Introduction to Orthogonal Transforms

With Applications in Data Processing and Analysis

A systematic, unified treatment of orthogonal transform methods for signal processing, data analysis, and communications, this book guides the reader from mathematical theory to problem solving in practice. It examines each transform method in depth, emphasizing the common mathematical principles and essential properties of each method in terms of signal decorrelation and energy compaction. The different forms of Fourier transform, as well as the Laplace, Z-, Walsh–Hadamard, slant, Haar, Karhunen–Loève, and wavelet transforms, are all covered, with discussion of how these transform methods can be applied to real-world problems. Numerous practical examples and end-of-chapter problems, supported by online Matlab and C code and an instructor-only solutions manual, make this an ideal resource for students and practitioners alike.

Ruye Wang is a Professor in the Engineering Department at Harvey Mudd College. Previously a Principal Investigator at the Jet Propulsion Laboratory, NASA, his research interests include image processing, computer vision, machine learning, and remote sensing.

Introduction to Orthogonal Transforms

With Applications in Data Processing and Analysis

RUYE WANG
Harvey Mudd College, California, USA

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521516884

First published 2012

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

ISBN 978-0-521-51688-4 Hardback

Additional resources for this publication at www.cambridge.org/9780521516884

To my parents

Cambridge University Press
978-0-521-51688-4 - Introduction to Orthogonal Transforms: With Applications in Data Processing and Analysis
Ruye Wang
Frontmatter
More information

Contents

Cambridge University Press
978-0-521-51688-4 - Introduction to Orthogonal Transforms: With Applications in Data Processing and Analysis
Ruye Wang
Frontmatter
More information

Preface

What is the book all about?

When a straight line standing on a straight line makes the adjacent angles equal to one another, each of the equal angles is right, and the straight line standing on the other is called a *perpendicular* to that on which it stands.
— Euclid, *Elements, Book 1, definition 10*

This is Euclid's definition for "perpendicular", which is synonymous with the word "orthogonal" used in the title of this book. Although the meaning of this word has been generalized since Euclid's time to describe the relationship between two functions as well as two vectors, as what we will be mostly concerned with in this book, they are essentially no different from two perpendicular straight lines, as discussed by Euclid some 23 centuries ago.

Orthogonality is of important significance not only in geometry and mathematics, but also in science and engineering in general, and in data processing and analysis in particular. This book is about a set of mathematical and computational methods, known collectively as the orthogonal transforms, that enables us to take advantage of the orthogonal axes of the space in which the data reside. As we will see throughout the book, such orthogonality is a much desired property that can keep things untangled and nicely separated for ease of manipulation, and an orthogonal transform can rotate a signal, represented as a vector in a Euclidean space, or more generally Hilbert space, in such a way that the signal components tend to become, approximately or accurately, orthogonal to each other. Such orthogonal transforms, typified by the most well-known Fourier transform, lend themselves well to various data processing and analysis needs, and therefore are used in a wide variety of disciplines and areas, including both social and natural sciences and engineering. The book also covers the Laplace and z-transforms, which can be considered as the extended versions of the Fourier transform for continuous and discrete functions respectively, and the wavelet transforms which may not be strictly orthogonal but which are still closely related to those that are.

In the last few decades the scales of data collection across almost all fields have been increasing dramatically due mostly to the rapid advances in technologies. Consequently, how best to make sense of the fast accumulating data has become more challenging than ever. Wherever a large amount of data is collected, from stock market indices in economics to microarray data in bioinformatics, from seismic data in geophysics to audio and video data in communication and broadcasting engineering, there is always the need to process, analyze, and compress the data in some meaningful way for the purposes of effective and efficient data transmission, interpretation, and storage by various computational methods and algorithms. The transform methods discussed in this book can be used as a set of basic tools for the data processing and the subsequent analysis, such as data mining, knowledge discovery, and machine learning.

The specific purpose of each data processing and analysis task at hand may vary from case to case. From a set of given data, one may desire to remove a certain type of noise, extract particular kinds of features of interest, and/or reduce the quantity of the data without losing useful information for storage and transmission. On the other hand, many operations needed for achieving these very different goals may all be carried out using the same mathematical tool of orthogonal transform, by which the data are manipulated and represented in such a way that the desired results can be achieved effectively in the subsequent stage. To address all such needs, this book presents a thorough introduction to the mathematical background common to these transform methods, and provides a repertoire of computational algorithms for these methods.

