7-2013

Representing and Inferring Visual Perceptual Skills in Dermatological Image Understanding

Rui Li

Representing and Inferring Visual Perceptual Skills in Dermatological Image Understanding

Rui Li

Ph.D. Program in Computing and Information Science

Rochester Institute of Technology

A dissertation submitted to Rochester Institute of Technology
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computing and Information Sciences
B. Thomas Golisano College of Computing and Information Sciences
Approved by Pengcheng Shi, Ph.D. Program Director

____

Signature Date

July 2013, Rochester, NY, USA

CERTIFICATE OF APPROVAL

DOCTORAL DISSERTATION

This is to certify that the Doctoral Dissertation of

Rui Li

has been examined and approved by the dissertation committee as complete and satisfactory for the dissertation requirement for the degree of Doctor of Philosophy in Computing and Information Sciences

May 2013, Rochester, NY, USA

Dr. Evelyn P. Rozanski

Dr. Justin Domke

Dr. Pengcheng Shi

Dr. Jeff Pelz

Chair of Dissertation Committee

Dr. Anne R. Haake

Dissertation Supervisor

Abstract

Experts have a remarkable capability of locating, perceptually organizing, identifying, and categorizing objects in images specific to their domains of expertise. Eliciting and representing their visual strategies and some aspects of domain knowledge will benefit a wide range of studies and applications. For example, image understanding may be improved through active learning frameworks by transferring human domain knowledge into image-based computational procedures, intelligent user interfaces enhanced by inferring dynamic informational needs in real time, and cognitive processing analyzed via unveiling the engaged underlying cognitive processes.

An eye tracking experiment was conducted to collect both eye movement and verbal narrative data from three groups of subjects with different medical training levels or no medical training in order to study perceptual skill. Each subject examined and described 50 photographical dermatological images. One group comprised 11 board-certified dermatologists (attendings), another group was 4 dermatologists in training (residents), and the third group 13 novices (undergraduate students with no medical training).

We develop a novel hierarchical probabilistic framework to discover the stereotypical and idiosyncratic viewing behaviors exhibited by the three expertise-specific groups. A hidden Markov model is used to

describe each subject's eye movement sequence combined with hierarchical stochastic processes to capture and differentiate the discovered eye movement patterns shared by multiple subjects' eye movement sequences within and among the three expertise-specific groups. Through these patterned eye movement behaviors we are able to elicit some aspects of the domain-specific knowledge and perceptual skill from the subjects whose eye movements are recorded during diagnostic reasoning processes on medical images. Analyzing experts' eye movement patterns provides us insight into cognitive strategies exploited to solve complex perceptual reasoning tasks. Independent experts' annotations of diagnostic conceptual units of thought in the transcribed verbal narratives are time-aligned with discovered eye movement patterns to help interpret the patterns' meanings. By mapping eye movement patterns to thought units, we uncover the relationships between visual and linguistic elements of their reasoning and perceptual processes, and show the manner in which these subjects varied their behaviors while parsing the images.

We further discover that inferred eye movement patterns characterize groups of similar temporal and spatial properties, and specify a subset of distinctive eye movement patterns which are commonly exhibited across multiple images. Based on the combinations of the occurrences of these eye movement patterns, we are able to categorize the images from the perspective of experts' viewing strategies in a novel way. In each category, images share similar lesion distributions and configurations. Our results show that modeling with multi-modal data, representative of physicians' diagnostic viewing behaviors and thought processes, is feasible and informative to gain insights into physicians' cognitive strategies, as well as medical image understanding.

Acknowledgements

Many people have helped me along the way to achieve this milestone. In particular, my advisor, Professor Anne R. Haake, who has assembled a wonderful research group where I was given the freedom to pursue the research topics interested me. I am grateful for that freedom and her indulgence in our lengthy discussions. This thesis would not have been possible without her encouragement, insight, and guidance. I would also like to thank Professor Evelyn P. Rozanski for welcoming me into her lab during my first two academic years at RIT. It was a refreshing experience to work with her who can look at problems from different perspectives.

I have also had the good fortune to work with two inspiring mentors Professor Pengcheng Shi and Professor Jeff Pelz. I have appreciated Professor Pengcheng shi's seemingly limitless supply of clever, and often unexpected, ideas ever since. When I wandered into human visual system, visual attention, and eye tracking technique as a sophomore, I was lucky to take Professor Jeff Pelz's course. My thesis committee, Professor Justin Domke, also provided thoughtful suggestions which continue to guide my research.

I owe thanks to other members of my research group. Notably, Dong Wang provided informative discussions about her study on the accuracy of our collected eye tracking data. Xuan Guo and Preethi

Vaidyanathan were great teammates in this work. I would also like to thank Professor Cecilia Ovesdotter Alm who provided numerous interesting discussions.

I cannot thank my parents enough for giving me the opportunity to freely pursue my interests, academic and otherwise. Finally, I thank my wife Qiao for being helpful, understanding, and patient over the past few years.

Contents

List of Figures

LIST OF FIGURES

List of Tables

1

Introduction

1.1 Current problems

People are different not only in how expert they are but also in their personal approaches to particular cognitive tasks. In order to evaluate the differences, eye movements, as both direct input and measurable output of real time signal processing in the brain, provide us an effective and reliable measure of human visual strategies and perceptual skill. This perceptual processes dynamically supports ongoing cognitive and behavioral activity while the viewer seeks specific information (1). Since high visual acuity is limited to the foveal region and resolution fades dramatically in the periphery, the brain has to strategically select the particular information needed for current cognitive tasks, and simultaneously monitor the rest at low resolution. As long as we can develop effective approaches which allow us to elicit and represent latent visual strategies and perceptual skill as implicit human capabilities, eye movements as valuable yet effortless resources will have a wide variety of applications (2, 3, 4, 5, 6, 7). In particular, in various domains of expertise where perceptual skill is paramount, experts' perceptual skill is

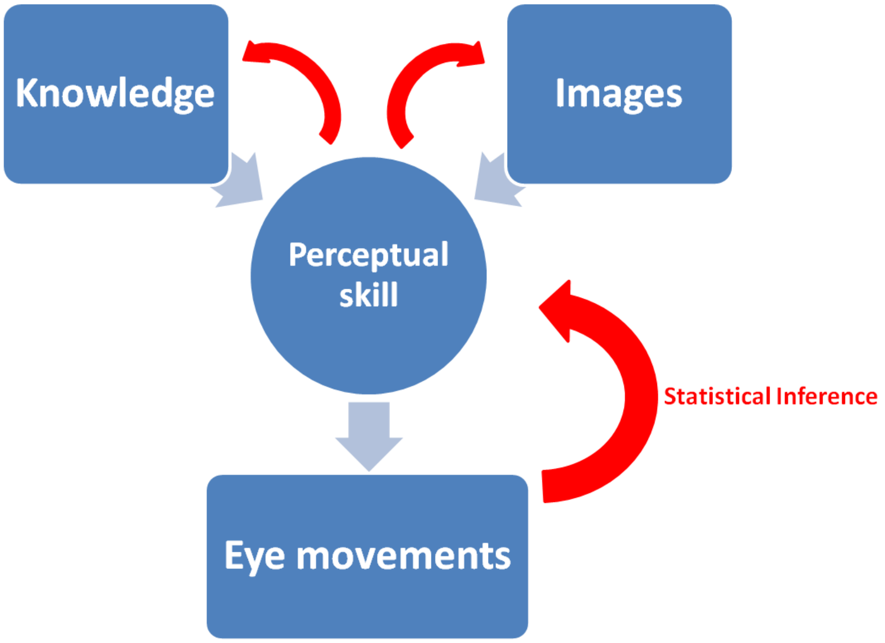


Figure 1.1: Scheme of our approach to investigate perceptual skill and image understanding. As previous studies suggested, experts' perceptual skill is based on their domain knowledge and expertise, as well as the current information gathered from the stimuli during perceptual processes. Some aspects of perceptual skill are manifested by experts' eye movements. Our approach is to view this structure as an inverse problem. By tracking experts' eye movements to follow along the attention paths, we can not only gain some insight into what interests experts but also how they perceive images in their domains of expertise. Furthermore, perceptual skill, as effective and robust indication of cognitive processing, allows us to discover some meaningful properties of domain-specific images.

considered to be more consistent and informative than traditional explicit human knowledge acquisition methods, such as manual markings, and annotations.

The eye is continuously moving to sample the visual stimuli in our environment. The brain incorporates memory, heuristics, and prior knowledge to reconstruct a 3D world from 2D projections via perceptual processes. This perception which is built from ongoing sensation and registration differs among individuals.

The research questions that interest us are: what approaches allow us to elicit latent nature of eye movements which is not directly observable, how it can inform us about experts in their domain of expertise, and in what way eye movement studies can advance image understanding based on the human visual system. Furthermore, we also attempt to demonstrate that during image-based diagnostic reasoning processes experts' eye movements are one of the best sources manifesting human cognitive strategies for medical image understanding, and the key to achieve this is how to maximize its potential while minimizing the processing load.

In our work we focus on medical images where domain knowledge and perceptual skill are in demand. We use eye movement data as an input for implicit measures of experts' perceptual skill to investigate some aspects of their visual strategies in the diagnostic reasoning processes and medical image understanding. We develop a computational approach to extract measures from eye movement data based on probabilistic modeling. In this way, we are able to use eye movement measures to explore attentional states and strategies across diagnostic reasoning tasks, and elicit perceptual skill and cognitive style from experts, as depicted in Figure 1.1.

1.2 Contributions

The way from a novice to expert can be characterized as a bumpy road of deliberate practice and effort (8). For facilitating novices in developing their skills and knowledge, a deep understanding of expertise and its unique differences to novices' knowledge and behavioral strategies is crucial, and will inform design of effective decision support systems, training programs, and so on.

We propose a novel hierarchical dynamical model which is capable of summarizing the stereotypical and idiosyncratic eye movement patterns from multiple expertise-specific groups of eye movement sequences. Since eye movement data are deployed sequentially, we use an autoregressive hidden Markov model (auto-HMMs) to account for the temporal-spatial nature of each subject's eye movement sequence. To characterize the patterned visual behaviors shared within multiple eye movement sequences of each expertise-specific group, as well as allow such information to be shared among multiple groups, we combine hierarchical beta processes to these auto-HMMs in a principled way. What's more, we interpret the discovered eye movement patterns by time-aligning them with standardized thought unit annotations (9).

Furthermore, we facilitate image understanding by incorporating experts' viewing strategies through an active learning paradigm. We combine multiple experts' strengths by summarizing their shared eye movement patterns and decode the patterns' semantic meanings. These results can also be applied as semantic labeling without manual annotation with respect to image understanding.

The major contributions of this work are as follows:

- Capture and describe the overall viewing strategies of subjects at various training levels during image-based diagnostic-reasoning processes through eye movement sequences. Results will enhance the understanding of perceptual skill in the medical domain and uncover the role of domain knowledge through eye movement pattern comparison between professionals and novices.

- Development, implementation, and evaluation of a computational approach to extract and characterize implicit perceptual skill through objective eye movement patterns on medical images. Perceptual skill has various manifestations. We profile stereotypical and idiosyncratic eye movement patterns exhibited among multiple subjects of each expertise-specific group. We demonstrate that the hierarchical beta processes are more appropriate, capturing common eye movement patterns shared among subjects while allowing subject-specific variability.

- Development of a new method for eliciting human perceptual skill to improve image understanding. We propose that the extracted perceptual skill, as an effortless yet valuable cognitive resource, can be combined into active learning methods of image understanding based on our approach. To further evaluating medical images, it is necessary to associate the experts' meaningful viewing behaviors to pixel-based information.

- Study and contribution to image retrieval and image understanding.

- Synthesis of established, relevant studies in computational cognitive science, focusing on visual attention modeling, with applications for image retrieval and image understanding.

1.3 Current and Future Publications

Some of my dissertation work have been published in the proceedings of peer-reviewed conferences such as ETRA2012 (10), CogSci2012 (11), and CVPR2013 (12). The computational modeling part of our work will be submitted to NIPS2013. We are also preparing one to two journal papers as a summary of our work. A journal paper has been submitted to the Journal of Cognitive Psychology and is pending review. My previous related work has also been published in (13, 14, 15, 16, 17, 18).

2

Background

2.1 Perception

Perception of the world depends on not only information arriving at sensory receptors but also the capacity of transforming and interpreting sensory information based on what one knows and has experienced. The interpretation, which is informed guess in effect, provided by the processes of perception cooperating with other cognitive processes enables one to respond to the environment effectively. In particular, perceptual processes make sense of the continuously-changing, chaotic sensory input from the external energy-filled environment and transfer it into stable, orderly mental images.

The effortless and automatical characterization of perception broadly refers to the overall process of apprehending objects and events in the external environment by sensing, understanding, and identifying them in order to prepare to respond to them. In this sense, the process of perception can be divided into three stages: sensation, perceptual organization, and identification/recognition of objects (8).

- At the sensation stage, physical energy is converted into the neural signals streaming into the brain. Retinal cells are activated strongly to edges and contrasts versus homogenous, unchanging stimulations.

- At the perceptual organization stage, an internal representation of an object is formed and a mental image of the external stimuli is created. This working description is computed by integrating past knowledge with the present evidence from senses and the stimuli within their perceptual context. Simple sensory features, such as colors, edges, and lines, are synthesized into the description which can be recognized in the later stage.

- Identification and recognition, as the third stage, assigns meaning to mental images.

These perceptual processes jointly give rise to a diagram of incoming information transformation, during which bottom-up processing occurs when the perceptual representation is derived from the information available in the sensory input, and top-down processing occurs when the perceptual representation is affected by one's prior knowledge, expectations, and other aspects of higher mental functioning.

From neuroscience perspective, a widely accepted theory is that there are two separate systems for visual processing in the brain which are the ventral and dorsal visual processing streams. The dorsal stream involves some neurons in the posterior parietal cortex selectively responding to object locations, while the ventral stream respond selectively to other properties such as object recognition and form representations. This neurophysiological viewpoint suggests that perceptual processing should be characterized on the basis of both spatial domain and feature domain.

Attention is a critical topic in perception studies in that focus of attention determines the portion of the sensory input from the external environment that will be readily available to perceptual processes. Moreover, visual attention is a strategic and effective image processing system through selecting and ignoring visual input based on current goals and past experience.

2.2 Visual Attention

Complex visual information available in real-world scenes or stimuli exceeds the processing capability of the human visual system. Consequently, human vision is an actively selective process in which the viewer seeks out specific information to support ongoing cognitive and behavioral activity, and filters out the rest (1). Human vision as an exquisite biological system maintains high resolution fovea subtending less than 2 visual degree of the visual field visible at any instant. It dynamically samples two dimensional information to reconstruct the three dimensional world with high resolution.

2.2.1 Visual psychophysics

Since high visual acuity is limited to the foveal region and resolution fades dramatically in the periphery, we move our eyes to bring a portion of the visual field into high resolution at the center of gaze. A series of fixations and saccades are used to describe such eye movements. Fixations occur when the gaze is held at a particular location, whereas saccades are rapid eye movements used to reposition the fovea to a new location. Studies have shown that visual attention is influenced by two main sources of input: bottom-up visual attention driven by low-level saliency stimulus features which are stimulus properties that are dis-

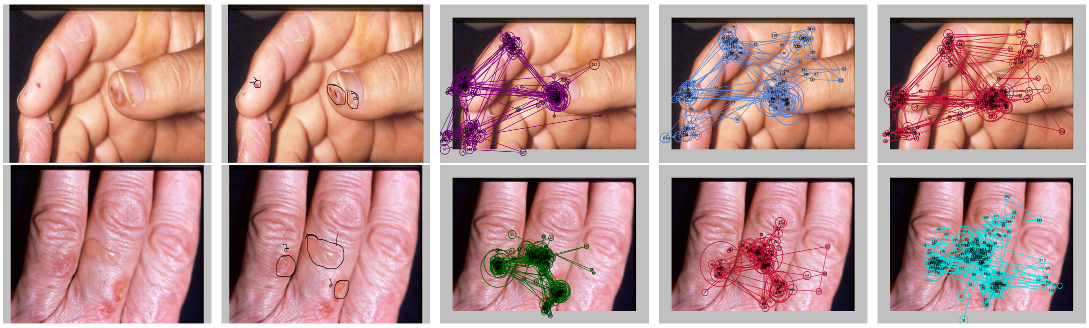


Figure 2.1: Two example dermatological images examined by the subjects. The images from left to right are the original images, the primary and secondary abnormalities marked and numbered by an experienced dermatologist and three subjects' complete eye movement sequences acquired during the inspection process superimposed onto the image, respectively. To visualize eye movement sequences, each circle center represents a fixation location and the radius is proportional to the fixation duration. A line connecting two fixations represents a saccade. Images used with permission from Logical Images, Inc.

tinctively different from their surroundings' (19, 20, 21) and top-down process in which cognitive processes, guided by the viewing task and scene context, influence visual attention (22, 23). In particular, growing evidence suggests that top-down information dominates the active scene viewing process and the influence of low-level salience guidance is minimal (23). These theoretical foundations provide us with the possibility to pursue this engaged cognitive processing based on observed eye movements.

There are a number of types of eye movements:

- *Saccade* denotes moving the eye from one location to another. The features of saccades are the saccade amplitude which is the length in degrees of visual angle, and the speed in degree per second.

- *Fixation* depicts the eye focusing on some target and keeping still. The features of fixations are the location (where the eye was fixating), the duration (how long did the eye fixate), and pupil dilation.
- *Smooth pursuit* represents the eye closely following a moving target at a relatively slow speed.
- *Vergence* means the inward or outward turning of eyes while focusing on a target to obtain single binocular vision.
- *Microsaccade* is tiny involuntary eye movement which typically occurs during long-duration fixation.

It is acknowledged that covert visual attention can be dissociated from eye movements (24). Nevertheless, saccades which direct gaze to a new location usually follow a shift of covert attention to this location, leading to speculation that covert attention serves to plan saccades (25). In particular, studies have shown that overt visual attention and covert visual attention are tightly coupled in complex information processing tasks, such as reading and scene perception (26). Thus, we can gain certain insight into the subjects' interests or problem-solving strategies through their eye movements. Both the number of fixations and their durations are commonly assumed to indicate the depth of information processing associated with the visual fields. On the other hand, saccade amplitudes, which are rarely considered in the analysis of eye movement data, may also have an important impact on some conclusions drawn from the visual processing (23, 27, 28). The visualization of some types of eye movements from dermatologists examining dermatological images is illustrated in Figure 2.1.

2.2.2 Modeling visual attention

The concept of the saliency map originally introduced in (19) is based on the Feature Integration Theory (29). A saliency map characterizes the bottom-up distinctiveness of a particular location relative to that of other locations in the scene through its conspicuousness. One derived computational model concerned with understanding people's visual attention deployments on natural images was developed by Itti et al. (30). They built a computational model to evaluate the saliency level of an image based only on extracted low-level visual features such as intensity, color and orientation. According to the computed saliency map, they attempted to predict people's visual attention allocation. The model has been tested over various images, and its performance is generally robust. Particularly in regards to artificial images, its performance is consistent with observations in human. Saliency map is useful for a variety of application (31, 32). Recent research extended from using only low-level visual features to compute the salient scene regions to investigate modeling approaches of multiple cognitive factors that influence visual attention. The main additional factors include one's expectations about where to find information as well as one's current information need (25). To formulate these three cognitive factors, image saliency was redefined in terms of the combination of both top-down and bottom-up cognitive influence and computed to predict users' viewing behaviors from the perspective of probability theory (33, 34), and users were found to adapt their visual search in order to optimize the expected information gain (35).

2.3 Perceptual Skill

Perceptual skill is considered to be the crucial cognitive factor accounting for the advantage of highly trained experts in many domains. Experts who benefit from training, domain knowledge and rich experience can perceive important relationships among multiple findings and identify promising abnormalities (36). Experts generate distinctively different perceptual representations when they view the same scenes as novices (37). Rather than passively ”photocopying” the visual information directly from retinas into minds, visual perception actively interprets the information by altering perceptual representations of the images based on experience and goals. Without guidance of perceptual expertise and domain knowledge, scenes cannot be interpreted effectively solely based on visual features. This motivates us to investigate how to formalize perceptual expertise and reasoning about image contents from the experts' points of view.

