$

$$= C\,(2\rho)^{-2m}, \tag{3.32}$$

where C is a constant independent of m. Next equate $(2\rho)^{-2m}$ with $2^{(\gamma-1)m}$ and solve for γ. This gives

$$\nu_m = C\,2^{(\gamma-1)m}, \tag{3.33}$$

where $\gamma = -1 - \log_2 \rho^2$. Note that $1/4 < \rho^2 < 1/2$, where the lower and upper bounds corresponds to the two extreme limits of the beta density: a point mass at $1/2$ or two point masses at 0 and 1, respectively. This implies $0 < \gamma < 1$. Combining (3.33) with (3.29) and after some algebra, we have

$$\begin{aligned} r(k) &= C(2^{(\gamma-1)\lceil \log_2(|k|+1)\rceil} - \beta|k|2^{(\gamma-2)J} \\ &\quad +\beta|k|2^{(\gamma-2)\lceil \log_2(|k|+1)\rceil}) \end{aligned} \tag{3.34}$$

for $|k| > 0$, where $\beta = \frac{1-2^{(\gamma-1)}}{1-2^{(\gamma-2)}}$. In the case $k = 0$, $r(0) = \mu_J^2 \rho^{2J}$. Finally, making use of the approximation $2^{\lceil \log_2(|k|+1)\rceil} \approx |k|$, for $|k| > 0$ we have

$$r(k) \approx C\left[(1-\beta)|k|^{(\gamma-1)} + \beta|k|2^{(\gamma-2)J}\right]. \tag{3.35}$$

For large J and $\gamma < 1$, the term $\beta|k|2^{(\gamma-2)J}$ is negligible, and hence the correlation function behaves like $|k|^{(\gamma-1)}$ and the power spectrum decays like $\frac{1}{|f|^\gamma}$ (see [67] for the relationship between autocorrelation functions and power spectrums of $1/f$ processes). That is, the SI-MMI model produces non-negative, stationary processes with $1/f$ characteristics. For more details on SI wavelet models and $1/f$ processes see [64].

3.7 Numerical Comparison of Wavelet-Based Intensity Estimators

Here we compare the performance of the new Bayesian estimation algorithm with several existing methods. To assess the performance of each method, four test intensity functions were used. These functions were the "Doppler," "Blocks," "HeaviSine," and "Bumps" test signals proposed in [12]. Each test function is 1024 samples long (*i.e.*, $N = 1024$, $J = 10$). These functions serve as benchmark tests for signal estimators, and they were designed to be representative of a variety of natural signal structures. We refer the reader to [12] for more information about the test functions. Since the intensity functions must be non-negative, each test function was shifted and scaled to obtain an intensity with a desired peak value and a minimum value of $\frac{1}{\text{peak value}}$. Realizations of counts are generated from each intensity using a standard Poisson random number generator [23]. We compare the performance of the simple estimator based on the raw counts (COUNT), a shift-invariant version of the cross-validation estimator[11] (CV) proposed in [44], the SI-MMI model estimator described in Section 3.6.1 with a three component beta-mixture model for the innovations with parameters[12] $s_1 = 1$, $s_2 = 100$, and $s_3 = 10000$, the square-root estimation methods using a shift-invariant version of the Haar wavelet transform (D2), and the square-root estimation method using a shift-invariant version of the Daubechies-8 (D8) wavelet. The method proposed in [14] is not compared since it is derived under a "burst-like" process model which is not appropriate for these test functions with the exception of the Bumps function.

The square-root method first computes the square-root of the counts, then treats the square-root data as though it were Gaussian, takes the shift-invariant discrete

[11] See footnote 8 in page 93.

[12] The mixing parameter $p_1 = 0.001$, and parameter p_2 (and hence p_3) is determined using the data-adaptive moment-matching method given in Section IV-C.

Table 3.1. *AMSE results for various test intensities and estimation algorithms. Peak intensity* = 8.

Intensity	*COUNT*	*CV*	*BAYES*	*D2*	*D8*
Doppler	0.1786	0.0588	0.0154	0.0548	0.0443
Blocks	0.1935	0.0617	0.0178	0.0700	0.0800
HeaviSine	0.1833	0.0552	0.0052	0.0294	0.0300
Bumps	0.5207	0.1877	0.1475	0.4570	0.4317

Table 3.2. *AMSE results for various test intensities and estimation algorithms. Peak intensity* = 128.

Intensity	*COUNT*	*CV*	*BAYES*	*D2*	*D8*
Doppler	0.0111	0.0047	0.0026	0.0095	0.0059
Blocks	0.0120	0.0040	0.0027	0.0077	0.0126
HeaviSine	0.0115	0.0039	0.0007	0.0036	0.0028
Bumps	0.0324	0.0171	0.0143	0.1046	0.0908

wavelet transform [61, 62], applies a soft-threshold nonlinearity to wavelet coefficients, and computes the inverse transform of the thresholded coefficients. After this processing, the result is squared to obtain an intensity estimate. For both square-root methods the universal threshold proposed in [12] was used.

We consider both the D2 and D8 wavelet (which can be applied in the case of Gaussian data) to demonstrate that the Haar-based method introduced here can outperform the square-root method even when more regular wavelets like the D8 are used. All methods employ a 5-scale wavelet transform. In practice, we could use full J-scale transforms, but their performances were roughly the same as that of the 5-scale transforms in the experiments, and the 5-scale transforms are more computationally efficient. Table 1 gives the average mean-square errors (AMSE) of the various methods for a peak intensity of 8. Table 2 gives the AMSE of each method for

a peak intensity of 128. The AMSE is estimated using 25 independent trials in each case, and each AMSE is normalized by the squared Euclidean norm of the underlying intensity function. Inspection of the tables shows that all methods offer significant improvements over the simple count estimator. Moreover, the SI-MMI based estimator outperforms all others in every case. We also note that similar tests and comparisons were made with the *shift-variant* versions of each estimator. As expected, the shift-variant estimators did not perform as well as their shift-invariant counterparts. However, the MMI model estimator outperformed the other shift-variant methods in all cases as well.

3.8 Application to Photon-Limited Imaging

In this section we apply the MMI models and estimation procedure to the problem of photon-limited imaging. Photon-limited imaging arises in many fields including medicine and astronomy. The fundamental problem in photon-limited imaging is the variability due to quantum effects in the emission and detection of photons. In many problems, the photon counts collected during image acquisition are well-modeled by a temporally homogeneous and spatially inhomogeneous Poisson process.

Assume that we detect photon emissions in a compact region of the plane. The photon emissions are the result of an underlying two-dimensional continuous intensity function. We are interested in estimating the intensity function from the photon detections. For practical reasons (computing and display), we seek an estimate of the intensity at a finite scale (resolution) represented by a 'pixelized' intensity. A crude estimate of the pixelized intensity is obtained by simply counting the number of photons detected in each square pixel region of the plane. This "count" image is highly variable due to the random nature of the photon emission process. However, lower resolution images, obtained by counting the number of photons detected in

larger square pixel regions of the plane, provide better (less variable) estimates of the low-resolution intensities. This illustrates the advantage of multiscale analysis in photon-limited imaging. Relatively reliable coarse-scale estimators of the intensity can be leveraged to obtain finer details using the multiscale Bayesian framework.

To illustrate the effectiveness of the framework in photon-limited imaging applications, we consider a simulated experiment and a real-world application to nuclear medicine imaging. Note that the MMI models and estimator are easily generalized to two dimensions. Specifically, we take the 2-d multiscale parameters to be the factors corresponding to the *multiplicative* refinement of a coarse scaling coefficient (intensity) into four finer scaling coefficients by first splitting it vertically (horizontally) into two halves, then next horizontally (vertically) splitting each half into two quarters. That is, if $\lambda_{k,l}$ is a 2-d intensity function, then we define $\lambda_{0,k,l} \equiv \lambda_{k,l}$ at the finest scale $j = 0$ and for coarser scales take

$$\begin{aligned}
\lambda_{j+1,k,l} &= \lambda_{j,2k,2l} + \lambda_{j,2k,2l+1} + \lambda_{j,2k+1,2l} + \lambda_{j,2k+1,2l+1}, \\
y^1_{j+1,k,l} &= \frac{\lambda_{j,2k,2l} + \lambda_{j,2k,2l+1}}{\lambda_{j,2k,2l} + \lambda_{j,2k,2l+1} + \lambda_{j,2k+1,2l} + \lambda_{j,2k+1,2l+1}}, \\
y^2_{j+1,k,l} &= \frac{\lambda_{j,2k,2l}}{\lambda_{j,2k,2l} + \lambda_{j,2k,2l+1}}, \\
y^3_{j+1,k,l} &= \frac{\lambda_{j,2k+1,2l}}{\lambda_{j,2k+1,2l} + \lambda_{j,2k+1,2l+1}}.
\end{aligned} \tag{3.36}$$

Note that in the 2-d case we have three sets of multiplicative innovations, one vertical set y^1 and two horizontal sets y^2 and y^3. In the analysis of count images, each scaling coefficient is the sum of four counts, and each wavelet coefficient is simply the difference of two counts. Hence, all the machinery developed for the one-dimensional case, based on sums and differences of pairs of counts, is immediately applicable to two (or even higher) dimensional data. Note that this 2-d multiscale analysis defined above differs from the standard 2-d Haar wavelet analysis [68], which involves

a vertical, horizontal and diagonal differences. We use the alternative 2-d analysis because, unlike the standard 2-d Haar analysis, it allows us to decouple the Poisson problem, just as in the 1-d case. In both experiments, the 5-scale Haar transform is employed and a three component beta-mixture density is used for the SI-MMI model of Section 3.6.1 with fixed shape parameters $s_1 = 1$, $s_2 = 100$, and $s_3 = 10000$. The mixing probability p_1 is fixed at 0.001, and p_2 (and p_3) is adapted to the data at each scale using the moment-matching method described in Section 3.5.3. We have found these choices of s_1, s_2, and s_3, combined with the flexibility of the data-adaptive p_2, to provide very good results for a wide-variety of imagery.

3.8.1 Photon-Limited Imaging Simulation

Figure 3.10 depicts a typical realization of a simulated photon-limited imaging application and the resulting estimates provided by the MMI and SI-MMI models. The maximum intensity in the image in Figure 3.10(a) is 60.00, and the average intensity is 25.40. Hence, this simulation models a fairly low intensity (low SNR) imaging problem. A realization of counts is generated from this intensity using a standard Poisson random number generator [23]. Note the visual improvement provided by the MMI and SI-MMI model estimates in Figure 3.10(c) and (d), respectively, in comparison to the count image Figure 3.10(b). Furthermore, the estimate based on the SI-MMI model appears to be better than that of the MMI model. In fact, in 25 independent trials of this experiment we estimated the average mean squared pixel error to be 25.28 for the count image, 7.36 for the MMI model estimator, and 4.01 for the SI-MMI model estimator.

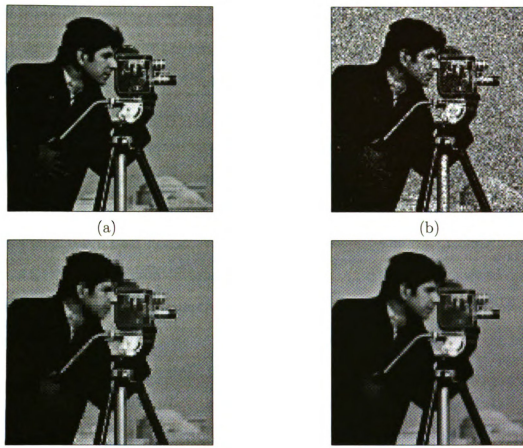


Figure 3.10. *Photon-limited image estimation using MMI models. (a) intensity function, (b) realization of Poisson counts (average squared pixel error = 24.80), (c) intensity estimate using MMI model (average squared pixel error = 7.36), (d) intensity estimate using SI-MMI model (average squared pixel error = 3.98).*

3.8.2 Application to Nuclear Medicine Imaging

Nuclear medicine imaging is a widely used commercial imaging modality [69]. Unlike many other medical imaging techniques, nuclear medicine imaging can provide both anatomical *and* functional information. However, nuclear medicine imaging has a much lower signal-to-noise ratio relative to other imaging techniques. Hence, improvements in image quality via optimized signal processing represent a significant opportunity to advance the state-of-the-art in nuclear medicine.

Nuclear medicine images are acquired by the following procedure [69]. Radioactive pharmaceuticals that are targeted for uptake in specific regions of the body are injected into the patient's bloodstream. As the radioactive pharmaceuticals decay, gamma rays are emitted from within the patient. Imaging the gamma ray emissions provides a mapping of the distribution of the pharmaceutical, and hence a mapping of the anatomy or physiologic function of the patient. Gamma rays are detected and spatially located using a gamma camera, which converts gamma rays into light. Photomultiplier tubes then detect and locate the emissions. The raw nuclear medicine data is an image of photon detections (counts). The raw data may be viewed directly or used for tomographic reconstruction. The major limitation of nuclear medicine imaging is the low-count levels acquired in typical studies, due in part to the limited level of radioactive dosage required to insure patient safety. Because of the variability of low-count images, it is very common to employ a post-filtering or estimation procedure to obtain a "better" estimate of the underlying intensity. An excellent discussion of the potential diagnostic benefits of various frequency domain processing methods is given in [5]. Advantages of multiscale methods over frequency domain methods for photon-limited imaging problems are discussed in [59].

To illustrate the potential of our multiscale Bayesian framework in nuclear medicine imaging, consider the spine and heart studies depicted in Figure 3.11. Figure 3.11(a) depicts the count image from a nuclear medicine spine study. The radiopharmaceutical used here is Technetium-99m labeled diphosphonate. In bone studies such as this, brighter areas indicate increased uptake of blood in areas where bone growth is occurring. This may reflect areas where bone damage has occurred. Functional changes in bone can be detected using nuclear medicine images before they will show up in x-ray images. The maximum count in this image is 178 in the "hot-spot" at the bottom of the spine. The maximum count in the upper portion of the spine is 75. Figure 3.11(b) shows the SI-MMI model estimate of the underlying intensity.

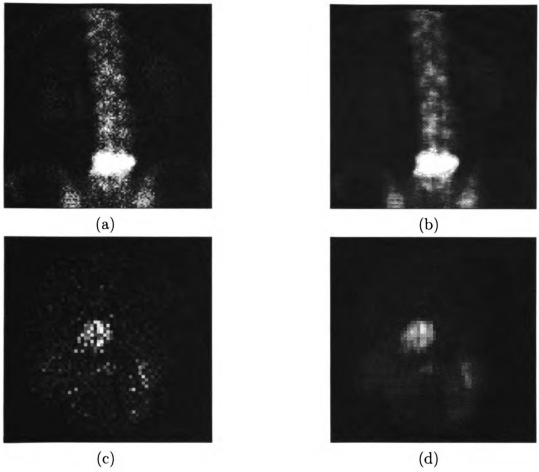


Figure 3.11. *Nuclear medicine image estimation using SI-MMI models. (a) Spine count image. (b) SI-MMI model estimate of underlying intensity. (c) Heart count image. (d) SI-MMI model estimate.*

Figure 3.11(c) depicts an image of a heart obtained from a nuclear medicine study. The image was obtained using the radiopharmaceutical Thallium-201, and the maximum count is 33. In this type of study, the radiopharmaceutical is injected into the bloodstream of the patient and moves into the heart wall in proportion to the local degree of blood perfusion. The purpose of this procedure is to determine if there is decreased blood flow to the heart muscle. Figure 3.11(d) shows the SI-MMI model estimate. In both studies, we see that the SI-MMI model estimator preserves the important image structure and has significantly lower variance compared to the raw

count image. The intensity estimates provided by the SI-MMI model may enable better diagnosis in clinical nuclear medicine.

CHAPTER 4

Emission Computed Tomography

In this chapter, we extend the MMI models and estimation approaches to emission tomography imaging. We introduce two extensions: one based on geometrical considerations of natural pixels of images, and the other based on a new multiscale inversion approach to the tomographic image reconstruction problem. The new models are used to represent the intensity sinogram of the projected images, from which their optimal estimation can be computed prior to reconstruction. We illustrate the performance of the new estimators with examples using synthetic and clinical data.

4.1 Preliminaries

The first explicit formula for the reconstruction of a function on a plane given its integrals over lines was first derived by Radon in 1917. Although it was not until the late sixties that practical applications of the Radon formula started to appear, first in radioastronomy and then in electron micrography [70], the practical computed tomography that it gave birth to has since become one of the most important tools in many branches of science for processing information that otherwise would be inaccessible. For example, in astronomy and geology, imaging of the Sun's and Earth's internal structures have been possible only by tomographical techniques [71]. It is perhaps in

medicine where we have witnessed the greatest impacts of this technology. Imaging the internal structure of human patients by noninvasive methods is frequently the safest and most expeditious method in determining their conditions, and thus, often the most desirable method for diagnosis.

Although our interest is in emission computed tomography (ECT) in general, in this chapter we focus exclusively on nuclear medicine tomography—more specifically, in single-photon emission tomography (SPECT)—for the following reasons. The problems encountered in nuclear medicine tomography typify many of the problems encountered in ECT at large; and, concentrating in SPECT in particular, permits us to be more specific in the treatment of the subject. Also, nuclear medicine tomography represents one of the most challenging applications due to typical low SNR regimes under which it takes place; therefore, the methods presented here are tested for robustness. Further, nuclear medicine tomography represents one very important application that directly affects many people's quality of life.

There exists a number of tomographical methods for diagnostic imaging. For example, x-ray computed tomography (x-ray CT), magnetic resonance imaging (MRI), positron emission tomography (PET), and SPECT. The physiological basis of emission computed tomography medical imaging (*e.g.*, PET and SPECT), opposed to the anatomic approach of transmission imaging (*e.g.*, x-ray CT and MRI), are unique in that measurements of functional abnormalities are readily possible from a single image.[1]

In nuclear medicine imaging a radiopharmaceutical, which is targeted for uptake by some specific organs, is administered to the patient. As the process of nuclear decay takes place in the radiopharmaceutical, gamma-ray photons are emitted in all directions. Those photons incident to an array of detectors placed in close proximity

[1]Although functional measurements in transmission-based imaging are possible, these require more complex procedures involving sequences of images obtained at controlled time-intervals that must be carefully registered [72, 73].

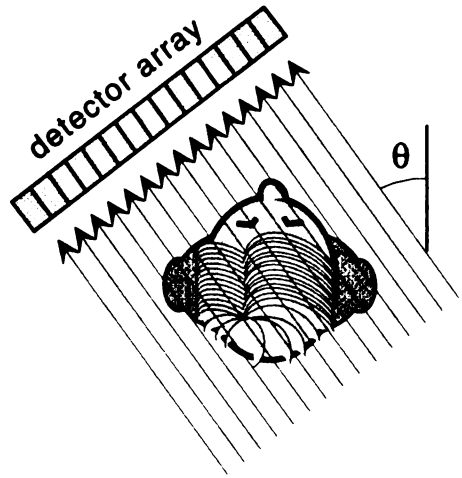


Figure 4.1. *Tomographic data collection geometry. The array of detectors collect and count the number of photons induced by the arriving gamma rays emanating from within the subject. The array is repositioned at equal angle intervals about the long axis of the subject so as to obtain a sufficient number of projections.*

to the patient and in the vicinity of the organs of interest are counted. The number of photon counts are then interpreted as a function of physiological activity or intensity.

For tomographical image reconstruction of a transversal slice of the body, measurements of photon counts are made at many angles about the long axis of the body. Other image geometries are also possible, in which case, the axes about which measurements are taken are also perpendicular to the images' planes. Figure 4.1 depicts a typical data measurement geometry.

In modern non-diffracting tomography (*e.g.*, all the tomographic methods referred to thus far), Radon-transform, algebraic, and iterative-based reconstruction algorithms are the main general approaches to image formation. Radon-transform based methods are almost exclusively used in nuclear medicine because algebraic and iterative reconstructions generally are more computationally intensive. One very popular and efficient algorithm for carrying out the reconstruction process is the *filtered-backprojection* (FBP) method, which in effect implements the inverse Radon

transform [74].

4.2 The Image Reconstruction Problem

A typical data gathering geometry for parallel-scanning tomography in SPECT is depicted in Figure 4.2. In nuclear medicine, the intensity $f(\mathbf{x})$ represents the concentration of radioactive material at a point $\mathbf{x}$ in space, and is typically the quantity of interest. $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ are orthogonal unit vectors which together define the *projection* plane on which the data collection process takes place. The plane contains the detector array, positioned according to the vector $\boldsymbol{\theta}$. The convergence point of a projection line (ray) on the plane is identified by s, the number of units along $\boldsymbol{\theta}$. Thus, we speak of the projection of f at $s\boldsymbol{\theta}$. The vector $\mathbf{x}$ and the projection rays lie on the $\boldsymbol{\eta}$-$\boldsymbol{\theta}$ plane.

The geometry in Figure 4.2 depicts a data collection process in which the object, and not the array of sensors, is rotated in order to capture the projection data. The two schemes are equivalent as far as the processing of the data is concerned, however, our choice simplifies some of the notation. During the process, the object f is rotated through angles $-\theta$ about $\boldsymbol{\zeta}$ in the right coordinate frame $(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\zeta})$[2].

The scanning process as just described is an idealization because it models a process that originates from zero volume of active material—the material "contained" in the $\boldsymbol{\eta}$-$\boldsymbol{\theta}$ plane. In order for the model to be meaningful, f must be interpreted as the average concentration of radioactive material within a neighborhood of this plane: let f_{vol} represent the true volumetric density of this material, and let a be the

[2]Note the slightly overloaded use of the theta symbol: in bold phase, it stands for the array of sensors' unit vector, and in plane phase, it denotes the relative rotation of this unit vector and the object. The distinction is clear and should not cause any problem.

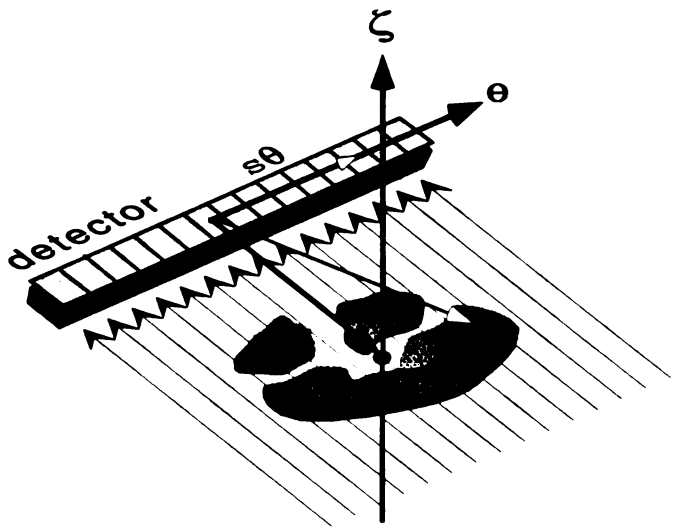


Figure 4.2. *Projection geometry for the intensity $f(\mathbf{x})$: The unit vector $\boldsymbol{\theta}$ defines the direction of the detector array on a plane perpendicular to $\boldsymbol{\zeta}$, while s identifies a sensor within the array. $f(\mathbf{x})$ is the density of radiopharmaceutical at $\mathbf{x}$.*

aperture gain (loss) of any one detector's collimator, then

$$f(\mathbf{x}, s) = \int \frac{a(\mathbf{x} - s\boldsymbol{\theta} + t\boldsymbol{\zeta})}{a(\mathbf{x} - s\boldsymbol{\theta})} f_{\text{vol}}(\mathbf{x} + t\boldsymbol{\zeta})\, dt$$

is the effective density of radiopharmaceutical which is responsible for the activity sensed by the detector at $s\boldsymbol{\theta}$. If, indeed, this density is a function of s as indicated, it would prove the poor selectivity of each collimator's aperture, resulting in blurring effects. In general, however, the aperture for rectangular collimators may be expressed as the product of the in-plane gain a_1, and the gain a_2 due to the displacement in the $\boldsymbol{\zeta}$-direction, *i.e.*, $a(\mathbf{x} - s\boldsymbol{\theta} + t\boldsymbol{\zeta}) = a_1(\mathbf{x} - s\boldsymbol{\theta})\, a_2(t\boldsymbol{\zeta})$. Thus,

$$f(\mathbf{x}) = \int a_2(t\boldsymbol{\zeta})\, f_{\text{vol}}(\mathbf{x} + t\boldsymbol{\zeta})\, dt \tag{4.1}$$

Clearly, the narrower the support for a_2, the finer the detail conveyed by f in the $\boldsymbol{\zeta}$-direction, however, at the expense of lower intensity and greater susceptibility to

noise.

As photons travel towards the array of detectors from within the subject, many are absorbed by nuclear interactions with the surrounding matter and only a fraction of the original number survive to be registered and counted. Let $\mu(\mathbf{z})\,dt$ be the probability that a photon reaching the point $\mathbf{z}$ will be absorbed within dt units from $\mathbf{z}$. Then, the number of photons $c_\theta(s)$ that during a time interval T reach the detector at $s\boldsymbol{\theta}$ due to the radioactivity at a point $\mathbf{x}$ with the object rotated an angle $-\theta$ is Poisson distributed:

$$c_\theta(s)|\lambda_\theta(s) \sim \text{Poisson}(\lambda_\theta(s)),$$

where the underlying intensity's projection $\lambda_\theta(s)$ is given by[3]

$$\lambda_\theta(s) = \int a(\mathbf{x} - s\boldsymbol{\theta}) f(\mathrm{R}_\theta(\mathbf{x})) e^{-\int_1^0 \mu(\mathrm{R}_\theta(t\mathbf{x}+(1-t)s\boldsymbol{\theta}))\,dt}\,d\mathbf{x}. \tag{4.2}$$

f is as defined in (4.1), μ is known as the *absorption loss* coefficient, $\mathrm{R}_\theta(\mathbf{x}) = e^{\mathbf{K}\theta}(\mathbf{x} - \boldsymbol{\eta}) + \boldsymbol{\eta}$, and $\mathbf{K} = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$. The time units have been chosen so that $T = 1$.

Since $e^{\mathbf{K}\theta}$ is the transformation that rotates a vector by an angle θ about $\boldsymbol{\zeta}$, $\mathrm{R}_\theta(\mathbf{y})$ transforms the vector $\mathbf{y}$ by rotating its $\mathbf{y} - \boldsymbol{\eta}$ component by this angle. Thus, $f(\mathrm{R}_\theta(\mathbf{x}))$ is a measure of the radioactivity at $\mathbf{x}$ after the object has been rotated by the angle $-\theta$ about $\boldsymbol{\zeta}$. The exponential $e^{\int_0^1 \mu(\mathrm{R}_\theta(t\mathbf{x}+(1-t)s\boldsymbol{\theta}))\,dt}$ gives the total absorption loss for the trajectory through the rotated object from $\mathbf{x}$ to $s\boldsymbol{\theta}$.

Recovering f from the set of projections $\lambda_\theta(s)$ represents a very difficult problem. There exists, however, an exact solution to (4.2) when the attenuation function μ is constant on a convex region that includes the support of f (assumed compact), and if the aperture gain at $s\boldsymbol{\theta}$ is zero everywhere off the perpendicular to $\boldsymbol{\theta}$ at $s\boldsymbol{\theta}$ [75].

