Ulf Grenander and Michael Miller

Pattern Theory

From Representation to Inference

OXFORD

PATTERN THEORY: FROM REPRESENTATION TO INFERENCE

Ulf Grenander and Michael I. Miller

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published in the United States
by Oxford University Press Inc., New York

First published 2007

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

Typeset by Newgen Imaging Systems (P) Ltd., Chennai, India
Printed in Great Britain
on acid-free paper by
CPI Bath

ISBN 0–19–850570–1 978–0–19–850570–9
ISBN 0–19–929706–1 978–0–19–929706–1

1 3 5 7 9 10 8 6 4 2

PATTERN THEORY

CONTENTS

1 INTRODUCTION

This book is to be an accessible book on patterns, their representation, and inference. There are a small number of ideas and techniques that, when mastered, make the subject more accessible. This book has arisen from ten years of a research program which the authors have embarked upon, building on the more abstract developments of metric pattern theory developed by one of the authors during the 1970s and 1980s. The material has been taught over multiple semesters as part of a second year graduate-level course in pattern theory, essentially an introduction for students interested in the representation of patterns which are observed in the natural world. The course has attracted students studying biomedical engineering, computer science, electrical engineering, and applied mathematics interested in speech recognition and computational linguistics, as well as areas of image analysis, and computer vision.

Now the concept of *patterns* pervades the history of intellectual endeavor; it is one of the eternal followers in human thought. It appears again and again in science, taking on different forms in the various disciplines, and made rigorous through mathematical formalization. But the concept also lives in a less stringent form in the humanities, in novels and plays, even in everyday language. We use it all the time without attributing a formal meaning to it and yet with little risk of misunderstanding. So, what do we really mean by a pattern? Can we define it in strictly logical terms? And if we can, what use can we make of such a definition?

These questions were answered by *General Pattern Theory*, a discipline initiated by Ulf Grenander in the late 1960s [1–5]. It has been an ambitious effort with the only original sketchy program having few if any practical applications, growing in mathematical maturity with a multitude of applications having appeared in biology/medicine and in computer vision, in language theory and object recognition, to mention but a few. Pattern theory attempts to provide an algebraic framework for describing patterns as structures regulated by rules, essentially a finite number of both local and global combinatory operations. Pattern theory takes a *compositional* view of the world, building more and more complex structures starting from simple ones. The basic rules for combining and building complex patterns from simpler ones are encoded via graphs and rules on transformation of these graphs.

In contrast to other dominating scientific themes, in Pattern Theory we start from the belief that *real world patterns are complex*: Galielaen simplification that has been so successful in physics and other natural sciences will not suffice when it comes to explaining other regularities, for example in the life sciences. If one accepts this belief it follows that complexity must be allowed in the ensuing representations of knowledge. For this, probabilities naturally enter, superimposed on the graphs so as to express the variability of the real world by describing its fluctuations as randomness. Take as a goal the development of algorithms which assist in the ambitious task of *image understanding* or *recognition*. Imagine an expert studying a natural scene, trying to understand it in terms of the awesome body of knowledge that is informally available to humans about the context of the scene: identify components, relate them to each other, make statements about the fine structure as well as the overall appearance. If it is truly the goal to create algorithmic tools which assist experts in carrying out the time-consuming labor of pattern analysis, while leaving the final decision to their judgment, to arrive at more than ad hoc algorithms the subject matter knowledge must be expressed precisely and as compactly as possible.

This is the central focus of the book: 'How can such empirical knowledge be represented in mathematical form, including both structure and the all important variability?' This task of presenting an organized and coherent view of the field of Pattern theory seems bewildering at best. But what are today's challenges in signal, data and pattern analysis? With the advent of geometric increases in computational and storage resources, there has been a dramatic increase in the solution of highly complex pattern representation and recognition problems. Historically books on pattern recognition present a diverse set of problems with diverse methods for building recognition algorithms, each approach handcrafted to the particular task. The complexity and diversity of patterns in the world presents one of the most significant challenges to the pedagogical approach to the teaching of Pattern theory. Real world patterns are often the results of evolutionary change, and most times cannot be selected by the practitioner to have particular properties. The representations require models using mathematics which span multiple fields in *algebra, geometry, statistics and statistical communications*.

Contrasting this to the now classical field of statistical communications, it might appear that the task seems orders of magnitude bigger than modelling signal ensembles in the communication environment. Thinking historically of the now classical field of statistical communications, the discipline can be traced back far, to Helmholtz and earlier, but here we are thinking of its history in the twentieth century. For the latter a small number of parameters may be needed, means, covariances, for Gaussian noise, or the spectral density of a signal source, and so on. The development of communication engineering from the 1920s on consisted in part of formalizing the observed, more or less noisy, signals. Statistical signal processing is of course one of the great success stories of statistics/engineering. It is natural to ask why. We believe that it was because the pioneers in the field managed to construct representations of signal ensembles, models that were realistic and at the same time tractable both analytically and computationally (by analog devices at the time). The classical signalling models: choose $s_0(t), s_1(t)$ to be orthogonal elements in L^2, with the noise model additive stationary noise with covariance representation via a complete orthonormal basis. Such a beautiful story, utilizing ideas from Fourier analysis, stationary stochastic processes, Toeplitz forms, and Bayesian inference! Eventually this resulted in more or less automated procedures for the detection and understanding of noisy signals: matched filters, optimal detectors, and the like. Today these models are familiar, they look simple and natural, but in a historical perspective the phenomena must have appeared highly complex and bewildering.