Perceptual skill has been studied across various domains where perceptual expertise is highly involved such as sports (2), chess (3), geo-spatial image analysis (4), airport security screening (5) and clinical diagnosis (6, 7, 38). Empirical perceptual studies of medical image-based diagnosis suggest that subjects vary their eye movement behaviors while they proceed in diagnosis on medical images. Furthermore, by analyzing whole sequences of fixation and saccadic eye movements from groups with different expertise levels, significant differences in visual search strategies between groups show that human expertise plays a great role in medical image examination. The nature of expert performance of four observer groups with different levels of expertise has been investigated (7). They compared multiple eye movement measures and suggested these distinctive variations among the observations of the better performance from higher expertise level are due to the consequences of experience and training. Eye movement studies on diagnostic pathology of light microscopy to identify distinctive viewing stereotypes for each level of experience have also been conducted (6). Their results suggest that eye movement monitoring could serve as a basis for the creation of innovative pathology training routines.

Although capturing perceptual skill is challenging, comprehension of the cognitive basis could benefit a wide range of research areas in medical informatics such as medical image retrieval, proactive human computer interaction, and training. We approach this challenge by working closely with medical specialists (dermatologists) using human-centered experimental approaches to observe and record their perceptual processing while inspecting medical images towards diagnosis. The inherent dynamic property and complexity of experts' diagnostic reasoning motivates our investigation into the temporal dynamics of this perceptual-conceptual-interleaving process.

Previous studies fill the gap between physicians' interpretations and the statistics of pixel values by experts' manual annotations on segmented images and mapping into a domain knowledge ontology so as to perform medical image analysis at a semantic level (39, 40). However, there exists great inter-variability between experts and inner-variability with which a single expert's performance changes from time to time also hinders this approach (41). Moreover experts' perception, as tacit knowledge, functions below the level of consciousness. The eye tracking technique allows researchers to study experts' subconscious image viewing behaviors by objectively measuring eye movements and is a promising way to address these challenges. Recently, more and more studies have tried to incorporate human perceptual skills into image understanding approaches, treating eye movements as a static process by directly mapping eye movement data into the image feature space or by weighting image segments. However, the facts that meaningful perceptual patterns sometimes exist only over time and that the observed eye movement data are noisy and inconsistent undermine the reliability and robustness of these methods. In particular, latent behaviors underlying these observable human behaviors is a critical intermediate step in terms of advancing image understanding, as shown in Figure 2.2 (c). One of the important contributions of our work is that we are trying to capture the spatial-temporal patterns existing in eye movement data.

2.3.1 Applications of perceptual skill

Recent empirical studies (42, 43, 44) suggest that eye movement patterns, as a promising resource of implicit relevance feedback, are inherently encoded with rich information about users' interests. Various computational approaches (43, 45, 46, 47) have been explored to elicit some aspects of cognitive processing information from users' eye movements while they were reading documents or viewing images, and achieved reasonable accuracy on relevance evaluation. Comprehension of this perceptual processing has implications for research in design of information systems such as multi-modal interfaces, clinical decision support, performance support, learning and medical training. Because of the ubiquity of the graphical information and image-rich tasks, results we present here could benefit a wide range of user modeling and interaction with novel interfaces that incorporate knowledge- or agent-based approaches.

Implicit relevance feedback in various forms, captured through unobtrusive observation of users' behaviors, is valuable to intelligent interfaces, since it can provide subconscious information with regard to users' informational needs or interests (48, 49, 50). The analysis of users' thought processes based on their verbal narratives is a powerful approach which can be used to monitor and understand users' dynamic behaviors (51, 52, 53), as well as to shed light on interpersonal interaction. In particular, annotations on verbal narratives of thinking processes provide concise yet expressive information regarding coordination between perceptual and conceptual processing of experts' thinking processes in medical domain (54). By instructing experts to express their diagnostic reasoning and decision-making using language, we can coordinate these multi-modal data to capture their dynamic information interests.

According to previous studies(55, 56), there exist commonalities in tasks such as diagnostic reasoning due to the remarkable number of regularities in human information processing, despite all the idiosyncrasies and individual differences. Idiosyncratic behavior differs from the average behavior of a population of observers, whereas stereotypical behavior closely complies with such an average behavior. If we can capture and describe the general strategies of knowledgeable and skilled doctors, we can apply this to help more novice students. In that way, the educational experience of the medical student should be improved, as they are in real need of basic strategies and principles of diagnostic-reasoning. The question is whether cognitive strategies can be captured and characterized through eye movement sequences during diagnostic reasoning processes. In order to differentiate between such types of behavior, previous methods make use of similarity measures that allow for the comparison between eye movement sequences of different observers (56). However, these methods are inevitably constrained by various limitations, which will be elaborated in the following section.

So as to capture medical specialists' (dermatologists) stereotypical and idiosyncratic visual behaviors from their eye movements, we use human-centered experimental approaches, which means incorporating human knowledge and skills into the procedure, to observe and record their reasoning processes while inspecting medical images. We then profile the shared time-evolving eye movement patterns among physicians through our computational model, and also time-align eye movement patterns with semantic group labels annotated by experts based on other dermatologists' verbal descriptions. We discuss the implications of integrating these multi-modal data towards understanding users' dynamic informational needs.

2.3.2 Evaluation of visual behaviors

Since the eye movement analysis method of summary fixation statistics is limited in terms of eliciting hidden knowledge from eye tracking data, the area of developing effective metrics for comparing and evaluating large amounts of fixation and saccadic eye movement data represented as scanpaths becomes more and more prominent (57, 58, 59, 60, 61). The basic idea of the current eye movement analysis methods generally needs to define a notion of "distance" between eye movement sequences first. They evaluate the similarity of the eye movement sequences by calculating the pairwise distances between them. These methods can be broadly categorized into two classes.

One class of these algorithms are based on predefined AOIs (61). A temporal sequence of AOIs is defined based on either dividing a scene in equally spaced bins or segmenting semantically meaningful regions in the scene. Then string-edit algorithms can be used to compare different sequences. These algorithms calculate the distance between two strings as the minimum number of edits required to transform one into the other. However, there are some issues: human intervention is still needed with respect to defining AOIs or specifying the size of the square regions and their locations; fixation durations are not taken into account; string editing comparison among multiple scanpaths fail to measure meaningful variations between scanpaths. A recent study develops a comprehensive pair-wise comparison approach in order to take more eye movement features into account

(62).

The other analysis methods are based on clustering techniques (59). Clusters of fixation points are first grouped via parametric or non-parametric clustering algorithms based on their relative locations (x-y coordinate). Foveal acuity is usually modeled as a 2D Gaussian distribution around a fixation location. Since the spatial resolution of the visual processing attenuates sharply from fovea vision to peripheral vision, Gaussian distribution is a reasonable approximation because of its light tails. After these clusters are labeled, we can measure what percentage of the image is selected for high acuity and foveal registration by summing up the number of pixels falling within this range and calculating as percentage of total number of pixels in the viewed image. The problems with this methods are that sequential information is ignored, the clusters are not always meaningful, and fixation durations or saccade information are still not taken into account.

To compensate for the above limitations, the Earth Mover's Distance (EMD) metric was proposed to measure the similarity of different visual behavior sequences (63). The similarity between eye movement sequences are viewed as a transportation problem by defining one sequence as a set of piles of earth and another sequence as a collection of holes and by setting the cost for a pile-hole pair to equal the ground distance between fixation in the two sequences. Studies show that this type of pair-wise comparison metric is very sensitive to data variance and performs particular poorly to deal with noise.

Some studies adopted HMMs to measure or profile eye movement sequences (38, 64, 65). The disadvantage of these approaches is that they either have to heuristically predefine the number of hidden states or use standard parametric model selection methods to identify a "best" single number, the strengths and weaknesses of which in this problem setting is unknown. Two alternatives to HMMs are AOI-based or clustering-based methods mentioned above.

Although comprehensive eye movement features are taken into account in recent studies (62), current pairwise comparison algorithms among multiple scanpaths are sensitive to data noise and minor variations between scanpaths. Furthermore, meaningful patterns may only exist over the time of complete processes, rather than comparing them piece by piece. This suggests a Markovian framework in which the model transitions among patterned eye movement behaviors, and these meaningful components are associated with perceptual expertise and domain knowledge.

Recently there has been significant interest in augmenting dynamic systems' capabilities of modeling time series by combining stochastic processes. The hierarchical Dirichlet process (HDP) based HMMs allow the number of hidden states to be learned from observations by treating transition distributions as realizations of the HDP over countably infinite state spaces (66, 67, 68). The infinite factorial HMM models a single time-series with emissions dependent on a feature with potentially infinite dimensionality which evolves with independent Markov processes (69). Beta process (BP) based HMMs model multiple time series and capture an infinite number of potential dynamical modes which are shared among the series using the Indian buffet process (IBP) by integrating over the latent BP (70). However, these approaches lack the capability of modeling multiple related but distinct families of time series. This modeling requirement in our problem scenario motivates us to develop a novel hierarchically-structured dynamic model which is capable of profiling stereotypical and idiosyncratic patterns from multiple expertise-specific groups of eye movement sequences.

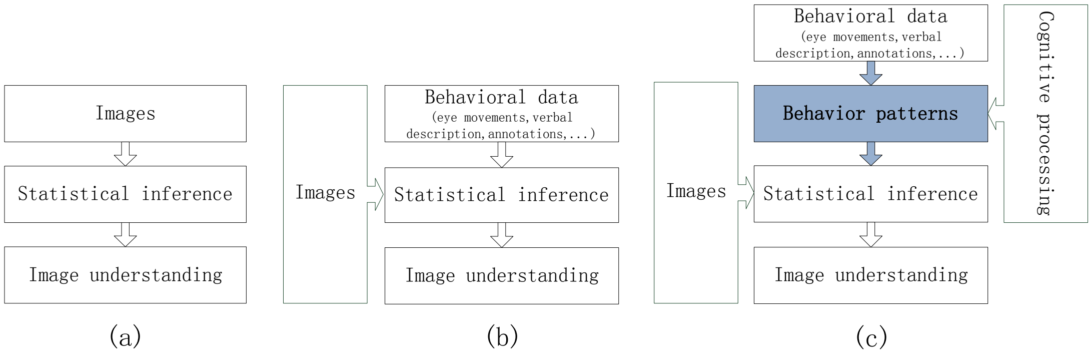


Figure 2.2: Diagram of three approaches to image understanding. In (a), early image understanding approaches attempt to interpret images solely based on statistical analysis of image pixel values. As shown in (b), recently researchers have incorporated human subjectivity into image understanding by essentially treating observed human behavioral data as weights added into selected image features or segments. Since this approach fails to consider the underlying behavior patterns and cognitive processing (domain knowledge, expertise, expectations...) that dominates observable human behaviors, it is hindered by the noisy and inconsistent nature of the observed behaviorial data. In (c), we propose that novel approaches to extract tacit knowledge from experts engaging in these observable behaviors will be a more effective approach to incorporate human capabilities. The extracted behavior patterns are not only more robust and consistent but also shed light on revealing latent cognitive processing. The dynamic nature of human behaviors involved in diagnostic reasoning is important. Our approach aims at not only capturing spatial information but also temporal characteristics of human behaviors.

3

Probabilistic Modeling Approaches

If we assume that components of human cognitive processing approximately follow the principles of probability theory, we can computationally reason backwards from observed human behavioral data to the engagement of particular cognitive functions as an inverse problem. In this way, the computational framework for probabilistic inference provides a general approach to understanding the deeper cognitive processing that is not directly observable to us based on sparse, noisy and ambiguous behavioral data.

To understand the computational basis of the knowledge-based diagnosis processes, there are several principles that must be addressed.

Persons may have different prior experience of what the world would be. This leads to their use of unique prior knowledge to guide a learning process. Bayesian methods of the probability theory allow us to incorporate various forms of prior knowledge into learning, inference and decision-making in a principled manner.

The basic notion called "prior" can be formulated as the probabilistic representation of human abstract knowledge regarding how they expect the world to be.

To uncover what are the forms and contents of persons' knowledge of the world and make comparison between them, there are two schools of approaches available. One typically represents the knowledge relevancy with simple probabilistic models based on numbers or parametric probability distributions without considering more structural representation. The other approaches the problem with structured logical and symbolic knowledge representation. The tools they normally use are graphs, grammars and system of logic. We realize that the integration of the two is absolutely essential to our particular study on perception-based diagnostic-reasoning. To combine these two knowledge representation strategies, we need to use structures and symbols as a way of representing the structure of human knowledge and then define probabilistic models over those structure representations. As Glenn Shafer and Judea Pearl pointed out, probability is not really about numbers, it is about the structure of reasoning. Graphical models as a general class of probabilistic models are used to define probabilities over structured knowledge representations. This modeling technique allows us to model not only how systems of knowledge can be applied to guide perceptual processing but also how they can be learned by various kinds of statistical inferences. The key characters of diagnostic reasoning processes can be depicted using several properties of the probabilistic graphical models:

- Levels of abstraction: Hierarchical probabilistic models allow us to use multiple levels of abstraction to represent human cognitive processing. We assume there are multiple levels of representation which are all linked by probability distribution, for example physicians' prior at the lowest level (a set of specific diagnostic cues) itself is generated by a distribution over distribution, some sort of prior on priors. And by doing this inference on multiple levels models, we can understand how physicians acquire the conceptual knowledge of the world as well as how they use it to guide their perceptual reasoning process.

- Infinity of learning: medical training is not just accumulation of more and more bits of knowledge, but qualitatively transformation as real cognitive growth. This requires us to have ways of building probabilistic models which in some sense are not constrained by initial structure but where the structure itself can do qualitative transformation as data come in and may grow in qualitative ways. We propose to use non-parametric probabilistic graphical models whose structures can keep evolving as more data are observed via assimilation-accommodation mechanism.
- Dynamics of learning: If the diagnostic-reasoning process can be characterized as dynamic systems, we can reinterpret these dynamic systems as stochastic processes that can be represented as non-parametric probabilistic graphical models.

According to these analysis, we review the statistical theories and methodologies upon which our contributions are based.

In Section 3.1, we discuss the theoretical rationale for the Bayesian approaches. In Section 3.2 we describe the exponential families which represent a family of functions and probability distributions extensively used as data models in the Bayesian approaches. In Section 3.3 we analyze some attractive properties of the exponential families in order to uncover the reasons why they are widely used in probabilistic modeling works. In Section 3.4 and Section 3.5, we propose conjugate priors as the other critical component of the Bayesian approaches. Through Section 3.6 to Section 3.8, we demonstrate several conjugate priors and their likelihoods, and discuss their mathematical and computational properties as well.

These examples will be served as some basic components in our modeling approach.

3.1 Exchangeability

In some cases the physical process giving rise to the observed data is known, a probabilistic topological structure can be determined accordingly. This probabilistic topological structure is composed of a set of interrelated random variables and their dependence. For instance, hidden Markov models (HMMs) are often derived from some known dynamical systems, and Markov random fields (MRFs) as a spatial generation from HMMs arise from the discretization of stochastic partial differential equations. However, in other learning cases where the generative process may be unknown or too complex to be characterized explicitly, some simple assumptions about the indistinguishability of different observations can lead naturally to a family of hierarchical, directed topological structures.

The concept of exchangeability serves as a critical theoretical base for various statistical approaches. Assume we are gathering data in an attempt to make predictions about future observations of the underlying random process. With the strong assumption of the data being independently distributed, we would treat every new data point individually with no need to predict future observations given the past, since:

$$p(y_1, ..., y_n) = \prod_{i=1}^{n} p(y_i) \tag{3.1}$$

implies that

$$p(y_{n+1}, ..., y_m | y_1, ..., y_n) = p(y_{n+1}, ..., y_m) \tag{3.2}$$

However, we could relax the assumption using a weaker one which often better describes the observations. The weaker assumption is exchangeability which states that the data order we observe is inconsequential.

Definition 3.1.1. *A sequence of random variables y_1, y_2,..., y_n is said to be finitely exchangeable if*

$$y_1, y_2, ..., y_n \overset{\circ}{=} y_{\pi(1)}, y_{\pi(2)}, ..., y_{\pi(n)} \tag{3.3}$$

for every permutation π on $\{1, ..., n\}$. Here, we use the notation $\overset{\circ}{=}$ to mean equality in distribution.

These variables are exchangeable in the sense that every permutation, or reordering, of their indices has equal probability. This definition suggests that independence implies exchangeability, but not vice versa. Since the computational problems are often involved with handling cases where data is continually accumulated or an upper bound is challenging, it would be useful to extend the notion for infinite sequences.

Definition 3.1.2. *A sequence y_1, y_2,... is infinitely exchangeable if every finite subsequence is finitely exchangeable.*

When no auxiliary information is available, this assumption is usually reasonable. Sometimes, we can take a further step to relax exchangeability by considering partially exchangeable data where some auxiliary information allows us to partition the data into exchangeable sets. An important result derived from the assumption of exchangeable data is de Finetti's theorem which states that an infinite sequence of random variables y_1, y_2,... is exchangeable if and only if there exists a random probability measure ν with respect to which y_1, y_2,... are conditionally independent identically distributed (i.i.d.) with distribution ν. The general form of the de Finetti theorem:

Theorem 3.1.1. *If y_1, y_2,... is an infinitely exgchangeable sequence of real-valued random variables with probability measure P, then there exists a probability measure μ defined on the space of all probability measures $\wp(\mathbb{R})$ on $\mathbb{R}$ such that*

$$P(y_1 \in A_1, ..., y_n \in A_n) = \int_{\wp(\mathbb{R})} \prod_{i=1}^{n} \nu(A_i)\mu(d\nu) \tag{3.4}$$

Furthermore, μ is the law of a probability measure ν, where ν is almost surely defined by the limiting empirical measure. Namely,

$$\nu(B) \overset{a.s.}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_B(y_i), \qquad \nu \sim \mu \tag{3.5}$$

where B ranges over all elements of the Borel σ-algebra. The measure μ is often referred to as the de Finetti measure.

De Finetti proved this in the case of binary random variables (71). There is also a simpler proof in more modern terms in (72) and (73). Generatively, the theorem states that if y_1, y_2, ... are infinitely exchangeable, then there exists a measure μ on measures such that:

$$\nu \sim \mu \tag{3.6}$$

$$y_i|\nu \overset{i.i.d.}{\sim} \nu \tag{3.7}$$

According to de Finetti theorem, the Bayesian perspective of the parameter yielding the observations i.i.d. is treated as a random quantity with some distribution μ. On the contrary, from the frequentist perspective the parameter yielding the observations i.i.d. is considered as a fixed unknown quantity. If we focus our attention onto the finite-dimensional parameter θ cases, we can invoke the following corollaries,

Corollary 3.1.1. *Assuming the required densities exist, and assuming the conditions of Theorem 2.5.1 hold, then exists a distribution function Q such that the joint density of y_1,..., y_n is of the form*

$$p(y_1, ..., y_n) = \int_\Theta \prod_{i=1}^{n} p(y_i|\vartheta)dQ(\vartheta) \tag{3.8}$$

with $p(\cdot|\vartheta)$ representing the density function corresponding to the finite-dimensional parameter $\vartheta \in \Theta$.

The above corollary explicitly indicates that the de Finetti theorem motivates the concept of a prior distribution $Q(\cdot)$ and a likelihood function $p(y|\cdot)$. In Bayesian statistics, this is known as a hierarchical model due to the layering by which observations depend on parameters, which are in turn related to hyper-parameters (73, 74).