[3]Since all but the unit vector $\boldsymbol{\zeta}$ considered here lie on the $(\boldsymbol{\eta}, \boldsymbol{\theta})$ plane, we restrict the notation to a two-coordinate format from now on.

Since in practice the absorption coefficient is generally unknown, it is common to ignore it and compensate for its effect by postprocessing the reconstructed image. This approach is computationally intensive and leads to suboptimal results as the effect from absorption is nonlinear and, consequently, cannot be fully accounted for independently of the reconstruction process [76]. For this reason, in applications in which the value of μ is small, it is often ignored altogether. For the remaining of this chapter, we only consider examples where it is safe to do so.[4] For this simplified model the solution is

$$f(\mathbf{x}) = \frac{1}{2\pi^2} \int_0^{\pi} \int \frac{\lambda'_{\theta}(s)}{(\mathbf{x} - \boldsymbol{\eta})^{\mathsf{T}} e^{\mathbf{K}\theta} \boldsymbol{\theta} - s} \, ds \, d\theta. \tag{4.3}$$

The superscript $^{\mathsf{T}}$ denotes transposition, and $\lambda'_{\theta}(s)$ stands for the derivative $\frac{d\lambda_{\theta}(s)}{ds}$. This expression is one incarnation of the celebrated inverse Radon transform and is the basis for the FBP reconstruction technique—see Section 4.3 and [75].

The assumption of having only those photons counted that strike the detector array at right angles implies the decoupling of the counting processes taking place among the sensors. This is an idealization that is invariably made in practice, but which results in blurred images. Much of this blurring, however, can be removed by high-pass filtering if noiseless estimates for the projections λ_{θ} are available. Since the actual sensors' aperture gain a acts as a spatial low-pass filter on the projections, the linear shift-invariant operation may be reversed by operating on the reconstructed image given that the inverse Radon transform and convolution operators commute. For simplicity's sake, we ignore blurring effects in the examples to follow and concentrate on the most detrimental factors in nuclear medicine tomography imaging: data variability due to its Poisson statistics.

[4] In Section 4.7, however, we show how to correct for the lack of symmetry in the effective attenuations associated with $\lambda_{\theta}(s)$ and $\lambda_{\theta+180^{\circ}}(s)$ when present.

ECT methods aim to recover the function f from a finite set of measurements $c_\theta(s)$. For example, for K detectors in the array and N projections, $s = -\frac{K}{2}, \ldots, \frac{K}{2} - 1$ and $\theta = 0, \frac{2\pi}{N}, \ldots, (N-1)\frac{2\pi}{N}$. Sometimes it is helpful to work with a normalized length for the array of sensors. We choose a length of 2 length-units per aperture (per array of sensors) so that now, $s = -1, -1 + \frac{2}{K}, \ldots, 1 - \frac{2}{K}$. Also, in the interest of brevity, we use the alternate indexing c_k^n, which denotes the counts from projection angle $\theta = n\frac{2\pi}{N}$ and sensor $s = \frac{2}{K}k - 1$. Here, $n = 0, \ldots, N-1$ and $k = 0, \ldots, K-1$, where it is always assumed $K = 2^J$ for some positive integer J.

The data vectors $\mathbf{c}^n \equiv [c_{K-1}^n, \ldots, c_0^n]^T$ arranged in matrix form $\mathbf{c} \equiv [\mathbf{c}^0, \ldots, \mathbf{c}^{N-1}]$ constitutes the *data sinogram.* The corresponding *intensity sinogram* is given by $\boldsymbol{\lambda} \equiv [\boldsymbol{\lambda}^0, \ldots, \boldsymbol{\lambda}^{N-1}]$, where each column vector is similarly defined as $\boldsymbol{\lambda}^n \equiv [\lambda_{K-1}^n, \ldots, \lambda_0^n]^T$. Note that

$$\lambda_k^n = \int_{2k/K-1}^{2(k+1)/K-1} \lambda_{\frac{2\pi}{N}n}(s)\, ds, \tag{4.4}$$

the highest resolution unnormalized Haar scaling coefficient at position (sensor) k of the projection at $\theta = \frac{2\pi}{N}n$. A similar relation holds for c_k^n, which is the total photon count in the $(\frac{2k}{K} - 1, \frac{2(k+1)}{K}K - 1)$ interval.

Each set of K measurements in a projection constitute a finite sampling of (4.2), and cannot strictly satisfy the Nyquist sampling requirement for the object f is of finite extent, and therefore, of infinite frequency support. Consequently, the quality of the output of any method that one may devise to manipulate the data is compromised. Since there exists no recourse to amend this common degradation, we assume the "K-sampling" to be fine enough that all high frequency features of interest may be reconstructed without distortion.

4.3 The Filtered-Backprojection Reconstruction Technique

The FBP method for inverting the Radon transform of an image may be deduced using the Fourier Slice Theorem [77]. Using two-dimensional polar notation for the object function, *i.e.*, writing $f(r,\phi)$ instead of $f(\mathbf{x})$,[5] and similarly for its Fourier transform, this theorem simply states that $\hat{f}(\omega,\theta) = \hat{\lambda}_\theta(\omega)$, that is, the Fourier Transform $\hat{f}$ may be written in terms of the projections' Fourier transforms.

Since the Fourier inversion formula can also be expressed as $f(r,\phi) = \frac{1}{4\pi^2}\int_0^\pi \int_{-\infty}^\infty \hat{f}(\omega,\theta)\,|\omega|\,e^{i\omega r\cos(\theta-\phi)}\,d\omega\,d\theta$, a simple substitution leads to

$$f(r,\phi) = \frac{1}{2\pi}\int_0^\pi \mathcal{T}\lambda_\theta(r\cos(\theta-\phi))\,d\theta, \tag{4.5}$$

where $\mathcal{T}\lambda_\theta(s) \equiv \frac{1}{2\pi}\int_{-\infty}^\infty \hat{\lambda}_\theta(\omega)\,|\omega|\,e^{i\omega s}\,d\omega$. This transformation represents a high-pass filter operation on the projections λ_θ. The frequency response of the filter is $|\omega|$. The integration (4.5) suggests that $f(r,\phi)$ is the result of a projection operation on $\mathcal{T}\lambda_\theta$, thus, the name filtered-backprojection method. This projection, however, is not equivalent to the projection process by which the set of λ_θ's are formed, for some non-linear weighting is implicitly applied to $\mathcal{T}\lambda_\theta$, but we shall not elaborate on this.

Practical Radon inversion algorithms often begin by discretizing the expression in (4.5). Although radically distinct in nature to algebraic approaches (ART), these methods may become similar in form [78]; but, most often, (4.5) is implemented by means of the FFT for computational efficiency reasons. When this is done, band-pass filters are chosen to replace the high-pass filter so that excessive noise amplification

[5]The various parameters relate as follows. Referring to expression (4.3), $(\mathbf{x}-\boldsymbol{\eta})^{\mathrm{T}}e^{\mathbf{K}\boldsymbol{\theta}}\boldsymbol{\theta}$ is a dot product of the vectors $\mathbf{x}-\boldsymbol{\eta}$ and $e^{\mathbf{K}\boldsymbol{\theta}}\boldsymbol{\theta}$. Their magnitudes are $|\mathbf{x}-\boldsymbol{\eta}| = r$ and 1, respectively. Thus, $(\mathbf{x}-\boldsymbol{\eta})^{\mathrm{T}}e^{\mathbf{K}\boldsymbol{\theta}}\boldsymbol{\theta} = r\cos(\nu)$, for some angle ν, and $\phi = \theta - \nu$.

will not occur past the object's main frequency band. These filters by necessity represent a compromise between recovering high resolution details and minimizing high-frequency noise.

4.4 Some Limiting Aspects of Conventional Reconstruction Methods

Classical reconstruction methods applied to SPECT produce undesirable and highly variable image reconstructions due to the Poisson statistics of the projection data. In practice, it is common to postprocess the raw reconstructions with a low-pass smoothing filter to improve the images. This approach often produces visually appealing results, but can lead to detrimental loss in resolution and fine detail structure. More advanced methods estimate the intensity sinogram $\boldsymbol{\lambda}$ from the noisy data. For example, Wiener filtering [79] approaches have been applied to the projections prior to reconstruction. Furthermore, (fully) Bayesian approaches to this problem have been proposed based on Markov random field models, *e.g.*, [80, 81]. However, these methods are very computationally expensive.

One possible remedy to these methods' complexity consists of simply estimating the individual intensity projections $\boldsymbol{\lambda}^n$ from the set of counts, and then applying an inversion method (*e.g.*, FBP) to obtain an approximation to the function f. This method is widely used in ECT [69, 75, 77], and in practice, the estimation is carried out by simply low-pass filtering the data. This approach, although straightforward, is highly suboptimal for the following reasons:

1. Due to the Poisson distribution of the data, the noise is signal dependent, and so, the linear filtering process cannot optimally remove it everywhere—see Chapter 3.

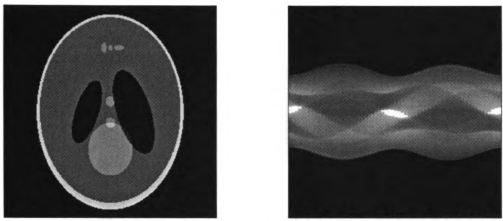
2. High frequency intensity details in the underlying signal projections are as attenuated as the noise.

3. Useful information that may exist in adjacent and nearby data projections is not exploited towards leveraging the estimation process.

4. The greater residual noise in the projection estimates due to points 1 and 3 leads to a degraded performance of any subsequent restoring operation, including the inverse Radon transformation. For example, a deblurring filter (essentially a high-pass filter) will emphasize noise in high-frequency bands.

One very simple approach to improving the quality of the estimates for the underlying intensity's projections is to impose the 1-d MMI model on each individual projection, and estimate accordingly. Doing this, however, induces the strong condition that any two consecutive projections $\boldsymbol{\lambda}^n$ and $\boldsymbol{\lambda}^{n+1}$ be independent. Since the actual projections' estimates are driven by the corresponding data $\mathbf{c}^n$ and $\mathbf{c}^{n+1}$, and since these data projections are correlated, the effect of this assumption is lessened. Nevertheless, given that in nuclear medicine the data's SNR is typically low, the unrealistic prior model tends to exert a strong influence on the estimator's outcome. Our own studies indicate that this does in fact occur, with the consequence of having highly detrimental circular artifacts appearing in the reconstructed images.

A step towards correcting for the projections' independence condition consists of imposing the 2-d MMI model on an image in the projection space (*i.e.*, sinogram.) In the following sections, we introduce two extensions to the basic 2-d MMI modeling approach introduced in the last chapter, and which lead to estimators that produce well-coordinated estimates of the intensity sinogram and improved reconstructed intensity image.



(a) (b)

Figure 4.3. *Shepp-Logan head phantom. (a) Phantom intensity. (b) Phantom intensity sinogram.*

4.5 First Approach to Multiscale Modeling and Estimation of ECT Intensities

One important characteristic of a sinogram $\boldsymbol{\lambda}$ is that point sources in the original object f correspond to sinusoids in the sinogram (thus, its name); each point corresponds to a unique amplitude-phase pair, but all have the same period. As a result, a very strong correlation exists among consecutive projections. This phenomenon is clearly displayed in Figure 4.3(b), which is the intensity sinogram of the Shepp-Logan head phantom image in Figure 4.3(a).

Correlation among neighboring pixels suggests that a MMI-like model may be applicable to the intensity sinograms as they appear to be characterized by smooth intensity variations with a few discontinuities. To understand the mechanism by which this characteristic emerges in sinograms corresponding to intensity images that are themselves well-modeled by the MMI paradigm, the concept of *natural pixels* (NP) is helpful. Originally developed for the incomplete data tomographic problem [78, 82], NPs give us the means to visualize the relationship of corresponding samples

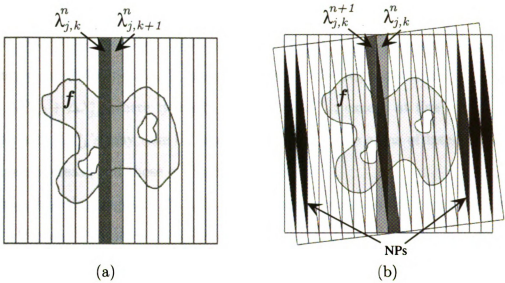


Figure 4.4. *Natural Pixels. (a) Projection samples at scale j are obtained by integrating the object over strips 2^j "sensors" wide. Thus, the highest resolution achievable is determined by the sensors' apertures, and correspond to samples at scale $j = 0$. (b) Natural pixels are delimited by the overlap of the integration strips corresponding to the same sample position k at two consecutive projection angles n and $n + 1$.*

in consecutive projections. We use this newly gained insight to extend the basic MMI approach of Chapter 3 to the ECT modeling and estimation problem.

4.5.1 The SI-MMI Method

In emission tomography, sinograms are 2-d positive functions, and so we may represent them at multiple scales just as we did the intensity images of Chapter 3. As before, the representation at scale j is denoted by $\boldsymbol{\lambda}_j$, where for $j = 0$, $\boldsymbol{\lambda}_0 \equiv \boldsymbol{\lambda}$. Clearly, $\boldsymbol{\lambda}_j$ may be obtained directly from the object f by simply projecting through wider swaths.

The set of projection strips belonging to two consecutive projections $\boldsymbol{\lambda}^n_j$ and $\boldsymbol{\lambda}^{n+1}_j$ at scale j are shown in Figure 4.4(b). The overlap of the two samples $\lambda^n_{j,k}$ and $\lambda^{n+1}_{j,k}$ delimit the natural pixel $\mathrm{NP}^n_{j,k}$ shown at the center of the figure. It is clear

that, under the smooth-intensity-variation-with-a-few-discontinuities assumption, the greatest contribution of the object to these samples comes from the $NP^n_{j,k}$ region, and only a small fraction $\Delta\lambda^n_{j,k}$ is contributed by the regions outside it. That is, $\lambda^n_{j,k} = NP^n_{j,k} + \Delta\lambda^n_{j,k}$ where typically $NP^n_{j,k} >> \Delta\lambda^n_{j,k}$; except at fine scales, where the residuals $\Delta\lambda^n_{j,k}$'s may take on significant values as the areas of integration they encompass increase and those of the NP's decrease.

In the object's sinogram, $\lambda^n_{j,k}$ and $\lambda^{n+1}_{j,k}$ appear as two horizontally adjacent pixels at the k^{th} row. The perturbation δ associated with these pixels can be written in terms of the natural pixel they define:

$$\delta = \frac{\Delta\lambda^n_{j,k} - \Delta\lambda^{n+1}_{j,k}}{\Delta\lambda^n_{j,k} + \Delta\lambda^{n+1}_{j,k} + 2NP^n_{j,k}} \approx 0. \tag{4.6}$$

This shows that the "horizontal" perturbations will tend to be concentrated around zero, and given the symmetry of the NPs, these will be symmetrically distributed as well. Note that for the perturbation to take on an extreme value of 1 or -1, it is necessary that either $\Delta\lambda^n_{j,k}$ or $\Delta\lambda^{n+1}_{j,k}$, but not both, assume the -NP value; a very unlikely situation as demonstrated by the geometry of integration.

Perturbation variates corresponding to differences between vertically adjacent samples in the intensity sinogram also behave in the desired manner. This is because the projections themselves enhance the smoothness characteristic of the original object through the integration process (see Fig. 4.4(a)). Specifically, since a projection sample is simply a scaled value of an average taken over an integration strip, a projection sample represents, within a fixed factor, the "typical" intensity[6] that the object assumes in the integration strip. Therefore, the perturbation δ associated with any two typical and adjacent samples of f behaves in similar manner to those perturbations associated with the object itself. Histograms for perturbations

[6]Note that these "typical" values need not ever occur.

between vertically adjacent pixels in synthetic and clinical data have verified this general behavior, and resemble the histograms in Figure 3.3.

The behavior of the perturbation (innovation) variates according to the desired prior does not on itself warrant the suitability of the MMI model, for the intensity variation in sinograms is arranged in a highly structured fashion, and so, the assumed independence of the innovations in the model may not apply. We have taken a practical approach and have sought to empirically justify the model's applicability. The steps taken are:

1. Form sinogram $\mathbf{c}$ from raw projection data $\{\mathbf{c}_k^n\}$.

2. Obtain optimal estimate $\widehat{\boldsymbol{\lambda}}$ of underlying intensity sinogram $\boldsymbol{\lambda}$ from data sinogram $\mathbf{c}$ using the 2-d SI-MMI-model based estimator of Chapter 3.

3. Reconstruct intensity image estimate $\widehat{f}$ from sinogram estimate $\widehat{\boldsymbol{\lambda}}$ by the FBP method.

We call this the sinogram-image-SI-MMI-model-based filtered-backprojection reconstruction method, or the SI-MMI approach to ECT, for short.

4.5.2 Example

To demonstrate the performance of the new approach to SPECT, we have applied it to two sets of data: the Shepp-Logan head phantom and real data from a human pelvic clinical study. The two data sets have been processed in three different ways so that comparisons can be made and a relative degree of performance can be determined. In Figure 4.5(b), an (unprocessed) reconstruction of the phantom data based on Poisson data projections is presented. Figure 4.5(c) shows a lowpass filtered[7] version of Fig. 4.5(b). The smoothing process removes high frequency noise at the expense

[7] A 2-d Butterworth filter with cut-off $\pi/3$ rad. was used in both examples of this section.

of some loss of high definition detail in the image. Finally, Figure 4.5(d) displays the reconstructed image using the new sinogram estimation method. In this new image, high definition features (*e.g.*, edges) are clearly preserved despite the high degree of noise reduction.

Figures 4.6(b), (c), and (d) of the pelvic bone study were obtained by the same processes used in the corresponding phantom figures. Note that in Figure 4.6(d), the high definition image structure becomes evident after processing according to the new approach. This is in contrast with the result in Figure 4.6(c), where the image is seen to be oversmoothed to achieve a similar degree of noise removal.

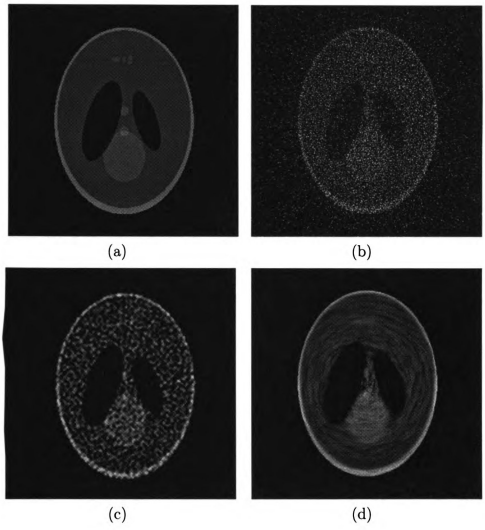


Figure 4.5. *Shepp-Logan head phantom image reconstruction. (a) Phantom. (b) Phantom reconstruction from noisy data. (c) Lowpass filtered image from (b). (d) Phantom reconstruction from MMI-processed sinogram.*

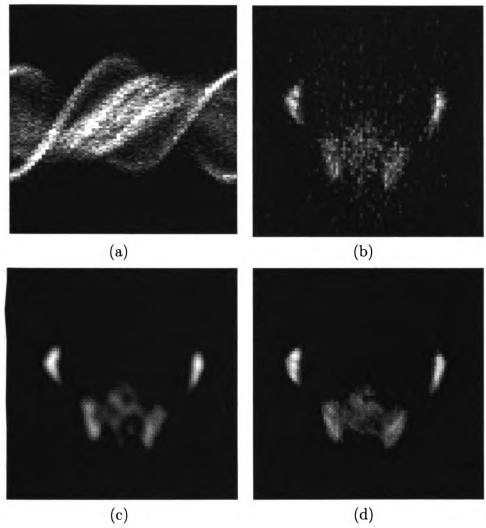


Figure 4.6. *Pelvic bone study image reconstruction. (a) Sinogram of pelvic bone. (b) Pelvic bone reconstruction from noisy data. (c) Lowpass filtered image from (b). (d) Pelvic bone reconstruction from MMI-processed sinogram.*

4.6 A New Multiscale-Based Tomographic Inversion Method

Despite the substantial improvements in quality and fidelity of reconstruction that the SI-MMI approach represents, we recognize that the always present sinusoidal structure of sinogram images has not been fully exploited by this method, when it could be used to leverage the reconstruction process. The second method to ECT to be introduced in Section 4.7 is based on the tomographic reconstruction approach developed here and which itself evolves from considerations of this sinusoidal extructure. As a result, the new inversion formula reconstructs the desired images not from their corresponding sinograms, but from their cumulative sinogram intensities. This will be especially important in the ECT problem, where for low SNR data sinograms the high variability of the data significantly undermines this structural constraint in sinograms. In Section 4.6.3 a fast implementation algorithm for the new Radon-inverse transform is also presented.

There exist other multiscale-base tomographic reconstruction techniques [83, 84, 85, 86], but none is statistically motivated and offers no advantage to the ECT reconstruction problem.

4.6.1 The Multiscale Reconstruction Formula

We begin with the inverse Radon transform (4.5), which may also be expressed in terms of the Hilbert transform[8] of the derivative $\lambda'_\theta(s)$ of the intensity projections:

$$f(r,\phi) = \frac{1}{2\pi}\int_0^\pi \mathcal{H}\lambda'_\theta(r\cos(\theta-\phi))\,d\theta. \tag{4.7}$$

[8]See Section C.1 of Appendix C for the definition of the Hilbert transform and some alternate representations.

This

the

riat

λ_6'

σ_1

in

wh

and

In

guarant

of finite

when app

are met.

Since g

with $\hat{g}_i \equiv l_i$

This is equivalent to (4.3). For simplicity, and without loss of generality, we assume the support of f to be contained in a disc of radius $r_{\max} < 1$. Additionally, differentiation of the intensity projections are assumed to exist everywhere in $[-1, 1]$, with $\lambda'_\theta(1) = \lambda'_\theta(-1) = 0$. Differentiation at the boundaries are of left and right type.

In Section C.1 we prove that $\mathcal{H}\lambda'_\theta(s) = -\frac{1}{\pi}\int_0^1 \{\lambda'_\theta(\sigma_2(\tau)) - \lambda'_\theta(\sigma_1(\tau))\}\frac{d\tau}{\tau}$ with $\sigma_1(\tau) = (-1 - s)\,\tau + s$ and $\sigma_2(\tau) = (1 - s)\,\tau + s$. Substituting this into (4.7) results in

$$f(r, \phi) = \frac{-1}{2\pi^2}\int_0^1 g(r, \phi; \tau)\frac{d\tau}{\tau} \tag{4.8}$$

where

$$g(r, \phi; \tau) \equiv \begin{cases} \int_0^\pi \{\lambda'_\theta(t(\tau)) - \lambda'_\theta(t(-\tau))\}\, d\theta, & \text{for } -1 \le \tau \le 1 \\ 0, & \text{otherwise} \end{cases} \tag{4.9}$$

and $t(\tau) \equiv \tau + (1 - |\tau|)\, r\cos(\theta - \phi)$ for all τ in $\mathbb{R}$.

In arriving at (4.8) we have made use of the Fubini-Tonelli Theorem [28], which guarantees the interchangeability of the integration operations as we assume f to be of finite energy. In the rest of this chapter, however, we generally make no special note when appealing to this theorem, and give as understood that the required conditions are met.

Since g is real and odd in τ, its Fourier transform $\hat{g}$ is imaginary and odd. Then, with $\hat{g}_i \equiv \mathrm{Im}\{\hat{g}\}$,

$$g(r, \phi; \tau) = \frac{-1}{2\pi}\int_{-\infty}^{\infty} \hat{g}_i(r, \phi; \omega)\sin(\omega\tau)\, d\omega, \tag{4.10}$$

and

$$\hat{g}_i(r,\phi;\omega) = -2\int_0^1 g(r,\phi;\tau)\sin(\omega\tau)\,d\tau. \tag{4.11}$$

Thus, (4.10) into (4.8) proves that

$$\begin{aligned} f(r,\phi) &= \frac{1}{4\pi^3}\int_{-\infty}^{\infty}\hat{g}_i(r,\phi;\omega)\int_0^1 \frac{\sin(\omega\tau)}{\tau}\,d\tau\,d\omega \\ &= \frac{1}{4\pi^3}\int_{-\infty}^{\infty}\hat{g}_i(r,\phi;\omega)\,\mathrm{Si}(\omega)\,d\omega, \end{aligned} \tag{4.12}$$

where $Si(\omega)$ denotes the sine integral.

Since $t(\tau)$ is differentiable everywhere except at $\tau = 0$,

$$\begin{aligned} \lambda_\theta'(t(\tau)) &= \frac{(\lambda_\theta \circ t)'(\tau)}{t'(\tau)} \\ &= \frac{(\lambda_\theta \circ t)'(\tau)}{1-\mathrm{sgn}(\tau)\,r\cos(\theta-\phi)} \end{aligned} \tag{4.13}$$

for $0 < |\tau| \leq 1$. Note that since $|r| < 1$, the ratio is well defined. Using (4.13) in (4.9), and recognizing that $g(r,\phi;0) = 0$, gives us

$$g(r,\phi;\tau) = \int_0^\pi \left\{ \frac{(\lambda_\theta \circ t)'(\tau)}{1-\mathrm{sgn}(\tau)\,r\cos(\theta-\phi)} + \frac{(\lambda_\theta \circ t)'(-\tau)}{1+\mathrm{sgn}(\tau)\,r\cos(\theta-\phi)} \right\} d\theta$$

for $0 < |\tau| \leq 1$, and zero otherwise. Once this expression is used in (4.11), an integration by parts over τ results in

$$\hat{g}_i(r,\phi;\omega) = 2\int_0^\pi \left\{ \int_0^1 \frac{\omega\cos(\omega\tau)\,(\lambda_\theta \circ t)(\tau)}{1-r\cos(\theta-\phi)}\,d\tau + \int_0^1 \frac{\omega\cos(\omega\tau)\,(\lambda_\theta \circ t)(-\tau)}{1+r\cos(\theta-\phi)}\,d\tau \right\} d\theta.$$

This expression accounts for the fact that $(\lambda_\theta \circ t)(1) = (\lambda_\theta \circ t)(-1) = 0$. Here, $\circ$ denotes the composition operation.