We believe the same to be true for pattern theory. The central challenge is the formalization of a small set of ideas for constructing the *representations of the patterns themselves* which accommodate variability and structure simultaneously. This is the point of view from which this book is written. Even though the field of pattern theory has grown considerably over the past 30 years, we have striven to emphasize its coherence. There are essentially two overarching principles. The first is the representation of regularity via graphs which essentially encode the rules of combination which allow for the generation of complex structures from simpler ones. The second is the application of transformations to generate from the exemplars entire orbits. To represent typicality probabilistic structures are superimposed on the graphs and the rules of transformation. Naturally then the conditional probabilities encode the regularity of the patterns, and become the central tool for studying pattern formation.

We have been drawn to the field of pattern theory from backgrounds in communication theory, probability theory and statistics. The overall framework fits comfortably within the source-channel view of Shannon. The underlying deep regular structures are descriptions of the source, which are hidden via the sensing channel. We believe that the principle challenge is the representation of the source of patterns, and for that reason the majority of the book is focused precisely on this topic. A multiplicity of channels or sensor models will be used throughout the book, those appropriate for the pattern class being studied. They are however studied superficially, drawn from the engineering literature and taken as given, but certainly studied more deeply elsewhere. The channel sensor models of course shape the overall performance of the inference algorithms; but the major focus of our work is on the development of stochastic models for the structural understanding of the variabilities of the patterns at the source. This also explains the major deviation of this pattern theory from that which has come to be known as pattern recognition. Only in the final chapters will pattern recognition algorithms be studied, attempting to answer the question of how well the algorithm can estimate (recognize) the source when seen through the noisy sensor channel.

1.1 Organization

Throughout this book we use methods from estimation, stochastic processes and information theory. Chapter 2 includes the basic stalwarts of statistics and estimation theory which should be familiar to the reader, including minimum-risk estimation, Fisher Information, hypothesis testing and maximum-likelihood, consistency, model order estinmation, and entropies.

Chapters 3–6 bring into central prominence the role of representation of patterns via conditioning structure. Chapter 3 examines discrete patterns represented via probabilistic directed acyclic graphs (DAGs) emphasizing the pivoting properties and conditional factorizations of DAGs which are familiar for Markov chains and random branching processes. This provides ample opportunity to construct and study the syntactic theory of Chomsky Languages via the classical formulation of graphs and grammatical transformation. Chapter 4 relaxes away from the pivoting property of directed graphs to the conditioning structure of neighborhoods in Markov random fields. Chapter 5 brings the added structure of of the Gaussian law for representing real-valued patterns via Gaussian fields. In this context entropy and maximum entropy distributions are examined in these three chapters as a means of representing conditioning information for representing patterns of regularity for speech, language, and image analysis. Chapter 6 presents the abstract representation of patterns via generators and probabilistic structures on the generators. The generator representation is explored as it provides a unified way of dealing with DAGs and random fields.

Chapters 7 and 8 begin examining in their own right the second central component of pattern theory, groups of geometric transformation applied to the representation of geometric objects. The patterns and shapes are represented as submanifolds of $\mathbb{R}^n$, including points, curves, surfaces and subvolumes. They are enriched via the actions of the linear matrix groups, studying the patterns as orbits defined via the group actions. In this context active models and deformable templates are studied. The basic fundamentals of groups and matrix groups are explored assuming that the typical engineering graduate student will not be familiar with their structure.

Chapter 9 makes the first significant foray into probabilistic structures in the continuum, studying random processes and random fields indexed over subsets of $\mathbb{R}^n$. Classical topics are examined in some detail including second order processes, covariance representation, and Karhunen-Loeve transforms. This is the first chapter where more significant understanding is required for understanding signal and patterns as functions in a Hilbert space.

Chapters 10 and 11 continue the major thrust into transformations and patterns indexed over the continuum. In this context, the finite dimensional matrix groups are studied as diffeomorphic actions on $\mathbb{R}^n$, as well their infinite dimensional analogues are established. It is in these chapters in which the substantial bridge between Pattern theory, mechanics, and differential geometry is established. The links come through the study of flows of diffeomorphisms. Chapter 10 focuses on the study of the finite dimensional matrix groups, and Chapter 11 on the infinite dimensional diffeomorphisms acting on manifolds of $\mathbb{R}^n$ as a Riemannian metric space. The metric is induced by the geodesic length between elements in the space defined through the Riemannian length of the flow connecting one point to another. Herein the classical equations of motion for the geodesics in the finite dimensional case are expanded to include the Euler formulation of the infinite dimensional case.

Chapter 12 expands this view to examine the orbit of imagery as a deformable template under diffeomorphic action; the orbit is endowed with the metric structure through the length minimizing geodesics connecting images. In this chapter the photometric orbit is studied as well, adding notions from transport to define a metric on the product space of geometric and photometric variation.