Corollary 3.1.2. *Given that the conditions of Corollary 2.5.1 hold, then the predictive density is given by*

$$p(y_{m+1}, ..., y_n|y_1, ..., y_m) = \int_\Theta p(y_i|\vartheta)dQ(\vartheta|y_1, ..., y_m) \tag{3.9}$$

where

$$dQ(\theta|y_1, ..., y_m) = \frac{\prod_{i=1}^{m} p(y_i|\theta)dQ(\theta)}{\int_\Theta \prod_{i=1}^{m} p(y_i|\vartheta)dQ(\vartheta)} \tag{3.10}$$

The form of the predictive density in Eq. (3.9) shows a core idea of Bayesian inference, which is that the prior belief $Q(\theta)$ is updated into a posterior belief $Q(\theta|y_{1,...,y_m})$ through an application of Bayes rule without changing our view of the existence of an underlying random parameter θ yielding the data i.i.d. In particular, the process of forming the posterior distribution in Eq. (3.10) from the prior by incorporating observations is a fundamental step in examining the predictive distribution.

The issue of tractable inference often leads to the use of conjugate priors. Besides mathematical convenience, conjugate priors allow us to encode and quantify prior knowledge as a set of fictional observations in the posterior distributions. The goal of flexibility in our models motivates us to adopt non-parametric methods in terms of making as few assumptions as possible. Another key issue of the Bayesian framework is in characterizing a likelihood distribution for how our data are given rise to condition a parameter value θ. This choice is often motivated by practical considerations that are typically related to those of choosing a prior distribution. As practitioners, we do not focus on a full analysis of model selection in this thesis, instead we use a combination of our insight on the process and our adherence to computational limitations to define a model.

3.2 Exponential Families

An exponential family of probability distributions is characterized by certain sufficient statistics which summarize the observations using a fixed number of values (75, 76, 77). To present a general form of exponential family, let x be a random variable with values from a sample space $\mathcal{X}$, which may be either continuous or discrete. The corresponding exponential family of densities is given by

$$p(x|\theta) = \nu(x)\exp\{\sum_{a\in\mathcal{A}} \theta_a \phi_a(x) - \Phi(\theta)\} \tag{3.11}$$

where $\{\phi_a | a \in \mathcal{A}\}$ is a set of statistics or potentials, $\theta \in \mathbb{R}^{|\mathcal{A}|}$ are the family's canonical parameters, and $\nu(x)$ is a non-negative reference measure. The parameter θ can be set either to fixed constants or latent random variables. The log partition function $\Phi(\theta)$ is defined to normalize $p(x|\theta)$:

$$\Phi(\theta) = \log \int_{\mathcal{X}} \nu(x) \exp\{\sum_{a \in \mathcal{A}} \theta_a \phi_a(x)\} dx \tag{3.12}$$

so that it integrates to one.

This construction is valid as long as the canonical parameters θ belong to the set Θ for which the log partition function is finite:

$$\Theta \triangleq \{\theta \in \mathbb{R}^{|\mathcal{A}|} | \Phi(\theta) < \infty\} \tag{3.13}$$

Since $\Phi(\theta)$ is a convex function, Θ is also convex. In particular, the exponential family is said to be regular when Θ is open. Many probability density functions belong to regular exponential families, including the Bernoulli, beta, Poisson, Gaussian and gamma densities (73, 74). We present a set of examples of such manipulations.

Bernoulli:

$$\begin{aligned} p(x|\theta) &= \theta^x (1-\theta)^{1-x} \qquad x \in \{0, 1\} \qquad (3.14) \\ \ln(p(x|\theta) &= x \ln \theta + (1-x) \ln(1-\theta) \qquad (3.15) \\ &= \ln(\frac{\theta}{1-\theta}) x + \ln(1-\theta) \qquad (3.16) \end{aligned}$$

Geometric:

$$\begin{aligned} p(x|\theta) &= (1-\theta)\theta^x \qquad x \in \{0, 1, 2, ...\} \qquad (3.17) \\ \ln p(x|\theta) &= \ln(\theta) x + \ln(1-\theta) \qquad (3.18) \end{aligned}$$

Poisson:

$$\begin{aligned} p(x|\theta) &= \frac{\theta^x e^{-\theta}}{x!} \qquad x \in \{0, 1, 2, ...\} \qquad (3.19) \\ \ln p(x|\theta) &= \ln(\theta) x - \theta - \ln x \qquad (3.20) \end{aligned}$$

Exponential:

$$p(x|\theta) = \theta e^{-\theta x} \qquad x > 0 \tag{3.21}$$

$$\ln p(x|\theta) = -\theta x + \ln \theta \tag{3.22}$$

3.3 Properties of canonical exponential families

In this section, we discuss some properties of exponential families as the motivation of its widely use through the notion of sufficiency and information theory. In particular, the following properties of the log partition function is critical in the study of exponential families (75, 76, 77).

Proposition 3.3.1. *The log partition function $\Phi(\theta)$ of Eq. 3.12 is convex (strictly so for minimal representations) and continuously differentiable over its domain Θ. Its derivatives are the cumulants of the sufficient statistics $\phi_a | a \in \mathcal{A}$, so that*

$$\frac{\partial \Phi(\theta)}{\partial \theta_a} = \mathbb{E}_\theta[\phi_a(x)] \triangleq \int_{\mathcal{X}} \phi_a(x) p(x|\theta) dx \tag{3.23}$$

$$\frac{\partial^2 \Phi(\theta)}{\partial \theta_a \partial \theta_b} = \mathbb{E}_\theta[\phi_a(x)\phi_b(x)] - \mathbb{E}_\theta[\phi_a(x)]\mathbb{E}_\theta[\phi_b(x)] \tag{3.24}$$

The log partition function $\Phi(\theta)$ is also called cumulant generating function of the exponential family for this reason, the convexity of which has important implications for other properties of exponential families (78, 79).

For problems of model selection and approximation, we need a similarity measure of probability distributions. To use the relative entropy or Kullback-Leibler (KL) divergence to measure an approximation accuracy, we need to introduce some information-theoretic concepts first (80).

Shannon's measure of entropy conveys the inherent uncertainty of a random variable x taking values within a finite space $\mathcal{X}$:

$$H(x) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{3.25}$$

where $p(x)$ is the associated probability mass function defining the law of x. The notion of entropy can be straightforwardly extended to jointly random variables $(x, y) \sim p(x, y)$, in which the joint entropy is defined as

$$H(x, y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \tag{3.26}$$

Similarly, a conditional entropy of a random variable x given y can be defined:

$$H(x|y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y)\ logp(x|y) \tag{3.27}$$

The joint entropy $H(x, y)$ is simply the sum of the entropy of y, $H(y)$, and the conditional entropy of x given y, $H(x|y)$, which has a nice interpretation with respect to conservation of uncertainty. The change in entropy of a random variable x after an observation y is given by the mutual information:

$$\begin{align} I(x; y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{3.28} \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y)(\log p(x) - \log p(x|y)) \tag{3.29} \\ &= H(x) - H(x|y) \tag{3.30} \end{align}$$

A nice property of mutual information is that it is symmetric in terms of $I(x; y)$ can also be seen as the change in entropy of y after observing x.

The above definitions can be extended to continuous random variables by considering differential entropy

$$h(x) = -\int_{\mathcal{X}} p(x) \log p(x) dx \tag{3.31}$$

and differential conditional entropy

$$h(x|y) = -\int_{\mathcal{X}}\int_{\mathcal{Y}} p(x,y)\log p(x|y)dxdy \tag{3.32}$$

However, differential entropy does not have the non-negative property as discrete entropy does.

The KL divergence between two probability distributions $p(x)$ and $q(x)$ equals

$$D(p||q) = \int_{\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}dx \tag{3.33}$$

Although KL divergence is not actually a distance metric because of its asymmetric property, it is informative for variational methods about a 'better' approximation. From this definition, mutual information can be interpreted as the KL divergence between a joint distribution of (x, y) and the distribution assuming they are independent:

$$D(p||q) = D(p(x,y)||p(x)p(y)) \tag{3.34}$$

The mutual information which is defined with respect to differential entropy is presented in (80).

Learning problems can be posed as a search for the best approximation of an empirically derived target density $\widetilde{p}(x)$. The KL divergence $D(p||q)$ is a natural measure of the accuracy of an approximation $q(x)$. The following moment-matching conditions elegantly characterize the optimal approximating density for exponential families:

Proposition 3.3.2. *Let $\widetilde{p}$ denote a target probability density, and p_θ an exponential family. The approximating density p_θ minimizing $D(\widetilde{p}||p_\theta)$ then has canonical parameters $\widehat{\theta}$ chosen to match the expected values of that family's sufficient statistics:*

$$\mathbb{E}_{\widehat{\theta}}[\phi_a(x)] = \int_{\mathcal{X}} \phi_a(x)\widetilde{p}(x)dx \qquad a \in \mathcal{A} \tag{3.35}$$

For minimal families, these optimal parameters $\widehat{\theta}$ are uniquely determined.

It is worth noting that for non-minimal families, while the optimal parameters are not unique, the resulting distribution p_θ is still the same, since all the different minimizers are just re-parameterizations of the same density.

The approximating density p_θ minimizing $D(\widetilde{p}||p_\theta)$ depends only on the potential functions' expected values under $\widetilde{p}(x)$, so that these statistics are sufficient to determine the closest approximation.

We normally observe L independent samples $\{x^{(l)}\}_{l=1}^{L}$ from a target density $\widetilde{p}(x)$ instead of that density explicitly itself. In this case, we define the empirical density of the samples as follows:

$$\widetilde{p}(x) = \frac{1}{L}\sum_{l=1}^{L}\delta(x, x^{(l)}) \tag{3.36}$$

Here, $\delta(x, x^{(l)})$ is the Dirac delta function for continuous $\mathcal{X}$, and the Kronecker delta for discrete $\mathcal{X}$. In such case, there is a correspondence between information projection and maximum likelihood (ML) parameter estimation as stated in the following proposition:

Proposition 3.3.3. *Let p_θ denote an exponential family with canonical parameters θ. Given L independent, identically distributed samples $\{x^{(l)}\}_{l=1}^{L}$, with empirical density $\widetilde{p}(x)$ as Eq. 3.36, the maximum likelihood estimate $\widehat{\theta}$ of the canonical parameters coincides with the empirical density's information projection:*

$$\widehat{\theta} = \arg\max_{\theta}\sum_{l=1}^{L}\log p(x^{(l)}|\theta) = \arg\min_{\theta} D(\widetilde{p}||p_\theta) \tag{3.37}$$

These optimal parameters are uniquely determined for minimal families, and characterized by the following moment matching conditions:

$$\mathbb{E}_{\widehat{\theta}}[\phi_a(x)] = \frac{1}{L}\sum_{l=1}^{L}\phi_a(x^{(l)}) \qquad a \in \mathcal{A} \tag{3.38}$$

From these results, we can see that certain statistics are sufficient to characterize the best exponential family approximation of a given target density. In principle, Prop. 3.3.2 and Prop. 3.3.3 suggest a straightforward procedure for learning exponential families: estimate appropriate sufficient statistics, and then find corresponding canonical parameters via convex optimization (75, 76, 79). However, significant difficulties may arise in practice. Sometimes the required statistics cannot be directly measured such as semi-supervised learning from partially labeled training data, or calculation of the corresponding parameters is intractable in some large complex models. Another constraint of these results is the selection of appropriate exponential families. In particular, since the chosen statistics are sufficient for parameter estimation, the learned model cannot capture aspects of the target distribution neglected by these statistics. These concerns motivate us to non-parametric methods which extend exponential families to learn richer and more flexible models.

Theorem 3.3.1. *Consider a collection of statistics* $\{\phi_a | a \in \mathcal{A}\}$*, whose expectations with respect to some target density* $\widetilde{p}(x)$ *are known:*

$$\int_{\mathcal{X}} \phi_a(x)\widetilde{p}(x)dx = \mu_a \qquad a \in \mathcal{A} \tag{3.39}$$

The unique distribution $\widehat{p}(x)$ *maximizing the entropy* $H(\widehat{p})$*, subject to these moment constraints, is then a member of the exponential family of Eq. 3.11, with* $\nu(x) = 1$ *and canonical parameters* $\widehat{\theta}$ *chosen so that* $\mathbb{E}_{\widehat{\theta}}[\phi_a(x)] = \mu_a$*.*

Previous propositions show that it is sufficient to characterize the best exponential family approximation of a given target density by certain statistics. This theorem indicates that if those statistics are the only available information about a target density, then the corresponding exponential family provides a natural model which imposes the fewest additional assumptions about the data generation process. Eq. 3.39 implicitly assumes the existence of some distribution satisfying the specified moment constraints. Also given insufficient moment constraints for non-compact continuous spaces, the maximizing density may be improper and have infinite entropy.

3.4 Incorporating prior knowledge

Besides the fact that exponential families use sufficient statistics to characterize the likelihood function of the parameters given observed training data, we normally have some prior knowledge about the process giving rise to the data. Bayesian methods allow a principled way of incorporating prior knowledge with likelihoods by treating the parameters of exponential family density functions as random variables. In particular, consistent incorporation of prior knowledge can dramatically improve the accuracy and robustness of the learned model when datasets are small (73).

Bayesian analysis begins with a prior distribution $p(\theta|\lambda)$ describing people's available knowledge about how the data are generated. Then an exponential family $p(x|\theta)$ with canonical parameters θ updates our belief. Given L i.i.d. observations $\{x^{(l)}\}_{l=1}^{L}$, the posterior distribution of the canonical parameters can be written as follows according to Bayes' rule:

$$p(\theta|x^{(1)}, ..., x^{(L)}, \lambda) = \frac{p(x^{(1)}, ..., x^{(L)}|\theta, \lambda)p(\theta|\lambda)}{\int_{\Theta} p(x^{(1)}, ...x^{(L)}|\theta, \lambda)p(\theta|\lambda)d\theta} \tag{3.40}$$

$$\propto p(\theta|\lambda) \prod_{l=1}^{L} p(x^{(l)}|\theta) \tag{3.41}$$

Since for minimal exponential families the canonical parameters are uniquely associated with expectations of that family's sufficient statistics, the posterior distribution of Eq. 3.40 describes our belief about the statistics which is likely to be exhibited by future observations.

When statistical models are used to predict future observations, the predictive likelihood of a new observation $\bar{x}$ can be written as follows given L i.i.d. observations:

$$p(\bar{x}|x^{(1)},...,x^{(L)},\lambda) = \int_{\Theta} p(\bar{x}|\theta)p(\theta|x^{(1)},...,x^{(L)},\lambda)d\theta \tag{3.42}$$

where the posterior distribution over parameters θ is as in Eq. 3.40. This predictive likelihood provides us typical predictions which are more robust than single parameter estimation by averaging over our posterior uncertainty in the parameter θ. However, the predictive likelihood computation is intractable for many practical models. In these cases, the parameters' posterior distribution is often approximated by a single maximum a posteriori (MAP):

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} p(\theta|x^{(1)},...,x^{(L)},\lambda) && (3.43) \\ &= \arg\max_{\theta} p(\theta|\lambda)\prod_{l=1}^{L} p(x^{(l)}|\theta) && (3.44) \end{aligned}$$

This approach is best justified when the training set size L is large, so that the posterior distribution is tightly concentrated.

A fully Bayesian analysis should also specify a prior distribution $p(\lambda)$ over the hyper-parameter λ. An empirical Bayesian approach (73), however, estimates the hyper-parameter λ by maximizing the training data's marginal likelihood:

$$\begin{aligned} \hat{\lambda} &= \arg\max_{\lambda} p(x^{(1)},...,x^{(L)}|\lambda) && (3.45) \\ &= \arg\max_{\lambda} \int_{\Theta} p(\theta|\lambda)\prod_{l=1}^{L} p(x^{(l)}|\theta) && (3.46) \end{aligned}$$

In situations where this optimization is intractable, we can optimize the predictive likelihood of a held-out dataset through cross-validation approaches (73).

It is useful to have compact ways of characterizing large datasets when computing the posterior distributions and predictive likelihoods. The following theorem shows that the notions of sufficiency can be used to simplify learning with prior knowledge (73).

Theorem 3.4.1. *Let $p(x|\theta)$ denote an exponential family with canonical parameters θ, and $p(\theta|\lambda)$ a corresponding prior density. Given L i.i.d. samples $\{x^{(l)}\}_{l=1}^{L}$, consider the following statistics:*

$$\Phi(x^{(1)}, ..., x^{(L)}) \triangleq \{\frac{1}{L}\sum_{l=1}^{L}\phi_a(x^{(l)})|a \in A\} \tag{3.47}$$

These empirical moments, along with the sample size L, are then said to be parametric sufficient for the posterior distribution over canonical parameters, so that

$$p(\theta|x^{(1)}, ...x^{(L)}, \lambda) = p(\theta|\Phi(x^{(1)}, ..., x^{(L)}), L, \lambda) \tag{3.48}$$

Equivalently, they are predictive sufficient for the likelihood of new data $\bar{x}$:

$$p(\bar{x}|x^{(1)}, ...x^{(L)}, \lambda) = p(\bar{x}|\Phi(x^{(1)}, ..., x^{(L)}), L, \lambda) \tag{3.49}$$

The significant compression provided by the above statistics makes exponential families particularly useful. This theorem also emphasizes the importance of selecting appropriate sufficient statistics, since other features of the data cannot affect subsequent model predictions.

3.5 Properties of conjugate priors

The motivation to introduce conjugate priors is that although Theorem 3.4.1 shows that statistical predictions $p(\bar{x}|x^{(1)}, ...x^{(L)}, \lambda)$ in exponential families $p(x|\theta)$

can be expressed in functions solely of the chosen sufficient statistics $\Phi(x^{(1)}, ..., x^{(L)})$, it neither provides us with an explicit representation of the posterior distribution $p(\theta|x^{(1)}, ...x^{(L)}, \lambda)$ over model parameters nor guarantee a tractable computation of the predictive likelihood $p(\bar{x}|x^{(1)}, ...x^{(L)}, \lambda)$. We, therefore, describe an expressive family of prior distributions which are analytically tractable.

Let $p(x|\theta)$ denote a family of probability densities parameterized by θ. A family of prior densities $p(\theta|\lambda)$ is said to be conjugate to $p(x|\theta)$ if, for any observation x and hyper-parameters λ, the posterior distribution $p(\theta|x, \lambda)$ remains in that family:

$$p(\theta|x, \lambda) \propto p(x|\theta)p(\theta|\lambda) \propto p(\theta|\bar{\lambda}) \tag{3.50}$$

In this case, the posterior distribution is compactly described by an updated set of hyper-parameters $\bar{\lambda}$. For exponential families parameterized as in Eq. 3.11, the general form of conjugate priors is as follows (73, 75):

$$p(\theta|\lambda) = \exp\{\sum_{a\in\mathcal{A}} \theta_a \lambda_0 \lambda_a - \lambda_0 \Phi(\theta) - \Omega(\lambda)\} \tag{3.51}$$

While this functional form duplicates the exponential family's, the interpretation is different: the density is over the space of parameters Θ, and determined by hyper-parameter λ. The conjugate prior is proper, or normalizable, when the hyper-parameters take values in the space Λ where the log normalization constant $\Omega(\lambda)$ is finite:

$$\Omega(\lambda) = \log \int_{\Theta} \exp\{\sum_{a\in\mathcal{A}} \theta_a \lambda_0 \lambda_a - \lambda_0 \Phi(\theta)\} d\theta \tag{3.52}$$

$$\Lambda \triangleq \{\lambda \in \mathbb{R}^{|\mathcal{A}|+1} | \Omega(\lambda) < \infty\} \tag{3.53}$$

Note that the dimension of the conjugate family's hyper-parameters λ is one larger than the corresponding canonical parameters θ.