We may express $\lambda_\theta \circ t$ separately in the intervals $[-1,0]$ and $[0,1]$ in terms of Haar wavelet expansions (see Chapter 2). With $\rho^-_{j,k}(r,\phi;\theta)$ and $\rho^+_{j,k}(r,\phi;\theta)$ denoting its wavelet coefficients corresponding to these intervals at scale j and shift k, the expansions are

$$(\lambda_\theta \circ t)(-\tau) = \sum_{j,k} \rho^-_{j,k}(r,\phi;\theta)\,\psi_{j,k}(\tau)$$
$$(\lambda_\theta \circ t)(\tau) = \sum_{j,k} \rho^+_{j,k}(r,\phi;\theta)\,\psi_{j,k}(\tau),$$

both for $\tau \in [0,1]$. Substituting these in the above integrals yields

$$\begin{aligned}\hat{g}_i(r,\phi;\omega) &= 2\int_0^\pi \left\{ \frac{\omega \sum_{j,k}\ \rho^+_{j,k}(r,\phi;\theta) \int_0^1 \psi_{j,k}(\tau)\cos(\omega\tau)\,d\tau}{1 - r\cos(\theta-\phi)} \right. \\ &\qquad \left. + \frac{\omega \sum_{j,k}\ \rho^-_{j,k}(r,\phi;\theta) \int_0^1 \psi_{j,k}(\tau)\cos(\omega\tau)\,d\tau}{1 + r\cos(\theta-\phi)} \right\} d\theta \\ &= 2\omega \sum_{\substack{j\le 0 \\ 0\le k\le 2^{-j}-1}} 2^{-j}\, Q_{j,k}(r,\phi) \int_0^1 \psi(2^{-j}\tau - k)\cos(\omega\tau)\,d\tau \\ &\qquad + 2\sin(\omega) \sum_{j\ge 1} 2^{-j}\, Q_{j,0}(r,\phi),\end{aligned} \tag{4.14}$$

where $-\omega \int_0^1 \psi(2^{-j}\tau - k)\cos(\omega\tau)\,d\tau = \sin(2^j k\omega) - 2\sin(2^j(k+1/2)\omega) + \sin(2^j(k+1)\omega)$ and

$$Q_{j,k}(r,\phi) \equiv 2^{j/2} \int_0^\pi \left\{ \frac{\rho^+_{j,k}(r,\phi;\theta)}{1 - r\cos(\theta-\phi)} + \frac{\rho^-_{j,k}(r,\phi;\theta)}{1 + r\cos(\theta-\phi)} \right\} d\theta. \tag{4.15}$$

The factor $2^{j/2}$ is required in this definition so that $Q_{j,k}$ is independent of scale for $j \ge 1$.

Now, (4.14) into (4.12) becomes

$$f(r,\phi) = \frac{1}{2\pi} \sum_{\substack{j\le 0 \\ 0\le k\le 2^{-j}-1}} \alpha_{j,k}\, Q_{j,k}(r,\phi) + \frac{1}{4\pi^2}\, Q_{1,0}(r,\phi), \tag{4.16}$$

where

$$\alpha_{j,k} \equiv \begin{cases} 7/2\pi & \text{for } j=0,\ k=0 \\ 3/\pi \cdot 2^{-2j} & \text{for } j<0,\ k=0 \\ 1/\pi \cdot \frac{-2^{-2j}}{k+3k^2+2k^3} & \text{for } j<0,\ 0<k<2^{-j}-1 \\ 1/2\pi \cdot \frac{2-3\cdot 2^j-4^j}{2-3\cdot 2^j+4^j}\, 2^{-j} & \text{for } j<0,\ k=2^{-j}-1. \end{cases} \tag{4.17}$$

Define the unnormalized Haar wavelet and scaling coefficients

$$\tilde{\theta}_{j,k}(r,\phi;\theta) \equiv \int_{t(2^j k)}^{t(2^j(k+1/2))} \lambda_\theta(t)\, dt - \int_{t(2^j(k+1/2))}^{t(2^j(k+1))} \lambda_\theta(t)\, dt \tag{4.18}$$

for $j \le 0$ and $0 \le k \le 2^{-j}-1$, and

$$\tilde{\lambda}_{0,0}(r,\phi;\theta) \equiv \int_{r\cos(\theta-\phi)}^{1} \lambda_\theta(t)\, dt. \tag{4.19}$$

These and the (normalized) wavelet coefficients in (4.15) are related according to

$$\rho^{+}_{j,k}(r,\phi;\theta) = \frac{2^{-j/2}\,\tilde{\theta}_{j,k}(r,\phi;\theta)}{1-r\cos(\theta-\phi)}$$

and

$$\rho^{-}_{j,k}(r,\phi;\theta) = \rho^{+}_{j,k}(r,\phi;\theta+\pi).$$

The last expression is due to the equalities $\lambda_\theta(-t) = \lambda_{\theta+\pi}(t)$ and $1+r\cos(\theta-\phi) = 1-r\cos(\theta+\pi-\phi)$, which hold for all intensity sinograms void of attenuation effects.

In Sect

mitigati

more co

for $j \leq$

Expre

struct the

the one co

Figure 4.7

the wavele

$\tilde{\theta}_{0,0}(r, \phi; \theta_1$

wavelet fun

Formula

on the tom

of projectio

other recons

sampling. C

appealing ap

agnostic mec

resolution rec

struction whi

task yet.

In Section 4.7, we make use of this equivalence to introduce a simple strategy for mitigating attenuation effects in real data. By assuming this condition, the following more compact notation for the coefficients $Q_{j,k}(r,\phi)$'s is possible:

$$Q_{j,k}(r,\phi) = \int_0^{2\pi} \frac{\tilde{\theta}_{j,k}(r,\phi;\theta)}{(1-r\cos(\theta-\phi))^2}\, d\theta \tag{4.20}$$

for $j \leq 0$ and $0 \leq k \leq 2^{-j}-1$, and

$$Q_{1,0}(r,\phi) = \int_0^{2\pi} \frac{\tilde{\lambda}_{0,0}(r,\phi;\theta)}{(1-r\cos(\theta-\phi))^2}\, d\theta. \tag{4.21}$$

Expression (4.16) together with (4.17), (4.20) and (4.21) give the means to reconstruct the intensity f in a scale-by-scale basis from the unnormalized wavelet (and the one coarse scale) coefficients (4.18) and (4.19) of the "warped" projections $\lambda_\theta \circ t$. Figure 4.7 depicts the wavelet $\psi_{0,0}$ superimposed on the projection λ_{θ_1}, as well as the wavelets $\psi_{-1,0}$ and $\psi_{-1,1}$ superimposed on the projection λ_{θ_2}. The coefficients $\tilde{\theta}_{0,0}(r,\phi;\theta_1)$, $\tilde{\theta}_{-1,0}(r,\phi;\theta_2)$ and $\tilde{\theta}_{-1,1}(r,\phi;\theta_2)$ are the coefficients associated with these wavelet functions.

Formula (4.16) may be used as a starting point for further theoretical analysis on the tomographic reconstruction process; however, since it requires a continuum of projections, it is of very limited practical use. As is conventionally done with other reconstruction methods, one may discretize formula (4.16) by straightforward sampling. Clearly, the reconstructions obtained in this manner will be a visually appealing approximation at best, but of poor reliability for delicate studies, *e.g.*, diagnostic medicine. In general, it seems very difficult to predict a priori the sampling resolution required to achieve a desired image quality, and to determine after reconstruction which features are real and which are artifacts, often proves to be a bigger task yet.

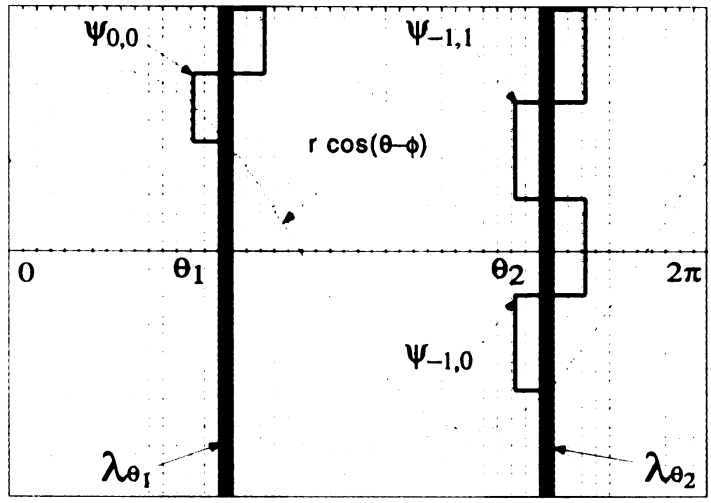


Figure 4.7. *Wavelet functions for image reconstruction. The contribution to the sinogram of an intensity image from a single point appears along a sinusoidal path. The point intensity can be determined from the unnormalized Haar wavelet coefficients of each projection. The associated wavelet functions are defined in intervals* $[r\cos(\theta-\phi), 1]$ *according to the* (r, ϕ) *location of the intensity point, and the projections' angles* θ*'s. Three such wavelet functions are depicted here over vertical axes.*

Below, we use the sampling theorem to represent the new reconstruction formula in terms of a finite number of projections. The resulting expression will also represent only an approximation. There exists, however, an important difference between our sampling approach and that of traditional methods used to account for the sampled nature of the data. In traditional methods, it is not only the data that is discretized by the sampling, but also the inverse Radon transform operator. This leads to errors on two counts: one, from the loss of information due to aliasing effects, and two, from the degradation of the conditions under which the inverse transformation holds true.

By using the sampling theorem to incorporate the sampled data into our otherwise perfect inverse transformation, the errors are limited to aliasing effects. However, for these types of errors there exist prefiltering steps that one can apply to the data to minimize their adverse effects.

Let $\Delta\theta \equiv 2\pi/N$ be the angular sampling interval. Then, assuming that only little

of the sinogram's frequency content in the θ direction falls above $N/4\pi$ cycles per radian (half the sampling frequency), we have

$$\tilde{\theta}_{j.k}(r,\phi;\theta) = \sum_{n=0}^{N-1} \tilde{\theta}_{j,k}(r,\phi;n\Delta\theta)\,\mathrm{Sa}\Big(\frac{\pi}{\Delta\theta}(\theta - n\Delta\theta)\Big)$$

and

$$\tilde{\lambda}_{0,0}(r,\phi;\theta) = \sum_{n=0}^{N-1} \tilde{\lambda}_{0,0}(r,\phi;n\Delta\theta)\,\mathrm{Sa}\Big(\frac{\pi}{\Delta\theta}(\theta - n\Delta\theta)\Big),$$

where $\mathrm{Sa}(x) \equiv \frac{\sin(x)}{x}$. Substituting these into (4.20) and (4.21) results in

$$Q_{j,k}(r,\phi) = \sum_{n=0}^{N-1} \tilde{\theta}_{j,k}(r,\phi;n\Delta\theta) \int_0^{2\pi} \frac{\mathrm{Sa}\big(\frac{\pi}{\Delta\theta}(\theta - n\Delta\theta)\big)}{(1 - r\cos(\theta-\phi))^2}\,d\theta \tag{4.22}$$

for $j \leq 0$ and $0 \leq k \leq 2^{-j} - 1$, and

$$Q_{1,0}(r,\phi) = \sum_{n=0}^{N-1} \tilde{\lambda}_{0,0}(r,\phi;n\Delta\theta) \int_0^{2\pi} \frac{\mathrm{Sa}\big(\frac{\pi}{\Delta\theta}(\theta - n\Delta\theta)\big)}{(1 - r\cos(\theta-\phi))^2}\,d\theta. \tag{4.23}$$

The integral $\int_0^{2\pi} \frac{\mathrm{Sa}\left(\frac{\pi}{\Delta\theta}(\theta-n\Delta\theta)\right)}{(1-r\cos(\theta-\phi))^2}\,d\theta$ defies a close-form solution, but an excellent approximation may be obtained if $N \geq 32$ and $r_{\max} = .96$. These conditions are easily met. In particular, studies often cite values $N \geq 64$, or even $N \geq 128$. The specification for $r_{\max}$ can always be met by simply padding the projection vectors with a few extra zeros. In fact, the condition for the number of projections N required can be further relaxed by simply making $r_{\max}$ yet smaller by this same procedure. For instance, if $N \geq 16$, an excellent approximation is achieved with $r_{\max} \leq .8$. In appendix C.2 we show under what conditions

$$\int_0^{2\pi} \frac{\mathrm{Sa}\big(\frac{\pi}{\Delta\theta}(\theta - n\Delta\theta)\big)}{(1 - r\cos(\theta-\phi))^2}\,d\theta \approx \frac{\Delta\theta}{(1 - r\cos(n\Delta\theta-\phi))^2} \tag{4.24}$$

of the singerup's frequency content in the k direction (the choice $\sqrt{3}/2$ comes from radian (half the sampling frequency), we have

$$\hat{\theta}^{\mathrm{Ny}}(r,\phi;\theta) = \sum_{n=0}^{N-1} \theta_n e^{i\pi r \cos(\phi - n\sqrt{3}\pi/N)}$$

and

$$\hat{\gamma}^{\mathrm{Ny}}(r,\phi;\theta) = \sum_{n=0}^{N-1} \gamma_n e^{i\pi r \cos(\phi - n\sqrt{3}\pi/N)}$$

where $\mathrm{Se}(x) \equiv \frac{\sin(x)}{x}$. Substituting these into (4.20) and (4.21) and

$$Q^{\mathrm{Ny}}_{\theta}(r,\phi) = \sum_{n=0}^{N-1} \theta_n \int_0^{2\pi} \mathrm{Se}\left(\frac{r}{2}(\phi - n\sqrt{3}\pi/N)\right) \frac{(1 - r\cos(\phi - n\sqrt{3}\pi/N))}{2} d\phi \tag{4.22}$$

for $\lambda \geq 0$ and $0 \leq k \leq 2\lambda - 1$, and

$$Q^{\mathrm{Ny}}_{\gamma}(r,\phi) = \sum_{n=0}^{N-1} \gamma_n \int_0^{2\pi} \mathrm{Se}\left(\frac{r}{2}(\phi - n\sqrt{3}\pi/N)\right) \frac{(1 - r\cos(\phi - n\sqrt{3}\pi/N))}{2} d\phi \tag{4.23}$$

The integral $\int_0^{2\pi} \mathrm{Se}\left(\frac{r}{2}(\phi - n\sqrt{3}\pi/N)\right) \frac{(1 - r\cos(\phi - n\sqrt{3}\pi/N))}{2} d\phi$ makes a nice form solution, but an excellent approximation may be obtained if $N \geq 32$ and $r_{\max} = .96$. There conditions are easily met. In particular, studies often use values $N \geq 64$, or even $N \geq 128$. The aperture for $r_{\max}$ can always be met by simply pushing the projection vectors with a low extra zeros. In fact, the condition for the number of projections N required can be further relaxed by simply making $r_{\max}$ not smaller by this same procedure. For instance, if $N \geq 16$, an excellent approximation is achieved with $r_{\max} \leq .8$. In appendix C.2 we show under what conditions

$$\int_0^{2\pi} \mathrm{Se}\left(\frac{r}{2}(\phi - n\sqrt{3}\pi/N)\right) \frac{(1 - r\cos(\phi - n\sqrt{3}\pi/N))}{2} d\phi \approx \frac{\sqrt{3}\pi}{N} \tag{4.24}$$

for $n = 0, 1, \ldots, N-1$. Since the approximation can always be made as good as needed, from now on we regard it as an equality.

Using (4.24) in (4.22) and (4.23), and in turn, substituting these into the reconstruction formula (4.16) we obtain

$$f(r,\phi) = \frac{1}{N}\sum_{n=0}^{N-1} \frac{\frac{1}{2\pi}\,\tilde{\lambda}^n_{0,0}(r,\phi) + \sum_{j\le 0}\sum_{0\le k\le 2^{-j}-1} \alpha_{j,k}\,\tilde{\theta}^n_{j,k}(r,\phi)}{(1 - r\cos(\phi - \frac{2\pi}{N}n))^2}, \tag{4.25}$$

where $\tilde{\theta}^n_{j,k}(r,\phi) \equiv \tilde{\theta}_{j,k}(r,\phi;\frac{2\pi}{N}n)$ and $\tilde{\lambda}^n_{0,0}(r,\phi) \equiv \tilde{\lambda}_{0,0}(r,\phi;\frac{2\pi}{N}n)$. That is,

$$\begin{aligned}\tilde{\theta}^n_{j,k}(r,\phi) = &\int_{2^j k+(1-2^j k)r\cos(\phi-\frac{2\pi}{N}n)}^{2^j(k+1/2)+(1-2^j(k+1/2))r\cos(\phi-\frac{2\pi}{N}n)} \lambda_{\frac{2\pi}{N}n}(t)\,dt \\ &- \int_{2^j(k+1/2)+(1-2^j(k+1/2))r\cos(\phi-\frac{2\pi}{N}n)}^{2^j(k+1)+(1-2^j(k+1))r\cos(\phi-\frac{2\pi}{N}n)} \lambda_{\frac{2\pi}{N}n}(t)\,dt\end{aligned} \tag{4.26}$$

for $j \le 0$ and $0 \le k \le 2^{-j} - 1$, and

$$\tilde{\lambda}^n_{0,0}(r,\phi) = \int_{r\cos(\phi-\frac{2\pi}{N}n)}^{1} \lambda_{\frac{2\pi}{N}n}(t)\,dt \tag{4.27}$$

(see (4.18) and (4.19)), and $\alpha_{j,k}$ as given in (4.17).

This formulation shows that the intensity image at any point (r,ϕ) may be obtained by filtering and adding the wavelet coefficients of each warped projection $\lambda_{\frac{2\pi}{N}n} \circ t$, and then averaging them together over the projection angles. One significant advantage of this method is that it bypasses altogether the need for the ramp filter $|\omega|$, characteristic of backprojection approaches. As already noted, these filters are known to enhance high-frequency noise in the projection data.

Another advantage of the proposed formulation surfaces when denoising and deblurring the projection data: multiscale-based filters and estimators that adapt to specific regions of the intensity image may be applied during the reconstruction process (4.25). This is a valuable feature not offered by conventional tomographic inver-

sion schemes. Since we are always free to coordinate any filtering done on consecutive projections, (4.25) also enables us to exploit the sinusoidal structure characteristic of sinograms. This could be useful when compensating against non-uniform attenuation effects, for example.

4.6.2 Computation of $\tilde{\theta}^n_{j,k}$ and $\tilde{\lambda}^n_{0,0}$

Formulas (4.26) and (4.27) do not lead to practical evaluation of the coefficients $\tilde{\theta}^n_{j,k}$ and $\tilde{\lambda}^n_{0,0}$ because the signals $\lambda_{\frac{2\pi}{N}n}$ are never available. Instead, in practical tomographic processes the set of values λ^n_k is produced. These are the integrals defined in (4.4), and represent samples of a scaled moving averaging process which may be described by

$$\begin{aligned}\lambda^n_t &\equiv \int_{2t/K-1}^{2(t+1)/K-1} \lambda_{\frac{2\pi}{N}n}(s)\,ds \\ &= \left(\mathrm{Sa}(\tfrac{\omega}{2})\, e^{-i\frac{K}{4}(K-1)\omega}\, \hat{\lambda}_{\frac{2\pi}{N}n}(\tfrac{K}{2}\omega)\right)^{\vee}\end{aligned} \tag{4.28}$$

for t in $[0, K-1]$. $(\cdot)^{\vee}$ denotes inverse Fourier transformation.

We would like to know under what conditions the sampling interval supports the "nominal" bandwidth[9] of the signals λ^n_t. This bandwidth could be no more than $\frac{\pi}{2}K$ radians per length-unit, for the sampling interval is $2/K$ length-units. In turn, this implies that the bandwidth for each of the signals $\lambda_{\frac{2\pi}{N}n}$ can be no more than $\frac{\pi}{4}K^2$ radians per length-unit, as dictated by (4.28). For example, if $K = 32$, the bandwidth of any $\lambda_{\frac{2\pi}{N}n}$ can be no more than 256π radians per length-unit, or equivalently, 256 cycles per aperture (per length of array of sensors).

These estimates have not accounted for the effect of the "envelope" $|\mathrm{Sa}(\frac{\omega}{2})|$ since its

[9]Strictly, no sampling interval could ever be small enough to satisfy the Nyquist criterion given that the projections are of bounded support. Nominal bandwidth refers to that which if "reconstructed" would represent an acceptable signal.

sion schemes. Since we are always free to coordinate any filtering done on consecutive projections, (4.25) also enables us to exploit the sinusoidal structure characteristic of sinograms. This could be useful when compensating various non-uniform attenuation effects, for example.

4.6.2 Computation of $\hat{\theta}^n_{\ell}$ and $\hat{\lambda}^n_{\ell}$

Formulas (4.26) and (4.27) do not lead to practical evaluation of the coefficients $\hat{\theta}^n_{\ell}$ and $\hat{\lambda}^n_{\ell}$ because the signals λ_{ξ_n} are never available. Instead, in practical tomographic processes the set of values λ^n_{ℓ} is produced. These are the integrals defined in (4.4), and represent samples of a scaled moving averaging process which may be described by

$$\lambda^n_{\ell} = \int [illegible] \lambda_{\xi_n}(\cdot)\,d\cdot$$

$$= \left(\mathrm{Sa}(\tfrac{\cdot}{2}) [illegible] \lambda_{\xi_n}\left(\tfrac{K}{2}\cdot\right)\right)^{\vee} \quad (4.28)$$

for ℓ in $[0, K-1]$. $(\cdot)^{\vee}$ denotes inverse Fourier transformation.

We would like to know under what conditions the sampling interval supports the "nominal" bandwidth[9] of the signals λ^n_{ℓ}. This bandwidth could be no more than $\frac{\pi}{2}K$ radians per length-unit, for the sampling interval is $2/K$ length-units. In turn, this implies that the bandwidth for each of the signals λ_{ξ_n} can be no more than $\frac{\pi}{4}K^2$ radians per length-unit, as dictated by (4.28). For example, if $K = 32$, the bandwidth of any λ_{ξ_n} can be no more than 256π radians per length-unit, or equivalently, 256 cycles per aperture (per length of array of sensors).

These estimates have not accounted for the effect of the "envelope" $|\mathrm{Sa}(\frac{\cdot}{2})|$ since its

[9]Strictly, no sampling interval could ever be small enough to satisfy the Nyquist criterion given that the projections are of bounded support. Nominal bandwidth refers to that which if "reconstructed" would represent an acceptable signal.

frequency contribution extends to infinity. However, beyond 64π radians per length-unit, its contribution, and therefore, that of $\lambda_{\frac{2\pi}{N}n}$, has been reduced by over 40 dB from its peak value at d.c. Consequently, the sampling need only satisfy the Nyquist requirement for, say, 64π radians per length-unit;[10] and this occurs for $K \geq 16$. Since this is the case for practically all tomographic studies of interest that we may be concerned with, we conclude that each λ_t^n can be recovered from the K samples λ_k^n. Thus,

$$\lambda_t^n = \sum_{k=0}^{K-1} \lambda_k^n \, \mathrm{Sa}\left(\tfrac{K\pi}{2}(t - \tfrac{2}{K}k)\right) \tag{4.29}$$

The projections $\lambda_{\frac{2\pi}{N}n}$ are finite in extent and so, we can think of them as consisting of zeros beyond the aperture's boundaries, if necessary. Then, any integration of $\lambda_{\frac{2\pi}{N}n}$ over an arbitrary interval $[\xi_1, \xi_2]$, where $\xi_1 \leq \xi_2$, can be computed as

$$\begin{aligned}
&\int_{\xi_1}^{\xi_2} \lambda_{\frac{2\pi}{N}n}(s)\, ds = \sum_{l=0}^{L-1} \left(\lambda^n_{\frac{K}{2}(1+\xi_1)+l} - \lambda^n_{\frac{K}{2}(1+\xi_2)+l}\right) \\
&= \sum_{l=0}^{L-1} \sum_{k=-\infty}^{\infty} \lambda_k^n \left\{ \mathrm{Sa}\Big(\pi\big[\big(\tfrac{K}{2}\big)^2(1+\xi_1) + \tfrac{K}{2}l - k\big]\Big) - \mathrm{Sa}\Big(\pi\big[\big(\tfrac{K}{2}\big)^2(1+\xi_2) + \tfrac{K}{2}l - k\big]\Big)\right\},
\end{aligned}$$

with $L = \lceil \frac{K}{2}(1 - \xi_1) \rceil$. Equivalently,

$$\begin{aligned}
\int_{\xi_1}^{\xi_2} \lambda_{\frac{2\pi}{N}n}(s)\, ds &= \sum_{l=0}^{L-1} \left(\hat{\lambda}^n_\omega \ \mathrm{rect}(\omega/2\pi) \left(e^{i\omega\left(\frac{K}{2}\right)^2(1+\xi_1)} - e^{i\omega\left(\frac{K}{2}\right)^2(1+\xi_2)} \right) \right)^{\vee} \Bigg|_{k=\frac{K}{2}l} \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{\lambda}^n_\omega \, \frac{\sin(\omega KL/4)}{\sin(\omega K/4)} \, e^{-i\frac{K(L-1)}{4}\omega} \left(e^{i\omega\left(\frac{K}{2}\right)^2(1+\xi_1)} - e^{i\omega\left(\frac{K}{2}\right)^2(1+\xi_2)} \right) d\omega,
\end{aligned}$$

where $\hat{\lambda}^n_\omega$ denotes the discrete-time Fourier transform of the extended projection

[10] Here, we assume that $\lambda_{\frac{2\pi}{N}n}$ has most of its energy concentrated below 64 cycles per aperture.

$(\cdots, 0, \lambda_0^n, \lambda_1^n, \cdots, \lambda_{N-1}^n, 0, \cdots)$. This expression leads to the desired result:

$$\int_{\xi_1}^{\xi_2} \lambda_{\frac{2\pi}{N}n}(s)\, ds = \left(\hat{\lambda}_\omega^n \frac{\sin(\omega KL/4)}{\sin(\omega K/4)}\right)^\vee \Big|_{k=\frac{K^2(1+\xi_1)-K(L-1)}{4}} - \left(\hat{\lambda}_\omega^n \frac{\sin(\omega KL/4)}{\sin(\omega K/4)}\right)^\vee \Big|_{k=\frac{K^2(1+\xi_2)-K(L-1)}{4}}. \quad (4.30)$$

Expression (4.30) allows us to compute the integral of $\lambda_{\frac{2\pi}{N}n}$ over any chosen interval by means of the fast Fourier transform. In general, the integration values obtained are only approximations, but this is of no concern for they may be made as accurate as necessary by simply extending the length of the FFT by zero padding.

The coefficients $\tilde{\theta}_{j,k}^n$ and $\tilde{\lambda}_{0,0}^n$ can, therefore, be computed accurately by applying (4.30) to (4.26) and (4.27). Unfortunately, (4.30) cannot be computed efficiently since the indicated inverse Fourier transformations must be recomputed for every change of the integration's lower limit ξ_1, as the parameter L is a function of it. Next, we introduce a new formulation that avoids the computation of these coefficients altogether, and which leads to a very efficient inversion algorithm.