The basic approach of the book is the combination of the theoretical derivation and practical implementation of each transform method considered. Certainly, many existing books touch upon the topics of both orthogonal and wavelet transforms, from either a mathematical or an engineering point of view. Some of them may concentrate on the theories of these methods, while others may emphasize their applications, but relatively few would guide the reader directly from the mathematical theories to the computational algorithms, and then to their applications to real data analysis, as this book intends to do. Here, deliberate efforts are made to bridge the gap between the theoretical background and the practical implementation, based on the belief that, to truly understand a certain method, one needs ultimately to be able to convert the mathematical theory into computer code for the algorithms to be actually implemented and tested. This idea has been the guiding principle throughout the writing of the book. For each of the methods covered, we will first derive the theory mathematically, then present the corresponding computational algorithm, and finally provide the necessary code segments in Matlab or C for the key parts of the algorithm. Moreover, we will also include some relatively simple application examples to illustrate the actual data-processing effects of the algorithm. In fact, every one of the transform methods considered in the book has been implemented in either Matlab and/or the C programming language and tested on real data. The complete programs are also made readily available on a website dedicated to the book at: www.cambridge.org/orthogonaltransforms. The reader is encouraged and expected to try these algorithms out by running the code on his/her own data.

Why orthogonal transforms?

The transform methods covered in the book are a collection of both old and new ideas ranging from the classical Fourier series expansion that goes back almost 200 years, to some relatively recent thoughts such as the various origins of the wavelet transform. While all of these ideas were originally developed with different goals and applications in mind, from solving the heat equation to the analysis of seismic data, they can all be considered to belong to the same family, based on the common mathematical framework they all share, and their similar applications in data processing and analysis. The discussions of specific methods and algorithms in the chapters will all be approached from such a unified point of view.

Before the specific discussion of each of the methods, let us first address a fundamental issue: why do we need to carry out an orthogonal transform to start with? A signal, as the measurement of a certain variable (e.g., the temperature of a physical process) tends to vary continuously and smoothly, as the energy associated with the physical process is most probably distributed relatively evenly in both space and time. Most such spatial or temporal signals are likely to be correlated, in the sense that, given the value of a signal at a certain point in space or time, one can predict with reasonable confidence that the signal at a neighboring point will take a similar value. Such everyday experience is due to the fundamental nature of the physical world governed by the principles of minimum energy and maximum entropy, in which any abruption and discontinuities, typically caused by an energy surge of some kind, are relatively rare and unlikely events (except in the microscopic world governed by quantum mechanics). On the other hand, from the signal processing viewpoint, the high signal correlation and even energy distribution are not desirable in general, as it is difficult to decompose such a signal, as needed in various applications such as information extraction, noise reduction, and data compression. The issue, therefore, becomes one of how the signal can be converted in such a way that it is less correlated and its energy less evenly distributed, and to what extent such a conversion can be carried out to achieve the goal.

Specifically, in order to represent, process, and analyze a signal, it needs to be decomposed into a set of components along a certain dimension. While a signal is typically represented by default as a continuous or discrete function of time or space, it may be desirable to represent it along some alternative dimension, most commonly (but not exclusively) frequency, so that it can be processed and analyzed more effectively and conveniently. Viewed mathematically, a signal is a vector in some vector space which can be represented by any of a set of different orthogonal bases all spanning the same space. Each representation corresponds to a different decomposition of the signal. Moreover, all such representations are equivalent, in the sense that they are related to each other by certain rotation in the space by which the total energy or information contained in the signal is conserved. From this point of view, all different orthogonal transform methods developed in the last 200 years by mathematicians, scientists, and engineers for various purposes can be unified to form a family of methods for the same general purpose.

While all transform methods are equivalent, as they all conserve the total energy or information of the signal, they can be very different in terms of how the total energy or information in the signal is redistributed among its components after the transform, and how much these components are correlated. If, after a properly chosen orthogonal transform, the signal is represented in such a way that its components are decorrelated and most of the signal information of interest is concentrated in a small subset of its components, then the remaining components could be neglected as they carry little information. This simple idea is essentially the answer to the question asked above about why an orthogonal transform is needed, and it is actually the foundation of the general orthogonal transform method for feature selection, data compression, and noise reduction. In a certain sense, once a proper basis of the space is chosen so that the signal is represented in such a favorable manner, the signal-processing goal is already achieved to a significant extent.