The following result verifies that the conjugate family of Eq. 3.51 satisfies the definition of Eq. 3.50, and provides an intuitive interpretation for the hyper-parameters:

Proposition 3.5.1. *Let $p(x|\theta)$ denote an exponential family with canonical parameters θ, and $p(\theta|\lambda)$ a family of conjugate priors defined as in Eq. 3.51. Given L independent samples $\{x^{(l)}\}_{l=1}^{L}$, the posterior distribution $p(\bar{x}|x^{(1)}, ...x^{(L)}, \lambda)$ remains in the same family:*

$$p(\theta|x^{(1)}, ..., x^{(L)}, \lambda) = p(\theta|\bar{\lambda}) \tag{3.54}$$

$$\bar{\lambda}_0 = \lambda_0 + L \qquad \bar{\lambda}_a = \frac{\lambda_0\lambda_a + \sum_{l=1}^{L} \phi_a(x^{(l)})}{\lambda_0 + L} \qquad a \in \mathcal{A} \tag{3.55}$$

Integrating over Θ, the log-likelihood of the observations can then be compactly written using the normalization constant of Eq. 3.52:

$$\log p(x^{(1)}, ...x^{(L)}|\lambda) = \Omega(\bar{\lambda}) - \Omega(\lambda) + \sum_{l=1}^{L} \log \nu(x^{(l)}) \tag{3.56}$$

In principle, Prop. 3.5.1 provides a framework for conjugate analysis with any exponential family. From this proposition, we can see that the posterior hyper-parameters $\bar{\lambda}_a$ are a weighted average of the prior hyper-parameters λ_a and the corresponding sufficient statistics of the observations. Conjugate priors can thus be effectively viewed as a set of synthetic pseudo-observations, where λ_a is interpreted as the average of $\phi_a(x)$ with respect to this synthetic data, and λ_0 is the effective size of this synthetic dataset which expresses confidence in these prior statistics and need not be integral. This interpretation often makes it easy to select an appropriate conjugate prior, since hyper-parameters correspond to sufficient statistics with intuitive meaning. In particular, when the number of observations L is large relative to λ_0, the posterior distribution is primarily determined by the observed sufficient statistics.

In the following sections, we briefly outline some probability density and mass functions, and the associated conjugate analysis used extensively in this thesis.

3.6 Multinomial and Bernoulli observations

Consider a random variable x taking one of K discrete, categorical values, so that $\mathcal{X} = \{1, ..., K\}$. Any probability mass function, or distribution, $p(x)$ is then parameterized by the probabilities $\pi_k \triangleq Pr[x = k]$ of the K discrete outcomes:

$$p(x|\pi_1, ..., \pi_K) = \prod_{k=1}^{K} \pi_k^{\delta(x,k)} \qquad \delta(x,k) \triangleq \begin{cases} 0 & x \neq k \\ 1 & x = k \end{cases} \tag{3.57}$$

Given L observations $\{x_{l=1}^{(l)}\}^L$, the multinomial distribution (73, 74, 81) gives the total probability of all possible length L discrete sequences taking those values:

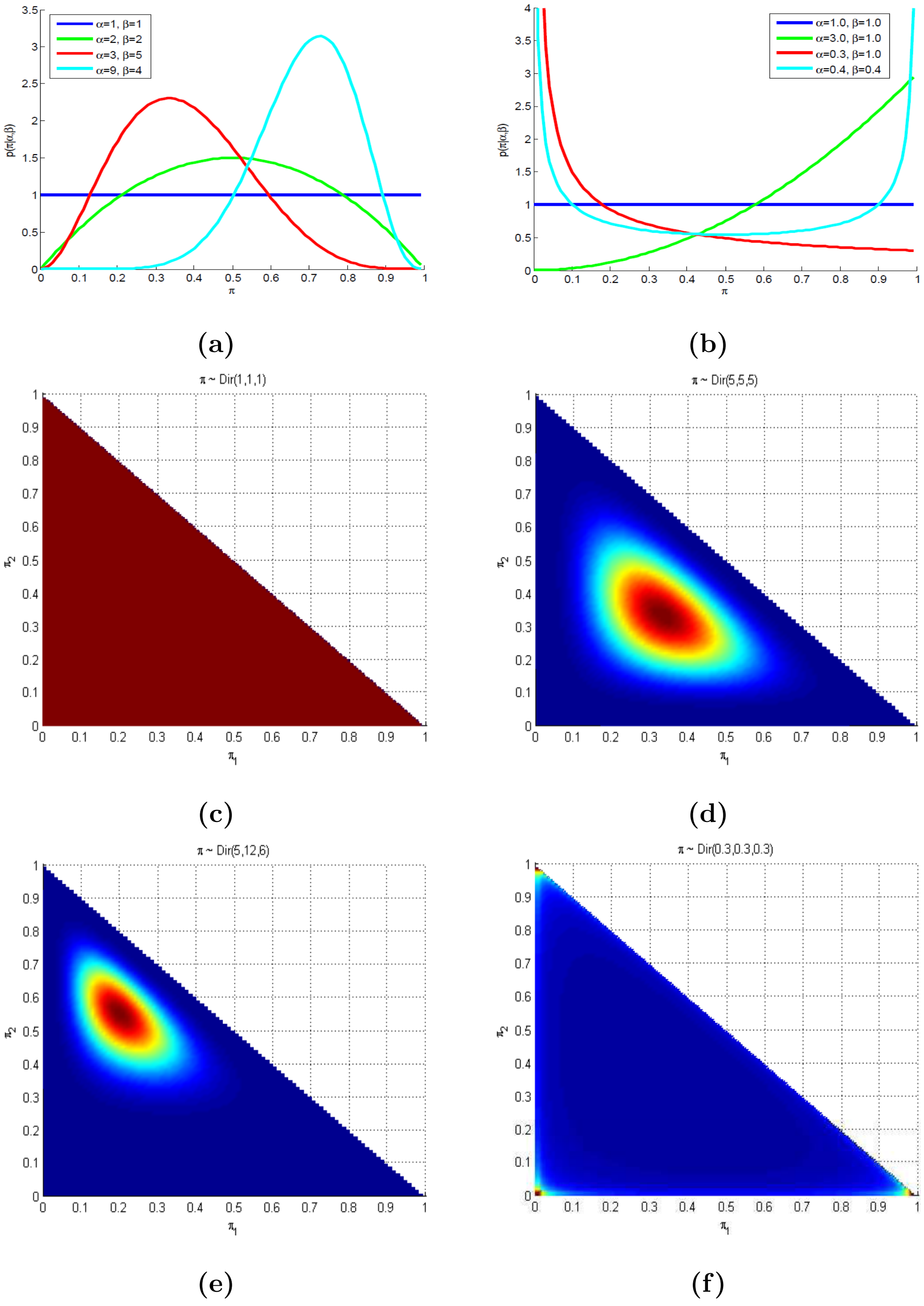
$$p(x^{(1)}, ..., x^{(L)}|\pi_1, ..., \pi_K) = \frac{L!}{\prod_k C_k!} \prod_{k=1}^{K} \pi_k^{C_k} \qquad C_k \triangleq \sum_{l=1}^{L} \delta(x^{(l)}, k) \tag{3.58}$$

When $K = 2$, this is known as the binomial distribution. Through comparison with Eq. 3.11, multinomial distributions define regular exponential families with sufficient statistics $\phi_k(x) = \delta(x, k)$ and canonical parameters $\theta_k = \log \pi_k$. In a minimal representation, only the first $(K-1)$ statistics are necessary. The multinomial distribution is valid when its parameters lie in the $(K-1)$-simplex:

$$\Pi_{K-1} \triangleq \{(\pi_1, ..., \pi_K)|\pi_k \geq 0, \sum_{k=1}^{K} \pi_k = 1\} \tag{3.59}$$

$$= \{(\pi_1, ..., \pi_{K-1}, 1 - \sum_{k=1}^{K-1} \pi_k)|\pi_k \geq 0, \sum_{k=1}^{K-1} \pi_k \leq 1\} \tag{3.60}$$

In particular, Π_{K-1} is the set defining the simplex. Note that this implicitly defines π_K as the complement of the probabilities of the other $(K-1)$ categories.



(a) **(b)**

(c) **(d)**

(e) **(f)**

Figure 3.1: Beta and Dirichlet distributions. (a)-(b) Beta densities with large parameters are unimodal while with small values favor biased binomial distributions. (c)-(f) Dirichlet densities on $\Pi_2 = (\pi_1, \pi_2, 1 - \pi_1 - \pi_2)$. A uniform prior, an unbiased and a biased unimodal priors, and a prior favoring sparse multinomial distributions.

Given L observations as in Eq. 3.58, Prop. 3.3.3 shows that the maximum likelihood estimates of the multinomial parameters $\pi = (\pi_1, ...\pi_K)$ equal the empirical frequencies of the discrete categories:

$$\hat{\pi} = \arg\max_{\pi} \sum_{l=1}^{L} \log\, p(x^{(l)}|\pi) = (\frac{C_1}{L}, ..., \frac{C_K}{L}) \tag{3.61}$$

When L is not much larger that K, the ML estimate may assign zero probability to some values and produce misleading predictions. Family of conjugate priors is useful in these situations.

The Dirichlet distribution (73, 74) is the conjugate prior for the multinomial exponential family. Adapting the general form of Eq. 3.51, the Dirichlet distribution with hyper-parameters $\alpha = (\alpha_1, ..., \alpha_K)$ can be written as follows:

$$p(\pi|\alpha) = \frac{\Gamma(\Sigma_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \qquad \alpha_k > 0 \tag{3.62}$$

Note that the Dirichlet distribution's normalization constant involves a ratio of gamma functions. By convention, the exponents are defined to equal $(\alpha_k = 1)$ so that the density's mean has the following simple form:

$$\mathbb{E}_{\alpha}[\pi_k] = \frac{\alpha_k}{\alpha_0} \qquad \alpha_0 \triangleq \sum_{k=1}^{K} \alpha_k \tag{3.63}$$

$Dir(\alpha)$ is used to denote a Dirichlet density with hyper-parameters α. The hyper-parameters α controls the density mean, shape, and sparsity of π. Samples can be drawn from a Dirichlet distribution by normalizing a set of K independent gamma random variables.

There is sometimes no prior knowledge distinguishing the categories, and the K hyper-parameters are set symmetrically as $\alpha_k = \frac{\alpha_0}{K}$. The variance of the multinomial parameters then equals

$$Var_{\alpha}[\pi_k] = \frac{K-1}{K^2(\alpha_0 + 1)} \qquad \alpha_k = \frac{\alpha_0}{K} \tag{3.64}$$

Because the variance is inversely proportional to α_0, it is known as the precision parameter.

When $K = 2$, the Dirichlet distribution is equivalent to the beta distribution (74). Denoting the beta density's two hyper-parmeters by α and β, let $\pi \sim Beta(\alpha, \beta)$ indicate that

$$p(\pi|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1} \qquad \alpha, \beta > 0 \tag{3.65}$$

By convention, samples from the beta density are the probability $\pi \in [0, 1]$ of the first category, while the two-dimensional Dirichlet distribution is equivalently expressed in terms of the probability vector $(\pi, 1-\pi)$. As in Eq. 3.62 and Eq. 3.63, the beta density's hyper-parameters can be interpreted as setting the prior mean and variance of the binomial parameter π.

In Figure. 3.1, several beta distributions are illustrated. When $\alpha = \beta = 1$, it assigns equal prior probability to all possible binomial parameters π. Larger hyper-parameters which correspond to smaller variances lead to unimodal priors concentrated on the chosen mean. To extend beta distribution into $K = 3$ multinomial categories, we also demonstrate examples of Dirichlet distributions, using the minimal 2-simplex representation of Eq. 3.59. As with the beta density, setting $\alpha_k = 1(\alpha_0 = K)$ defines a uniform prior on the simplex, while larger precisions lead to unimodal priors. Smaller values of the hyper-parameters ($\alpha_k < 1$) favor sparse multinomial distributions which assign most of their probability mass to a subset of the categories.

Consider a set of L observations $\{x^{(l)}\}_{l=1}^{L}$ from a multinomial distribution $p(x|\pi)$ with Dirichlet prior $p(\pi|\alpha)$. The posterior distribution is also Dirichlet via conjugacy:

$$\begin{aligned} p(\pi|x^{(1)},...,x^{(L)},\alpha) &\propto p(\pi|\alpha)p(x^{(1)},...,x^{(L)}|\pi) \qquad (3.66)\\ &\propto \prod_{k=1}^{K} \pi_k^{\alpha_k+C_k-1} \propto Dir(\alpha_1+C_1,...,\alpha_K+C_K) \qquad (3.67)\end{aligned}$$

C_k is the number of observations of category k as in Eq. 3.58. If L is sufficiently large, the mean of this posterior distribution provides a useful summary statistic as in Eq. 3.63. α_k is equivalent to a number of pseudo-observations of category k, and the precision α_0 is the total size of the pseudo-dataset.

The predictive likelihood of future observation $\bar{x}$ is calculated using the Dirichlet normalization constant of Eq. 3.62:

$$p(\bar{x}=k|x^{(1)},...,x^{(L)},\alpha) = \frac{C_k+\alpha_k}{L+\alpha_0} \tag{3.68}$$

C_k is the number of times category k was observed in the previous L observations. These observation counts provide easily updated sufficient statistics which allow rapid predictive likelihood evaluation. Comparing this prediction to that of Eq. 3.61, the raw frequencies underlying the ML estimate have been smoothed by the pseudo-counts contributed by the Dirichlet prior.

3.7 Gaussian observations

Consider a continuous-valued random variable x taking values in d-dimensional Euclidean space $\mathcal{X}=\mathbb{R}^d$. A Gaussian or normal distribution (73, 74, 81) with mean μ and covariance matrix Λ then has the following form:

$$p(x|\mu,\Lambda) = \frac{1}{(2\pi)^{d/2}|\Lambda|^{1/2}}\exp\{-\frac{1}{2}(x-\mu)^T\Lambda^{-1}(x-\mu)\} \tag{3.69}$$

This distribution denoted by $\mathcal{N}(\mu, \Lambda)$ is normalizable if and only if Λ is positive definite. Given L independent Gaussian observations $\{x^{(l)}\}_{l=1}^{L}$, the joint likelihood is

$$p(x^{(1)}, ..., x^{(L)}|\mu, \Lambda) \propto |\Lambda|^{-L/2} \exp\{-\frac{1}{2}\sum_{l=1}^{L}(x^{(l)} - \mu)^T \Lambda^{-1}(x^{(l)} - \mu)\} \tag{3.70}$$

Gaussian densities define a regular exponential family with canonical parameters proportional to the Gaussian's information parameterization $(\Lambda^{-1}, \Lambda^{-1}\mu)$. The maximum likelihood estimates of the Gaussian's parameters based on the dataset are the sample mean and covariance:

$$\hat{\mu} = \frac{1}{L}\sum_{l=1}^{L} x^{(l)} \qquad \hat{\Lambda} = \frac{1}{L}\sum_{l=1}^{L}(x^{(l)} - \hat{\mu})(x^{(l)} - \hat{\mu})^T \tag{3.71}$$

The sample mean and covariance provide sufficient statistics.

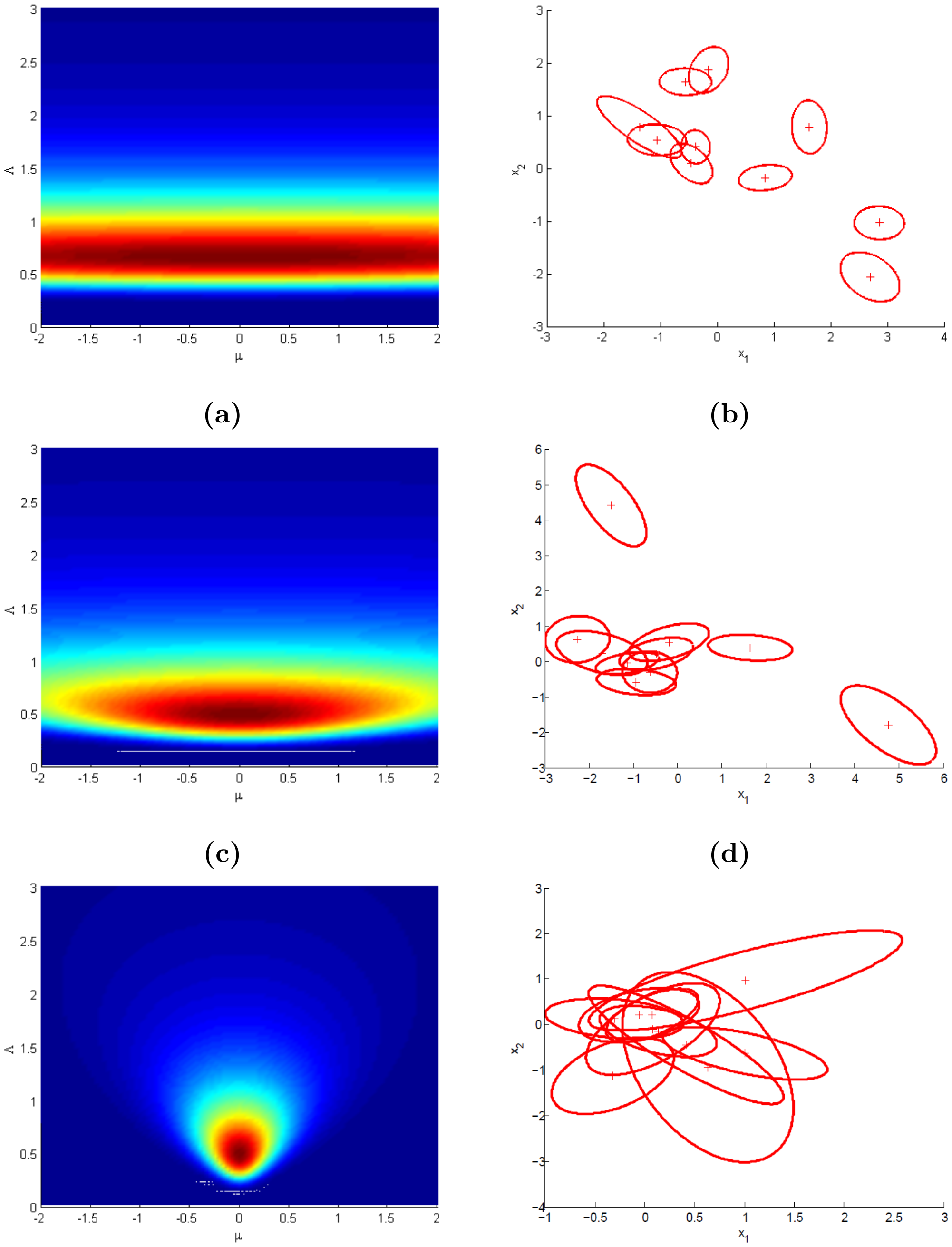
Any distribution satisfying certain spherical symmetries has a representation as a continuous mixture of Gaussian densities for some prior on that Gaussian's covariance matrix. The conjugate prior for the covariance matrix of a Gaussian distribution with known mean is the inverse-Wishart distribution which is a multivariate generalization of the scaled inverse-χ^2 density (74). The d-dimensional inverse-Wishart density with covariance parameter Δ and ν degrees of freedom equals

$$p(\Lambda|\nu, \Delta) \propto |\Lambda|^{-(\frac{\nu+d+1}{2})} \exp\{-\frac{1}{2}tr(\nu\Delta\Lambda^{-1})\} \tag{3.72}$$

This density is denoted by $\mathcal{W}(\nu, \Delta)$. An inverse-Wishart prior is proper when $\nu > d$ and skewed towards larger covariances (74). Its mean and mode equal

$$\mathbb{E}_{\nu}[\Lambda] = \frac{\nu}{\nu - d - 1}\Delta \qquad \nu > d + 1 \tag{3.73}$$

$$\arg\max_{\Lambda} \mathcal{W}(\Lambda; \nu, \Delta) = \frac{\nu}{\nu + d + 1}\Delta \tag{3.74}$$



(a) (b)

(c) (d)

(e) (f)

Figure 3.2: Normal-inverse-Wishart distributions. (a) Joint probability density of a scalar normal-inverse-χ^2 distribution $\mathcal{NW}(0.1, 0, 8, 1)$. (b) Covariance ellipses corresponding to ten samples from a 2D normal-inverse-Wishart distribution $\mathcal{NW}(0.1, 0, 8, I_2)$. (c)-(d) Density and samples of $\mathcal{NW}(0.3, 0, 4, 1)$. (e)-(f) Density and samples of $\mathcal{NW}(2, 0, 4, 1)$.

The degrees of freedom ν, which can be viewed as a precision parameter, is interpreted as the size of a pseudo-dataset with sample covariance Δ.