4.6.3 A Fast Multiscale Radon-Inverse Transform Algorithm

Define the n^{th} *cumulative projection* as

$$\bar{\lambda}^n(s) \equiv \int_s^1 \lambda_{\frac{2\pi}{N}n}(t)\, dt \quad (4.31)$$

for $-1 \leq s \leq 1$, and every $n = 0, \ldots, N-1$, and the associated (normalized) quantity

$$\nu_{j,k}^n(r,\phi) \equiv \frac{\bar{\lambda}^n(2^j k + (1-2^j k)\, r\cos(\phi - \frac{2\pi}{N}n))}{\left(1 - \left[2^j k + (1-2^j k)\, r\cos(\phi - \frac{2\pi}{N}n)\right]\right)^2} \quad (4.32)$$

for $j \leq 0$ and $0 \leq k \leq 2^{-j} - 1$. Note that $\bar{\lambda}^n(r\cos(\phi - \frac{2\pi}{N}n)) = \tilde{\lambda}_{0,0}^n(r,\phi)$, or equivalently, $\left(1 - r\cos(\phi - \frac{2\pi}{N}n)\right)^2 \nu_{0,0}^n(r,\phi) = \tilde{\lambda}_{0,0}^n(r,\phi)$, with $\tilde{\lambda}_{0,0}^n(r,\phi)$ as defined in

(4.27).

The normalized wavelet coefficients in (4.26) corresponding to the warped projections $\lambda_{\frac{2\pi}{N}n} \circ t$ may now be expressed as

$$\frac{\tilde{\theta}^n_{j,k}(r,\phi)}{(1-r\cos(\phi-\frac{2\pi}{N}n))^2} = (1-2^j k)^2 \cdot \nu^n_{j,k}(r,\phi) \\ - (1-2^j(k+1/2))^2 \cdot 2\nu^n_{j,k+1/2}(r,\phi) + (1-2^j(k+1))^2 \cdot \nu^n_{j,k+1}(r,\phi).$$

Substituting this into (4.25), the following reconstruction formula results.

$$f(r,\phi) = \frac{1}{N}\sum_{n=0}^{N-1}\left\{\frac{1}{2\pi}\,\nu^n_{0,0}(r,\phi) + \sum_{j\le 0}\sum_{k=0}^{2^{-j}-1}\alpha_{j,k}\left[(1-2^j k)^2\,\nu^n_{j,k}(r,\phi)\right.\right. \\ \left.\left. - 2(1-2^j(k+1/2))^2\,\nu^n_{j,k+1/2}(r,\phi) + (1-2^j(k+1))^2\,\nu^n_{j,k+1}(r,\phi)\right]\right\}. \quad (4.33)$$

In practice, we are always interested in a discretized reconstruction formulation for computer implementation purposes, and so, the reconstructions are necessarily of finite resolution; thus, it suffices to work with only a finite number of scales. With $J \le 0$, we denote by f_J the intensity image reconstructed from the $-J$ coarsest scales in (4.33). The scale-limited inversion formula may be written in simpler form if we first gather the $\nu^n_{j,k}(r,\phi)$ terms with equal value for all allowed combinations of parameters j and k. For example, it is always true that $\nu^n_{-3,k}(r,\phi)\big|_{k=1} = \nu^n_{-2,k+1/2}(r,\phi)\Big|_{k=0} = \nu^n_{-3,k+1}(r,\phi)\big|_{k=0}$. Doing this yields

$$f_J(r,\phi) = \frac{1}{N}\sum_{n=0}^{N-1}\sum_{k=0}^{2^{-J+1}-1}(1-2^{J-1}k)^2\,\beta_{J,k}\;\nu^n_{J-1,k}(r,\phi) \quad (4.34)$$

where

$$\beta_{J,0} = 4^{1-J}/\pi \quad (4.35)$$

(4.27).

The normalized wavelet coefficients in (4.29) corresponding to the warped bivariate functions $\sqrt{2^{2j}}\,\phi \circ t$ may now be expressed as

$$\frac{\tilde{q}_{\lambda}^{2j}(v,\phi)}{(1-\varepsilon\cos(\phi-\frac{2\pi}{N}n))^2} = (1-2^{j}\lambda_{\varepsilon}\ldots$$

$$-(1-2^{j}(\lambda+1/2)\ldots$$

Substituting this into (4.28), the following expression formula results:

$$-3(1-2^{j}(\lambda+1/2)\ldots \tag{4.33}$$

In practice, we are always interested in a discretized interpolation formula for computer implementation purposes, and $\ldots$ discretizations are necessarily of finite resolution; thus, it suffices to work with only a finite number of scales. With $\lambda \geq 0$, we denote by λ_1 the interesting range corresponding to $\ldots$ one may scales in (4.33). The scale-limited inversion formula may be written in simpler form if we first gather the $\tilde{s}_{\lambda}^{2j}(v,\phi)$ terms with equal values for all allowed combinations of resolutions λ and k. For example, it is always true that $\tilde{s}_{\lambda}^{2j+1}(v,\phi)\big|_{\lambda=0} = \tilde{s}_{0}^{2j+1}(v,\phi)\big|_{\lambda=0}$. Doing this yields:

$$\ldots \tag{4.34}$$

where

$$\ldots \tag{4.35}$$

and, for $1 \leq k \leq 2^{-J+1} - 1$,

$$\beta_{J,k} = \begin{cases} -2\,\alpha_{J,(k-1)/2} & \text{for } k \text{ odd} \\ -2\,\alpha_{J+m^*,(2^{-m^*}k-1)/2} & \text{for } k \text{ even.} \\ \quad + \sum_{j=J}^{J+m^*-1} \alpha_{j,2^{-j+J-1}k} + \alpha_{j,2^{-j+J-1}k-1} & \end{cases} \tag{4.36}$$

Here, $m^* \equiv \max\{m \in \mathbb{N} \cup \{0\} \mid 2^{-m}k \in \mathbb{N}\}$.

The reconstruction formula (4.34) represents a fast multiscale inversion algorithm for the Radon transform, and to our knowledge, is the most efficient algorithm today. Its complexity is less than that of fast-Fourier-transform based approaches by a factor of $\log K$. This estimate assumes that all the coefficients are precomputed to reconstruction, an easily satisfied condition.

Figure 4.8 displays the general behavior of the coefficients $(1-2^{J-1})^2\,\beta_{J,k}$ used in (4.34). The plots correspond to the case of $J = -5$, but the fast decay behavior of the coefficients as a function of k is observed whenever $J < 0$. From Figure 4.8 (b), it is evident that at $k = 8$, the factor $(1-2^{J-1})^2\,\beta_{J,k}$ has been reduced by over 50 dB from its value at $k = 0$. This and the fact that the data $\nu^n_{J-1,k}(r,\phi)$ is non-increasing with k, justifies the simpler reconstruction formula

$$f_J(r,\phi) = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{k=0}^{k_0-1} (1 - 2^{J-1}k)^2\,\beta_{J,k}\;\nu^n_{J-1,k}(r,\phi), \tag{4.37}$$

where k_0 is a small positive integer, say, $k_0 = 8$. For values of J much smaller than -5, however, k_0 needs to be increased accordingly.

As mentioned before, an important advantage of the new formulation is that it allows us to selectively reconstruct an image at any desired scale supported by the data. Also, since the algorithm does not assume a discretized integration process in the data collection stage, it is better suited for real-world applications. In fact, our experiments have shown that for simulated data where the projections are the result

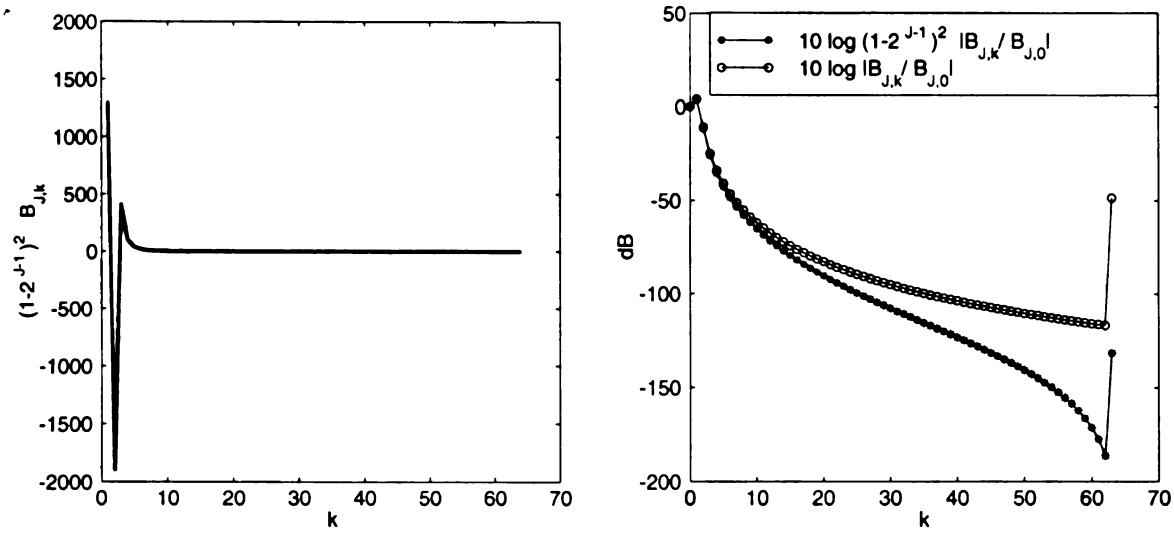


Figure 4.8. *Tomography reconstruction-formula coefficients. (a)* $(1 - 2^{J-1})^2\ \beta_{J,k}$ *versus* k. *(b)* $10\log|\beta_{J,k}/\beta_{J,0}|$ *and* $10\log(1 - 2^{J-1})^2\ |\beta_{J,k}/\beta_{J,0}|$ *versus* k. *In both cases (a) and (b),* $J = -5$. *These graphs reveal the rapid decay of the coefficients' magnitude as* k *increases.*

of a simple counting process, not a true integration operation, the new formulation does not perform as well as conventional discretized algorithms. Here, however, we are mostly interested in ECT applications, to which the fast tomographic inversion formula is especially well suited as we show next.

4.7 Second Approach to Multiscale Modeling and Estimation of ECT Intensities

The SI-MMI modeling approach to ECT applications relies on the proximity of adjacent pixels in the intensity sinogram. In Section 4.5, it was argued that such proximity typically exists in the vertical direction (same projection vector) due to the proximity often encountered among adjacent samples in the intensity image f, which when inte-

grated together, only reinforces this characteristic. On the other hand, the proximity of pixels in the horizontal direction was justified by the large overlap of the integration swaths corresponding to consecutive pixels. The regions of overlap defined the natural pixels.

We also noted that the inequality $\mathrm{NP}^n_{j,k} >> \Delta\lambda^n_{j,k}$, which validated (4.6), could not be justified for narrow integration swaths (high-resolution scale projections). Only as the increment angle between projections is significantly reduced, does the difference $\mathrm{NP}^n_{j,k} - \Delta\lambda^n_{j,k}$ become large. Therefore, the SI-MMI method can only be justified for low resolution projections when the angular sampling is also coarse. Certainly, in these cases, it would be suboptimal to have to work with coarse scale representations even though high-resolution projection data are available.

To circumvent this restriction, we introduce a new approach to modeling, estimating, and reconstructing the intensity image from its Poisson data sinogram.

4.7.1 The CSI-MMI Method

We define a discretized version of the cumulative projection (4.31) as $\bar{\boldsymbol{\lambda}}^n \equiv \left[\bar{\lambda}^n_{K-1}, \ldots, \bar{\lambda}^n_0\right]^T$, with elements $\bar{\lambda}^n_k \equiv \bar{\lambda}^n(\frac{2}{K}k - 1)$ for $k = 0, \ldots, K-1$. Then, we let the *cumulative sinogram image* (CSI) corresponding to these projections be $\bar{\boldsymbol{\lambda}} \equiv [\bar{\boldsymbol{\lambda}}^0, \ldots, \bar{\boldsymbol{\lambda}}^{N-1}]$.

The cumulative projections making up the CSI represent integrations of the object f over increasingly wider swaths, starting at the top end of the projection vectors. Therefore, the natural pixels associated with horizontally adjacent samples (pixels) increasingly approximate more accurately the value of these samples as we move from top to bottom in the CSI. Consequently, the intensity profiles $(\bar{\lambda}^0_k, \bar{\lambda}^1_k, \ldots, \bar{\lambda}^{N-1}_k)$ become smoother as we choose smaller values of k. In fact, in an ideal tomographic projection process, integration over the entire object is the same regardless of projection angle, and so, $\bar{\lambda}^0_0 = \cdots = \bar{\lambda}^{N-1}_0$, *i.e.*, a constant intensity profile.

From this, we may model the intensity versus angle corresponding to each row of the CSI by a 1-d MMI model, each with a different prior. For example, the prior corresponding to $(\bar{\lambda}^0_{k+1}, \bar{\lambda}^1_{k+1}, \ldots, \bar{\lambda}^{N-1}_{k+1})$ would have parameters that reflect the greater variability of the intensity than, say, the intensity $(\bar{\lambda}^0_k, \bar{\lambda}^1_k, \ldots, \bar{\lambda}^{N-1}_k)$. In terms of the various beta density components of the mixture prior for the perturbation variate δ introduced in Chapter 3, those representing high variability, *e.g.*, those defined by small values of the shape parameter s_i, will have higher contribution to the model for high values of k than for those of low values of k.

The modeling of the CSI's intensity variation in the vertical direction, *i.e.*, $\bar{\boldsymbol{\lambda}}^n$, proves more problematic than for the case of the standard (non-cumulative) sinogram. This is because in the cumulative projections, the consecutive pixels may not be modeled by independent innovation variates—as is characteristic of MMI-based models—given their strong dependence due to their construction, particularly for low values of k. We have carried out numerous experiments where we have neglected this fact in a pragmatic pursuit of the "what if", just to be reminded by the very poor results that in this case the fact is not to be ignored.

The need for a suitable model for the individual cumulative projections resulting from ECT studies arises only from the necessity to estimate their underlying intensities before reconstruction. In Section 4.5 we saw the great impact that a good model can make to the tomographic reconstruction process. However, because the multiscale-based tomographic reconstruction approach developed in the previous section only makes use of the cumulative sinogram—as opposed to the standard sinogram—the Poisson-distributed data is an excellent representation of the underlying intensity themselves. To justify this, we make the following observations:

1. Much of the CSI Poisson data have high SNR due to the rapid accumulation of high counts obtained by construction.

2. CSI data characterized by low counts enter the reconstruction process with little weight due (i) to its relative low magnitude, and (ii) to the weight given by the coefficients of the reconstruction formula (4.34) (See also Figure 4.8.)

Modeling the CSI's intensity by the data has the advantage of simplicity, but more importantly, it helps to maintain a higher degree of truth in the reconstructed image. Moreover, due to the accumulation process by which the CSI is formed, the coordination of the estimation of the underlying intensity in the vertical direction—as k varies—is assured despite the fact that we only impose the 1-d MMI model in the horizontal direction—as the angle varies. With this model at hand, the underlying intensity can be estimated using the Bayesian estimator introduced in the last Chapter. We call this approach to modeling, estimation, and reconstruction the CSI-MMI method.

4.7.2 Example

Figure 4.9 is a pelvic bone study obtained using the CSI-MMI method. Figure 4.9 (c) shows the pelvic bone reconstructed image obtained from raw data using the multiscale-based inversion approach of Section 4.6. In Figure 4.9 (d), the underlying intensity of the same raw data used in (c) has been estimated and reconstructed by the CSI-MMI method. For purposes of comparison, we have repeated the images of Figures 4.6 (b) and (d) in Figures 4.9 (a) and (b), respectively. These images correspond to the first pelvic study presented in the example of Section 4.5.

The CSI-MMI processed image in Figure 4.9 (d) reveals a greater amount of detail than does the SI-MMI processed image of Figure 4.9 (b). In particular, note that what appears as two bright blobs in the upper left of the SI-MMI image, it becomes better represented in the result obtained by the CSI-MMI method. In this case, the features preserve the structure found in the unprocessed image of Figure (a). Likewise, the

three bright areas easily distinguishable in the lower left of Figure 4.9 (d) are almost impossible to pick out in the SI-MMI image, although their existence is well suggested by a narrow band of bright pixels.

The images reconstructed directly from raw data using the FBP algorithm and the new multiscale reconstruction method display greater detail than either image in Figures 4.9 (b) and (d); however, the variability of intensity of their pixels is large enough that makes it very difficult to determine the true structure within the images. Clearly, this is especially true of Figure 4.9 (a), demonstrating the great advantage of the multiscale inversion algorithm alone.

In addition to the above differences of image quality among the set, there exist several other features in the CSI-MMI image that are hardly discernible in the SI-MMI and the raw filtered backpropagated images of Figures 4.9 (a) and (b). One example is the very prevalent small dark region near the center of the image which lies on the diagonal going from the lower left to the upper right corners. While in Figures 4.9 (a) the region is well defined, the variability of pixels surrounding it makes it an unreliable feature. In contrast to the virtual disappearance of this feature in the SI-MMI image, the region is well defined in the image constructed with the new multiscale approach.

Although we deem the CSI-MMI methods superior to the SI-MMI approach from the above results, we believe that in practical nuclear medicine and other emission tomgraphy studies, there is a real advantage by working with both type of processed images that we have introduced in this chapter. While the CSI-MMI method offers greater fidelity of images, the SI-MMI-based images displays the greater structures of objects in a more visually appealing manner without unduly sacrificing too much small detail.

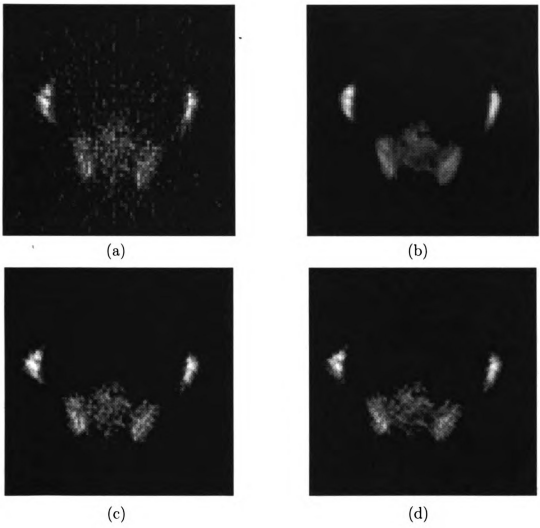


Figure 4.9. *Second pelvic bone study image reconstruction. (a) Pelvic bone from noisy data using a FBP-based reconstruction. (b) Pelvic bone from SI-MMI-processed sinogram using a FBP-based reconstruction. (c) Pelvic bone from raw data using the multiscale-based reconstruction of Section 4.6. (d) Pelvic bone from CSI-MMI-processed sinogram using the multiscale-based reconstruction.*

CHAPTER 5

Conclusions

We have introduced a view of the multiscale structure of processes that succinctly expresses the limitations of models constructed from limited-scale information. At the same time, this view has made apparent the potential advantages of modeling within the multiscale framework for estimation purposes. For this, we introduced a unifying approach to characterizing models and estimators alike. The measures of *anomy*, *accuracy*, *precision*, and *resolution power* were introduced in order to characterize models and quantify their degree of goodness as estimators. While formally defined using the information-theoretic concept of distance, the new concepts reflect our common-day sense of the terms, and so, they are appealing to use and general in nature. Also, the new characterization of models and estimators has given a clearer view on interplay between scale (space/frequency) resolution and information resolution (resolution power).

We have given general guidelines for obtaining linear transformations to construct Gaussian multiscale models that are potentially better suited for estimation purposes. Much work needs to be done in this respect, but a wide range of possibilities has been opened by the idea of constructing transformations that are statistically motivated by using the criteria set forth here. We argued the various advantages of the Bayesian estimators within the multiscale framework, and showed how this framework facilitates

the formulation of practical priors.

We have developed a Bayesian approach for Poisson intensity estimation, and introduced a novel MMI prior model for intensity functions based on a *multiplicative* innovations structure. The MMI captures many of the key features of real-world intensity functions and provides an excellent match to the Poisson distribution. The MMI model facilitates a multiscale Bayesian estimation procedure that proceeds in a natural fashion from coarse-to-fine resolutions. The estimator has a simple closed expression and can be implemented in $O(N)$ operations, where N is the dimension of the finest resolution of the discretized intensity. The issue of choosing the parameters of the prior model was addressed, and a simple moment-matching method was proposed for fitting the parameters to a given set of data. The MMI model was extended to a shift-invariant (stationary) model called the SI-MMI model, and it was shown that the SI-MMI model has a $1/f$ correlation structure that is more regular than that of the MMI model. We developed a fast version of the SI-MMI which leads to a very efficient algorithm of complexity $O(N)$.

We have illustrated the performance of the multiscale Bayesian estimator by comparing its performance to other wavelet-based approaches on several benchmark problems. We have also studied the application of the framework to photon-limited imaging problems, and examined its potential to improve the quality of nuclear medicine images. The framework also appears to have potential in other applications such as network tomography [4] and network traffic modeling and synthesis.

We have introduced two new approaches to the estimation and reconstruction of emission computed tomography images. The SI-MMI method is a simple, computationally efficient, extension to the 2-d MMI approach which similarly incorporates a Bayesian estimator in a very natural way. Its MMI-based prior model is well justified based on the sinogram formation process. The superiority of the SI-MMI over the commonly used FFT-based-reconstruction post-filtering approach was demonstrated

with comparative examples using real and simulated data.

The second method introduced was the CSI-MMI method. This method is based on the new multiscale approach to tomographic reconstruction also presented here. The reconstruction method was statistically motivated on the ECT problem and facilitates the robust reconstruction of ECT images. The new Radon-inversion transform formulation allows incorporating some local filtering within the reconstruction steps, which leads to a more efficient "one-step" process. In future work, we plan to exploit this feature to develop methods to mitigate attenuation and blurring effects. We developed a fast algorithm that implements the reconstruction more efficiently than FFT-based inversion methods. The new inversion approach has the advantage over more traditional reconstruction methods of allowing efficient reconstruction at any desired scale. Fitting the reconstruction scale to the highest resolution supported by the data is important to the efficiency of reconstruction, and to the accuracy and precision of the final image. In future work we will investigate further the application of the new reconstruction method to simulated data, which from our very limited trials, proves not to perform as well as FFT-based inversion approaches.

CSI-MMI method estimates the cumulative sinogram image by imposing 1-d MMI priors to each of its rows, and adapting them individually using the data. The robustness of the data column-wise in the cumulative sinogram justifies this one dimensional approach and helps the reconstructed image stay true. Also, the one dimensional approach in conjunction with the new inversion method makes the CSI-MMI method a highly computationally efficient alternative to the image formation process from photon-limited projection data.

APPENDICES

APPENDICES

APPENDIX A

Appendix to Chapter 2

A.1 Proof of Expression (2.2)

Recall that

$$\hat{h}_{j,k}(\omega) \equiv \frac{1}{\xi_0}\mathbf{1}\left(\frac{\omega - j\xi_0}{\xi_0}\right) e^{i\,k u_0(\omega - j\xi_0)} \quad \text{for all } j,k \in \mathbb{Z}, \tag{A.1}$$

and that f is an $L^2(\mathbb{R})$ function. Thus, by Plancherel formula, $\hat{f}$ is also in $L^2(\mathbb{R})$. Since

$$\sum_j \mathbf{1}\left(\frac{\omega - j\xi_0}{\xi_0}\right)\mathbf{1}\left(\frac{\omega - \frac{2\pi}{u_0}k - j\xi_0}{\xi_0}\right) = 1 \tag{A.2}$$

whenever $k = 0$, and zero otherwise, with $u_0\,\xi_0 = 2\pi$

$$\begin{aligned}\hat{f}(\omega) &= \frac{2\pi}{u_0\xi_0}\sum_{j,k}\mathbf{1}\left(\frac{\omega - j\xi_0}{\xi_0}\right)\mathbf{1}\left(\frac{\omega - \frac{2\pi}{u_0}k - j\xi_0}{\xi_0}\right)\hat{f}(\omega - \frac{2\pi}{u_0}k) \\ &= \frac{1}{\xi_0}\sum_j \mathbf{1}\left(\frac{\omega - j\xi_0}{\xi_0}\right)\int_{\mathbb{R}}\hat{f}(\omega')\mathbf{1}\left(\frac{\omega' - j\xi_0}{\xi_0}\right)\frac{2\pi}{u_0}\sum_k \delta(\omega - \omega' - \frac{2\pi}{u_0}k)\,d\omega'.\end{aligned} \tag{A.3}$$

Using the Poisson formula and the fact that every Fourier series' component is individually integrable [19], the expression may be rewritten as

$$\begin{aligned}
&\frac{1}{\xi_0}\sum_j \mathbf{1}\left(\frac{\omega - j\xi_0}{\xi_0}\right)\int_{\mathbb{R}} \hat{f}(\omega')\mathbf{1}\left(\frac{\omega' - j\xi_0}{\xi_0}\right)\sum_k e^{ik u_0(\omega-\omega')}\,d\omega' \\
&= \frac{1}{\xi_0}\sum_{j,k} \mathbf{1}\left(\frac{\omega - j\xi_0}{\xi_0}\right) e^{ik u_0\omega}\int_{\mathbb{R}} \hat{f}(\omega')\mathbf{1}\left(\frac{\omega' - j\xi_0}{\xi_0}\right) e^{-ik u_0\omega'}\,d\omega' \qquad \text{(A.4)} \\
&= \sum_{j,k}\langle \hat{f}, \hat{h}_{j,k}\rangle \hat{h}_{j,k}(\omega).
\end{aligned}$$

A.2 On the Monotonicity of $I_\rho(X;\Lambda)$

In this section we show that the average mutual information $I_\rho(X;\Lambda)$ is a strictly decreasing function of ρ, and thus, bijective and invertible; consequently, that ρ is uniquely determined for each value of $I_\rho(X;\Lambda) = \mathcal{P}$. We assume an absolutely continuous model, one with absolutely continuous densities.

By defintion

$$I_\rho(X;\Lambda) = \int p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) \log \frac{p_\rho(\mathbf{x}|\boldsymbol{\lambda})}{p_\rho(\mathbf{x})}\,d\boldsymbol{\lambda}\,d\mathbf{x},$$

where $p_\rho(\mathbf{x}|\boldsymbol{\lambda}) \equiv \mathcal{U}_\rho(\mathbf{x} - \boldsymbol{\lambda})$, *i.e.*, a uniform density with support of which is the N-dimensional cube $C(\boldsymbol{\lambda};\rho)$ centered at $\boldsymbol{\lambda}$ and sides 2ρ, and

$$\begin{aligned}
p_\rho(\mathbf{x}) &= \int p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})\,d\boldsymbol{\lambda} \\
&= \int \mathcal{U}_\rho(\mathbf{x} - \boldsymbol{\lambda})p(\boldsymbol{\lambda})\,d\boldsymbol{\lambda} \\
&= (p_{\boldsymbol{\lambda}} * \mathcal{U}_\rho)(\mathbf{x}).
\end{aligned}$$

Here, $*$ denotes convolution. It is not difficult to show that if $\mathbf{y}_\rho$ is an independent *rv* distributed according to $\mathcal{U}_\rho$, then $\boldsymbol{\lambda} + \mathbf{y}_\rho \sim p_\rho$ [30].