What is in the chapters?

The purpose of the first two chapters is to establish a solid mathematical foundation for the thorough understanding of the topics of the subsequent chapters, which each discuss a specific type of transform method. Chapter 1 is a brief summary of the basic concepts of signals and linear time-invariant (LTI) systems. For readers with an engineering background, much of this chapter may be a quick review that could be scanned through or even skipped. For others, this chapter serves as an introduction to the mathematical language by which the signals and systems will be described in the following chapters.

Chapter 2 sets up the stage for all transform methods by introducing the key concepts of the vector space, or more strictly speaking the Hilbert space, and the linear transformations in such a space. Here, a usual N-dimensional space can be generalized in several aspects: (1) the dimension N of the space may be extended to infinity, (2) a vector space may also include a function space composed of all continuous functions satisfying certain conditions, and (3) the basis vectors of a space may become uncountable. The mathematics needed for a rigorous treatment of these much-generalized spaces is likely to be beyond the comfort zone of most readers with a typical engineering or science background, and it is therefore also beyond the scope of this book. The emphasis of the discussion here is not mathematical rigor, but the basic understanding and realization that many of the properties of these generalized spaces are just the natural extensions of those of the familiar N-dimensional vector space. The purpose of such discussions is to establish a common foundation for all transform methods so that they can all be studied from a unified point of view, namely, that any given signal, either continuous or discrete, with either finite or infinite duration, can be treated as a vector in a certain space and represented differently by any of a variety of orthogonal transform methods, each corresponding to one of the orthogonal bases that span the space. Moreover, all of these different representations are related to each other by rotations in the space. Such basic ideas may also be extended to non-orthogonal (e.g., biorthogonal) bases that are used in wavelet transforms. All transform methods considered in later chapters will be studied in light of such a framework. Although it is highly recommended for the reader to at least read through the materials in the first two chapters, those who feel it is difficult to thoroughly follow the discussions could skip them and move on to the following chapters, as many of the topics could be studied relatively independently, and one can always come back to learn some of the concepts in the first two chapters when needed.

In Chapters 3 and 4 we study the classical Fourier methods for continuous and discrete signals respectively. Fourier's theory is mathematically beautiful and is referred to as "mathematical poem"; and it has great significance throughout a wide variety of disciplines, in practice as well as in theory. While the general topic of the Fourier transform is covered in a large number of textbooks in various fields, such as engineering, physics, and mathematics, a not-so-conventional approach is adopted here to treat all Fourier-related methods from a unified point of view. Specifically, the Fourier series expansion, the continuous- and discrete-time Fourier transforms (CTFT and DTFT), and the discrete Fourier transform (DFT) will be considered as four different variations of the same general Fourier transform, corresponding to the four combinations of the two basic categories of signals: continuous versus discrete, periodic versus non-periodic. By doing so, many of the dual and symmetrical relationships among these four different forms and between time and frequency domains of the Fourier transform can be much more clearly and conveniently presented and understood.

Chapter 5 discusses the Laplace and z-transforms. Strictly speaking, these transforms do not belong to the family of orthogonal transforms, which convert a one-dimensional (1-D) signal of time t into another 1-D function along a different variable, typically frequency f or angular frequency $\omega = 2\pi f$. Instead, the Laplace converts a 1-D continuous signal from the time domain into a function in a two-dimensional (2-D) complex plane $s = \sigma + j\omega$, and the z-transform converts a 1-D discrete signal from the time domain into a function in a 2-D complex plane $z = e^s$. However, as these transforms are respectively the natural extensions of the CTFT and DTFT, and are widely used in signal processing and system analysis, they are included in the book as two extra tools in our toolbox.

Chapter 6 discusses the Hartly and sine/cosine transforms, both of which are closely related to the Fourier transform. As real transforms, both Hartly and sine/cosine transforms have the advantage of reduced computational cost when compared with the complex Fourier transform. If the signal in question is real with zero imaginary part, then half of the computation in its Fourier transform is redundant and, therefore, wasted. However, this redundancy is avoided by a real transform such as the cosine transform, which is widely used for data compression, such as in the image compression standard JPEG.

Chapter 7 combines three transform methods, the Walsh-Hadamard, slant, and Haar transforms, all sharing some similar characteristics (i.e., the basis functions associated with these transforms all have square-wave-like waveforms). Moreover, as the Haar transform also possesses the basic characteristics of the wavelet transform method, it can also serve as a bridge between the two camps of the orthogonal transforms and the wavelet transforms, leading to a natural transition from the former to the latter.