If a multivariate Gaussian's mean and covariance are both unknown, the normal-inverse-Wishard distribution provides an appropriate conjugate prior. The covariance matrix is assigned an inverse-Wishard prior $\Lambda \sim \mathcal{W}(\nu, \Delta)$ following Eq. 3.72. Conditioned on Λ, the mean $\mu \sim \mathcal{N}(\vartheta, \Lambda/\kappa)$. ϑ is the expected mean, for which there are κ pseudo-observations on the scale of observations $x \sim \mathcal{N}(\mu, \Lambda)$. The joint prior distribution denoted by $\mathcal{NW}(\kappa, \vartheta, \nu, \Lambda)$ takes the following form:

$$p(\mu, \Lambda|\kappa, \vartheta, \nu, \Delta) \propto |\Lambda|^{-(\frac{\nu+d}{2}+1)} \exp\{-\frac{1}{2}tr(\nu\Delta\Lambda^{-1}) - \frac{\kappa}{2}(\mu-\vartheta)^T\Lambda^{-1}(\mu-\vartheta)\} \tag{3.75}$$

Fig. 3.2 illustrates three sets of normal-inverse-χ^2 density when $d = 1$. Note that the mean and variance are dependent, so there is greater uncertainty in the mean value for larger underlying variances. This scaling is often, but not always, appropriate, and is necessary if conjugacy is desired. Fig. 3.2 also shows the three sets of ten Gaussian distributions drawn from the corresponding priors. As was shown in these figures, normal-inverse-Wishard distributions are a lot like Gaussian distributions, except that whereas the Gaussian distributions go from $-\infty$ to ∞, normal-inverse-Wishard distributions go from 0 to ∞.

Consider a set of L observations $\{x^{(l)}\}_{l=1}^{L}$ from a multivariate Gaussian distribution $\mathcal{N}(\mu, \Lambda)$ with normal-inverse-Wishart prior $\mathcal{NW}(\kappa, \vartheta, \nu, \Delta)$. The posterior distribution $p(\mu, \Lambda|x^{(1)}, ..., x^{(l)}, \kappa, \vartheta, \nu, \Delta)$ is also normal-inverse-Wishart and compactly described by a set of updated hyper-parameters $\mathcal{NW}(\bar{\kappa}, \bar{\vartheta}, \bar{\nu}, \bar{\Delta})$ via conjugacy. Through multiplication of Eq. 3.70 and Eq. 3.75 and manipulation of the quadratic form in Eq. 3.75, these posterior hyper-parameters equal:

$$\bar{\kappa}\bar{\vartheta} = \kappa\vartheta + \sum_{l=1}^{L} x^{(l)} \tag{3.76}$$

$$\bar{\kappa} = \kappa + L \tag{3.77}$$

$$\bar{\nu}\bar{\Delta} = \nu\Delta + \sum_{l=1}^{L} x^{(l)} x^{(l)^T} + \kappa\vartheta\vartheta^T - \bar{\kappa}\bar{\vartheta}\bar{\vartheta}^T \tag{3.78}$$

$$\bar{\nu} = \nu + L \tag{3.79}$$

$$\tag{3.80}$$

By caching the observations' sum and using Cholesky decompositions (82, 83)to calculate the sum of observation outer products, we can efficiently compute these posterior parameters.

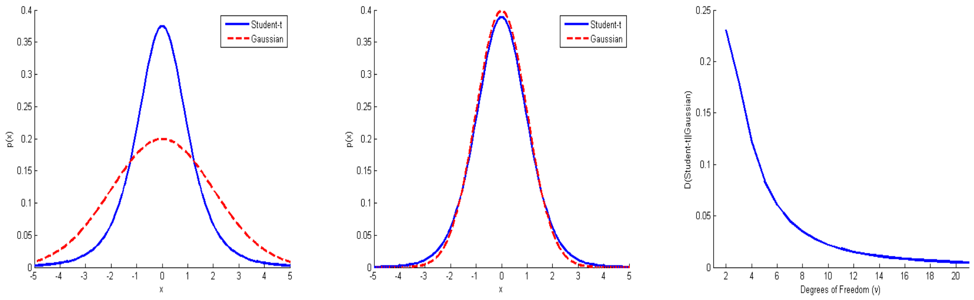
The predictive likelihood of a new observation $\bar{x}$ is multivariate Student-t with $(\bar{\nu} - d + 1)$ degrees of freedom by integrating over the parameters of the normal-inverse-Wishart posterior distribution. Assuming $\bar{\nu} > (d+1)$, this posterior density has finite covariance, and can be approximated by a moment-matched Gaussian:

$$p(\bar{x}|x^{(1)}, ...x^{(L)}, \kappa, \vartheta, \nu, \Delta) \approx \mathcal{N}(\bar{x}; \bar{\vartheta}, \frac{(\bar{\kappa}+1)\bar{\nu}}{\bar{\kappa}(\bar{\nu} - d - 1)}\bar{\Delta}) \tag{3.81}$$

The predictive likelihood depends on regularized estimates of the mean and covariance of previous observations. As illustrated in Fig. 3.3, although Student-t distributions have heavier tails than Gaussian, the KL divergence plot in (c) shows that the Gaussian approximation is accurate unless $\bar{\nu}$ is very small.

3.8 Linear observations

The normal linear regression model is one in which the observations (responses) $y_i \in \mathbb{R}^d$ can be described as a linear combination of a set of known regressors



(a) **(b)** **(c)**

Figure 3.3: Approximation of Student-t predictive distributions by a Gaussian with moments matched as in Eq. 3.81. One-dimensional Gaussian and heavier-tailed Student-t densities are compared with $\nu = 4$ in (a) and $\nu = 10$ in (b) degrees of freedom. For moderate ν, the Gaussian approximation becomes very accurate as shown in (c) plotted KL divergence depending on ν.

$x_i \in \mathbb{R}^n$ with errors accounted for by additive Gaussian noise (84):

$$y_i = x_{i1}a_1 + \cdots + x_{in}a_n + e_i \qquad e_i \sim \mathcal{N}(0, \Lambda) \tag{3.82}$$

By Combining N response vectors into a matrix $Y = [y_1 ... y_N]$, the regressors into a matrix $X = [x_1 ... x_N]$, and the noise terms into $E = [e_1 ... e_N]$ we can compactly write:

$$Y = AX + E \tag{3.83}$$

where $A = [a_1 ... a_N]$ is the design matrix.

The conjugate prior on the set of design matrix A and the noise covariance Λ is the matrix normal-inverse-Wishart prior. This distribution places a conditionally matrix normal prior on A given Λ:

$$p(A|\Lambda, M, K) = \frac{|K|^{\frac{d}{2}}}{|2\pi\Lambda|} \exp\{-\frac{1}{2} tr((A - M)^T \Lambda^{-1} (A - M)K)\} \tag{3.84}$$

and an inverse-Wishart prior on Λ

$$\Lambda \sim \mathcal{W}(\nu, \Delta) \tag{3.85}$$

Consider a set of observations $D = \{X, Y\}$, the posterior distribution of $\{A, \Lambda\}$ can be decomposed as the product of posterior A as $\mathcal{MN}(A; S_{yx}S_{xx}^{-1}, \Lambda, S_{xx})$ with $S_{xx} = XX^T + K$, $S_{yx} = YX^T + MK$, and $S_{yy} = YY^T + MKM^T$ and the marginal posterior of Λ as $\mathcal{W}(\nu + N, \Delta + S_{y|x})$ where $S_{y|x} = S_{yy} - S_{yx}S_{xx}^{-1}S_{yx}^T$.

3.9 Probabilistic graphical models

Probabilistic graphical models (85, 86, 87, 88, 89) as a powerful, flexible framework are motivated from several perspectives. Large collections of random variables are involved in many practical applications. This makes direct application of the classic exponential families and their priors become typically infeasible. For example, a generic high-dimensional discrete model of the joint distribution of 100 binary variables has $2^{100} \approx 10^{30}$ parameters. Even if this density could be stored and manipulated, reliable parameter estimation would require an unrealistically massive computation and dataset. Probabilistic graphical models allow us to decompose multivariate, joint distributions into a set of local interactions among small subsets of variables. These local relationships produce conditional independencies which lead to efficient inference and learning algorithms. Furthermore, probabilistic graphical models are also a general class of probabilistic models that can be used to infer latent variables from impoverished data. Latent variables in cognitive science can refer to any deeper laying causes that are not directly observable to us, such as attentional states, knowledge representations, contents of memory, and brain states.

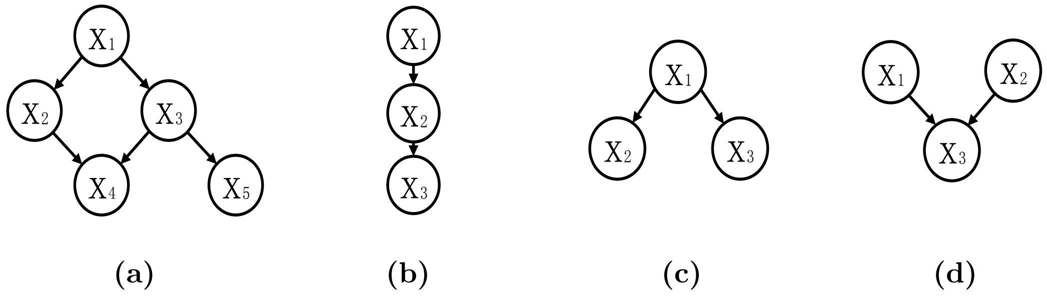


Figure 3.4: Four examples of Bayesian networks.

A graphical model allows us to specify certain local statistical dependencies between the random variables including both the latent variables and observed data as well as develop efficient inference and learning techniques such as belief propagation (87), and for advances in variational methods (77). Many classical models such as the hidden Markov model (HMM) (90) can be formulated within the graphical model framework. The inference and learning algorithms developed specifically for these models such as the forward-backward algorithm (90), Viterbi decoding (91), and Kalman filtering (92) can be derived as special cases of generic graphical model inference and learning algorithms.

There are different families of graphical models, including directed Bayesian networks, undirected Markov random fields, and factor graphs. We will focus on directed graphical models, in which the statistical dependency between random variables is based on directional relationships. These models are also known as Bayesian networks and belief networks.

3.9.1 Graph theory review

We briefly review some definitions from graph theory in order to describing graphical models subsequently. A graph $\mathfrak{G} = (\nu, \xi)$ consists of a set of nodes or vertices ν, and a corresponding set of edges ξ. Each edge $(i, j) \in \xi$ connects two distinct nodes $i, j \in \nu$. For directed graphs, and edge (i, j) connects a parent vertex i to its child j, and is pictorially represented by and arrow pointing from i to j (see Figure 3.4). The set of all parents $\Gamma(j)$ of node j is then given by

$$\Gamma(j) \triangleq \{i \in \nu | (i, j) \in \xi\} \tag{3.86}$$

Given a graph $\mathfrak{G} = (\nu, \xi)$, graphical models represent probability distributions by associating each node $i \in \nu$ with a random variable $x_i \in \mathfrak{X}_\mathrm{i}$. The structure of the joint distribution $p(x)$, where $x \triangleq \{x_i | i \in \nu\}$ takes values in the joint sample space $\mathfrak{X} = \mathfrak{X}_i \times \cdots \times \mathfrak{X}_N$, can be decomposed based on the corresponding edges.

3.9.2 Directed Bayesian networks

Bayesian networks associate each node $i \in \nu$ with a random variable x_i, and decompose $p(x)$ via the conditional density of each child node i given its parents $\Gamma(i)$:

$$p(x) = \prod_{i \in \nu} p(x_i | x_{\Gamma(i)}) \tag{3.87}$$

For nodes i without parents ($\Gamma(i) = \emptyset$), we define $p(x_i | x_{\Gamma(i)}) = p(x_i)$. This factorization is consistent whenever $\mathfrak{G}$ is a directed acyclic graph, so that its edges specify a valid partial ordering of the random variables (87, 93, 94). For example, the directed graph of Figure 3.4 (a) implies the following conditional densities:

$$p(x) = p(x_1)p(x_2|x1)p(x_3|x_1)p(x_4|x_2, x_3)p(x_5|x_3) \tag{3.88}$$

Bayesian networks effectively define a causal generative process, beginning with nodes without parents and proceeding from parent to child throughout the graph.

The Markov properties of directed Bayesian networks are that a random variable x_i is conditionally independent of the remaining process given its parents

$x_{\Gamma(i)}$, children $x_j | i \in \Gamma(i)$, and its children's parents. Exponential families usually provide convenient parameterizations of the conditional densities composing a Bayesian network. For small network, it is relatively easy to visually read off the independence relationships from a network. Besides, one can also use the joint distribution to derive all the conditional independence relations. For example, consider the graphical model in Figure 3.4. To see whether x_2 and x_3 are independent conditional on x_1 such that $p(x_2, x_3|x_1) = p(x_2|x_1)p(x_3|x_1)$, we can use the product rule to re-write the conditional probability $p(x_2, x_3|x_1)$ into its joint distribution form, and replace the joint distribution in the numerator by the factorization that the network implies:

$$\begin{aligned} p(x_2, x_3|x_1) &= \frac{p(x_2, x_3, x_1)}{p(x_1)} && (3.89) \\ &= \frac{p(x_2|x_1)p(x_3|x_1)p(x_1)}{p(x_1)} && (3.90) \\ &= p(x_2|x_1)p(x_3|x_1) && (3.91) \end{aligned}$$

Therefore, the conditional independence holds in this case.

3.9.3 Exchangeability via graphical models

Assuming the distribution Q has a parameterized density $q(\cdot|\lambda)$, the de Finetti theorem in Corollary 3.1.1 implies the following hierarchical Bayesian model:

$$p(x_1, \cdots, x_n, \theta|\lambda) = q(\theta|\lambda) \prod_{i=1}^{n} p(x_i|\theta) \qquad (3.92)$$

This equation has a directed graphical representation based on Equation 3.87, which is shown in Figure 3.5. This figure contains both an explicit representation of the graphical model, and an equivalent representation using plate notation to compactly represent the n observations x_i. It can be directly proved by using the Markov blanket concept that this set of random variables is yielded conditionally i.i.d. given θ from the graphical model.

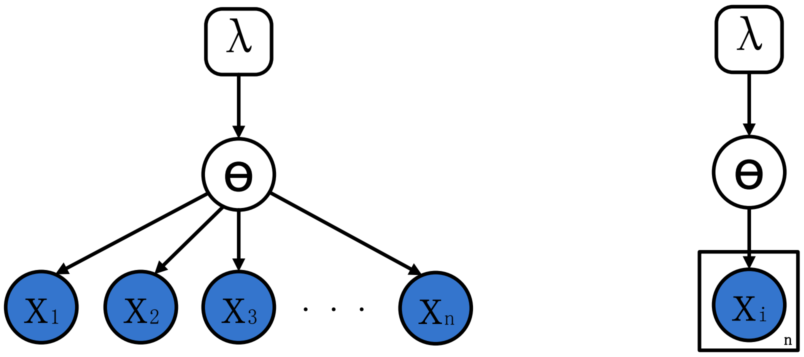


Figure 3.5: Graphical representation of the hierarchical Bayesian model of n exchangeable random variables implied by de Finetti's theorem. Each observation is an independent sample from a density parameterized by θ, which itself has a prior distribution with hyper-parameter λ. Left: An explicit representation of the graphical model. Right: A compact representation using a plate to denote n replicates of the observations x_i.

3.9.4 Hidden Markov models

Directed graphical models provide a unified theoretical framework for a family of hidden Markov models (HMMs) which are widely used to model temporal stochastic processes (90, 95, 96, 97). Let $y = \{y_t\}_{t=0}^{T-1}$ denote observations of a temporal process collected at T discrete time points. We assume that each observation y_t is independently sampled conditioned on an underlying hidden state x_t. If we further assume that these states $x = \{x_t\}_{t=0}^{T-1}$ evolve according to a first-order temporal Markov process, the joint distribution equals

$$p(x, y) = p(x_0)p(y_0|x_0)\prod_{t=1}^{T-1} p(x_t|x_{t-1})p(y_t|x_t) \tag{3.93}$$

Figure 3.7 shows a directed graphical representation of this density. In later chapters, we extend this model to develop methods for modeling eye movements.

Let π_j denote the state-specific transition distribution for state j. Then, the Markovian structure on the state sequence dictates that for all $t > 1$

$$x_t | x_{t-1} \sim \pi_{x_{t-1}} \tag{3.94}$$

The state at the first time step is distributed according to an initial transition distribution π^0:

$$x_1 \sim \pi^0 \tag{3.95}$$

Given the state x_t, the observation y_t is conditionally independent of the observations and states at other time steps. The observation is simply generated as

$$y_t | x_t \sim F(\theta_{x_t}) \tag{3.96}$$

for an indexed family of distributions $F(\cdot)$ where θ_i are the emission parameters for state i, assuming there exists a density associated with $F(\cdot)$.

Models equivalent to HMMs were independently developed and widely used in different domains, such as speech recognition and control theories (90, 97). All these disparate approaches can be unified via graphical models. Furthermore, graphical models provide possibilities for advances in inference and learning methods to be transferred between various domains (86, 93, 98).

3.9.5 Forward-backward algorithm

As a classical model, hidden Markov model (HMM) (90) has hand-tailored learning algorithms that can be described within the more general framework of inference on a graphical model. The forward-backward algorithm provides an efficient message-passing scheme for computing node marginals of interest for problems of filtering $p(x_n|y_1, ..., y_n)$, prediction $p(x_{n+m}|y_1, ..., y_n)$, and smoothing

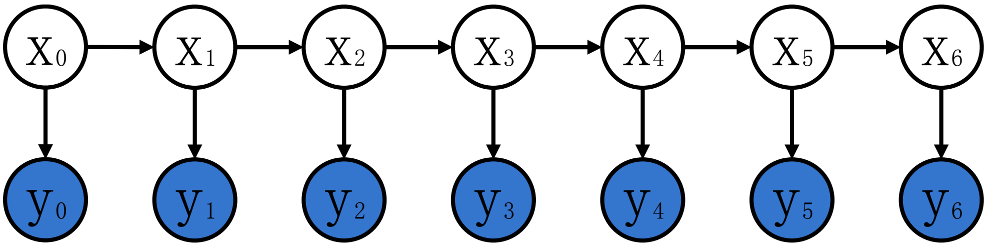


Figure 3.6: Directed graphical model of a hidden Markov model (HMM) for $T = 7$ samples of a temporal process. The hidden states x_t capture dependencies among the observations y_t.

$p(x_n|y_1, ..., y_N),\ N > n$. This classical algorithm has straightforward connections with the belief propagation algorithm. We define a set of forward messages:

$$\alpha_n(x_n) \triangleq p(y_1, ..., y_n, x_n) \tag{3.97}$$

and backward messages:

$$\beta_n(x_n) \triangleq p(y_{n+1}, ..., y_N|x_n) \tag{3.98}$$

For the problem of filtering:

$$p(x_n|y_1, ..., y_n) = \frac{p(x_n, y_1, ..., y_n)}{p(y_1, ..., y_n)} \tag{3.99}$$

$$= \frac{\alpha_n(x_n)}{\sum_x \alpha_n(x_n)} \tag{3.100}$$

For the problem of prediction:

$$p(x_{n+m}|y_1, ..., y_n) = \frac{p(x_{n+m}, y_1, ..., y_n)}{p(y_1, ..., y_n)} \tag{3.101}$$

$$= \frac{\sum_{x_{n+m-1}} p(x_{n+m}|x_{n+m-1}) \cdots \sum_{x_n} p(x_{n+1}|x_n)\alpha_n(x_n)}{\sum_x \alpha_n(x_n)} \tag{3.102}$$

which is equivalent to propagating the forward message without incorporating the missing observation: $y_{n+1}, ..., y_{n+m}$. The problem of smoothing:

$$\begin{aligned} p(x_n|y_1, ..., y_N) &= \frac{p(y_1, ..., y_N|x_n)p(x_n)}{p(y_1, ..., y_N)} && (3.103)\\ &= \frac{\alpha_n(x_n)\beta_n(x_n)}{\sum_x \alpha_m(x)\beta_m(x)} \qquad for\,\forall\, m && (3.104)\end{aligned}$$

Based on the conditional independencies implied by the graph of Figure 3.7, we can derive the recursions for these forward and backward messages, which are utilized by the inference algorithms in the thesis. For the forward message,

$$\begin{aligned} \alpha_{n+1}(x_{n+1}) &= p(y_{n+1}|x_{n+1})p(x_{n+1}, y_1, ..., y_n) && (3.105)\\ &= p(y_{n+1}|x_{n+1}) \sum_{x_n} p(x_{n+1}|x_n)\alpha_n(x_n) && (3.106)\end{aligned}$$

The backward recursion is derived as

$$\begin{aligned} \beta_n(x_n) &= \sum_{x_{n+1}} p(y_{n+1}, ..., y_N, x_{n+1}|x_n) && (3.107)\\ &= \sum_{x_{n+1}} p(y_{n+1}|x_{n+1})p(x_{n+1}|x_n)\beta_{n+1}(x_{n+1}) && (3.108)\end{aligned}$$

The forward initial condition and the backward final condition are given by:

$$\begin{aligned} \alpha_1(x_1) &= p(y_1, x_1) = p(y_1|x_1)\pi^0(x_1) && (3.109)\\ \beta_N(x_N) &= 1 && (3.110)\end{aligned}$$

The forward-backward algorithm can be derived as a special case of the belief propagation by converting the directed graph to its undirected form.