A well known relation for mutual information establishes that $I_\rho(X;\Lambda) = H_\rho(X) -$

$H_\rho(X|\Lambda)$. In this case, $H_\rho(X)$ is the entropy of $p_\rho(\mathbf{x})$, or equivalently, the entropy of the *rv* $\boldsymbol{\lambda}+\mathbf{y}_\rho$. Thus, we write $H(\boldsymbol{\lambda}+\mathbf{y}_\rho) \equiv H_\rho(X)$, when it is clear that $H(\boldsymbol{\lambda}+\mathbf{y}_\rho)$ must be understood as a functional over the sample space $\Lambda \times C(\boldsymbol{\lambda};\rho)$. The conditional entropy $H_\rho(X|\Lambda)$ is given by

$$\begin{aligned} H_\rho(X|\Lambda) &= \int p_\rho(\mathbf{x}|\boldsymbol{\lambda}) p(\boldsymbol{\lambda}) \log \frac{1}{p_\rho(\mathbf{x}|\boldsymbol{\lambda})}\, d\mathbf{x}\, d\boldsymbol{\lambda} \\ &= \int p(\boldsymbol{\lambda}) \int \mathcal{U}_\rho(\mathbf{x}-\boldsymbol{\lambda}) \log \frac{1}{\mathcal{U}_\rho(\mathbf{x}-\boldsymbol{\lambda})}\, d\mathbf{x}\, d\boldsymbol{\lambda} \\ &= \int_{\mathbb{R}^N} p(\boldsymbol{\lambda}) \int_{C(\boldsymbol{\lambda};\rho)} \frac{1}{|C|} \log |C|\, d\mathbf{x}\, d\boldsymbol{\lambda}. \end{aligned}$$

This entropy evaluates to $\log(2\rho)^N$, which is the entropy of $\mathbf{y}_\rho$. Consequently, $I_\rho(X;\Lambda) = H(\boldsymbol{\lambda}+\mathbf{y}_\rho) - H(\mathbf{y}_\rho)$.

Let $\rho_2 > \rho_1 > 0$, and consider the difference of entropies $H(\boldsymbol{\lambda}+\mathbf{y}_{\rho_1}) - H(\boldsymbol{\lambda}+\mathbf{y}_{\rho_2})$, where $\mathbf{y}_{\rho_1}$, $\mathbf{y}_{\rho_2}$, and $\boldsymbol{\lambda}$ are all independent. Clearly,

$$\begin{aligned} H(\boldsymbol{\lambda}+\mathbf{y}_{\rho_1}) - H(\boldsymbol{\lambda}+\mathbf{y}_{\rho_2}) &= (H(\boldsymbol{\lambda}+\mathbf{y}_{\rho_1}) - H(\boldsymbol{\lambda}+\mathbf{y}_{\rho_1}|\boldsymbol{\lambda})) \\ &\quad - (H(\boldsymbol{\lambda}+\mathbf{y}_{\rho_2}) - H(\boldsymbol{\lambda}+\mathbf{y}_{\rho_2}|\boldsymbol{\lambda})) + H(\boldsymbol{\lambda}+\mathbf{y}_{\rho_1}|\boldsymbol{\lambda}) - H(\boldsymbol{\lambda}+\mathbf{y}_{\rho_2}|\boldsymbol{\lambda}) \\ &= I(\boldsymbol{\lambda}+\mathbf{y}_{\rho_1};\boldsymbol{\lambda}) - I(\boldsymbol{\lambda}+\mathbf{y}_{\rho_2};\boldsymbol{\lambda}) + H(\mathbf{y}_{\rho_1}) - H(\mathbf{y}_{\rho_2}). \end{aligned}$$

In order to arrive at this expression we have made use of the fact that for any two independent *rv*'s λ and y, $H(\lambda+y|\lambda) = H(y|\lambda) = H(y)$. Subtracting $H(\mathbf{y}_{\rho_1}) - H(\mathbf{y}_{\rho_2})$ from every term in this expression, and writing $I_{\rho_1}(X;\Lambda)$ for $H(\boldsymbol{\lambda}+\mathbf{y}_{\rho_1}) - H(\mathbf{y}_{\rho_1})$, and $I_{\rho_2}(X;\Lambda)$ for $H(\boldsymbol{\lambda}+\mathbf{y}_{\rho_2}) - H(\mathbf{y}_{\rho_2})$, we obtain

$$I_{\rho_1}(X;\Lambda) - I_{\rho_2}(X;\Lambda) = I(\boldsymbol{\lambda}+\mathbf{y}_{\rho_1};\boldsymbol{\lambda}) - I(\boldsymbol{\lambda}+\mathbf{y}_{\rho_2};\boldsymbol{\lambda}).$$

By hypothesis, out of the *rv*'s $\mathbf{y}_{\rho_1}$ and $\mathbf{y}_{\rho_2}$, the latter induces greater uncertainty into its corresponding sum with $\boldsymbol{\lambda}$, and so, the information that $\boldsymbol{\lambda}$ conveys about

$\boldsymbol{\lambda}+\mathbf{y}_{\rho_2}$ is less than that conveyed about $\boldsymbol{\lambda}+\mathbf{y}_{\rho_1}$. Thus, $I(\boldsymbol{\lambda}+\mathbf{y}_{\rho_1};\boldsymbol{\lambda}) > I(\boldsymbol{\lambda}+\mathbf{y}_{\rho_2};\boldsymbol{\lambda})$ and, therefore,

$$I_{\rho_1}(X;\Lambda) > I_{\rho_2}(X;\Lambda).$$

A.3 An Alternate Interpratation for Precision

By definition

$$I_\rho(X;\Lambda) = \int p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) \log \frac{p_\rho(\mathbf{x}|\boldsymbol{\lambda})}{p_\rho(\mathbf{x})}\, d\boldsymbol{\lambda}\, d\mathbf{x},$$

where $p_\rho(\mathbf{x}|\boldsymbol{\lambda})$ is a uniform density the support of which is the N-dimensional cube $C(\boldsymbol{\lambda};\rho)$ centered at $\boldsymbol{\lambda}$ and sides 2ρ, and

$$\begin{aligned} p_\rho(\mathbf{x}) &= \int p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})\, d\boldsymbol{\lambda} \\ &= \frac{1}{|C|}\int_{C(\mathbf{x};\rho)} p(\boldsymbol{\lambda})\, d\boldsymbol{\lambda} \\ &= \frac{1}{|C|}\mathrm{P}(\boldsymbol{\lambda} \in C(\mathbf{x};\rho)). \end{aligned} \tag{A.5}$$

Here, $|C|$ stands for the volume of $C(\mathbf{x};\rho)$, *i.e.*, $|C| = (2\rho)^N$, and $\mathrm{P}(\cdot)$ denotes probability mass.

Writing $I_\rho(X;\Lambda)$ in terms of these expressions yields

$$\begin{aligned} I_\rho(X;\Lambda) &= \int_{\mathbb{R}^N}\int_{C(\mathbf{x};\rho)} \frac{p(\boldsymbol{\lambda})}{|C|} \log \frac{1}{\mathrm{P}(\boldsymbol{\lambda} \in C(\mathbf{x};\rho))}\, d\boldsymbol{\lambda}\, d\mathbf{x} \\ &= \frac{-1}{|C|}\int_{\mathbb{R}^N} \mathrm{P}(\boldsymbol{\lambda} \in C(\mathbf{x};\rho)) \log \mathrm{P}(\boldsymbol{\lambda} \in C(\mathbf{x};\rho))\, d\mathbf{x} \\ &= \frac{-1}{|C|}\sum_{\mathbf{k}}\int_{C(2\rho\mathbf{k};\rho)} \mathrm{P}(\boldsymbol{\lambda} \in C(\mathbf{x};\rho)) \log \mathrm{P}(\boldsymbol{\lambda} \in C(\mathbf{x};\rho))\, d\mathbf{x}, \end{aligned}$$

where the summation is over all vectors $\mathbf{k} = (k_1, \ldots, k_N)$ with $k_1, \ldots, k_N$ integer

$\lambda + \Sigma_{\mathbf{w}}$ is less than that removed about $\lambda + \Sigma_{\mathbf{w}}$. Thus $\ell(\lambda + \Sigma_{\mathbf{w}}; \mathbf{Y}) > \ell(\lambda + \Sigma_{\mathbf{w}}; \mathbf{Y})$ and, therefore,

$$\ell_{\mathbf{w}}(\mathbf{Y}; \lambda) > \ell_{\mathbf{w}}(\mathbf{Y}; \lambda).$$

A.3 An Alternate Interpretation for Precision

By definition

$$\ell_{\mathbf{w}}(\mathbf{Y}; \lambda) = \int p_{\mathbf{w}}(x|\lambda) \ln p(\mathbf{Y}|x) \frac{p(x|\lambda)}{p(x)} \, dx$$

where $p_{\mathbf{w}}(x|\lambda)$ is a uniform density over the simplex $\mathcal{C}(x;p)$ or the N-dimensional cube $C(x;p)$ centered at λ and sides $2p$, and

$$p(x) = \int p(x|\lambda) p(\lambda) \, d\lambda$$

$$= \frac{1}{|C|} \int_{C(x;p)} p(\lambda) \, d\lambda \tag{A.5}$$

$$= \frac{1}{|C|} P(\mathbf{Y} \in C(x;p))$$

Here, $|C|$ stands for the volume of $C(x;p)$, i.e., $|C| = (2p)^N$, and $P(\cdot)$ denotes probability mass.

Writing $\ell_{\mathbf{w}}(\mathbf{Y}; \lambda)$ in terms of these expressions yields

$$\ell_{\mathbf{w}}(\mathbf{Y}; \lambda) = \int_{\mathbb{R}^N} \int_{C(x;p)} \frac{p(\mathbf{Y})}{|C|} \log \frac{p(\mathbf{Y} \in C(x;p))}{1} \, dy \, dx$$

$$= \frac{1}{|C|} \int_{\mathbb{R}^N} \int_{C(x;p)} P(\mathbf{Y} \in C(x;p)) \log P(\mathbf{Y} \in C(x;p)) \, dy \, dx$$

$$= \frac{1}{|C|} \sum_{\mathbf{k}} \int_{C(x;p)} P(\mathbf{Y} \in C(x;p)) \log P(\mathbf{Y} \in C(x;p)) \, dy \, dx$$

where the summation is over all vectors $\mathbf{k} = (k_1, \ldots, k_N)$ with $k_1, \ldots, k_N$ integers

numbers. So,

$$I_\rho(X;\Lambda) = \frac{-1}{|C|}\sum_{\mathbf{k}}\int_{C(0;\rho)} \mathrm{P}(\boldsymbol{\lambda}\in C(\mathbf{x}+2\rho\mathbf{k};\rho))\log \mathrm{P}(\boldsymbol{\lambda}\in C(\mathbf{x}+2\rho\mathbf{k};\rho)))\,d\mathbf{x}.$$

Clearly, $\mathrm{P}(\boldsymbol{\lambda}\in C(\mathbf{x}+2\rho\mathbf{k};\rho))$ are probability masses on $\mathbf{k}$ parameterized by $\mathbf{x}$, *i.e.*, $\sum_{\mathbf{k}}\mathrm{P}(\boldsymbol{\lambda}\in C(\mathbf{x}+2\rho\mathbf{k};\rho)) = 1$ and $\mathrm{P}(\boldsymbol{\lambda}\in C(\mathbf{x}+2\rho\mathbf{k};\rho))\geq 0$ for all $\mathbf{x}$ and $\mathbf{k}$. Thus, the entropy $H_{\rho,\mathbf{x}}$ of the discrete *rv*'s $\boldsymbol{\xi}_{\rho,\mathbf{x}}(\boldsymbol{\lambda}) \equiv \lfloor\frac{\boldsymbol{\lambda}+\rho-\mathbf{x}}{2\rho}\rfloor$ with probability distributions

$$\mathrm{P}(\boldsymbol{\xi}_{\rho,\mathbf{x}}=\mathbf{k}) = \int_{C(\mathbf{x}+2\rho\mathbf{k};\rho)} p(\boldsymbol{\lambda})\,d\boldsymbol{\lambda} \tag{A.6}$$

are

$$H_{\rho,\mathbf{x}} \equiv -\sum_{\boldsymbol{\xi}_{\rho,\mathbf{x}}}\mathrm{P}(\boldsymbol{\xi}_{\rho,\mathbf{x}})\log\mathrm{P}(\boldsymbol{\xi}_{\rho,\mathbf{x}}). \tag{A.7}$$

Rewriting the mutual information in terms of these entropies,

$$I_\rho(X;\Lambda) = \frac{1}{|C|}\int_{C(0;\rho)} H_{\rho,\mathbf{x}}\,d\mathbf{x}. \tag{A.8}$$

This expression says that $I_\rho(X;\Lambda)$ is an average of entropies corresponding to discrete *rv*'s obtained by partitioning the N-dimensional space into cubes of sides 2ρ each, and assigning them probability distributions corresponding to the volume of $p(\boldsymbol{\lambda})$ "under" each cube. The entire partition set of cubes are shifted by $\mathbf{x}\in C(0;\rho)$ to obtain the uncountable set of distinct discrete *rv*'s over which the averaging process is carried out. This justifies the alternate and intuitive interpretation given in Section 2.3.2 for a model's precision, and which we restate below. Its equivalence with the original definition is clear for in both cases we have $\mathcal{P} = I_\rho(X;\Lambda) = I(X;\Lambda)$.

number. So,

$$I_k^*(X;Y) = \frac{1}{|\mathcal{C}|}\sum_{\mathbf{c}\in\mathcal{C}_0} \log \frac{P(Y \in C(x+2\mathbf{c};\rho))}{\sum_{\mathbf{c}'\in\mathcal{C}_0} P(Y \in C(x+2\mathbf{c}';\rho))} \log P(Y \in C(x+2\mathbf{c};\rho))\,dx$$

Clearly, $P(Y \in C(x+2\mathbf{c};\rho))$ are probability masses on k points centered by x, i.e. $\sum_{\mathbf{c}} P(Y \in C(x+2\mathbf{c};\rho)) = 1$ and $P(Y \in C(x+2\mathbf{c};\rho)) \geq 0$ for all x and k. Thus, the entropy H_{Y^*} of the discrete r.v's $C(\cdot;Y) = \{C(x+2\mathbf{c};\rho)\}$ with probability distributions

$$P(C_{Y^*} = k) = \int_{C(x+2\mathbf{c};\rho)} \mu(Y)\,dY \tag{A.6}$$

are

$$H_{Y^*} = -\sum_{\mathbf{c}\in\mathcal{C}_0} P(C_{Y^*}) \log P(C_{Y^*}) \tag{A.7}$$

Rewriting the mutual information in terms of these entropies

$$I_k^*(X;Y) = \frac{1}{|\mathcal{C}|}\sum_{\mathbf{c}\in\mathcal{C}_0} H_{Y^*}\,dx \tag{A.8}$$

This expression says that $I_k^*(X;Y)$ is an average of entropies corresponding to discrete r.v's obtained by partitioning the N-dimensional space into cubes of side 2ρ each, and assigning them probability distributions corresponding to the volume of $p(Y)$ "under" each cube. The cube partition act of cubes are shifted by $x \in C(0;\rho)$ to obtain the consonant sub-set of distinct discrete r.v's over which the averaging process is carried out. This justifies the elementa and heuristic interpretation given in Section 2.3.2 for a model's precision, and which we restate below. Its equivalence with the original definition is clear for in both cases we have $P = I_k^*(X;A) = I(X;A)$.

$\mathcal{P}$ is the entropy of the *rv* $\boldsymbol{\lambda}$ when quantized to the number of bits determined by $I(X;\Lambda)$, averaged over all possible quantization schemes.

A.4 Only an MA Model has Anomy of Zero

To prove that only an MA model has anomy of zero, we let

$$\alpha_{\rho,\boldsymbol{\lambda}}(t) \equiv -\int p_\rho(\mathbf{x}|\boldsymbol{\lambda}) \log(t\, p(\mathbf{x}|\boldsymbol{\lambda}) + (1-t)\, p_\rho(\mathbf{x}|\boldsymbol{\lambda}))\, d\mathbf{x}$$

for all values of t in $[0,1]$, and show that $\mathcal{N} = \int(\alpha_{\rho,\boldsymbol{\lambda}}(1) - \alpha_{\rho,\boldsymbol{\lambda}}(0))p(\boldsymbol{\lambda})\, d\boldsymbol{\lambda} > 0$, unless $p(\mathbf{x}|\boldsymbol{\lambda}) = p_\rho(\mathbf{x}|\boldsymbol{\lambda})$.

By repeated differentiation of $\alpha_{\rho,\boldsymbol{\lambda}}(t)$ with respect to t, we obtain

$$\alpha^{(n)}_{\rho,\boldsymbol{\lambda}}(t) = (n-1)! \int p_\rho(\mathbf{x}|\boldsymbol{\lambda}) \left(\frac{p_\rho(\mathbf{x}|\boldsymbol{\lambda}) - p(\mathbf{x}|\boldsymbol{\lambda})}{t\, p(\mathbf{x}|\boldsymbol{\lambda}) + (1-t)\, p_\rho(\mathbf{x}|\boldsymbol{\lambda})} \right)^n d\mathbf{x}.$$

Suppose that $p_\rho(\mathbf{x}|\boldsymbol{\lambda}) \neq p(\mathbf{x}|\boldsymbol{\lambda})$. Then, it is clear that $\alpha^{(2)}_{\rho,\boldsymbol{\lambda}}(t) > 0$ for all t. Since $\alpha'_{\rho,\boldsymbol{\lambda}}(0) = 0$, $\alpha'_{\rho,\boldsymbol{\lambda}}(t) > 0$ for all $t > 0$. Thus, $\alpha_{\rho,\boldsymbol{\lambda}}(1) > \alpha_{\rho,\boldsymbol{\lambda}}(0)$ and, consequently, $\mathcal{N} > 0$. When $p(\mathbf{x}|\boldsymbol{\lambda}) = p_\rho(\mathbf{x}|\boldsymbol{\lambda})$, $\alpha_{\rho,\boldsymbol{\lambda}}(1) = \alpha_{\rho,\boldsymbol{\lambda}}(0)$, and the desired result is obtained.

A.5 Equivalence of Anomy Definitions

We want to show that

$$D(p_\rho(\mathbf{x})\|p(\mathbf{x})) = D(p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})\|p(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})).$$

Let $\mathcal{W}$ be a linear mapping $\boldsymbol{\lambda} \mapsto \boldsymbol{\lambda}_1$, where $\boldsymbol{\lambda} \in \mathbb{R}^N$, $\boldsymbol{\lambda}_1 \in \mathbb{R}^{\frac{N}{2}}$, and N a positive

even number. We only consider those mappings $\mathcal{W}$ for which the N by N matrix

$$A \equiv \begin{pmatrix} \mathrm{I} & \mathbf{0} \\ \mathcal{W} & \end{pmatrix},$$

is invetible. Here, I and $\mathbf{0}$ are the $\frac{N}{2}$ by $\frac{N}{2}$ identity and zero matrices.

Denote by $\boldsymbol{\lambda}^t$ and $\boldsymbol{\lambda}^b$ the first (*top*) and last (*bottom*) $N/2$ elements of $\boldsymbol{\lambda}$, respectively. Then, assuming all the following ratios to be well defined,

$$\frac{p_\rho(\mathbf{x}|\boldsymbol{\lambda}_1)}{p(\mathbf{x}|\boldsymbol{\lambda}_1)} = \frac{p_\rho(\mathbf{x})}{p(\mathbf{x})}\,\frac{p_\rho(\boldsymbol{\lambda}_1|\mathbf{x})}{p(\boldsymbol{\lambda}_1|\mathbf{x})} = \frac{p_\rho(\mathbf{x})}{p(\mathbf{x})}\,\frac{p_\rho(\boldsymbol{\lambda}^t,\boldsymbol{\lambda}_1|\mathbf{x})}{p(\boldsymbol{\lambda}^t,\boldsymbol{\lambda}_1|\mathbf{x})},$$

where the last equality is obtained by multiplying and dividing by $p(\boldsymbol{\lambda}^t|\boldsymbol{\lambda}_1,\mathbf{x})$. Recall that in general, meaning is given to $p_\rho(\mathbf{x}|\boldsymbol{\phi})$ by the integral $\int p_\rho(\mathbf{x}|\boldsymbol{\lambda},\boldsymbol{\phi})\,p(\boldsymbol{\lambda}|\boldsymbol{\phi})\,d\boldsymbol{\lambda}$ $\left(= \int p_\rho(\mathbf{x}|\boldsymbol{\lambda})\,p(\boldsymbol{\lambda}|\boldsymbol{\phi})\,d\boldsymbol{\lambda}.\right)$ Since $\begin{pmatrix}\boldsymbol{\lambda}^t \\ \boldsymbol{\lambda}_1\end{pmatrix} = A\begin{pmatrix}\boldsymbol{\lambda}^t \\ \boldsymbol{\lambda}^b\end{pmatrix} = A\,\boldsymbol{\lambda}$,

$$\frac{p_\rho(\mathbf{x}|\boldsymbol{\lambda}_1)}{p(\mathbf{x}|\boldsymbol{\lambda}_1)} = \frac{p_\rho(\mathbf{x})}{p(\mathbf{x})}\,\frac{p_\rho(\boldsymbol{\lambda}|\mathbf{x})}{p(\boldsymbol{\lambda}|\mathbf{x})} = \frac{p_\rho(\mathbf{x}|\boldsymbol{\lambda})}{p(\mathbf{x}|\boldsymbol{\lambda})}. \tag{A.9}$$

Since the derivation could have equally been started from this last ratio of densities and concluded that removal of information from the conditioning of the densities is inconsequential,[1]

$$\frac{p_\rho(\mathbf{x})}{p(\mathbf{x})} = \frac{p_\rho(\mathbf{x}|\boldsymbol{\lambda})}{p(\mathbf{x}|\boldsymbol{\lambda})}. \tag{A.10}$$

[1]Note that by choosing $\mathcal{W} = [\mathbf{0}, \mathrm{I}]$ and $\boldsymbol{\lambda}^b$ a constant vector independent of $\mathbf{x}$, the desired result is explicitly obtained.

Now, from the definition of the information-theoretic distance $D(\cdot\|\cdot)$ and (A.10)

$$\begin{aligned} D(p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})\|p(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})) &= \int p_\rho(\mathbf{x}|\boldsymbol{\lambda})\, p(\boldsymbol{\lambda}) \log \frac{p_\rho(\mathbf{x}|\boldsymbol{\lambda})}{p(\mathbf{x}|\boldsymbol{\lambda})}\, d\mathbf{x}\, d\boldsymbol{\lambda} \\ &= \int p_\rho(\mathbf{x}) \log \frac{p_\rho(\mathbf{x})}{p(\mathbf{x})}\, d\mathbf{x} = D(p_\rho(\mathbf{x})\|p(\mathbf{x})). \end{aligned} \tag{A.11}$$

A.6 Proof of Expression (2.35)

We shall prove that for a sufficiently smooth (slowly varying) density function $p(\boldsymbol{\lambda})$ relative to $p(\mathbf{x}|\boldsymbol{\lambda})$ the parameterized mutual information $I_\rho(X;\Lambda)$ can be well approximated by $H(\Lambda) - \log|C|$. We present two different proofs, starting with the simplest. Although the proofs are essentially the same, they begin from the two different interpretations given to the quantity $I_\rho(X;\Lambda)$. The first of these interpretations was introduced in Section 2.3.2; the second interpretation was introduced in Section A.3 of this appendix.

First Proof

We know that $I_\rho(X;\Lambda)$ may be written as the difference $H_\rho(X) - H_\rho(X|\Lambda)$, where $H_\rho(X) \equiv -\int p_\rho(\mathbf{x}) \log p_\rho(\mathbf{x})\, d\mathbf{x}$ and $H_\rho(X|\Lambda) \equiv -\int p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) \log p_\rho(\mathbf{x}|\boldsymbol{\lambda})\, d\mathbf{x}\, d\boldsymbol{\lambda}$. In Appendix A.2 we showed that $H_\rho(X|\Lambda) = \log|C|$, where $|C| = (2\rho)^N$ and N is the size of the signals $\mathbf{x}$ and $\boldsymbol{\lambda}$. Now, since

$$p_\rho(\mathbf{x}) = \int p_\rho(\mathbf{x}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})\, d\boldsymbol{\lambda},$$

by the assumption of smoothness of $p(\boldsymbol{\lambda})$ the 2ρ width-per-dimension of $p_\rho(\mathbf{x}|\boldsymbol{\lambda}) = \mathcal{U}_\rho(\mathbf{x} - \boldsymbol{\lambda})$ is of the order of the width-per-dimension of $p(\mathbf{x}|\boldsymbol{\lambda})$.Therefore, $p(\boldsymbol{\lambda})$ may

be regarded constant within the N-dimensional cube of sides 2ρ, and

$$p_\rho(\mathbf{x}) = \frac{1}{|C|}\int_{C(\mathbf{x};\rho)} p(\boldsymbol{\lambda})\, d\boldsymbol{\lambda} \approx p_{\boldsymbol{\lambda}}(\mathbf{x}).$$

Therefore,

$$H_\rho(X) \approx -\int p_{\boldsymbol{\lambda}}(\mathbf{x}) \log p_{\boldsymbol{\lambda}}(\mathbf{x})\, d\mathbf{x} = H(\Lambda).$$

Second Proof

From Appendix A.3 we known that $I_\rho(X;\Lambda)$ can be written as

$$I_\rho(X;\Lambda) = \frac{1}{|C|}\int_{C(0\,;\,\rho)} H_{\rho,\mathbf{x}}\, d\mathbf{x},$$

where

$$H_{\rho,\mathbf{x}} \equiv -\sum_{\boldsymbol{\xi}_{\rho,\mathbf{x}}} \mathrm{P}(\boldsymbol{\xi}_{\rho,\mathbf{x}}) \log \mathrm{P}(\boldsymbol{\xi}_{\rho,\mathbf{x}})$$

and

$$\mathrm{P}(\boldsymbol{\xi}_{\rho,\mathbf{x}} = \mathbf{k}) = \int_{C(\mathbf{x}+2\rho\mathbf{k}\,;\,\rho)} p(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}.$$

From the assumption of smoothness of $p(\boldsymbol{\lambda})$ the 2ρ width-per-dimension of $p_\rho(\mathbf{x}|\boldsymbol{\lambda}) = \mathcal{U}_\rho(\mathbf{x}-\boldsymbol{\lambda})$ is of the order of the width-per-dimension of $p(\mathbf{x}|\boldsymbol{\lambda})$. Therefore, $p(\boldsymbol{\lambda})$ may be regarded constant within the N-dimensional cube of sides 2ρ, and

$$\mathrm{P}(\boldsymbol{\xi}_{\rho,\mathbf{x}} = \mathbf{k}) \approx p_{\boldsymbol{\lambda}}(\mathbf{x} + 2\rho\mathbf{k})|C|.$$

Then,

$$H_{\rho,\mathbf{x}} \approx -\sum_{\mathbf{k}} p_{\boldsymbol{\lambda}}(\mathbf{x} + 2\rho\mathbf{k})|C| \log\left(p_{\boldsymbol{\lambda}}(\mathbf{x} + 2\rho\mathbf{k})|C|\right).$$

Applying this result to the expression for $I_\rho(X;\Lambda)$ gives

$$
\begin{aligned}
I_\rho(X;\Lambda) &\approx -\sum_{\mathbf{k}} \int_{C(0;\rho)} p_{\boldsymbol{\lambda}}(\mathbf{x}+2\rho\mathbf{k}) \log p_{\boldsymbol{\lambda}}(\mathbf{x}+2\rho\mathbf{k})\, d\boldsymbol{\lambda} \\
&\qquad\qquad - \log|C| \sum_{\mathbf{k}} \int_{C(0;\rho)} p_{\boldsymbol{\lambda}}(\mathbf{x}+2\rho\mathbf{k})\, d\boldsymbol{\lambda} \\
&\approx -\sum_{\mathbf{k}} \int_{C(\mathbf{x}+2\rho\mathbf{k};\rho)} p(\boldsymbol{\lambda}) \log p(\boldsymbol{\lambda})\, d\boldsymbol{\lambda} \\
&\qquad\qquad - \log|C| \sum_{\mathbf{k}} \int_{C(\mathbf{x}+2\rho\mathbf{k};\rho)} p(\boldsymbol{\lambda})\, d\boldsymbol{\lambda} \\
&= -\int_{\mathbb{R}^N} p(\boldsymbol{\lambda}) \log p(\boldsymbol{\lambda})\, d\boldsymbol{\lambda} - \log|C| \int_{\mathbb{R}^N} p(\boldsymbol{\lambda})\, d\boldsymbol{\lambda},
\end{aligned}
$$

which is the desired result.