Chapters 13–15 extend from the pure representations of shapes to the Bayes estimation of shapes and their parametric representation. Classical minimum-mean-squared error and maximum a-posteriori estimators of shapes are explored in these chapters as viewed through various remote sensing models. Chapter 13 focuses on estimating the pose of rigid objects; chapter 14 focuses on accommodating photometric variability superimposed on the geometric variability of rigid pose. Chapter 15 focuses on information bounds for quantifying estimation accuracy of the matrix group, comparing mean-squared error bounds with capacity and rate-distortion bounds for codebooks.

Chapters 16 and 17 turn from the estimation of finite dimensional matrix groups to the study of the estimation of infinite dimensional shape in the newly emergent field of Computational Anatomy. Chapter 15 focuses on estimating landmark and image based shape metrics in volumes, with Chapter 16 focusing on submanifolds and on the inference of disease and hypothesis testing in Computational Anatomy.

The last two Chapters 18 and 19 conclude on inference, exploring random sampling approaches for estimation of model order and parametric representing of shapes. Chapter 18 reviews jump and diffusion processes and their use in random sampling of discrete and continuum spaces. Chapter 19 examines a series of problems in object recognition.

We have made an attempt to keep the theory at a consistent level. The mathematical level is a reasonably high one, first-year graduate level, with a background of at least one good semester course in probability and a solid background in mathematics. We have, however, been able to avoid the use of measure theory.

Appendices outlining proofs, theorems and solutions to exercises together with a comphrehive list of figures, tables and plates are freely available on an accompanying website www.oup.com/uk/academic/companion/mathematics/patterntheory

In this book plates 1–16 appear between pages 180–181, plates 17–34 appear between pages 372–373, and plates 35–53 appear between 564–565.

2 THE BAYES PARADIGM, ESTIMATION AND INFORMATION MEASURES

ABSTRACT The basic paradigm is the Bayesian setup, given is the source of parameters $X \in \mathcal{X}$ which are seen through a noisy channel giving observations $Y \in \mathcal{Y}$. The posterior distribution determines the bounds on estimation of X given Y, the risk associated with estimating it, as well as a characterization of the information in the observation in Y about X.

2.1 Bayes Posterior Distribution

The basic set up throughout is we are given a model of the *source* of possible objects $X \in \mathcal{X}$. These are observed through a *noisy channel* giving observations $Y \in \mathcal{Y}$. The source $X \in \mathcal{X}$ is modelled with distribution and density $P_X(dx) = p(x)dx$, $\int_{\mathcal{X}} p(x)dx = 1$. Generally the source can only be observed with loss of information due to observational noise or limited accuracy in the sensor. The mapping from the input source $X \in \mathcal{X}$ to the observed output $Y \in \mathcal{Y}$ expresses the physics of the sensing channels; the data $Y \in \mathcal{Y}$ will in general contain multiple components corresponding to several sensors $Y = \left(Y_1, Y_2, \dots\right)$. The observation process is characterized via a statistical transition law, transferring $X \to Y$ $P_{Y|X}(\cdot|\cdot) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$, summarizing completely the transition law mapping the input model parameters X to the output Y, the likelihood of Y given X.

This Bayesian paradigm, separating the source from the channel is what has become the modern view of communications developed out of Shannon's theory of communications [6]. Figure 2.1, clearly delineates the separation of the source of information and the channel through which the messages are observed. To infer the transmitted message at the output of the channel, the observation Y must be processed optimally. The inference engine is a decoder, working to recover properties of the original message from the source.

Given such a communications *source/channel* decomposition, we shall be interested in both specifying optimal procedures for inferring properties of pattern generation systems given the observable measurements, and quantifying the information content and information gain of the observation system. Pattern deduction becomes an enterprise consisting of constructing the source and channel models, and essentially has several parts: (i) selection and fitting of parameters parametrizing the models representing the patterns, and (ii) construction of the family of probability models representing the knowledge about the pattern classes. For this we shall examine classical minimum-risk estimators, such as minimum-mean-squared-error (MMSE) estimators, maximum-aposteriori and likelihood (MAP, MLE). For constructing the models we shall examine various forms and principles of entropy and mutual information.

At the most fundamental level, the posterior distribution represents the information contained in the observables about the underlying imagery. All provably optimal structured methods of inference and information gathering fundamentally involve the posterior density or distribution of the random variables $X \in \mathcal{X}$ given the observed deformed image $Y \in \mathcal{Y}$.

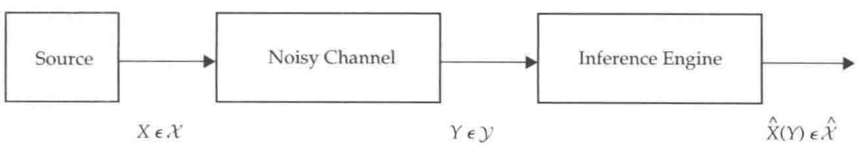

Figure 2.1 Shannon's source channel model for communications systems