3.9.6 Viterbi Algorithm

Given a set of HMM parameters, the Viterbi algorithm (91) provides an efficient dynamic programming approach to computing the most likely state sequence to

Compute the MAP hidden Markov model state sequence $\hat{x}_1, ..., \hat{x}_N$ as follows:

1. Initialize minimum path sum to state $x_1 = k$ for each $k \in \{1, ..., K\}$:

$$\mathcal{S}_1(x_1 = k) = -\log\, \pi^0(x_1 = k) - \log\, p(y_1|x_1 = k)$$

2. For $n = 2, ..., N$ and for each $k \in \{1, ..., K\}$, calculate the minimum path sum to state $x_n = k$:

$$\mathcal{S}_n(x_n = k) = -\log\, p(y_n|x_n = k) + \min_{x_{n-1}}\{\mathcal{S}_{n-1}(x_{n-1} - \log\, p(x_n = k|x_{n-1}))\}$$

and let

$$x^*_{n-1}(x_n) = \arg\min_{x_{n-1}}\{\mathcal{S}_{n-1}(x_{n-1}) - \log\, p(x_n - k|x_{n-1})\}$$

3. Compute

$$\min_{x_1,...,x_N} -\log\, p(x_1, ..., x_N|y_1, ..., y_N) = \min_{x_N} \mathcal{S}_N(x_N)$$

and set

$$\hat{x}_N = \arg\min_{x_N} \mathcal{S}_N(x_N)$$

4. Iteratively set, for $n \in \{N - 1, ..., 1\}$

$$\hat{x}_n = x^*_n(\hat{x}_{n+1})$$

Table 3.1: Algorithm 1. Viterbi hidden Markov model decoding.

have generated an observation sequence $y_1, ..., y_N$:

$$\hat{x} = \max_{x_1,...,x_N} \pi^0(x_1)p(y_1|x_1)\prod_{n=2}^{N} p(x_n|x_{n-1})p(y_n|x_n) \tag{3.111}$$

$$= \min_{x_1,...,x_N} [-\log \pi^0(x_1) - \log p(y_1|x_1) + \sum_{n=2}^{N} - \log p(x_n|x_{n-1}) - \log p(y_n|x_n)] \tag{3.112}$$

The Viterbi algorithm works on the dynamic programming principle that the minimum cost path to $x_n = k$ is equivalent to the minimum cost path to node x_{n-1} plus the cost of a transition from x_{n-1} to $x_n = k$, which is incurred by observation y_n given $x_n = k$.

Viterbi decoding reduces the computation complexity to $O(K^2N)$ instead of the brute force $O(K^N)$. Algebraically, the Viterbi algorithm is closely related to the max-product (min-sum) algorithm that operates by distributing the maximization (minimization) operators over the elements of the product (sum) in Equation 3.111. It is worth to note that choosing the MAP sequence is not necessarily equivalent to choosing the maximum marginal independently at each node:

$$\hat{x} = \max\ p(x_n|y_1, ..., y_N) \tag{3.113}$$

Actually, such a maximum marginal sequence may not even be a feasible sequence for the HMM.

3.10 Inference and Learning Methods

In order to make efficient estimation on these complex models, we need approximated methods of learning and inference, instead of exact methods which become difficult to carry out by complete enumeration of all hypotheses and evaluation of their probabilities (99). Guidance comes from how processing the dynamics in the brain, such as belief propagation, expectation-maximization (EM), Markov chain Monte Carlo (MCMC), and sequential Monte Carlo (particle filtering). In particular, sequential Monte Carlo methods are used to simulate human reasoning process with great success.

Inference and learning of graphical models can be posed on the basis of some canonical computational tasks in many cases. We summarize the random variables composing the graphical models into four sets for the sake of discussion: observations y, latent variables x, parameter θ, and hyper-parameter λ.

3.10.1 Inference

Assuming the graph's parameters θ are fixed and known, the posterior distribution $p(x|y,\theta)$ fully captures available information about the hidden variables x. For most realistic graphs the joint sample space $\mathcal{X}$ is too large to characterize in exact methods. So approximate methods to infer statistics summarizing this posterior density is necessary.

The joint density $p(x|y,\theta)$ can be effectively summarized by the following posterior marginal distribution:

$$p(x_i|y,\theta) = \int_{\mathcal{X}_{\mathcal{V}\setminus i}} p(x|y,\theta)dx_{\mathcal{V}\setminus i} \qquad i \in \mathcal{V} \tag{3.114}$$

Here, $\mathcal{V}\setminus i$ denotes all nodes except x_i. The mean of this conditional density is the Bayes' least sequares estimate (81, 100). The mode of this conditional density is the maximizer of the posterior marginals (MPM) (101) by minimizing the expected number of mis-classified variables. The variance or entropy of this conditional density measure the posterior uncertainty in these estimates (87, 102, 103).

An alternative method is to infer hidden variables via a global MAP estimate:

$$\hat{x} = \arg\max_{x} p(x|y, \theta) \tag{3.115}$$

It shows that when observations are noisy or ambiguous, MAP estimation is often less robust than the MPM estimation (101).

3.10.2 Learning

Given observations y a straightforward parameter learning method is to determine a single MAP parameter estimate:

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} p(\theta|y, \lambda) & (3.116) \\ &= \arg\max_{\theta} p(\theta|\lambda) \int_{\mathcal{X}} p(x, y|\theta) dx & (3.117) \end{aligned}$$

The difficult part of this optimization is the marginalization over hidden variables x. Inference problem analogous to the posterior marginal computation of Eq. 3.114 is also required when learning with hidden variables.

When the parameters themselves are of interest, characterizations of their posterior uncertainty are useful. Given some decomposition $\theta = \{\theta_a | a \in \mathcal{A}\}$ of the joint parameter space, the posterior marginal distribution of these parameters, and the corresponding hidden variables, equal

$$p(\theta_a|y, \lambda) = \int_{\mathcal{X}} \int_{\Theta_{\mathcal{A}\backslash a}} p(x|y, \theta) p(\theta|y, \lambda) d\theta_{\mathcal{A}\backslash a} dx \qquad a \in \mathcal{A} \tag{3.118}$$

$$p(x_i|y, \lambda) = \int_{\Theta} \int_{\mathcal{X}_{\nu\backslash i}} p(x|y, \theta) p(\theta|y, \lambda) dx_{\nu\backslash i} d\theta \qquad i \in \mathcal{V} \tag{3.119}$$

Here, θ_a parameterizes an individual potential function in undirected graphs, or the conditional distribution of a single variable in a directed graphs. Integrating over all parameters and hidden variables, the observations' marginal likelihood can be recovered:

$$p(y|\lambda) = \int_{\mathcal{X}} \int_{\Theta} p(x, y|\theta) p(\theta|\lambda) d\theta dx \tag{3.120}$$

The marginal likelihood is used in Bayesian approaches to model selection, classification problems to determine the most likely explanation of the given observations, and empirical Bayesian estimate of the prior distribution.

Exact inference and learning for many graphical models arising in practice is computationally intractable. For example, given N variables of the posterior marginal computation, each taking one of K discrete states, this expression leads to a summation containing K^{N-1} terms, which for arbitrary graphs is NP hard (99). Optimization to compute the MAP is equally challenging (104). A high-dimensional integration of continuous $\mathcal{X}$ is usually also intractable. There are two other principle class of estimation methods providing approximate solutions to learning and inference tasks, which are variational methods (77, 102, 105) and Monte Carlo methods. We will focus on Monte Carlo methods which explore posterior distributions via efficient numerical simulations.

3.10.3 Monte Carlo Integration

The success of Monte Carlo methods is due to that the aim of inference is not always to find the most probable explanation for the observed human behaviors, which is essentially the peak of a probability distribution. While this most probable hypothesis may be of interest, and some inference methods do locate it, in cognition modeling it is the whole distribution that is of interest. The most probable outcome from a human is often not a typical outcome from that human. Similarly, the most probable hypothesis given some observed human behaviors may be atypical of the whole set of reasonably-plausible hypotheses.

Monte Carlo methods use random samples to simulate probabilistic models (74, 106, 107). Although they are guaranteed to yield arbitrarily precise estimates with sufficient computation, efficient algorithm design is necessary in practice in order to obtain reliable, accurate estimates at a tractable computational cost.

Let $p(x)$ denote some target density with sample space $\mathcal{X}$. The expected value $\mathbb{E}_p[f(x)]$ of an appropriately chosen function can be used to express various inference tasks such as the calculation of marginal densities and sufficient statistics. Suppose that $p(x)$ is difficult to analyze explicitly, but that we can draw L independent samples $\{x^{(l)}\}_{l=1}^{L}$ from it. The desired statistic can be approximated as follows (107):

$$\mathbb{E}_p[f(x)] = \int_{\mathcal{X}} f(x)p(x)dx \tag{3.121}$$

$$\approx \frac{1}{L}\sum_{l=1}^{L} f(x^{(l)}) = \mathbb{E}_{\tilde{p}}[f(x)] \tag{3.122}$$

Here, $\tilde{p}$ is the empirical density corresponding to the L samples. This estimate is unbiased, and converges to $\mathbb{E}_p[f(x)]$ almost surely as $L \to \infty$. Furthermore, its error is asymptotically Gaussian, with variance determined by $\mathbb{E}_p[f^2(x)]$ rather than the dimensionality of the sample space (107).

3.10.4 Kernel Density Estimation

In some cases of Monte Carlo methods, an explicit estimate $\hat{p}(x)$ is desired instead of a summary statistic as in Eq. 3.121. The advantage of non-parametric density estimators is that they allow the complexity of the estimated density to grow as more samples are observed. Given L independent samples $\{x^{(l)}\}_{l=1}^{L}$, the corresponding kernel density estimate can be written as follows:

$$\hat{p}(x) = \sum_{l=1}^{L} \omega^{(l)} N(x; x^{(l)}, \Lambda) \tag{3.123}$$

This estimator uses a Gaussian kernel function to smooth the raw sample set by placing more probability mass in regions with many samples. Although there are other kernels are possible instead of a Gaussian, we mainly use it in this thesis. If these samples are drawn from the target density $p(x)$, the weights are set uniformly to $\omega^{(l)} = 1/L$. The kernel density estimate of Eq. 3.123 depends on the bandwidth Λ of the Gaussian kernel function. There is extensive literature on methods for automatic bandwidth selection ranging from the simple "rule of thumb" method to more sophisticated cross-validation schemes (108).

3.10.5 The Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm constructs an ergodic Markov chain by defining a valid proposal distribution $q(\cdot|\cdot)$ and evaluation of the target distribution p up to a normalization constant. Since we are able to evaluate $p(x)$, the key is how to sample from this distribution. The proposal distribution must satisfy some weak conditions (107).

Given a previous sample $x^{(t-1)}$:

1. Sample $x' \sim q(x'|x^{(t-1)})$.

2. Determine the acceptance probability:

$$\rho(x'|x^{(t-1)}) = \min\{\frac{p(x')q(x^{(t-1)}|x')}{p(x^{(t-1)q(x'|x^{(t-1)}))})}, 1\}$$

3. Sample

$$x^{(t)} \sim \rho(x'|x^{(t-1)})\delta_{x'} + (1 - \rho(x'|x^{(t-1)}))\delta_{x^{(t-1)}},$$

where δ_x is a Dirac mass at x.

Table 3.2: Algorithm 2. Metropolis-Hastings algorithm.

As long as $p(x^{(0)}) > 0$ the chain defined in Algorithm 3.3 will have $p(x^{(t)}) > 0$ for all t. So, the acceptance probability $\rho(y|x)$ which is defined only when $p(x) > 0$ is valid. The detailed balance condition is also necessary for the Markov chain defined by the Metropolis-Hastings algorithm.

Proposition 3.10.1. *Let $T(x_a|x_b) = p(x_{n+1} = x_a|x_n = x_b)$ be the transition distribution for a given Markov chain. If $T(x_a|x_b)$ satisfies detailed balance:*

$$T(x_a|x_b)\pi_{x_b} = T(x_b|x_a)\pi_{x_a} \tag{3.124}$$

then the chain defined by this transition distribution has stationary distribution π. A Markov chain satisfying detailed balance is said to be reversible with respect to π.

It is straightforward to show that the transition distribution defined by Algorithm 3.3 satisfies detailed balance. According to the transition distribution, the sampling chain transitions from x_a to a sample $x_b \sim q(x_b|x_a)$ with probability $\rho(x_b|x_a)$ and stay at x_a otherwise. Thus the transition distribution is a weighted mixture of the proposal distribution and a Dirac mass at x_a:

$$T(x_b|x_a) = \rho(x_b|x_a)q(x_b|x_a) + (1 - \int \rho(z|x_a)q(z|x_a)dz)\delta_{x_a} \tag{3.125}$$

We can analyze each term of the transition distribution separately to check the detailed balance condition. Thus the chain generated by the Metropolis-Hastings algorithm also define a Markov chain with π a stationary distribution.

To show that the Markov chain converges to π which is the unique invariant distribution for this chain and this distribution is reachable from all initial states, we invoke some mild conditions under which the chain is both aperiodic and irreducible (74, 107). A sufficient condition for the Metropolis-Hastings Markov

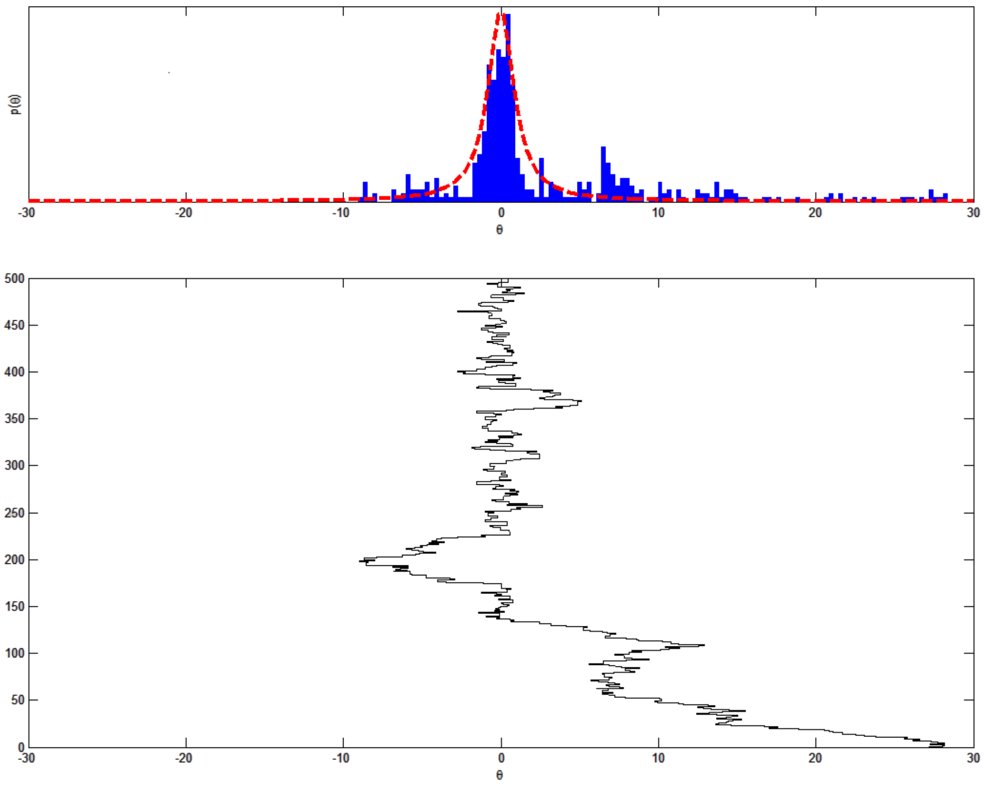


Figure 3.7: Demonstration of Metropolis-Hastings algorithm. 500 samples were drawn from the Cauchy distribution. The upper panel shows the theoretical density in the dashed red line and the histogram shows the distribution of the samples. The lower panel shows the sequence of samples of one chain.

chain to be aperiodic is for events $x^{(t)} = x^{(t-1)}$ to occur with some positive probability. That is

$$P[\pi(x^{(t-1)})q(y|x^{(t-1)}) \leq \pi(y)q(x^{(t-1)}|y)] < 1 \tag{3.126}$$

Furthermore, if

$$q(y|x) > 0 \qquad \forall(x, y) \in \mathcal{X} \times \mathcal{X} \tag{3.127}$$

then the Metropolis-Hastings Markov chain is irreducible. Thus, any Metropolis-

Hastings algorithm defined with a proposal distribution that satisfies the above conditions will eventually yield samples from the stationary distribution π and

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} f(x^{(t)}) = \int_{X} f(x)\pi(x)dx \tag{3.128}$$

The rate of convergence to the stationary distribution is extensively discussed in (74, 106). Generally, the burn-in period is challenging to quantify except by conservative bounds. Convergence can be affected by the initialization of the Markov chain to a great extent, so it is common to run multiple chains from different initializations in practice. Multi-modal target distributions with low valleys between the modes can be problematic by causing poorly mixing chains to stay in one region of the state space for long periods of time. It is also important to cleverly engineer proposal distribution.

3.10.6 Gibbs Sampling

Gibbs sampler, which is a special case of Metropolis-Hastings algorithm, is particularly well suited to the problems of which state spaces have internal topological structure in terms of probabilistic dependence among variables (101, 109, 110). Let $x = (x_1, ..., x_N)$ denote decomposition of the joint sample space into N variables. Gibbs sampler assume that it is tractable to sample from the conditional distribution of one of these variable given the other $(N-1)$. At iteration t, a particular variable $i(t)$ is selected for re-sampling and the rest are held constant:

$$x_i^{(t)} \sim p(x_i|x_j^{(t-1)}, j \neq i) \qquad i = i(t) \tag{3.129}$$

$$x_j^{(t)} = x_j^{(t-1)} \qquad j \neq i(t) \tag{3.130}$$

If these sampling updates are iterated so that all variables are re-sampled infinitely, mild conditions ensure $x^{(t)}$ will converge in distribution to a sample from $p(x)$ as $t \to \infty$ (107, 111).

Although there exist polynomial bounds on the time complexity to mix to the target equilibrium distribution (107, 111), it can be difficult to guarantee or diagnose convergence in high-dimensional models (106). Besides the practical techniques to improve the rate of convergence discussed in Metropolis-Hastings algorithm, one can consider blocked Gibbs samplers or randomly permuting the order in which variables are re-sampled (107, 111, 112).

```
1. Initialize $x_{0,1:n}$
2. for $i = 0$ to $N - 1$:
    - Sample   $x_1^{(i+1)} \sim p(x_1|x_2^{(i)}, x_3^{(i)}, ..., x_n^{(i)})$
    - Sample   $x_2^{(i+1)} \sim p(x_2|x_1^{(i+1)}, x_3^{(i)}, ..., x_n^{(i)})$
                                  $\vdots$
    - Sample   $x_j^{(i+1)} \sim p(x_j|x_1^{(i+1)}, ..., x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, ..., x_n^{(i)})$
                                  $\vdots$
    - Sample   $x_n^{(i+1)} \sim p(x_n|x_1^{(i+1)}, x_2^{(i+1)}, ..., x_{n-1}^{(i+1)})$
```

Table 3.3: Algorithm 3. Gibbs sampler.