In Chapter 8 we discuss the Karhunen-Loeve transform (KLT), which can be considered as a capstone of all previously discussed transform methods, and the associated principal component analysis (PCA), which is popularly used in many data-processing applications. The KLT is the optimal transform method among all orthogonal transforms in terms of the two main characteristics of the general orthogonal transform method, namely the compaction of signal energy and the decorrelation among all signal components. In this regard, all orthogonal transform methods can be compared against the optimal KLT for an assessment of their performances.

We next consider in Chapter 9 both the continuous- and discrete-time wavelet transforms (CTWT and DTWT), which differ from all orthogonal transforms discussed previously in two main aspects. First, the wavelet transforms are not strictly orthogonal, as the bases used to span the vector space and to represent a given signal may not be necessarily orthogonal. Second, the wavelet transform converts a 1-D time signal into a 2-D function of two variables, one for different levels of details or scales, corresponding to different frequencies in the Fourier transform, and the other for different temporal positions, which is totally absent in the Fourier or any other orthogonal transform. While redundancy is inevitably introduced into the 2-D transform domain by such a wavelet transform, the additional second dimension also enables the transform to achieve both temporal and frequency localities in signal representation at the same time (while all other transform methods can only achieve either one of the two localities). Such a capability of the wavelet transform is its main advantage over orthogonal transforms in some applications such as signal filtering.

Finally, in Chapter 10 we introduce the basic concept of multiresolution analysis (MRA) and Mallat's fast algorithm for the discrete wavelet transform (DWT), together with its filter bank implementation. Similar to the orthogonal transforms, this algorithm converts a discrete signal of size N into a set of DWT coefficients also of size N, from which the original signal can be perfectly reconstructed; i.e., there is no redundancy introduced by the DWT. However, different from orthogonal transforms, the DWT coefficients represent the signal with temporal as well as frequency (levels of details) localities, and can, therefore, be more advantageous in some applications, such as data compressions.

Moreover, some fundamental results in linear algebra and statistics are also summarized in the two appendices at the back of the book.

Who are the intended readers?

The book can be used as a textbook for either an undergraduate or graduate course in digital signal processing, communication, or other related areas. In such a classroom setting, all orthogonal transform methods can be systematically studied following a thorough introduction of the mathematical background common to these methods. The mathematics prerequisite is no more than basic calculus and linear algebra. Moreover, the book can also be used as a reference by practicing professionals in both natural and social sciences, as well as in engineering. A financial analyst or a biologist may need to learn how to effectively analyze and interpret his/her data, a database designer may need to know how to compress his data before storing them in the database, and a software engineer may need to learn the basic data-processing algorithms while developing a software tool in the field. In general, anyone who deals with a large quantity of data may desire to gain some basic knowledge in data-processing, regardless of his/her backgrounds and specialties. In fact the book has been developed with such potential readers in mind. Owing possibly to personal experience, I always feel that self-learning (or, to borrow a machine learning terminology, "unsupervised learning") is no less important than formal classroom learning. One may have been out of school for some years but still feel the need to update and expand one's knowledge. Such readers could certainly study whichever chapters of interest, instead of systematically reading through each chapter from beginning to end. They can also skip certain mathematical derivations, which are included in the book for completeness (and for those who feel comfortable only if the complete proof and derivations of all conclusions are provided). For some readers, neglecting much of the mathematical discussion for a specific transform method should be just fine if the basic ideas regarding the method and its implementation are understood. It is hoped that the book can serve as a toolbox, as well as a textbook, from which certain transform methods of interest can be learned and applied, in combination with the reader's expertise in his/her own field, to solving the specific data-processing/analysis problems at hand.

About the homework problems and projects

Understanding the transform methods and the corresponding computational algorithms is not all. Eventually they all need to be implemented and realized by either software or hardware, specifically by computer code of some sort. This is why the book emphasizes the algorithm and coding as well as theoretical derivation, and many homework problems and projects require certain basic coding skills, such as some knowledge in Matlab. However, being able to code is not expected of all readers. Those who may not need or wish to learn coding can by all means skip the sections in the text and those homework problems involving software programming. However, all readers are encouraged to at least run some of the Matlab functions provided to see the effects of the transform methods. (There are a lot of such Matlab m-files on the website of the book. In fact, all functions used to generate many of the figures in the book are provided on the site.) If a little more interested, the reader can read through the code to see how things are done. Of course, a step further is to modify the code and use different parameters and different datasets to better appreciate the various effects of the algorithms.