We note that the approximations made here are reminiscent to those often used within the context of high resolution quantizers in computing their mean-square errors (see for example [21].)

A.7 A Sufficient Condition for the A/P Accuracy Inequality

As usual, we let $|C(\boldsymbol{\lambda}_j;\rho_j)| = (2\rho_j)^{N_j}$ be the volume of the cube $C(\boldsymbol{\lambda}_j;\rho_j)$ centered at $\boldsymbol{\lambda}_j$, of sides twice the EPR ρ_j, and of dimension $N_j = N_0/2^j$. For brevity of notation, in the next few steps below we use the notation particular to scales $j = 0$ and $j+1 = 1$, but the development applies to any two consecutive scales in general.

From definition (2.24) and equality (2.25)

$$
\mathcal{N}_1 = -\log|C(\boldsymbol{\lambda}_1;\rho_1)| - \int_{\mathbb{R}^{N_1}} \frac{p(\boldsymbol{\lambda}_1)}{|C(\boldsymbol{\lambda}_1;\rho_1)|} \int_{C(\boldsymbol{\lambda}_1;\rho_1)} \log p(\mathbf{x}_1|\boldsymbol{\lambda}_1)\, d\mathbf{x}_1\, d\boldsymbol{\lambda}_1.
$$

If $\mathcal{W}$ is an orthonormal transformation such that[2]

$$\begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\lambda}_1 \end{pmatrix} = \mathcal{W}\boldsymbol{\lambda}_0 \qquad \text{and} \qquad \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{x}_1 \end{pmatrix} = \mathcal{W}\mathbf{x}_0,$$

where $\boldsymbol{\theta}_1$ and $\mathbf{w}_1$ are simply the "error" vectors $\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}_1$ and $\mathbf{x}_0 - \mathbf{x}_1$, respectively, the above inner-integral can be expressed as

$$\begin{aligned} &\int_{C(\boldsymbol{\lambda}_1;\rho_1)} \log p(\mathbf{x}_1|\boldsymbol{\lambda}_1)\, d\mathbf{x}_1 \\ &\qquad = \int_{C(\boldsymbol{\lambda}_1;\rho_1)} \log \int_{\mathbb{R}^{N_1}} p(\boldsymbol{\theta}_1|\boldsymbol{\lambda}_1) \sum_{\mathbf{k}} \int_{C(\boldsymbol{\theta}_1+2\rho_1\mathbf{k};\rho_1)} p(\mathbf{x}_1, \mathbf{w}_1|\boldsymbol{\lambda}_1, \boldsymbol{\theta}_1)\, d\mathbf{w}_1\, d\boldsymbol{\theta}_1\, d\mathbf{x}_1, \end{aligned}$$

where the summation is over all vectors $\mathbf{k} = (k_0, \ldots, k_{N_1-1})$ whose elements are integers. $C(\boldsymbol{\theta}_1 + 2\rho_1\mathbf{k}; \rho_1)$ are the cubes centered at $\boldsymbol{\theta}_1 + 2\rho_1\mathbf{k}$ with sides $2\rho_1$. Refer to Figure A.1 for a representation of the various components in these expressions. Note that in the figure the point $(\boldsymbol{\theta}_1, \boldsymbol{\lambda}_1)$ is the same as the point $\boldsymbol{\lambda}_0$.

Now, define

$$\alpha(\mathbf{x}_1|\boldsymbol{\lambda}_1, \boldsymbol{\theta}_1) \equiv \frac{\sum_{\mathbf{k}\neq 0} \int_{C(\boldsymbol{\theta}_1+2\rho_1\mathbf{k};\rho_1)} p(\mathbf{x}_1, \mathbf{w}_1|\boldsymbol{\lambda}_1, \boldsymbol{\theta}_1)\, d\mathbf{w}_1}{\int_{C(\boldsymbol{\theta}_1;\rho_1)} p(\mathbf{x}_1, \mathbf{w}_1|\boldsymbol{\lambda}_1, \boldsymbol{\theta}_1)\, d\mathbf{w}_1}.$$

With the aid of the Jensen's inequality we obtain

$$\begin{aligned} \int_{C(\boldsymbol{\lambda}_1;\rho_1)} \log p(\mathbf{x}_1|\boldsymbol{\lambda}_1)\, d\mathbf{x}_1 &\geq |C(\boldsymbol{\lambda}_1;\rho_1)| \log |C(\boldsymbol{\lambda}_1;\rho_1)| \\ &\quad + \int_{C(\boldsymbol{\lambda}_1;\rho_1)} \int_{\mathbb{R}^{N_1}} p(\boldsymbol{\theta}_1|\boldsymbol{\lambda}_1) \log(1 + \alpha(\mathbf{x}_1|\boldsymbol{\lambda}_1, \boldsymbol{\theta}_1))\, d\boldsymbol{\theta}_1\, d\mathbf{x}_1 \\ &\quad + \frac{1}{|C(\boldsymbol{\lambda}_1;\rho_1)|} \int_{C(\boldsymbol{\lambda}_1;\rho_1)} \int_{\mathbb{R}^{N_1}} p(\boldsymbol{\theta}_1|\boldsymbol{\lambda}_1) \int_{C(\boldsymbol{\theta}_1;\rho_1)} \log p(\mathbf{x}_1, \mathbf{w}_1|\boldsymbol{\lambda}_1, \boldsymbol{\theta}_1)\, d\mathbf{w}_1\, d\boldsymbol{\theta}_1\, d\mathbf{x}_1. \end{aligned}$$

[2]For later convenience, we have departed from writing in the customary order the projection vectors $\boldsymbol{\lambda}_1$ and $\mathbf{x}_1$ first, and the "errors" $\boldsymbol{\theta}_1$ and $\mathbf{w}_1$ second.

If $\mathcal{W}$ is an orthonormal transformation such that

$$\begin{pmatrix} \theta' \\ \gamma' \end{pmatrix} = \mathcal{W}\theta \qquad \text{and} \qquad \begin{pmatrix} w' \\ \sigma' \end{pmatrix} = \mathcal{W}w,$$

where θ' and w' are simply the vectors $(\theta_1, \ldots, \theta_k)$ and $(w_1, \ldots, w_k)$ corresponding to the above inner-integral can be expressed as

$$\int_{C(Y'\in\gamma_m)} \log p(x|_{Y'}) \, dx$$

$$= \int_{C(Y'\in\gamma_m)} \log \int_{\mathbb{R}^k} p(\theta|_{Y'}) \left(\frac{1}{\sqrt{\quad}} \right) \qquad (x \in \mathbb{R}^k \text{ and } \theta \in \mathbb{R}^k)$$

where the summation is over all vectors k [illegible] integers. $C(\theta + 2\pi k, \sigma)$ are the cubes centered at $\theta + 2\pi k$ with sides 2σ. Refer to Figure A.1 for a representation of the [illegible] in three dimensions. Note that in the figure the point (θ', Y') is [illegible] point γ_m.

Now, define

$$o(x|_{Y'}, \theta') \equiv \frac{\sum_{m=1}^{\infty} \int_{C(Y'\in\gamma_m)} p(x|_{Y'}, w', |Y', \theta'|) \, dw'}{\int_{C(Y'\in\gamma_m)} p(x|_{Y'}, w', |Y', \theta'|) \, dw'}$$

With the aid of the Jensen's inequality we obtain

$$\int_{C(Y'\in\gamma_m)} \log p(x|_{Y'}) \, dx \geq |C(Y'\in\gamma_m)| \log k \, C(Y'\in\gamma_m)|$$

$$+ \int_{C(Y'\in\gamma_m)} p(\theta|_{Y'}) \log \frac{}{} + o(x|_{Y'}, \theta')] \, d\theta' \, dx$$

$$= \frac{|C(Y'\in\gamma_m)|}{I} \int_{C(Y'\in\gamma_m)} \int_{\mathbb{R}^k} p(\theta|_{Y'}) \int_{C(\theta'\in\gamma_m)} \log p(x'_{|_{Y'}}, w'|_{Y'}, \theta') \, dw' \, d\theta' \, dx$$

The later conclusions we have digressed from stating in the existence only the previous means γ_1 and γ_2, thus, and the "errors" θ'_1 and w'_1 around.

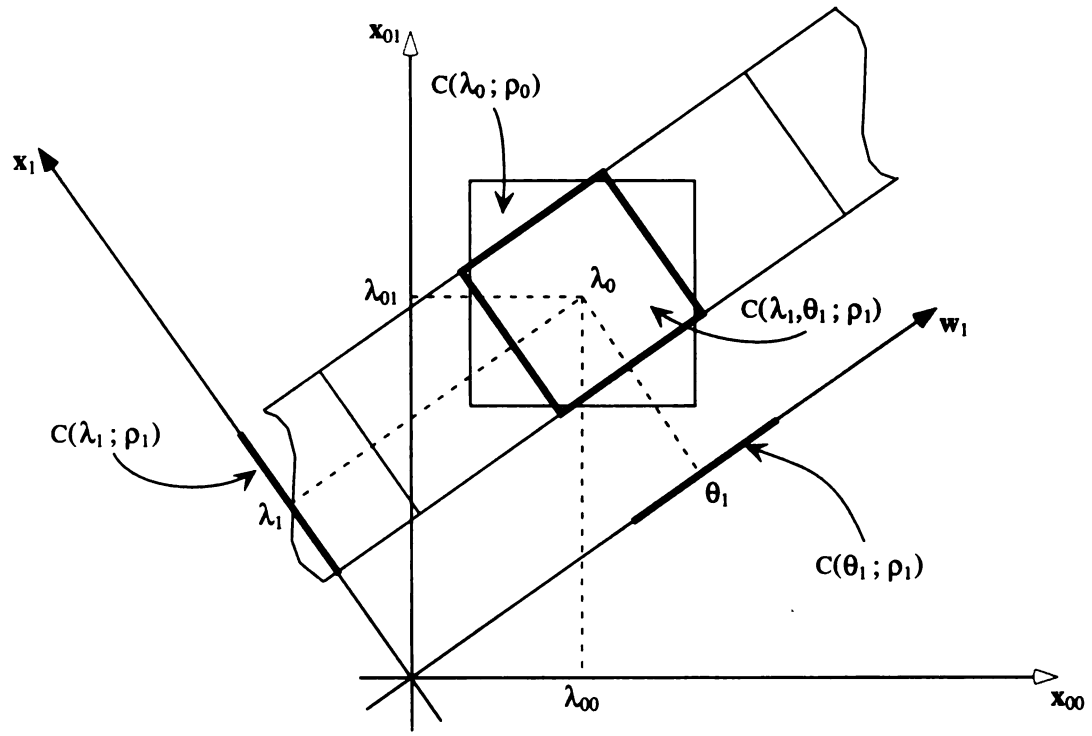


Figure A.1. *Simplified geometric representation of EPR cubes at consecutive scales. For simplicity of notation only those cubes corresponding to scales $j = 0$ and $j + 1 = 1$ are represented (e.g., $C(\boldsymbol{\lambda}_0;\rho_0)$ and $C(\boldsymbol{\lambda}_1;\rho_1)$). Here, $\mathbf{x}_0 = (\mathbf{x}_{0,0}^T, \mathbf{x}_{0,1}^T)^T$. By integrating $p(\mathbf{x}_0|\boldsymbol{\lambda}_0)$ over adjacent copies of $C(\boldsymbol{\lambda}_1, \theta_1; \rho_1)$ as shown, a useful relation between the anomies $\mathcal{N}_0$ and $\mathcal{N}_1$ can be obtained (See text for details.)*

Then,

$$\begin{aligned}\mathcal{N}_1 \leq & -\log|C(\boldsymbol{\lambda}_1;\rho_1)|^2 \\ & - \int_{\mathbb{R}^{N_1}} \frac{p(\boldsymbol{\lambda}_1)}{|C(\boldsymbol{\lambda}_1;\rho_1)|} \int_{C(\boldsymbol{\lambda}_1;\rho_1)} \int_{\mathbb{R}^{N_1}} p(\boldsymbol{\theta}_1|\boldsymbol{\lambda}_1) \log(1+\alpha(\mathbf{x}_1|\boldsymbol{\lambda}_1,\boldsymbol{\theta}_1))\, d\boldsymbol{\theta}_1\, d\mathbf{x}_1\, d\boldsymbol{\lambda}_1 \\ & - \int_{\mathbb{R}^{N_1}} \frac{p(\boldsymbol{\lambda}_1)}{|C(\boldsymbol{\lambda}_1;\rho_1)|^2} \int_{C(\boldsymbol{\lambda}_1;\rho_1)} \int_{\mathbb{R}^{N_1}} p(\boldsymbol{\theta}_1|\boldsymbol{\lambda}_1) \int_{C(\boldsymbol{\theta}_1;\rho_1)} \log p(\mathbf{x}_1, \mathbf{w}_1|\boldsymbol{\lambda}_1, \boldsymbol{\theta}_1)\, d\mathbf{w}_1\, d\boldsymbol{\theta}_1\, d\mathbf{x}_1\, d\boldsymbol{\lambda}_1,\end{aligned}$$

or

$$\begin{aligned}\mathcal{N}_1 \leq & -\log|C(\boldsymbol{\lambda}_1,\boldsymbol{\theta}_1;\rho_1)| - \int_{\mathbb{R}^{N_0}} \frac{p(\boldsymbol{\lambda}_0)}{|C(\boldsymbol{\lambda}_1;\rho_1)|} \int_{C(\boldsymbol{\lambda}_1;\rho_1)} \log(1+\alpha(\mathbf{x}_1|\boldsymbol{\lambda}_0))\, d\mathbf{x}_1\, d\boldsymbol{\lambda}_0 \\ & - \int_{\mathbb{R}^{N_0}} \frac{p(\boldsymbol{\lambda}_0)}{|C(\boldsymbol{\lambda}_1,\boldsymbol{\theta}_1;\rho_1)|} \int_{C(\boldsymbol{\lambda}_1,\boldsymbol{\theta}_1;\rho_1)} \log p(\mathbf{x}_0|\boldsymbol{\lambda}_0)\, d\mathbf{x}_0\, d\boldsymbol{\lambda}_0.\end{aligned}$$

Since

$$\mathcal{N}_0 = -\log|C(\boldsymbol{\lambda}_0;\rho_0)| - \int_{\mathbb{R}^{N_0}} \frac{p(\boldsymbol{\lambda}_0)}{|C(\boldsymbol{\lambda}_0;\rho_0)|} \int_{C(\boldsymbol{\lambda}_0;\rho_0)} \log p(\mathbf{x}_0|\boldsymbol{\lambda}_0)\, d\mathbf{x}_0\, d\boldsymbol{\lambda}_0,$$

$$\begin{aligned}\mathcal{N}_0 - \mathcal{N}_1 \geq \log\left(\frac{\rho_1}{\rho_0}\right)^{N_0} &+ \int_{\mathbb{R}^{N_0}} \frac{p(\boldsymbol{\lambda}_0)}{(2\rho_1)^{N_1}} \int_{C(\boldsymbol{\lambda}_1;\rho_1)} \log(1+\alpha(\mathbf{x}_1|\boldsymbol{\lambda}_0))\, d\mathbf{x}_1\, d\boldsymbol{\lambda}_0 \\ &+ \int_{\mathbb{R}^{N_0}} p(\boldsymbol{\lambda}_0) \left\{ \frac{1}{(2\rho_1)^{N_0}} \int_{C(\boldsymbol{\lambda}_1,\boldsymbol{\theta}_1;\rho_1)} \log p(\mathbf{x}_0|\boldsymbol{\lambda}_0)\, d\mathbf{x}_0 \right. \\ &\qquad \left. - \frac{1}{(2\rho_0)^{N_0}} \int_{C(\boldsymbol{\lambda}_0;\rho_0)} \log p(\mathbf{x}_0|\boldsymbol{\lambda}_0)\, d\mathbf{x}_0 \right\} d\boldsymbol{\lambda}_0.\end{aligned}$$

So, for a model to satisfy the A/P conditions, it suffices that for all scales

$$\log\left(\frac{\rho_j}{\rho_{j+1}}\right)^{N_j} < \int_{\mathbb{R}^{N_j}} p(\boldsymbol{\lambda}_j) \left\{ \mathrm{A}_{j+1}(\alpha) + \mathrm{A'}_j(p) - \mathrm{A}_j(p) \right\} d\boldsymbol{\lambda}_j, \tag{A.12}$$

where

$$\begin{aligned}\mathrm{A}_j(p) &\equiv \frac{1}{(2\rho_j)^{N_j}} \int_{C(\boldsymbol{\lambda}_j;\rho_j)} \log p(\mathbf{x}_j|\boldsymbol{\lambda}_j)\, d\mathbf{x}_j, \\ \mathrm{A'}_j(p) &\equiv \frac{1}{(2\rho_{j+1})^{N_j}} \int_{C(\boldsymbol{\lambda}_{j+1},\boldsymbol{\theta}_{j+1};\rho_{j+1})} \log p(\mathbf{x}_j|\boldsymbol{\lambda}_j)\, d\mathbf{x}_j\end{aligned}$$

and

$$\mathrm{A}_{j+1}(\alpha) \equiv \frac{1}{(2\rho_{j+1})^{N_{j+1}}} \int_{C(\boldsymbol{\lambda}_{j+1};\rho_{j+1})} \log(1+\alpha(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_j))\, d\mathbf{x}_{j+1}$$

are the averages of $\log p(\mathbf{x}_j|\boldsymbol{\lambda}_j)$ over $C(\boldsymbol{\lambda}_j;\rho_j)$ and $C(\boldsymbol{\lambda}_{j+1},\boldsymbol{\theta}_{j+1};\rho_{j+1})$, and the average of $\log(1+\alpha(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_j))$ over $C(\boldsymbol{\lambda}_{j+1};\rho_{j+1})$, respectively.

A.8 Coarse-Scale-Data Limited Models

We want to prove the equivalence between A/P and CSDL models. The following depicts the conditions associated with CSDL models.

The CSDL Model Conditions

j-scale Model		j-scale CSDL Model
$p(\mathbf{x}_j\|\boldsymbol{\lambda}_j)p(\boldsymbol{\lambda}_j)$		$p(\mathbf{x}_{j+1}\|\boldsymbol{\lambda}_j)p(\boldsymbol{\lambda}_j)$
$\mathcal{P}_j$	$\mathcal{P}_j > \mathcal{P}_{j+1,j}$	$\mathcal{P}_{j+1,j}$
$\mathcal{A}_j$	$\mathcal{A}_j < \mathcal{A}_{j+1,j}$	$\mathcal{A}_{j+1,j}$

Although the conditions have been stated in terms of data belonging to one coarser scale than that of the intensity's, it equally applies to observations obtained at any number of higher (coarser) scales above the scale of the intensity.

Recall that all the models under consideration satisfy $p(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_j) = p(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_{j+1})$. Motivated by this relation, we define $p_\rho(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_j) \equiv p_\rho(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_{j+1})$, which is a logical choice. Clearly,

$$\mathcal{P}_{j+1,j} \equiv I(X_{j+1};\Lambda_j) = I(X_{j+1};\Lambda_{j+1}) = \mathcal{P}_j.$$

Since similarly $I_{\tilde{\rho}}(X_{j+1};\Lambda_j) = I_{\tilde{\rho}}(X_{j+1};\Lambda_{j+1})$,

$$I_{\tilde{\rho}}(X_{j+1};\Lambda_{j+1}) = \mathcal{P}_{j+1,j} = \mathcal{P}_j = I_\rho(X_{j+1};\Lambda_{j+1})$$

and so, $\tilde{\rho} = \rho$. On the other hand we have

$$\begin{aligned}
\mathcal{N}_{j+1,j} &\equiv \int p_{\tilde{\rho}}(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_j)\, p(\boldsymbol{\lambda}_j) \log \frac{p_{\tilde{\rho}}(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_j)}{p(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_j)}\, d\mathbf{x}_{j+1}\, d\boldsymbol{\lambda}_j \\
&= \int p_{\tilde{\rho}}(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_{j+1})\, p(\boldsymbol{\lambda}_{j+1}) \log \frac{p_{\tilde{\rho}}(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_{j+1})}{p(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_{j+1})}\, d\mathbf{x}_{j+1}\, d\boldsymbol{\lambda}_{j+1} \\
&= \int p_{\rho}(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_{j+1})\, p(\boldsymbol{\lambda}_{j+1}) \log \frac{p_{\rho}(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_{j+1})}{p(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_{j+1})}\, d\mathbf{x}_{j+1}\, d\boldsymbol{\lambda}_{j+1} = \mathcal{N}_j.
\end{aligned}$$

Thus, $\mathcal{P}_{j+1,j} = \mathcal{P}_j$ and $\mathcal{A}_{j+1,j} = \mathcal{A}_j$, establishing that if a model satisfies the A/P conditions, it also satisfies the CSDL model conditions, and vice versa. Therefore, the two sets of conditions are equivalent.

Consequently, for models for which **(1)** $\mathrm{E}[\mathbf{x}_j|\boldsymbol{\lambda}_j] = \boldsymbol{\lambda}_j$, **(2)** $p(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_j) = p(\mathbf{x}_{j+1}|\boldsymbol{\lambda}_{j+1})$, and **(3)** $H(\boldsymbol{\lambda}_j|\mathbf{x}_{j+1}) > H(\boldsymbol{\lambda}_j|\mathbf{x}_j)$ hold at every valid scale—the same three conditions adopted in Section 2.4.1—finer scale patterns in the data convey the required information for a more precise model, but modeling based solely on this new information necessarily produces a less accurate representation.

APPENDIX B

Appendix to Chapter 3

B.1 Posterior Distributions

In this section we show that

$$f(\boldsymbol{\lambda}_j|\mathbf{c}_0) = f(\boldsymbol{\lambda}_j|\mathbf{c}_j) \tag{B.1}$$

and

$$f(\delta_{j,k}|\mathbf{c}_0) = f(\delta_{j,k}|c_{j-1,2k}, c_{j-1,2k+1}). \tag{B.2}$$

These are proved in the following two propositions[1]

Proposition 1 $p(\boldsymbol{\lambda}_j|\mathbf{c}_0) = p(\boldsymbol{\lambda}_j|\mathbf{c}_j)$ for $j = 0, \cdots, J$

Proof: Since the same information contained in the set $\{\mathbf{c}_0\}$ is contained in $\{\mathbf{c}_0, \mathbf{c}_j\}$,

[1]Throughout this dissertation numerous has been the instances where use of a table of integrals and other mathematical expressions has been made. Our main source of information in this respect has been the tables compounded by Gradshteyn and Ryzhik (see [87]).

APPENDIX B

Appendix to Chapter 3

B.1 Posterior Distributions

In this section we show that

$$p(\chi_j|\omega, \ldots) \propto p(\chi_j|\phi) \qquad (B.1)$$

and

$$p(\phi|\chi_{1:n}, \omega) = p(\phi|\chi_{1:n}) \qquad (B.2)$$

These are proved in the following two propositions.

Proposition 1 $p(\chi_j|\omega) = p(\chi_j|\phi_j)$ for $j = 0, \ldots, n$

Proof: Since the prior information contained in the set $\{\omega\}$ is contained in $\{\omega, \phi\}$

---

[1]Throughout this discussion nothing has been the mathematics which are of a form of integrals and other mathematical expressions has been made. One main source of information in this regard has been the tables compounded by Gradshteyn and Ryzhik [GR 80].

applying Bayes' theorem $f(\boldsymbol{\lambda}_j|\mathbf{c}_0)$ may be written as

$$f(\boldsymbol{\lambda}_j|\mathbf{c}_0) = f(\boldsymbol{\lambda}_j|\mathbf{c}_0, \mathbf{c}_j) = p(\mathbf{c}_0|\mathbf{c}_j, \boldsymbol{\lambda}_j)\frac{p(\boldsymbol{\lambda}_j|\mathbf{c}_j)}{p(\mathbf{c}_0|\mathbf{c}_j)}.$$

Thus, it suffices to show that $p(\mathbf{c}_0|\mathbf{c}_j, \boldsymbol{\lambda}_j) = p(\mathbf{c}_0|\mathbf{c}_j)$ for $j = 1, \cdots, J$.