The Gibbs sampler's use of partitioned state spaces is ideally suited for inference in graphical models (87, 101, 102, 109). We focus on using Gibbs sampling to estimate posterior distributions for directed graphical model parameters. Hidden variables are sampled given fixed parameters. Conditioned on these hidden variables, conjugate priors $p(\theta|\lambda)$ typically allow individual parameters to be tractably re-sampled (94, 110, 113). Statistics of the joint posterior $p(x, \theta|y, \lambda)$ can be estimated by alternatively sampling $x^{(t)} \sim p(x|\theta^{(t-1)}, y)$ and $\theta^{(t)} \sim p(\theta|x^{(t)}, y, \lambda)$.

4

Elicitation of Perceptual and Conceptual Tacit Knowledge

We designed and conducted an eye tracking experiment to investigate the conceptual and perceptual processing involved in subjects' medical image inspection (11). Preliminary statistical analysis of performance (in terms of time spent and diagnostic correctness), eye tracking, as well as verbal narrative data indicates significant difference between expertise-specific groups in those aspects.

4.1 Subjects

Subjects recruited for the eye tracking experiment belong to three groups based on their dermatology training level including eleven board-certified dermatologists (attending physicians), four dermatologists in training (residents) and thirteen undergraduate students who were lay people (novices). We also recruited physician assistant students who served as "trainees" in order to motivate der-

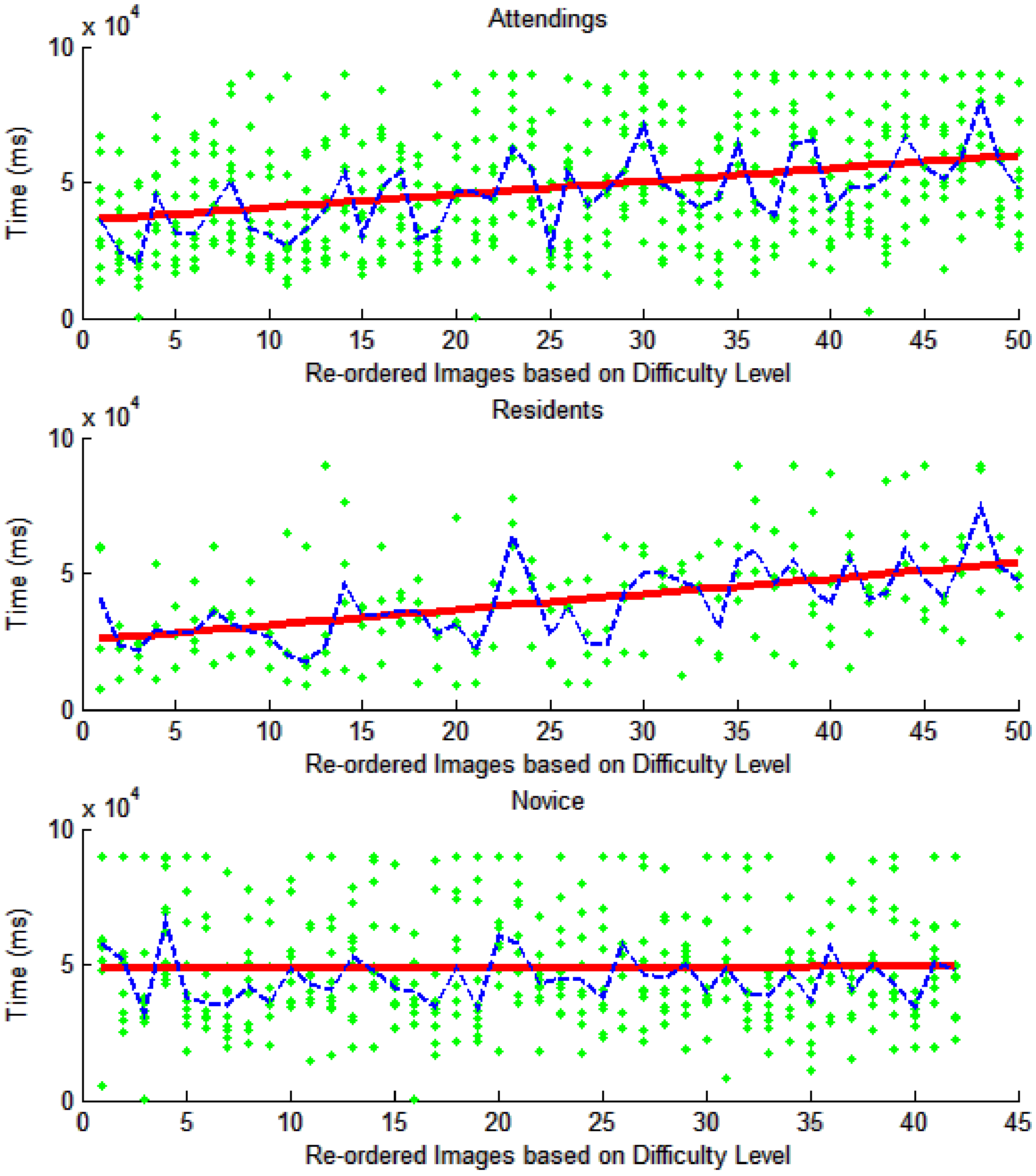


Figure 4.1: Scatterplots of time duration versus diagnostic difficulty level of each image. Each green point denotes the time duration of each subject on the corresponding image. The red line is the linear regression line in each expertise-specific dataset. The blue line represents the medians. The top panel shows 11 attendings' durations on 50 images. The middle panel shows 4 residents' durations on 50 images. The bottom panel shows 13 novices' duration on 43 images. Some jitter has been added on the vertical axis to reduce overlap and facilitate display.

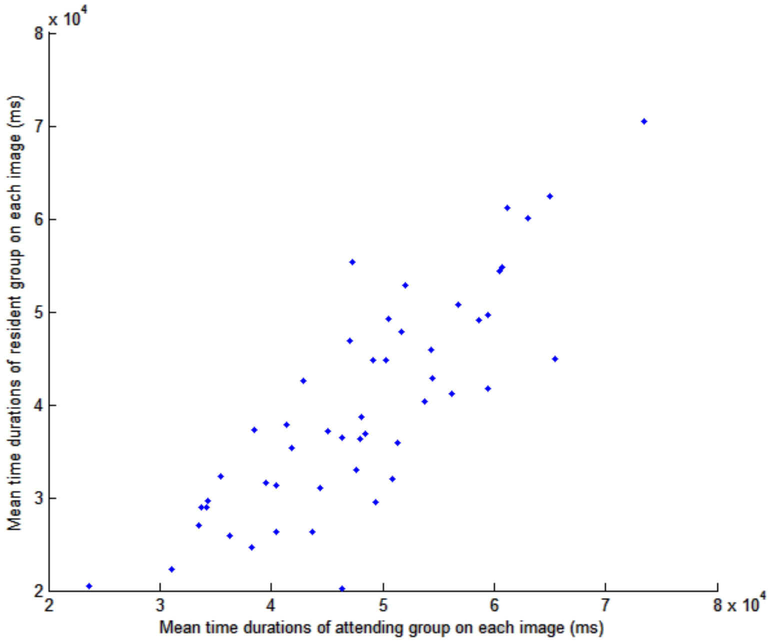


Figure 4.2: Scatterplot of mean time durations of attending group versus resident group on all 50 images.

matologists to verbalize their diagnostic reasoning using the Master-Apprentice scenario (114), which is known to be effective for eliciting tacit knowledge.

4.2 Apparatus

A SMI (Senso-Motoric Instruments) eye tracking apparatus was applied to display the stimuli at a resolution of 1680x1050 pixels for the collection of eye movement data and recording of verbal descriptions. The eye tracker was running at 50 Hz sampling rate and with reported accuracy of 0.5^o visual angle. The subjects viewed the medical images binocularly at a distance of about 60 cm. The experiment was conducted in an eye tracking laboratory with ambient light.

4.3 Materials and Procedure

A set of 50 dermatological images, each representing a different diagnosis, was selected for the study. These images were collected from the database of Logical Images Inc. and our collaborating author, Cara Calvilli MD. These images were presented to subjects on a monitor with 1280×1024 display. Medical professionals were instructed to examine and describe each image to the students while working towards diagnosis, as if teaching. Viewing time limit on each image is 90 sec. The professional groups (both attendings and residents) were instructed not only to view the medical images and work onto a diagnosis, but also to describe what they saw as well as their thought processes leading them to the diagnosis to the student sitting besides them as if they were in a training process, according to a modified Master-Apprentice scenario. The instruction is as follows:

Over the next hour we will show you 50 images representing a range of dermatological diagnoses. We would like you to examine each image and describe it to the trainee sitting next to you. When you have finished describing it, please offer your diagnosis or differential diagnosis of the projected disorder. Please describe the image as if you were teaching the trainee to make a diagnosis based on the image. It will be important to describe the image verbally rather than by pointing to areas of the image. Even if you reach a diagnosis quickly please describe the image characteristics that have allowed you to reach that diagnosis. We have asked the trainee to simply listen rather than interact or ask questions. We will be recording your eye movements and verbal descriptions as you examine the images. Each image will be shown for one and a half minutes and will advance automatically, although you can advance them using the space bar if you would like to work at a faster pace. Try to work at a pace that is natural to you and don't be concerned about whether the diagnosis is correct. We will be happy to *review any of the diagnoses with you at the end of the study if you would like.*

The novice group was instructed to describe the disease to a physician as detailed as possible in order to facilitate the diagnosis. Over the next hour, we will show you 42 medical images. The instruction for the novice group is as follows:

We like to understand how people look at bio medical images. We would like you to examine each image and describe it as if you are describing it over the phone to a dermatologist who cannot see the images but has to diagnose it. When you finish describing an image, please say 'Next' and we will proceed to the next image. Try to work at a pace that is natural to you, and try not to ask any questions during the experiment. We will be recording your eye movements and verbal descriptions as you examine the images. First we will need to do a calibration of your eye positions on the monitor. You will see a series of circles with a black dot in the center. Simply look at the center until the circle moves to a different position on the screen. We will repeat this calibration after each set of 6 images.

Both eye movements and verbal narratives were recorded for the viewing durations controlled by each subject. The experiment started with a 12-point calibration and the calibration was validated after every 10 images. As long as the deviations of the subjects' fixations from the target points are no more than 1 degree visual angle, we accept it. The audio recordings of the verbal narratives from the dermatologists were transcribed and annotated, as described below.

4.4 Performance Analysis

In order to compare and evaluate the performances of our three expertise-specific groups, we measured their time durations and diagnostic correctness scores on

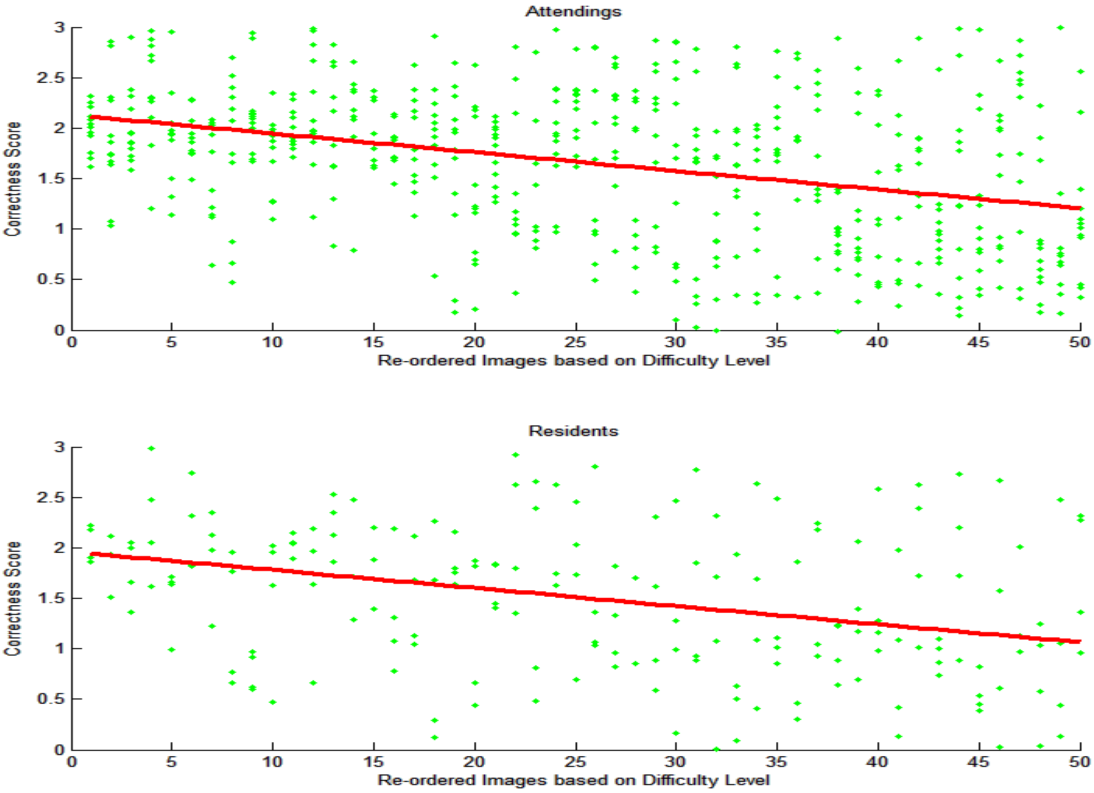


Figure 4.3: Scatterplots of diagnostic correctness scores versus diagnostic difficulty level of each image. The criterion includes identification of primary morphology, differential diagnosis, and final diagnosis. The scores range from 0 to 3, and 3 means all the three criterion are correct. Each green point denotes the diagnosis correctness of each subject on the corresponding image. The red line is the linear regression line in each expertise-specific dataset. The top panel shows 11 attendings' durations on 50 images. The bottom panel shows 4 residents' durations on 50 images. There is no diagnosis for novice group. Some jitter has been added on the vertical axis to facilitate display.

each image.

We first evaluate the correlations between the diagnostic difficulty level of each image and the subjects' time durations for each expertise-specific group separately. Figure 4.1 shows that there are significant positive correlations between the median time durations and the diagnostic difficulty levels for both attending group and resident group (for attending group Spearman's $\rho = 0.68, p < 0.001$, and for resident group $\rho = 0.73, p < 0.001$). This suggests that both attending and resident groups tend to spend longer time on more difficult cases and less time on easier cases, as one would predict.

However, the novice group's time durations are not correlated with the diagnostic difficulty levels. There are alternative possible explanation for this lack of correlation. The first is that the novice group lacks medical knowledge, so would be expected to be less discriminating in recognizing the diagnostic difficulty. The second is because they received different task instructions than the professional groups in our experimental design. While professionals are instructed to describe the image to a medical student during their working towards a diagnosis, novices are instructed to describe the image to a doctor to facilitate diagnosis.

The above discussion considers only the correlation between time durations and images for each expertise-specific group, rather than the time durations spent on the same image made by both attending group and resident group. In Figure 4.2, there is significant positive time duration correlation between attending group and resident group ($r(49) = 0.61, p < .001$). This finding confirms that both groups tend to spend longer time on some images and less time on others. A closer examination of Figure 4.2 reveals that attending group tends to spend a longer time than resident group on the same image. This suggests that attending group tends to be more thorough on examining images or more careful to render diagnosis hypotheses.

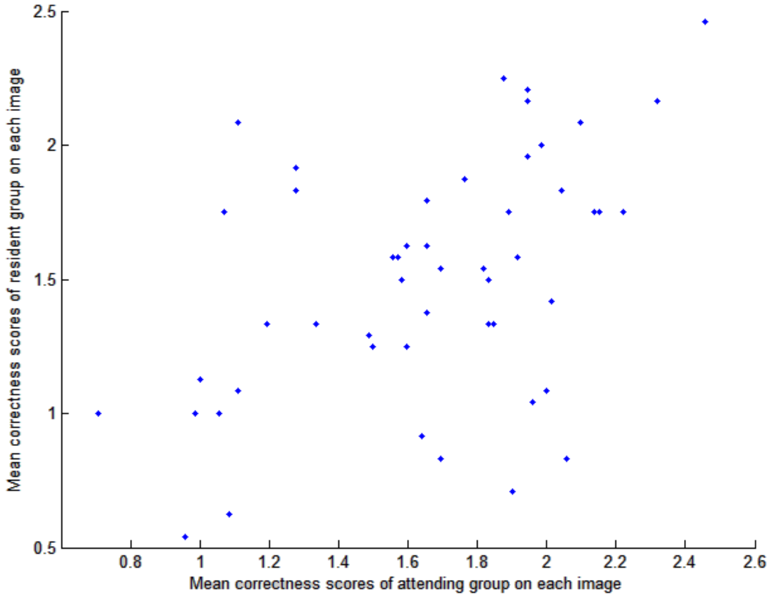


Figure 4.4: Scatterplot of diagnosis correctness scores of attending group versus resident group on all 50 images.

We recruited three additional dermatologists to evaluate the diagnostic correctness based on the transcribed verbal narratives. The criterion includes identification of primary morphology, differential diagnosis, and final diagnosis. The scores range from 0 to 3, and 3 means all the three criterion are correct. In Figure 4.4, the correctness scores decrease for both attending and resident groups (for attending group Spearman's $\rho = -0.53, p < 0.001$, and for resident group $\rho = -0.61, p < 0.001$). This suggests that both attending and resident groups tend to achieve higher scores on easier cases and lower scores on difficult cases. In particular, Figure 4.3 shows that the relationship of the mean correctness scores between attending and resident groups is positive, but not particularly strong as indicated by a correlation $r(49) = 0.32, p < .001$. Attending group tends to achieve better scores than resident group. This result indicates significant disagreement between the two groups on the correct diagnosis. From Figure 4.3, we

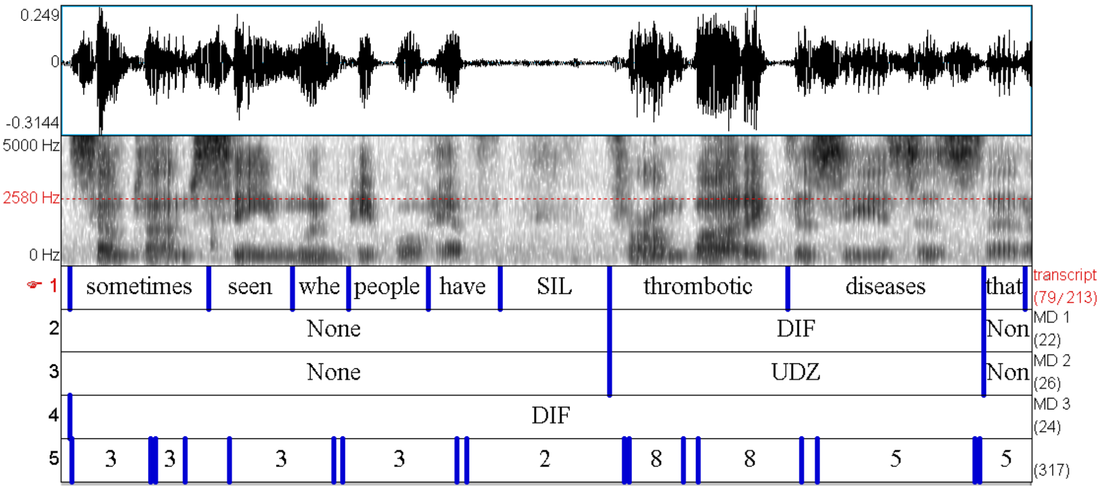


Figure 4.5: Time alignment of annotation and eye movement patterns. A time-aligned visualization of verbal descriptions, thought unit annotations, and eye movement patterns in Praat. A verbal transcript was split at the word level. Three annotations were provided by two different dermatologists, one of whom completed the annotation twice. The fifth tier shows eye movement pattern labels, which will be analyzed in the next chapter.