Back to Euclid

Finally, let us end by again quoting Euclid, this time, a story about him.

> A youth who had begun to study geometry with Euclid, when he had learned the first proposition, asked, "What do I get by learning these things?" So Euclid called a slave and said "Give him three pence, since he must make a gain out of what he learns."

Surely, explicit efforts are made in this book to discuss the practical uses of the orthogonal transforms and the mathematics behind them, but one should realize that, after all, the book is about a set of mathematical tools, just like those propositions in Euclid's geometry, out of learning which the reader may not be able to make a direct and immediate gain. However, in the end, it is the application of these tools toward solving specific problems in practice that will enable the reader to make a gain out of the book; much more than three pence, hopefully.

Acknowledgments

I am in debt to two of my colleagues, Professors John Molinder and Ellis Cumberbatch for their support and help with the book project. In addition to our discussions regarding some of the topics in the book, John provided the application example of orthogonal frequency division modulation discussed in section 5.8, together with the Matlab code that is used in a homework problem. Also, Ellis read through the first two chapters of the manuscript and made numerous suggestions for the improvement of the coverage of the topics in these two chapters. All such valuable help and support are greatly appreciated.

Notation

General notation

iff	if and only if
$j = \sqrt{-1} = e^{j\pi/2}$	imaginary unit
$\overline{u + jv} = u - jv$	complex conjugate of $u + jv$
$\mathrm{Re}(u + jv) = u$	real part of $u + jv$
$\mathrm{Im}(u + jv) = v$	imaginary part of $u + jv$
$\|u + jv\| = \sqrt{u^2 + v^2}$	magnitude (absolute value) of a complex value $u + jv$
$\angle(u + jv) = \tan^{-1}(v/u)$	phase of $u + jv$
$\boldsymbol{x}_{n\times 1}$	an n by 1 column vector
$\overline{\boldsymbol{x}}$	complex conjugate of $\boldsymbol{x}$
$\boldsymbol{x}^{\mathrm{T}}$	transpose of $\boldsymbol{x}$, a 1 by n row vector
$\boldsymbol{x}^* = \overline{\boldsymbol{x}}^{\mathrm{T}}$	conjugate transpose of matrix $\boldsymbol{A}$
$\|\|\boldsymbol{x}\|\|$	norm of vector $\boldsymbol{x}$
$\boldsymbol{A}_{m\times n}$	an m by n matrix of m rows and n columns
$\overline{\boldsymbol{A}}$	complex conjugate of matrix $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	inverse of matrix $\boldsymbol{A}$
$\boldsymbol{A}^{\mathrm{T}}$	transpose of matrix $\boldsymbol{A}$
$\boldsymbol{A}^* = \overline{\boldsymbol{A}}^{\mathrm{T}} = \overline{\boldsymbol{A}^{\mathrm{T}}}$	conjugate transpose of matrix $\boldsymbol{A}$
$\mathbb{N}$	set of all positive integers including 0
$\mathbb{Z}$	set of all real integers
$\mathbb{R}$	set of all real numbers
$\mathbb{C}$	set of all complex numbers
$\mathbb{R}^N$	N-dimensional Euclidean space
$\mathbb{C}^N$	N-dimensional unitary space
$\mathcal{L}^2$	space of all square-integrable functions
l^2	space of all square-summable sequences
$x(t)$	a function representing a continuous signal
$\boldsymbol{x} = [\ldots, x[n], \ldots]^{\mathrm{T}}$	a vector representing a discrete signal
$\dot{x}(t) = dx(t)/dt$	first order time derivative of $x(t)$
$\ddot{x}(t) = dx^2/dt^2$	second order time derivative of $x(t)$
f	frequency (cycle per unit time)
$\omega = 2\pi f$	angular frequency (radian per unit time)

Throughout the book, angular frequency ω will be used interchangeably with $2\pi f$, whichever is more convenient in the context of the discussion.

As a convention, a bold-faced lower case letter $\boldsymbol{x}$ is typically used to represent a vector, while a bold-faced upper case letter $\boldsymbol{A}$ represents a matrix, unless noted otherwise.