Define sequences of even and odd elements of $\mathbf{c}_{j-1}$ and $\boldsymbol{\lambda}_{j-1}$: $\mathbf{c}^e_{j-1} = (c_{j-1,0}, c_{j-1,2}, \cdots, c_{j-1,N/2^{j-1}-2})$ and $\mathbf{c}^o_{j-1} = (c_{j-1,1}, c_{j-1,3}, \cdots, c_{j-1,N/2^{j-1}-1})$, and similarly for $\boldsymbol{\lambda}^e_{j-1}$ and $\boldsymbol{\lambda}^o_{j-1}$. Clearly,

$$p(\mathbf{c}_{j-1}|\mathbf{c}_j, \boldsymbol{\lambda}_j) = \begin{cases} \frac{p(\mathbf{c}^e_{j-1}, \mathbf{c}^o_{j-1}|\boldsymbol{\lambda}_j)}{p(\mathbf{c}_j|\boldsymbol{\lambda}_j)} & \text{if } \mathbf{c}^o_{j-1} = \mathbf{c}_j - \mathbf{c}^e_{j-1} \geq \underline{0} \\ 0 & \text{otherwise.} \end{cases} \tag{B.3}$$

Then, with the aid of the total probability theorem we may write for the non-trivial case

$$\begin{aligned} p(\mathbf{c}_{j-1}|\mathbf{c}_j, \boldsymbol{\lambda}_j) &= \frac{\int p(\mathbf{c}^e_{j-1}, \mathbf{c}_j - \mathbf{c}^e_{j-1}|\boldsymbol{\lambda}_j, \boldsymbol{\lambda}^e_{j-1})\, p(\boldsymbol{\lambda}^e_{j-1}|\boldsymbol{\lambda}_j)\, d\boldsymbol{\lambda}^e_{j-1}}{p(\mathbf{c}_j|\boldsymbol{\lambda}_j)} \\ &= \frac{\int \prod_k p(c_{j-1,2k}|\lambda_{j-1,2k})\, \prod_k p(c_{j,k} - c_{j-1,2k}|\lambda_{j,k} - \lambda_{j-1,2k})\, p(\boldsymbol{\lambda}^e_{j-1}|\boldsymbol{\lambda}_j)\, d\boldsymbol{\lambda}^e_{j-1}}{\prod_k p(c_{j,k}|\lambda_{j,k})} \\ &= \frac{\int \prod_k e^{-\lambda_{j-1,2k}} \frac{(\lambda_{j-1,2k})^{c_{j-1,2k}}}{(c_{j-1,2k})!} \prod_k e^{-\lambda_{j,k}+\lambda_{j-1,2k}} \frac{(\lambda_{j,k}-\lambda_{j-1,2k})^{c_{j,k}-c_{j-1,2k}}}{(c_{j,k}-c_{j-1,2k})!}\, p(\boldsymbol{\lambda}^e_{j-1}|\boldsymbol{\lambda}_j)\, d\boldsymbol{\lambda}^e_{j-1}}{\prod_k e^{-\lambda_{j,k}} \frac{(\lambda_{j,k})^{c_{j,k}}}{(c_{j,k})!}} \\ &= \int \prod_k \binom{c_{j,k}}{c_{j-1,2k}} \left(\frac{\lambda_{j-1,2k}}{\lambda_{j,k}}\right)^{c_{j-1,2k}} \left(1 - \frac{\lambda_{j-1,2k}}{\lambda_{j,k}}\right)^{c_{j,k}-c_{j-1,2k}} p(\boldsymbol{\lambda}^e_{j-1}|\boldsymbol{\lambda}_j)\, d\boldsymbol{\lambda}^e_{j-1}, \end{aligned}$$

where all the indicated products are over $k = 0$ through $N/2^j - 1$. In these expressions we have exploited the conditional independence of the data scaling coefficients at any specific scale given their corresponding intensity scaling coefficients.

Since the innovation variates are given by $y_{j,k} = \frac{\lambda_{j-1,2k}}{\lambda_{j,k}}$ (see (3.11)), the conditional densities within the integral signs may also be written as $p_{\boldsymbol{\lambda}^e_{j-1}|\boldsymbol{\lambda}_j}(\boldsymbol{\lambda}^e_{j-1}|\boldsymbol{\lambda}_j) =$

$\frac{1}{\prod_k \lambda_{j,k}} p_{\mathbf{y}_j}\left(\left(\frac{\lambda_{j-1,2k}}{\lambda_{j,k}}\right)_k \mid (\lambda_{j,k})_k\right)$, where $\mathbf{y}_j = (y_{j,0}, y_{j,1}, \cdots, y_{j,N/2^j-1})$. Then, with a change of variables and recalling the mutual independence of $\mathbf{y}_j$ and $\boldsymbol{\lambda}_j$, we obtain the equivalent expression

$$p(\mathbf{c}_{j-1}|\mathbf{c}_j, \boldsymbol{\lambda}_j) = \int \prod_k \binom{c_{j,k}}{c_{j-1,2k}} (y_{j,k})^{c_{j-1,2k}} (1-y_{j,k})^{c_{j,k}-c_{j-1,2k}} p(\mathbf{y}_j)\, d\mathbf{y}_j. \quad \text{(B.4)}$$

The absence of $\boldsymbol{\lambda}_j$ in this expression indicates that $\mathbf{c}_{j-1}|\mathbf{c}_j$ is, at least explicitly, independent of $\boldsymbol{\lambda}_j$. However, one may argue that $\boldsymbol{\lambda}_j$ intervenes only functionally to define the probability mass $p(\mathbf{c}_{j-1}|\mathbf{c}_j, \boldsymbol{\lambda}_j)$. That is, if we denote the right side of (B.4) by $g(\mathbf{c}_{j-1}, \mathbf{c}_j)$, we must still verify whether $g(\mathbf{c}_{j-1}, \mathbf{c}_j) = p(\mathbf{c}_{j-1}|\mathbf{c}_j)$. We may achieve this as follows.

$$\begin{aligned} p(\mathbf{c}_{j-1}|\mathbf{c}_j) &= \int p(\mathbf{c}_{j-1}|\mathbf{c}_j, \boldsymbol{\lambda}_j)\, p(\boldsymbol{\lambda}_j|\mathbf{c}_j)\, d\boldsymbol{\lambda}_j \\ &= \int g(\mathbf{c}_{j-1}, \mathbf{c}_j)\, p(\boldsymbol{\lambda}_j|\mathbf{c}_j)\, d\boldsymbol{\lambda}_j = g(\mathbf{c}_{j-1}, \mathbf{c}_j). \end{aligned}$$

Therefore, $p(\mathbf{c}_{j-1}|\mathbf{c}_j, \boldsymbol{\lambda}_j) = p(\mathbf{c}_{j-1}|\mathbf{c}_j)$.

Using this result, one arrives at the desired final conclusion $p(\mathbf{c}_0|\mathbf{c}_j, \boldsymbol{\lambda}_j) = p(\mathbf{c}_0|\mathbf{c}_j)$ using induction: Let l be any integer such that $0 \le l \le j-1$, and assume that $p(\mathbf{c}_{j-l}|\mathbf{c}_j, \boldsymbol{\lambda}_j) = p(\mathbf{c}_{j-l}|\mathbf{c}_j)$. Clearly,

$$\begin{aligned} &p(\mathbf{c}_{j-l-1}|\mathbf{c}_j, \boldsymbol{\lambda}_j) \\ &\quad = \int \sum_{\mathbf{c}_{j-l}} p(\mathbf{c}_{j-l-1}|\mathbf{c}_{j-l}, \boldsymbol{\lambda}_{j-l}, \mathbf{c}_j, \boldsymbol{\lambda}_j) p(\mathbf{c}_{j-l}|\boldsymbol{\lambda}_{j-l}, \mathbf{c}_j, \boldsymbol{\lambda}_j) p(\boldsymbol{\lambda}_{j-l}|\mathbf{c}_j, \boldsymbol{\lambda}_j)\, d\boldsymbol{\lambda}_{j-l} \\ &\quad = \sum_{\mathbf{c}_{j-l}} \int p(\mathbf{c}_{j-l-1}|\mathbf{c}_{j-l}, \boldsymbol{\lambda}_{j-l}) p(\mathbf{c}_{j-l}|\boldsymbol{\lambda}_{j-l}, \mathbf{c}_j, \boldsymbol{\lambda}_j) p(\boldsymbol{\lambda}_{j-l}|\mathbf{c}_j, \boldsymbol{\lambda}_j)\, d\boldsymbol{\lambda}_{j-l} \end{aligned}$$

Now, since $p(\mathbf{c}_{j-1}|\mathbf{c}_j, \boldsymbol{\lambda}_j) = p(\mathbf{c}_{j-1}|\mathbf{c}_j)$ and the assumption,

$$\begin{aligned} p(\mathbf{c}_{j-l-1}|\mathbf{c}_j, \boldsymbol{\lambda}_j) &= \sum_{\mathbf{c}_{j-l}} p(\mathbf{c}_{j-l-1}|\mathbf{c}_{j-l}) p(\mathbf{c}_{j-l}|\mathbf{c}_j, \boldsymbol{\lambda}_j) \\ &= \sum_{\mathbf{c}_{j-l}} p(\mathbf{c}_{j-l-1}|\mathbf{c}_{j-l}, \mathbf{c}_j) p(\mathbf{c}_{j-l}|\mathbf{c}_j) \;=\; p(\mathbf{c}_{j-l-1}|\mathbf{c}_j). \end{aligned}$$

Taking $l = j - 1$ completes the proof. Note that we made use of the fact that given scaling coefficients at a given scale, knowledge of the corresponding scaling coefficients at a coarser scale (greater value of j) does not alter the likelihood of an event.

Proposition 2 $f(\delta_{j,k}|\mathbf{c}_0) = f(\delta_{j,k}|c_{j-1,2k}, c_{j-1,2k+1})$ for $j = 1, \cdots, J$

Proof: In order to keep the mathematical notation as clean as possible in these and subsequent derivations, we introduce the following simplifying notations: $c_1 = c_{j-1,2k}$ and $c_2 = c_{j-1,2k+1}$, $\mathbf{c}^*_{j-1} = \mathbf{c}_{j-1} \backslash (c_1, c_2)$, and $\mathbf{y}^*_j = \mathbf{y}_j \backslash y_{j,k}$. Now, since $y_{j,k} = \frac{1}{2}(1 + \delta_{j,k})$ (see (3.10)), it suffices to show that $p(y_{j,k}|\mathbf{c}_0) = p(y_{j,k}|c_1, c_2)$ for $j = 1, \cdots, J$. This may be done as follows.

By the total probability and Bayes' theorems,

$$\begin{aligned} p(y_{j,k}|\mathbf{c}_0) &= \int p(\mathbf{y}_j, \boldsymbol{\lambda}_j|\mathbf{c}_0)\, d\mathbf{y}^*_j\, d\boldsymbol{\lambda}_j \\ &= \int \frac{p(\mathbf{c}_0|\mathbf{y}_j, \boldsymbol{\lambda}_j)}{p(\mathbf{c}_0)} p(\mathbf{y}_j, \boldsymbol{\lambda}_j)\, d\mathbf{y}^*_j\, d\boldsymbol{\lambda}_j. \end{aligned}$$

Since the information conveyed by $\{\mathbf{y}_j, \boldsymbol{\lambda}_j\}$ is the same as that conveyed by $\{\boldsymbol{\lambda}_{j-1}\}$, $p(\mathbf{c}_0|\mathbf{y}_j, \boldsymbol{\lambda}_j) = p(\mathbf{c}_0|\boldsymbol{\lambda}_{j-1}) = p(\boldsymbol{\lambda}_{j-1}|\mathbf{c}_0)p(\mathbf{c}_0)/p(\boldsymbol{\lambda}_{j-1})$. Using Proposition 1, this becomes $p(\mathbf{c}_0|\mathbf{y}_j, \boldsymbol{\lambda}_j) = p(\boldsymbol{\lambda}_{j-1}|\mathbf{c}_{j-1})p(\mathbf{c}_0)/p(\boldsymbol{\lambda}_{j-1}) = p(\mathbf{c}_{j-1}|\boldsymbol{\lambda}_{j-1})p(\mathbf{c}_0)/p(\mathbf{c}_{j-1})$. Upon

substitution in the above integral

$$\begin{aligned}
p(y_{j,k}|\mathbf{c}_0) &= \int \frac{p(\mathbf{c}_{j-1}|\boldsymbol{\lambda}_{j-1})}{p(\mathbf{c}_{j-1})} p(\mathbf{y}_j, \boldsymbol{\lambda}_j)\, d\mathbf{y}_j^*\, d\boldsymbol{\lambda}_j \\
&= \int \frac{p(\mathbf{c}_{j-1}|\mathbf{y}_j, \boldsymbol{\lambda}_j)}{p(\mathbf{c}_{j-1})} p(\mathbf{y}_j, \boldsymbol{\lambda}_j)\, d\mathbf{y}_j^*\, d\boldsymbol{\lambda}_j \\
&= \int p(c_1, c_2|y_{j,k}, \lambda_{j,k}) \frac{p(\mathbf{c}_{j-1}^*|\mathbf{y}_j^*, \boldsymbol{\lambda}_j)}{p(\mathbf{c}_{j-1})} p(\mathbf{y}_j, \boldsymbol{\lambda}_j)\, d\mathbf{y}_j^*\, d\boldsymbol{\lambda}_j \\
&= p(y_{j,k}) \int p(c_1, c_2|y_{j,k}, \lambda_{j,k}) \frac{p(\mathbf{c}_{j-1}^*|\boldsymbol{\lambda}_j)}{p(\mathbf{c}_{j-1})} p(\boldsymbol{\lambda}_j)\, d\boldsymbol{\lambda}_j \\
&= p(y_{j,k})\, y_{j,k}^{c_1} (1 - y_{j,k})^{c_2} \int \frac{\lambda^{c_1+c_2} e^{-\lambda}}{c_1!\, c_2!} \frac{p(\mathbf{c}_{j-1}^*|\boldsymbol{\lambda}_j)}{p(\mathbf{c}_{j-1})} p(\boldsymbol{\lambda}_j)\, d\boldsymbol{\lambda}_j.
\end{aligned}$$

Thus, $p(y_{j,k}|\mathbf{c}_0) = p(y_{j,k}|c_1, c_2)$.

B.2 Optimal Estimation of the Multiplicative Innovation

In this section we find a closed form for the optimal estimate $\widehat{\delta}_{j,k}$ of the innovation coefficient $\delta_{j,k}$. For the sake of simplicity, in the few steps that follow we will disregard the indices j and k, and simply write δ for $\delta_{j,k}$ and similarly for other quantities.

The minimum mean square error (mmse) optimal estimate of the innovation coefficient δ, given all the information available in $\mathbf{c}_0$ is given by

$$\widehat{\delta} \equiv \mathrm{E}[\delta|\mathbf{c}_0] = \int_{-1}^{1} \delta\, p(\delta|\mathbf{c}_0)\, d\delta = \int_{-1}^{1} \delta\, p(\delta|c_1, c_2)\, d\delta \tag{B.5}$$

in accordance with (B.2). Applying Bayes' theorem to this expression we obtain

$$\widehat{\delta} = \frac{\int_{-1}^{1} \delta\, p(c_1, c_2|\delta) p(\delta)\, d\delta}{\int_{-1}^{1} p(c_1, c_2|\delta) p(\delta)\, d\delta} = \frac{\int_{-1}^{1} \int_{0}^{\infty} \delta\, p(c_1, c_2|\delta, \lambda) p(\lambda|\delta)\, d\lambda\, p(\delta)\, d\delta}{\int_{-1}^{1} \int_{0}^{\infty} p(c_1, c_2|\delta, \lambda) p(\lambda|\delta)\, d\lambda\, p(\delta)\, d\delta}. \tag{B.6}$$

Here, $p(\lambda|\delta) = p(\lambda)$ due to the independence of $\lambda_{j,k}$ and $\delta_{j,k}$. Also, notice that c_1 and

substitution in the above integral

$$
\begin{aligned}
p(\theta_{ij}|\sigma_0^2) &= \int \frac{p(\sigma_0^{-2}|\theta_{ij})}{p(\sigma_0^{-2})} p(\theta_{ij}|\gamma_{ij}) p(\gamma_{ij}) \, d\gamma_{ij} \\
&= \int \frac{p(\sigma_0^{-2}|\theta_{ij})}{p(\sigma_0^{-2})} p(\sigma_0^{-2}|\gamma_{ij}) p(\gamma_{ij}) \, d\gamma_{ij} \\
&= \int p(\sigma_0^{-2}|\theta_{ij}) \frac{p(\sigma_0^{-2}|\gamma_{ij})}{p(\sigma_0^{-2})} p(\gamma_{ij}) \, d\gamma_{ij} \\
&= p(\theta_{ij}) \int p(\sigma_0^{-2}|\gamma_{ij}) \frac{p(\sigma_0^{-2}|\gamma_{ij})}{p(\sigma_0^{-2})} p(\gamma_{ij}) \, d\gamma_{ij} \\
&= p(\theta_{ij}) \int_0^1 p(\gamma_{ij}) (1 - p_0)^2 \, d\gamma_{ij} = p(\theta_{ij}) (1 - p_0)^2
\end{aligned}
$$

Thus, $p(\theta_{ij}|\sigma_0) = p(\theta_{ij}|\sigma_0^2)$.

B.2 Optimal Estimation of the Multiplicative Innovation

In this section we find a closed form for the optimal estimate $\hat{\phi}_{ij}$ of the innovation coefficient ϕ_{ij}. For the sake of simplicity in the new steps that follow we will disregard the indices i and j, and simply write ϕ for ϕ_{ij} and similarly for other quantities.

The minimum mean square error (mmse) optimal estimate of the innovation coefficient ϕ, given all the data available to us is given by

$$
\hat{\phi} \equiv \mathrm{E}[\phi|\sigma] = \int_0^1 \phi \, p(\phi|\sigma) \, d\phi = \int_0^1 \phi \, p(\phi|\sigma, \sigma_0^2) \, d\phi \tag{B.5}
$$

in accordance with (B.2). Applying Bayes' theorem to this expression we obtain

$$
\hat{\phi} = \frac{\int_0^1 \phi \, p(\sigma, \sigma_0^2|\phi) p(\phi) \, d\phi}{\int_0^1 \phi \, p(\sigma, \sigma_0^2|\phi) p(\phi) \, d\phi} = \frac{\int_0^1 \int_0^\infty \phi \, p(\sigma, \sigma_0^2|\phi, \gamma) p(\gamma|\phi) \, p_\gamma(\phi) \, d\gamma \, p(\phi) \, d\phi}{\int_0^1 \int_0^\infty \phi \, p(\sigma, \sigma_0^2|\phi, \gamma) p(\gamma|\phi) \, p_\gamma(\phi) \, d\gamma \, p(\phi) \, d\phi} \tag{B.6}
$$

Here $p(\gamma|\phi) = p_\gamma(\gamma)$ due to the independence of γ_{ij} and ϕ_{ij}. Also, notice that σ_0 and

c_2 only depend on λ and δ by way of their respective statistical means $\lambda_1 = \frac{1}{2}\lambda(1+\delta)$ and $\lambda_2 = \frac{1}{2}\lambda(1-\delta)$ (See (3.11) and (3.12)). Given λ_1 and λ_2, λ and δ do not convey any further information on the behavior of c_1 and c_2. Moreover, given λ_1 and λ_2, c_1 and c_2 are independent. Therefore,

$$\begin{aligned}\widehat{\delta} &= \frac{\int_{-1}^{1}\int_{0}^{\infty}\delta\, p\left(c_1, c_2 \left| \lambda\tfrac{1+\delta}{2}, \lambda\tfrac{1-\delta}{2}\right.\right) p(\lambda)\, d\lambda\, p(\delta)\, d\delta}{\int_{-1}^{1}\int_{0}^{\infty} p\left(c_1, c_2 \left| \lambda\tfrac{1+\delta}{2}, \lambda\tfrac{1-\delta}{2}\right.\right) p(\lambda)\, d\lambda\, p(\delta)\, d\delta} \\ &= \frac{\int_{-1}^{1}\int_{0}^{\infty}\delta\, \frac{e^{-\lambda(1+\delta)/2}}{c_1!}\left(\lambda\tfrac{1+\delta}{2}\right)^{c_1}\frac{e^{-\lambda(1-\delta)/2}}{c_2!}\left(\lambda\tfrac{1-\delta}{2}\right)^{c_2} p(\lambda)\, d\lambda\, p(\delta)\, d\delta}{\int_{-1}^{1}\int_{0}^{\infty}\frac{e^{-\lambda(1+\delta)/2}}{c_1!}\left(\lambda\tfrac{1+\delta}{2}\right)^{c_1}\frac{e^{-\lambda(1-\delta)/2}}{c_2!}\left(\lambda\tfrac{1-\delta}{2}\right)^{c_2} p(\lambda)\, d\lambda\, p(\delta)\, d\delta}.\end{aligned}$$

It is now possible to factor out every λ-dependent term from the integrals in δ. Doing this, the resulting integrals in λ in the numerator and denominator cancel out leading to

$$\widehat{\delta} = \frac{\int_{-1}^{1}\delta\,(1+\delta)^{c_1}(1-\delta)^{c_2}\, p(\delta)\, d\delta}{\int_{-1}^{1}(1+\delta)^{c_1}(1-\delta)^{c_2}\, p(\delta)\, d\delta}. \tag{B.7}$$

Substituting the beta mixture model (3.13) into this expression and carrying out the integration we obtaine the desired result:

$$\widehat{\delta}_{j,k} = d_{j,k}\frac{\sum_i p_i \frac{B(s_i+c_1, s_i+c_2)}{B(s_i,s_i)\,(2s_i+c)}}{\sum_i p_i \frac{B(s_i+c_1, s_i+c_2)}{B(s_i,s_i)}}. \tag{B.8}$$

APPENDIX C

Appendix to Chapter 4

C.1 The Hilbert Transform as a Continuous Averaging Process

The Hilbert transform of an $L^2(\mathbb{R})$ function f is defined as

$$\mathcal{H}f(t) \equiv \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(\tau)}{t-\tau}\, d\tau, \tag{C.1}$$

which is the convolution of f and $\frac{1}{\pi t}$. For functions with finite support, it is possible to give a more intuitive meaning to this operation. In particular, if $\text{supp}(f) = [a, b]$, where $a < b$, then

$$\mathcal{H}f(t) = -\frac{b-a}{\pi} \int_0^1 m_{\sigma_1,\sigma_2}(\tau)\, d\tau, \tag{C.2}$$

where

$$m_{\sigma_1,\sigma_2}(\tau) \equiv \frac{f(\sigma_2(\tau)) - f(\sigma_1(\tau))}{\sigma_2(\tau) - \sigma_1(\tau)}, \tag{C.3}$$

APPENDIX C

Appendix to Chapter 4

C.1 The Hilbert Transform as a Continuous Averaging Process

The Hilbert transform of an $L^2(\mathbb{R})$ function f is defined as

$$\mathcal{H}f(x) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(t)}{x-t}\, dt. \tag{C.1}$$

which is the convolution of f and $\frac{1}{\pi x}$. For functions with finite support, it is possible to give a more intuitive meaning to this operation. In particular, if $\operatorname{supp}(f) = [a, b]$, where $a < b$, then

$$\mathcal{H}f(x) = -\frac{b-a}{\pi} \int_0^1 m_{\sigma_x, \sigma_x}(\tau)\, d\tau, \tag{C.2}$$

where

$$m_{\sigma_x, \sigma_x}(\tau) \equiv \frac{f(\sigma_x(\tau)) - f(\sigma_x(\tau))}{\sigma_x(\tau) - \sigma_x(\tau)}. \tag{C.3}$$

and $\sigma_1(\tau) \equiv (a-t)\tau + t$ and $\sigma_2(\tau) \equiv (b-t)\tau + t$. Note that $\sigma_1(0) = \sigma_2(0) = t$, while $\sigma_1 \longrightarrow a$ and $\sigma_2 \longrightarrow b$ linearly as $\tau \longrightarrow 1$.

Clearly, if the derivative of f exists everywhere in (a,b), $m_{\sigma_1,\sigma_2}(\tau)$ is the average derivative of f in the interval $(\sigma_1(\tau), \sigma_2(\tau)) \subset [a,b]$, *i.e.*,

$$m_{\sigma_1,\sigma_2}(\tau) = \frac{1}{\sigma_2(\tau) - \sigma_1(\tau)} \int_{\sigma_1(\tau)}^{\sigma_2(\tau)} f'(\sigma)\, d\sigma. \tag{C.4}$$

Proof: Substituting C.3 into C.2, and replacing the definitions for $\sigma_1(\tau)$ and $\sigma_2(\tau)$, we obtain

$$\mathcal{H}f(t) = -\frac{1}{\pi} \int_0^1 f((b-t)\tau + t)\, \frac{d\tau}{\tau} + \frac{1}{\pi} \int_0^1 f((a-t)\tau + t)\, \frac{d\tau}{\tau}$$

Now, in the first integral let $u = (a-t)\tau + t$, and in the second integral let $u = (b-t)\tau + t$. In both cases we have that $\frac{d\tau}{\tau} = \frac{du}{u-t}$, which upon substitution we get the desired result C.1:

$$\mathcal{H}f(t) = -\frac{1}{\pi} \int_t^b f(u)\, \frac{du}{u-t} + \frac{1}{\pi} \int_t^a f(u)\, \frac{du}{u-t}.$$

C.2 Proof of Expression (4.24)

In this section, we investigate the nature of the approximation

$$\int_0^{2\pi} \frac{\mathrm{Sa}\left(\frac{N}{2}\theta - \pi n\right)}{(1 - r\cos(\theta - \phi))^2}\, d\theta \approx \frac{2\pi/N}{(1 - r\cos(\phi - \frac{2\pi}{N}n))^2} \tag{C.5}$$

for $n = 0, 1, \ldots, N-1$, and find the conditions under which it is adequate.

Let $\mathrm{rect}(x)$ be the function that is 1 in $[-\frac{x}{2}, \frac{x}{2}]$, and zero otherwise. Then, the

Fourier transform of $\mathrm{Sa}(\frac{N}{2}\theta)$ is $\frac{2\pi}{N}\mathrm{rect}(\frac{\omega}{N})$. Thus, $\mathrm{Sa}\left(\frac{N}{2}\theta\right) = \frac{1}{N}\int_{-N/2}^{N/2} e^{i\omega\theta}\,d\omega$, and

$$\begin{aligned}\int_0^{2\pi} \frac{\mathrm{Sa}\left(\frac{N}{2}\theta - \pi n\right)}{(1 - r\cos(\theta-\phi))^2}\,d\theta &= \frac{1}{N}\int_{-N/2}^{N/2} e^{-i\omega\frac{2\pi}{N}n}\int_0^{2\pi}\frac{e^{i\omega\theta}}{(1 - r\cos(\theta-\phi))^2}\,d\theta\,d\omega \\ &= \frac{1}{2\pi}\int_{-N/2}^{N/2} e^{i\omega\frac{2\pi}{N}n}\int_{-\infty}^{\infty}\frac{\frac{2\pi}{N}\,\mathrm{rect}(\frac{\theta-\pi}{2\pi})}{(1 - r\cos(\theta-\phi))^2}\,e^{-i\omega\theta}\,d\theta\,d\omega,\end{aligned}$$

or

$$\int_0^{2\pi} \frac{\mathrm{Sa}\left(\frac{N}{2}\theta - \pi n\right)}{(1 - r\cos(\theta-\phi))^2}\,d\theta = \frac{1}{2\pi}\int_{-N/2}^{N/2}\left(\frac{\frac{2\pi}{N}\,\mathrm{rect}(\frac{\theta-\pi}{2\pi})}{(1 - r\cos(\theta-\phi))^2}\right)^{\wedge} e^{i\omega\frac{2\pi}{N}n}\,d\omega, \qquad \text{(C.6)}$$

where as usual, $(\cdot)^{\wedge}$ denotes Fourier transformation. Clearly, if the "bandwidth" of $\frac{\frac{2\pi}{N}\,\mathrm{rect}(\frac{\theta-\pi}{2\pi})}{(1-r\cos(\theta-\phi))^2}$ were less or equal to $N/2$, the right integral of (C.6) would reverse its Fourier transformation, giving (C.5) with equality. This, however, can never occur since $\frac{\frac{2\pi}{N}\,\mathrm{rect}(\frac{\theta-\pi}{2\pi})}{(1-r\cos(\theta-\phi))^2}$ has compact support, and so, its bandwidth is never finite. Note that C.6 is simply a restatement of the Nyquist Sampling Theorem and the reconstruction series which derives from it [88].

Although we must content ourselves with an approximation, we can choose the parameters in (C.5) such that its two sides are as close as desired. To see how this is possible, first note that the bandwidth of $\frac{\frac{2\pi}{N}\,\mathrm{rect}(\frac{\theta-\pi}{2\pi})}{(1-r\cos(\theta-\phi))^2}$ does not change with N, and so, the right-side integral of (C.6) monotonically approaches $\frac{\frac{2\pi}{N}\,\mathrm{rect}(\frac{\theta-\pi}{2\pi})}{(1-r\cos(\theta-\phi))^2}$ as N increases. Variations on ϕ only affect the phase of the Fourier transform and does not modify the bandwidth. On the other hand, as the value of r goes from 0 to 1, $\frac{\frac{2\pi}{N}\,\mathrm{rect}(\frac{\theta-\pi}{2\pi})}{(1-r\cos(\theta-\phi))^2}$ goes from a constant to a U shape function on its support $[0, 2\pi]$, drastically increasing its high-frequency content. A detailed analysis would show that this increase is monotonic because the energy increase in the signal continuously adds towards building up the singularities at the interval's boundaries.

In order to find values for r and N that make (C.5) a good approximation, numerical solutions were obtained. The Fourier magnitude of $\frac{\frac{2\pi}{N}\,\mathrm{rect}(\frac{\theta-\pi}{2\pi})}{(1-r\cos(\theta-\phi))^2}$ for four

Fourier transform of $\mathrm{Sa}(\frac{N}{2}\theta)$ is $\frac{2\pi}{N}\mathrm{rect}(\frac{\omega}{N})$. Thus $\mathrm{Sa}(\frac{N}{2}\theta) = \frac{N}{2\pi}\int_{-\pi N}^{\pi N} e^{j\omega\theta}\,d\omega$ and

$$\int_0^{2\pi} \frac{\mathrm{Sa}(\frac{N}{2}(\theta-\pi n))}{(1-r\cos(\theta-\phi))^2}\,d\theta = \frac{N}{2\pi}\int_{-\pi N}^{\pi N} e^{-j\omega\pi n}\int_0^{2\pi} \frac{e^{j\omega\theta}}{(1-r\cos(\theta-\phi))^2}\,d\theta\,d\omega$$

$$= \frac{1}{2\pi}\int_{-\pi N}^{\pi N} \frac{\frac{2\pi}{N}\,\mathrm{rect}(\frac{\omega}{N})}{(1-r\cos(\phi))^2}\,e^{-j\omega\pi n}\int_0^{2\pi} e^{j\omega\theta}\,d\theta\,d\omega$$

or

$$\int_0^{2\pi} \frac{\mathrm{Sa}(\frac{N}{2}\theta-\pi n)}{(1-r\cos(\theta-\phi))^2}\,d\theta = \frac{1}{2\pi}\int_{-\pi N}^{\pi N} \left(\frac{\frac{2\pi}{N}\,\mathrm{rect}(\frac{\omega}{N})}{(1-r\cos(\phi))^2}\right) e^{j\omega\pi n}\,d\omega \qquad (C.6)$$

where as usual, $(\cdot)^*$ denotes Fourier transformation. Clearly in the "bandwidth" of $\frac{2\pi\,\mathrm{rect}(\frac{\omega}{2\pi N})}{(1-r\cos(\phi))^2}$ were less or equal to $N/2$, the right integral of (C.6) would reverse to its Fourier transformation, giving (C.5) with aliasing. However, this never occur since $\frac{2\pi\,\mathrm{rect}(\frac{\omega}{2\pi N})}{(1-r\cos(\phi))^2}$ has compact support, and so it cannot be in L^2 entirely. Note that C.6 is simply a restatement of the Nyquist Sampling Theorem and the reconstruction series which derives from it [ ].

Although we must content ourselves with an approximation, we can choose the parameters in (C.5) such that its two sides are as close as desired. To see how this is possible, first note that the bandwidth of $\frac{2\pi\,\mathrm{rect}(\frac{\omega}{2\pi N})}{(1-r\cos(\phi))^2}$ does not change with N, and so, the right-side integral of (C.6) monotonically approaches $\frac{2\pi\,\mathrm{rect}(\frac{\omega}{2\pi N})}{(1-r\cos(\phi))^2}$ as N increases. Variations on ϕ only affect the place of the Fourier transform and does not modify the bandwidth. On the other hand, as the value of r goes from 0 to 1, $\frac{2\pi\,\mathrm{rect}(\frac{\omega}{2\pi N})}{(1-r\cos(\phi))^2}$ goes from a constant to a U shape function on its support $[0, 2\pi]$, drastically increasing its high-frequency content. A detailed analysis would show that this increase is monotonic because the energy increase in the signal continuously adds towards building up the singularities at the interval's boundaries.

In order to find values for r and N that make (C.6) a good approximation, numerical solutions were obtained. The Fourier magnitude of $\frac{2\pi\,\mathrm{rect}(\frac{\omega}{2\pi N})}{(1-r\cos(\phi))^2}$ for four

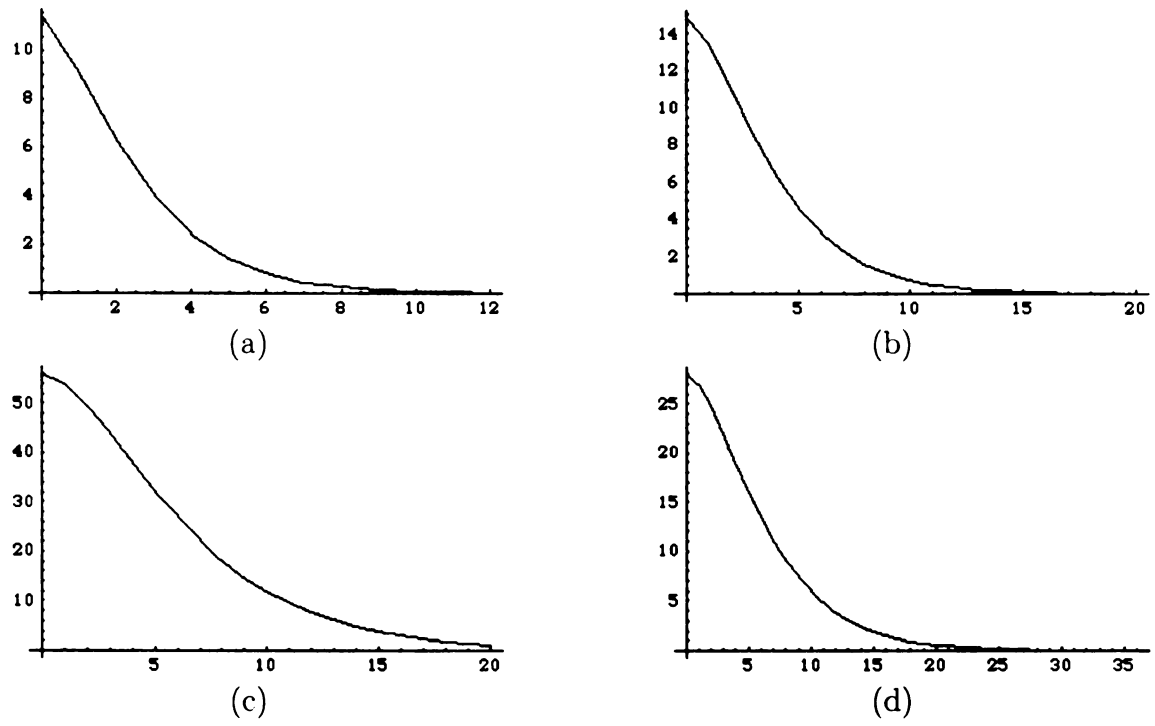


Figure C.1. *Magnitude frequency content of* $\frac{\frac{2\pi}{N}\,\mathrm{rect}(\frac{\theta-\pi}{2\pi})}{(1-r\cos(\theta-\phi))^2}$ *(abscissa in radians/s) for* $\phi = 0$ *and (a)* $N = 16$, $r = .8$, *(b)* $N = 32$, $r = .9$, *(c)* $N = 32$, $r = .96$, *and (d)* $N = 64$, $r = .96$,

combinations of N and r are shown in Figure C.1.

It is seen from these graphs that the right-hand side integration operation of C.6 would change very little if it were over $\mathbb{R}$ rather than over $[-\frac{N}{2}, \frac{N}{2}]$; although Figure C.1(c) probably corresponds to an acceptable upper limit for r when $N = 32$. The decay in the frequency response of $\frac{\frac{2\pi}{N}\,\mathrm{rect}(\frac{\theta-\pi}{2\pi})}{(1-r\cos(\theta-\phi))^2}$ at the lower and upper limits of integration $-\frac{N}{2}$ and $\frac{N}{2}$ in each graph is as follows: (a) 38 dB, (b) 53 dB, (c) 30 dB, and (d) 67 dB.

(a)

(b)

(c)

(d)

Figure C.1: Magnitude frequency response of $\frac{\omega_0^2}{s^2+\frac{\omega_0}{Q}s+\omega_0^2}$ (see text in continuation) for $\delta = 0$ and (a) $N = 10, r = -8$, (b) $N = 22 > -\delta \cdot (r) \lambda = 13, r = -30$, and (d) $N = 64, r = -30$.

combinations of N and r are shown in Fig. no. C.1.

It is seen from these graphs that the right-hand side recognition operation of C.6 would change very little if it were over R rather than over $[-\frac{\omega}{2}, \frac{\omega}{2}]$, although Figure C.1(c) probably corresponds to an acceptable upper limit for r when $N = 32$. The decay in the frequency response of $\frac{\omega_0^2}{s^2+\frac{\omega_0}{Q}s+\omega_0^2}$ at the lower and upper limits of integration $-\frac{\omega}{2}$ and $\frac{\omega}{2}$ in each graph is as follows: (a) 25 dB, (b) 33 dB, (c) 30 dB, and (d) 57 dB.

BIBLIOGRAPHY

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] D. L. Snyder and M. I. Miller, *Random Point Processes in Time and Space.* New York: Springer-Verlag, 1991.

[2] D. L. Snyder, A. M. Hammoud, and R. L. White, "Image recovery from data acquired with a charge-coupled-device camera," *J. Opt. Soc. Am. A*, vol. 10, no. 5, pp. 1014–1023, 1993.

[3] H. Heffes and D. M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE J. Selected Areas Comm.*, vol. 4, pp. 856–868, 1986.

[4] Y. Vardi, "Network tomography: Estimating source-destination traffic intensities from link data," *J. Amer. Statist. Assoc.*, vol. 91, pp. 365–377, 1996.

[5] M. A. King, R. B. Schwinger, P. W. Doherty, and B. C. Penney, "Two-dimensional filtering of SPECT images using the Metz and Wiener filters," *J. Nuc. Med.*, vol. 25, pp. 1234–1240, 1984.

[6] G. A. O'Driscoll and et al., "Differences in cerebral activation during smooth pursuit and saccadic eye movements using positron-emission tomography," *BIOLOGICAL PSYCHIATRY*, vol. 44, no. 8, pp. 685–689, 1998.

[7] E. E. Smith, J. Jonides, and et al., "Components of verbal working memory: Evidence from neuroimaging," *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES*, vol. 95, no. 3, pp. 876–882, 1998.

[8] R. Dume and P. Jansen, "Computer tomography of barrels with radioactive contents," *Nuclear Engineering and Design*, vol. 130, pp. 89–102, Sep. 1991.

[9] J. Frank, *Three-dimensional Electron Microscopy of Macromolecular Assemblies.* San Diego: Academic Press, 1996.

[10] F. Abramovich, T. Sapatinas, and B. W. Silverman, "Wavelet thresholding via a Bayesian approach," *Tech. Rep., Math. Dept., Univ. Bristol*, 1996.

[11] H. A. Chipman, E. D. Kolaczyk, and R. E. McCulloch, "Adaptive Bayesian wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 92, pp. 1413–1421, 1997.

[12] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, pp. 1200–1224, Dec. 1995.

[13] I. M. Johnstone and B. W. Silverman, "Wavelet threshold estimators for data with correlated noise," *Journal of the Royal Stastistical Society, Series B*, vol. 59, pp. 319–351, 1997.

[14] E. D. Kolaczyk, "Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds," *Statistica Sinica,* under revision, 1997.

[15] B. Vidakovic, "Nonlinear wavelet shrinkage with Bayes rules and Bayes factors," *J. Amer. Statist. Assoc.*, vol. 93, pp. 173–179, 1998.

[16] R. D. Nowak, R. L. Gregg, T. G. Cooper, and J. E. Siebert, "Removing Rician noise in MRI via wavelet-domain filtering," in *submitted to Annual Meeting of Intl. Soc. Magn. Reson. Med.*, 1998.

[17] P. G. Hewitt, *Conceptual Physics.* Boston: Little, Brown & Company, 1985.

[18] E. L. Lehmann and G. Casella, *Theory of Point Estimation.* New York: 2nd Edition, Springer-Verlag, 1998.

[19] G. H. Hardy and W. W. Rogosinski, *Fourier Series.* New York: Dover Publications, Inc., 1999 republication of the original 1956 edition.

[20] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 674–693, July 1989.

[21] S. Mallat, *A Wavelet Tour of Signal Processing.* San Diego: Academic Press, 1998.

[22] M. Frazier, *An Introduction to Wavelets Through Linear Algebra.* New York: Springer, 1999.

[23] C. P. Robert, *The Bayesian Choice.* New York: Springer-Verlag, 1994.

[24] M. D. Srinath, P. K. Rajasekaran, and R. Viswanathan, *Introduction to Statistical Signal Processing with Applications.* Englewood Cliffs, N.J.: Prentice-Hall, 1996.

[25] H. Stark and J. W. Woods, *Probability, Random Processes, and Estimation Theory for Engineers.* Englewood Cliffs, N.J.: Prentice-Hall, 1994 Second Edition.

[26] S. Wolf, *Guide to Electronic Measurements and Laboratory Practice.* Englewood Cliffs, NJ: Prentice-Hall, 1973.

[27] R. G. Gallager, *Information Theory and Reliable Communication.* New York: John Wiley and Sons, Inc., 1968.

[28] G. B. Folland, *Real Analysis —Modern Techniques and Their Applications.* New York: Wiley-Interscience, 1984.

[29] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication.* Urbana, Illinois: Univ. of Illinois Press, 1963.

[30] A. Papoulis, *Probability, Random Variables, and Stochastic Processes.* New York: McGraw-Hill, 1984.

[31] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: John Wiley & Sons, Inc., 1991.

[32] T. K. Moon and W. C. Stirling, *Mathematical Methods and Algorithms for Signal Processing.* Upper Saddle River, NJ: Prentice Hall, 2000.

[33] R. T. Odgen, *Essential Wavelets for Statistical Applications and Data Analysis.* Boston: Birkhäuser, 1997.

[34] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Processing*, to appear in Special Issue on Theory and Applications of Filter Banks and Wavelets, 1998.

[35] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[36] G. P. Nason and B. W. Silverman, "The stationary wavelet transform and some statistical applications," in *Lecture Notes in Statistics: Wavelets and Statistics*, vol. New York: Springer-Verlag, pp. 281–299, 1995.

[37] N. Weyrich and G. T. Warhola, "De-noising using wavelets and cross-validation," *Technical Report AFIT/EN/TR/94-01, Department of Mathematics and Statistics, Air Force Institute of Technology, Ohio*, 1994.

[38] G. P. Nason, "Wavelet regression by cross-validation," *Technical Report 447, Department Statistics, Standford University, California*, 1994.

[39] G. P. Nason, "Wavelet shrinkage using cross-validation," *Journal of the Royal Statistical Society, Series B*, vol. New York: Springer-Verlag, pp. 58:463–479, 1996.

[40] R. D. Nowak, "Optimal signal estimation using cross-validation," *IEEE Signal Processing Letters*, vol. 4, no. 1, pp. 23–25, 1997.

[41] F. Abramovich and Y. Benjamini, *Thresholding of wavelet coefficients as multiple hypotheses testing procedure.* New York: Springer-Verlag, 1995.

[42] H. C. Andrews and B. R. Hunt, *Digital Image Restoration.* Englewood Cliffs, New Jersey: Prentice Hall, 1977.

[43] D. T. Kuan, A. A. Sawchuk, T. C. Strand, and P. Chavel, "Adaptive noise smoothing filter for images with signal-dependent noise," *IEEE Trans. Pattern Anal. Machine Intelligence*, vol. 7, pp. 165–177, 1985.

[44] R. D. Nowak and R. G. Baraniuk, "Wavelet-domain filtering for photon imaging systems," *Proc. SPIE, Wavelet Applications in Signal and Image Processing V*, vol. 3169, pp. pp. 55–66, August 1997.

[45] R. Nowak, R. Hellman, D. Nowak, and R. Baraniuk, "Wavelet domain filtering for nuclear medicine imaging," in *Proc. IEEE Med. Imaging Conf.*, pp. 279–290, 1996.

[46] J.-C. Pesquet, H. Krim, and E. Hamman, "Bayesian approach to best basis selection," in *IEEE Int. Conf. on Acoust., Speech, Signal Proc. — ICASSP '96*, (Atlanta), pp. 2634–2637, 1996.

[47] E. P. Simoncelli and E. H. Adelson, "Noise removal via Bayesian wavelet coring," in *IEEE Int. Conf. on Image Proc. — ICIP 1996*, (Switzerland), September 1996.

[48] N. L. Johnson, S. Kotz, and A. W. Kemp, *Univariate Discrete Distributions.* New York: John Wiley and Sons, 1992.

[49] L. L. Scharf, *Statistical Signal Processing. Detection, Estimation, an Time Series Analysis.* Reading, MA: Addison-Wesley, 1991.

[50] B. Mandelbrot, "Intermittant turbulence in self-similar cascades: Divergence of high moments and dimensions of the carrier," *J. Fluid Mech.*, 1974.

[51] S. Lovejoy and D. Schertzer, *Multifractals in rain.* Cambridge, U.K.: Cambridge Press, 1995.

[52] C. J. G. Evertsz and B. B. Mandelbrot, *Multifractal measures.* New York: Springer-Verlag, 1992.

[53] R. D. Mauldin, W. D. Sudderth, and S. C. Williams, "Polya trees and random distributions," *Ann. Stat.*, vol. 20, pp. 1203–1221, 1992.

[54] M. Lavine, "Some aspects of Polya tree distributions for statistical modeling," *Ann. Stat.*, vol. 20, pp. 1222–1235, 1992.

[55] S. LoPresto, K. Ramchandran, and M. T. Orchard, "Image coding based on mixture modeling of wavelet coefficients and a fast estimation-quantization framework," in *Data Compression Conference '97*, (Snowbird, Utah), pp. 221–230, 1997.

[56] E. D. Kolaczyk, "Bayesian multi-scale models for Poisson processes," *Technical Report 468, Department of Statistics, University of Chicago*, 1998.

[57] P. Diaconis and B. Efron, "Computer-intensive methods in statistics," *Scientific American*, pp. 116–129, May 1983.

[58] A. M. Zoubir and B. Boashash, "The bootstrap and its application in signal processing," *Signal Processing Magazine*, vol. 15, no. 1, pp. 56–76, 1998.

[59] R. D. Nowak and R. G. Baraniuk, "Wavelet-based filtering for photon imaging systems," *IEEE Trans. Image Processing*, submitted April 1997.

[60] E. Simoncelli, W. Freeman, E. Adelson, and D. Heeger, "Shiftable multiscale transforms," *IEEE Trans. Inform. Theory*, vol. 38, pp. 587–607, 1992.

[61] R. Coifman and D. Donoho, "Translation invariant de-noising," in *Lecture Notes in Statistics: Wavelets and Statistics*, vol. New York: Springer-Verlag, pp. 125–150, 1995.

[62] M. Lang, H. Guo, J. E. Odegard, C. S. Burrus, and R. O. Wells, "Noise reduction using an undecimated discrete wavelet transform," *IEEE Signal Processing Letters*, vol. 3, no. 1, pp. 10–12, 1996.

[63] J.-C. Pesquet, H. Krim, and H. Carfantan, "Time-invariant orthonormal wavelet representations," *tSP*, vol. 44, no. 8, pp. 1964–1970, 1996.

[51] S. Lovejoy and D. Schertzer, *Multifractals in geo*. Cambridge, U.K.: Cambridge Press, 1995.

[52] C. J. C. Evers and B. B. Mandelbrot, *Multifractal measures*. New York: Springer-Verlag, 1992.

[53] R. D. Mauldin, W. D. Sudderth, and S. C. Williams, "Polya trees and random distributions," *Ann. Stat.*, vol. 20, pp. 1203–1221, 1992.

[54] M. Lavine, "Some aspects of Polya tree distributions for statistical modeling," *Ann. Stat.*, vol. 20, pp. 1222–1235, 1992.

[55] S. LoPresto, K. Ramchandran, and M. T. Orchard, "Image coding based on mixture modeling of wavelet coefficients and a fast estimation-quantization framework," in *Data Compression Conf.*, Snowbird, Utah, pp. 221–230, 1997.

[56] E. D. Kolaczyk, "Bayesian multiscale models for Poisson processes," *Technical Report 456, Department of Statistics, Univ. of Chicago*, 1995.

[57] P. Diaconis and B. Efron, "Computer-intensive methods in statistics," *Scientific American*, pp. 116–129, May 1983.

[58] A. M. Zoubir and B. Boashash, "The bootstrap and its application in signal processing," *Signal Processing Magazine*, vol. 15, no. 1, pp. 56–76, 1998.

[59] R. D. Nowak and R. G. Baraniuk, "Wavelet-based filtering for photon imaging systems," *IEEE Trans. Image Processing*, submitted, April 1997.

[60] E. Simoncelli, W. Freeman, E. Adelson, and D. Heeger, "Shiftable multiscale transforms," *IEEE Trans. Inform. Theory*, vol. 38, pp. 587–607, 1992.

[61] R. Coifman and D. Donoho, "Translation invariant de-noising," in *Lecture Notes in Statistics: Wavelets and Statistics*, vol. *New York: Springer-Verlag*, pp. 125–150, 1995.

[62] M. Lang, H. Guo, J. E. Odegard, C. S. Burrus, and R. O. Wells, "Noise reduction using an undecimated discrete wavelet transform," *IEEE Signal Processing Letters*, vol. 3, no. 1, pp. 10–12, 1996.

[63] J.-C. Pesquet, H. Krim, and H. Carfantan, "Time-invariant orthonormal wavelet representations," *ref?*, vol. 44, no. 8, pp. 1964–1970, 1996.

[64] R. Nowak, "Shift invariant wavelet-based statistical models and $1/f$ processes," in *Proc. IEEE Digital Signal Processing Workshop,* paper no.83, Bryce Canyon, UT, 1998.

[65] A. P. Pentland, "Fractal-based description of natural scenes," *IEEE Trans. Pattern Anal. Machine Intell.*, pp. 661–674, July 1984.

[66] B. J. West and M. F. Shlesinger, "On the ubiquity of $1/f$ noise," *Intl. J. Modern Physics (B)*, vol. 3, no. 6, pp. 795–819, 1989.

[67] G. W. Wornell, *Signal processing with fractals: a wavelet-based approach.* Upper Saddle River, NJ: Prentice-Hall, 1996.

[68] I. Daubechies, *Ten Lectures on Wavelets.* Philidelphia: SIAM CBMS-NSF Series in Applied Mathematics, no. 61, 1992.

[69] J. A. Sorenson and M. E. Phelps, *Physics in Nuclear Medicine.* New York: Grune & Stratton, 1987.

[70] I. M. Gelfand and S. G. Gindikin, eds., *Mathematical Problems of Tomography.* Rhode Island: American Mathematical Society, 1990.

[71] B. V. Jackson, P. L. Hick, M. Kojima, and A. Yokobe, "Heliospheric tomography using interplanetary scintillation observations," in *Proc. of the EGS XXI meeting,* paper no. 15.02, The Hague, Netherlands, 6-10 May, 1996.

[72] X. Descombes, F. Kruggel, and D. Y. von Cramon, "Spatio-temporal FMRI analysis using Markov random fields," *IEEE Trans. Medical Imaging*, vol. 17, no. 6, pp. 1028–1039, 1998.

[73] X. Hu, T. H. Le, T. Parrish, and P. Erhard, "Retrospective estimation and correlation of physiological fluctuation in functional MRI," *Magn. Reason. Med.*, vol. 34, pp. 201–212, 1995.

[74] A. K. Jain, *Fundamentals of Digital Image Processing.* Englewood Cliffs, N.J.: Prentice Hall, 1989.

[75] F. Natterer, *The Mathematics of Computerized Tomography.* Stuttgart: John Wiley & Sons Ltd and B G Teubner, 1986.

[76] T. F. Budinger and G. T. Gullberg, "Transverse section reconstruction of gamma-ray emitting radionuclides in patients," *Reconstruction Tomography in Diagnostic Radiology and Nuclear Medicine*, University Park Press, Baltimore, MD, 1976.

[77] A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging.* New York: IEEE Press, 1988.

[78] M. Bhatia, W. C. Karl, and A. S. Willsky, "Tomographic reconstruction and estimation based on multiscale natural-pixe base," *IEEE Trans. Image Processing*, vol. 6, no. 3, pp. 463–477, 1997.

[79] J. M. Links, R. W. Jeremy, S. M. Dyer, T. L. Frank, and L. C. Becker, "Wiener filtering improves quantification of regional myocardial perfusion with thallium-201 SPECT," *J. Nuclear Medicine*, vol. 31, pp. 1230–1236, 1990.

[80] T. Hebert and R. Leahy, "A generalized EM algorithm for 3-d Bayesian reconstruction from Poisson data using Gibbs priors," *IEEE Trans. Medical Imaging*, vol. 8, no. 2, pp. 194–202, 1989.

[81] S. S. Saquib, C. A. Bouman, and K. Sauer, "A non-homogeneous MRF model for multiresolution bayesian estimation," *IEEE Int'l Conf. on Image Proc.*, 1996.

[82] M. H. Buonocore, W. R. Brody, and A. Macovski, "A natural pixel decomposition for two-dimensional image reconstruction," *IEEE Trans. Biomed. Eng.*, vol. BME-28, no. 2, pp. 69–78, 1981.

[83] J. DeStefano and T. Olson, "Wavelet localization of the radon transform in even dimensions," *in IEEE-SP Int. Symp. Time-Frequency and Time-Scale Anal.*, pp. 137–140, Oct. 1992.

[84] B. Sahiner and A. E. Yagle, "Limited angle tomography using the wavelet transform," *in IEEE Nucler Science Symp. and Medical Imaging Conf.*, pp. 219–222, Oct. 1993.

[85] Z. Wu, "MAP image reconstruction using wavelet decomposition," *Lec. Notes Comput. Sci.*, vol. 687, pp. 354–371, 1993.

[86] F. Rashid-Farrokhi, K. J. R. Liu, C. A. Berenstein, and D. Walnut, "Wavelet-based multiresolution local tomographiy," *IEEE Trans. on Image Processing*, vol. 6, pp. 1412–1430, Oct. 1997.

[87] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products.* New York: Academic Press, 1965 fourth edition.

[88] A. Papoulis, *The Fourier Integral and its Applications.* New York: McGraw-Hill, 1962.