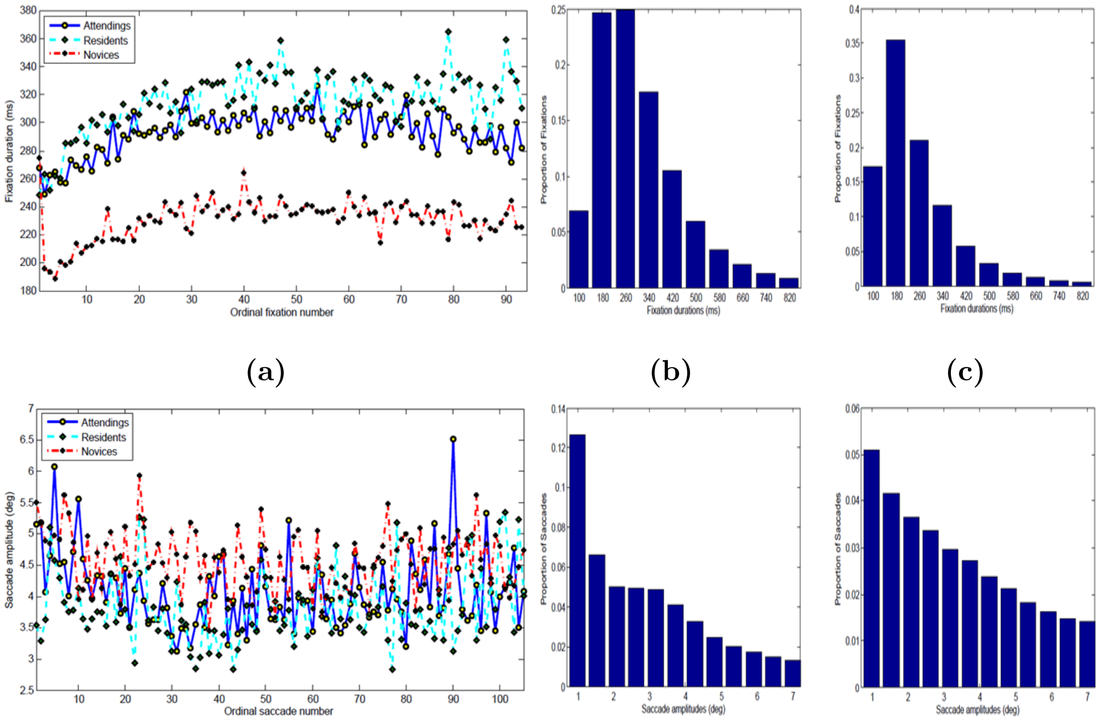
can also see that a partial explanation for the poor correlation may be restricted range with relatively few datapoints in the lower left quadrant. This is likely to reflect that both attending and resident groups achieve relatively high scores in general.

4.5 Eye Movement Data Analysis

Analysis of both fixation duration and saccade amplitude are conducted as a function of ordinal fixation number for the three expertise-specific groups to determine whether the two eye movement events, which are used as eye movement observation features, change over the time course of diagnosis and whether the differences as a function of expertise levels might be revealed at ordinal time points as shown in Figure 4.6. The first 20 fixations show a significant linear trend for all three groups (ANOVA: $F(19, 200) = 1.4$, $p < 0.01$; $F(19, 60) = 2.92$, $p < 0.01$ and $F(19, 240) = 1.98$, $p < 0.01$ respectively) and both attending group and resident group have significantly longer fixation durations than lay persons ($F(2, 273) = 12.5$, $p < 0.001$), which is also revealed through the histogram of fixation duration distribution as shown in Figure 4.6 b and c. Similar analysis on saccade amplitudes of the three expertise-specific groups shows that the first 20 saccade amplitudes of both dermatologists and residents follows a significant linear trend ($F(19, 200) = 1.24, p < 0.01$; and $F(19, 60) = 1.19$, $p < 0.01$). There was no effect for the novice group's average saccade amplitudes.

These shorter fixation durations and longer saccade amplitudes at the initial stage suggest both attending and residents started examining images with a quick image scan. After that, fixation durations became longer and saccade amplitudes decreased, suggesting a more thorough examination on some particular small regions. On the other hand, novice group's fixation durations increased at the initial stage but there is no statistically significant change for their saccades.

These descriptive statistical analysis indicate the difference between the professionals and the novice group, but they cannot tell us how the experts approach the task, not to mention the viewing strategies which experts bring into the cognitive processing. We thus apply our model on these time series data to reveal the subtlety of the behavior patterns varying over time in the next chapter.



(a) (b) (c)

(d) (e) (f)

Figure 4.6: (a) The average fixation durations by ordinal fixation number over the course of diagnosis for attending (blue), resident (cyan), and novice (red) groups. (b) The average saccade amplitudes by ordinal saccade number for the three groups. (c)-(d) The histograms of the fixation durations for professionals, and novices. (e)-(f) The histograms of the saccade amplitudes for professionals and novices.

4.6 Verbal Narrative Analysis

An annotation study was conducted on the experts' transcripts to investigate the verbalized cognitive processes of dermatologists on their paths toward a diagnosis (17). After transcribing the experts' narration of the images, independent experts identified conceptual units of thought (corresponding to particular steps or information in the diagnostic process) in the transcripts. These *thought units* were subsequently time-aligned with the recorded speech and eye movement patterns in the speech analysis tool Praat (115). Two highly trained dermatologists annotated transcribed verbal descriptions with these thought units. A *thought unit* is a single word or group of words that receives a descriptive label based on its semantic role in the diagnostic process. Nine basic thought units, provided by a dermatologist, were used for annotation. The provided thought unit labels are patient demographics (DEM), body location (LOC), configuration (CON), distribution (DIS), primary morphology (PRI), secondary morphology (SEC), differential diagnosis (DIF), final diagnosis (Dx), and recommendations (REC). Words not belonging to a thought unit were designated as 'None'. This time-alignment is illustrated in Figure 4.5

5

Hierarchical Dynamic Model

We briefly review non-homogenous Poisson process in order to derive completely random measures. Based on this framework, we can analyze the class of stochastic processes we will use extensively in our modeling work.

5.1 Poisson Processes

A stochastic process $\{X_t : t \in T\}$ is a collection of random variables (81). The variables X_t take values in the state space $\mathcal{X}$. The set T is called the index set, and can be discrete $T = 0, 1, 2, ...$ or continuous $T = [0, \infty)$.

The Poisson process describes counting occurrences of events over time, such as radioactive decay, telephone calls arriving at a switchboard, traffic accidents, page view requests to a website, etc. The Poisson process is constructed based on the Poisson distribution. It is written as $X \sim Poisson(\lambda)$ that X has a Poisson distribution with parameter λ, if

$$\mathbb{P}(X = x) = p(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x = 0, 1, 2, ... \tag{5.1}$$

In particular, $\mathbb{E}(X) = \lambda$ and $\mathbb{V}(X) = \lambda$. There are two useful properties of a Poisson distribution: if $X \sim Poisson(\lambda)$, $Y \sim Poisson(\nu)$, and $X \perp Y$, then $X + Y \sim Poisson(\lambda + \nu)$; if $N \sim Poisson(\lambda)$ and $Y|N = n \sim Binomial(n, p)$, then the marginal distribution of Y is $Y \sim Poisson(\lambda p)$.

The Poisson process with rate $\lambda > 0$ is a counting process $\{X_t : t \in [0, \infty)\}$ with state space $\mathcal{X} = \{0, 1, 2, ...\}$, where X_t is the number of events that occur in the time interval $[0, t]$, which fulfill the following conditions (We will write $X(t)$ instead of X_t):

- $X(0) = 0$
- For any $0 = t_0 < t_1 < t_2 < \cdots < t_n$, the increments $X(t_1) - X(t_0), X(t_2) - X(t_1), \cdots, X(t_n) - X(t_{n-1})$ are independent.
- The probability of the number of events that occur in a given interval depends only on the length of the interval and not on its location.
- $\mathbb{P}(X(t + h) - X(t) = 1) = \lambda h + o(h)$
- $\mathbb{P}(X(t + h) - X(t) \geq 2) = o(h)$

The second condition is the independent increment assumption which states that the number of events occurring in disjoint time intervals are independent. The third condition is the stationary increment assumption which states that the probability of a given interval $X(t + s) - X(t)$ is the same for all values of t. The last two conditions means that the probability of an event in $[t, t + h]$ is approximately $h\lambda(t)$ while the probability of more than one event is small.

Theorem 5.1.1. *Then number of events occurring in an interval of length t is a Poisson random variable with mean* λt

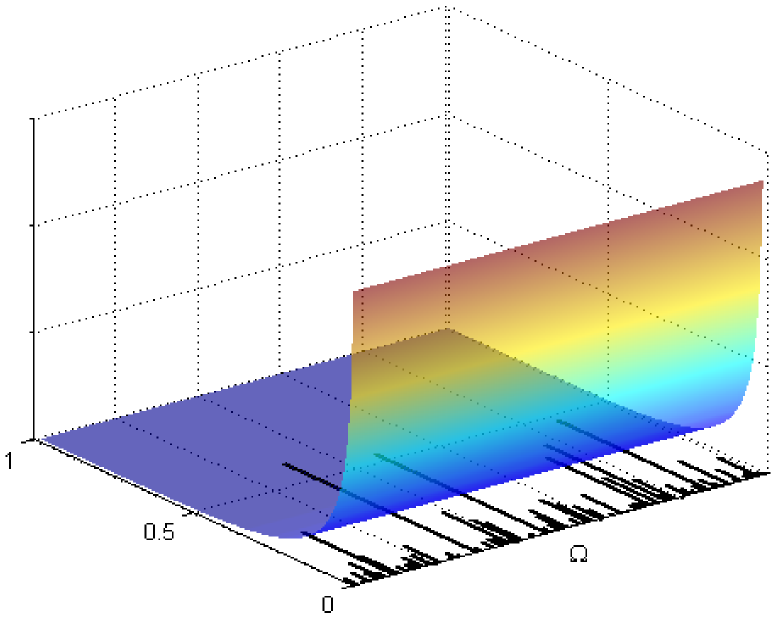


Figure 5.1: The construction of a completely random measure on Ω from a non-homogeneous Poisson process on $\Omega \times \Re$. In this example, Ω is a bounded interval. The colored surface illustrates the rate density for the Poisson process in Equation 5.3 which is the product of a uniform distribution (base measure) B_0 on Ω and an improper beta distribution on $(0, 1)$. Sampling the Poisson process gives rise to the non-zero endpoints in the plane, and these endpoints are connected by line segments to the Ω-axis interval to form the random measure $B = \sum_{i=1}^{\infty} p_i \delta_{\omega_i}$.

To prove this theorem, we first construct a representation based on the binomial distribution by breaking the interval $[0, t]$ into n non-overlapping sub-intervals of length t/n and considering the number of these sub-intervals that contain an event. The second and third conditions of a Poisson process imply that each interval contains an event with the same probability $\lambda t/n$ which means that $X(t)$ is binomial distributed with parameters n and $p \approx \lambda t/n$. Second, when $n \to \infty$, the binomial distribution converges towards the Poisson distribution with the mean $np = \lambda t$. Third, since the last Poisson process condition implies that P(2 or more events) $\to 0$, $X(t)$ is a Poisson distributed random variable with mean λt.

$\{X_t : t \in [0, \infty)\}$ is a non-homogeneous (non-stationary) Poisson process with intensity function $\lambda(t), t \geq 0$ if

- $X(0) = 0$
- For any $0 = t_0 < t_1 < t_2 < \cdots < t_n$, the increments $X(t_1) - X(t_0), X(t_2) - X(t_1), \cdots, X(t_n) - X(t_{n-1})$ are independent.
- $\mathbb{P}(X(t+h) - X(t) = 1) = \lambda(t)h + o(h)$
- $\mathbb{P}(X(t+h) - X(t) \geq 2) = o(h)$

For the non-homogeneous Poisson process, the probability of the number of events that occur in a given interval depends on both the length of the interval and its location. In particular,

Theorem 5.1.2. *If X_t is a non-homogeneous Poisson process with intensity function $\lambda(t)$, then $X(s+t) - X(s) \sim Poisson(m(s+t) - m(s))$ where $m(t) = \int_0^t \lambda(s)ds$. in particular, $X(t) \sim Poisson(m(t))$. Hence, $\mathbb{E}(X(t)) = m(t)$ and $\mathbb{V}(X(t)) = m(t)$.*

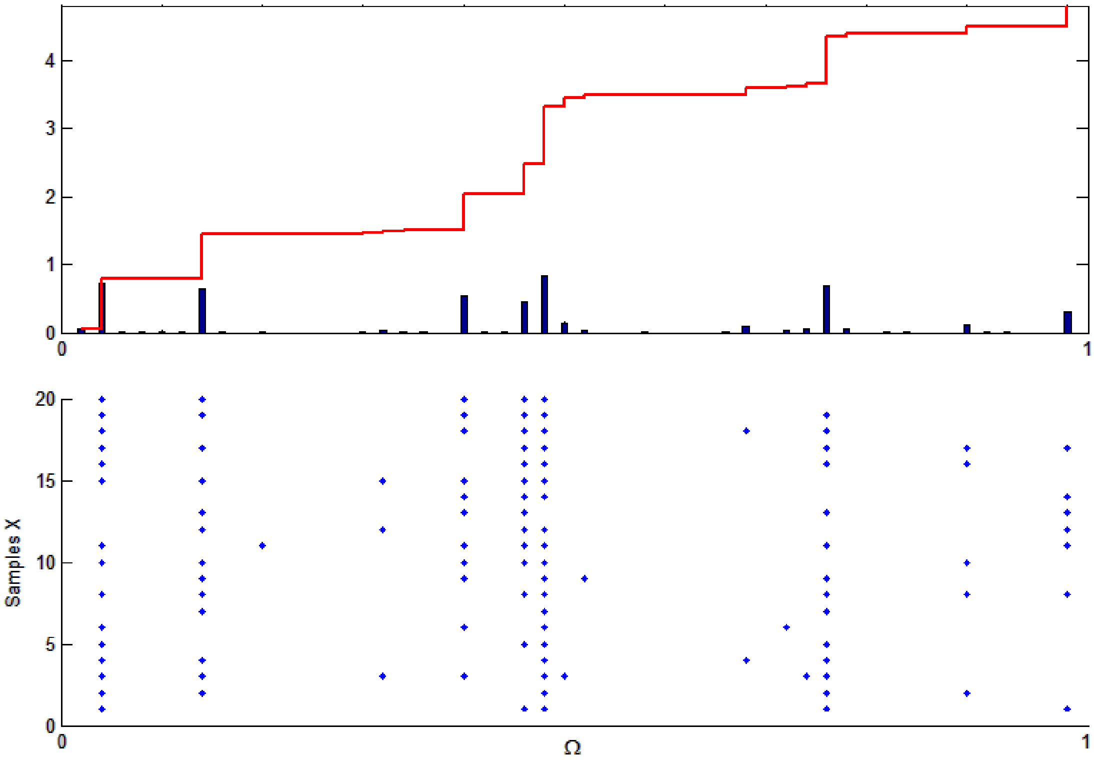


Figure 5.2: The top panel shows a measure B sampled from a beta process (blue), along with its cumulative distribution function (red). The horizontal axis is Ω. The bars of the blue segments are drawn from a Poisson process. The bottom panel shows 20 samples from a Bernoulli process with base measure B, one per line. Samples are represented as sets of points, obtained by calculating each point ω independently with probability $B(\{\omega\})$.

5.2 Completely Random Measures

A construction based on non-homogeneous Poisson process is significant to represent instances of the general family of random measures known as completely random measures (116, 117) in terms of modeling and computation.

A completely random measure G on a probability space $(\Omega, \mathfrak{R})$ is a random measure such that, for any measurable disjoint sets $A_1, \ldots, A_n$, the random variables $G(A_1), \ldots, G(A_n)$ are independent and $G(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} G(A_i)$. Complete random measure can be derived from an underlying non-homogeneous Poisson process as shown in Figure 5.1. Let $\nu(d\omega, dp)$ denote a measure on the product space $\Omega \times \mathfrak{R}$, such that $\nu(\Omega \times \mathfrak{R}) = \infty$. Draw a sample $\{(\omega_i, p_i)\}$ from this Poisson process. This sample yields a measure on Ω as follows:

$$G = \sum_{i=1}^{\infty} p_i \delta_{\omega_i} \tag{5.2}$$

$\{\omega_i\}$ are referred as the atoms of the measure G and $\{p_i\}$ as the weights. Since the Poisson process assigns independent mass to disjoint sets, this measure is completely random.

5.3 Beta and Bernoulli Processes

The beta process was first defined for applications in survival analysis (118). In (119), its definition has been relaxed onto more general spaces instead of a distribution on cumulative hazard functions over the positive real line.

Definition 5.3.1. *A positive random measure B on a space Ω is a Lévy process, if the masses $B(S_1)$,...,$B(S_k)$ assigned to disjoint subsets S_1,...,S_k of Ω are independent. The Lévy-Khinchine theorem implies that a positive Lévy process is uniquely characterized by its Lévy measure, a measure on $\Omega \times \mathbb{R}^+$.*

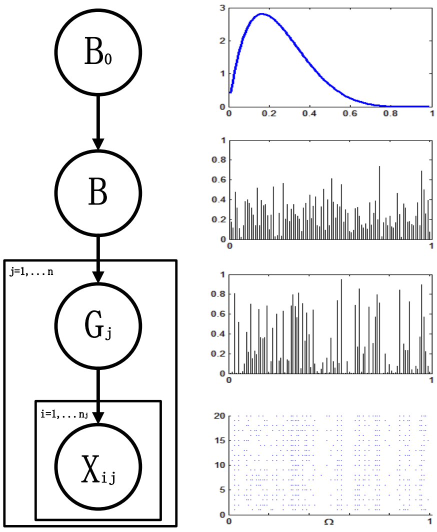


Figure 5.3: The left panel shows the graphical model for the hierarchical beta process. The right panel demonstrates example samples from this model with $c_0 = c_j = 2$ and B_0 beta-distributed on $[0, 1]$. From top to bottom are a beta process sample B from B_0, a beta process sample G_j from B, and 20 Bernoulli process samples $X_{1,j}, ..., X_{20,j}$ from G_j.

A beta process, as a special case of a Lévy measure can be defined as:

Definition 5.3.2. *A beta process $B \sim BP(c, B_0)$ is a positive Lévy process whose Lévy measure depends on two parameters: c is a positive function over Ω called the concentration function, and B_0 is a fixed measure on Ω, called the base measure. In the special case where c is a constant it will be called the concentration parameter. $\gamma = B_0(\Omega)$ is called the mass parameter.*

There are two cases with regard to whether base measure B_0 is continuous or discrete.

- If B_0 is continuous, the Lévy measure of the beta process is

$$\nu(d\omega, dp) = c(\omega)p^{-1}(1-p)^{c(\omega)-1}dpB_0(d\omega) \tag{5.3}$$

on $\Omega \times [0,1]$. As a function of p, $\nu(d\omega, dp)$ is a degenerate beta distribution. To draw $B \sim BP(c, B_0)$, draw a set of points $(\omega_i, p_i) \in \Omega \times [0,1]$ from a non-homogeneous Poisson process with base measure ν (see Figure 5.2), and let $B = \sum_i p_i \delta_{\omega_i}$. B is discrete, and the pairs (ω_i, p_i) correspond to the location $\omega_i \in \Omega$ and weight $p_i \in [0,1]$ of its atoms, respectively. Although the expectation of B is finite as long as B_0 is finite, B is a countably infinite sum because the Poisson process generates infinitely many points in terms of $\nu(\Omega \times [0,1]) = \infty$.

- If B_0 is discrete of the form $B_0 = \sum_i q_i \delta_{\omega_i}$, B has atoms at the same locations $B = \sum_i p_i \delta_{\omega_i}$ with $p_i \sim Beta(c(\omega_i)q_i, c(\omega_i)(1-q_i))$ which requires $q_i \in [0,1]$.

The Bernoulli process is defined as:

Definition 5.3.3. *Let B be a measure on Ω. A Bernoulli process is defined with hazard measure B as $X \sim BeP(B)$, as the Lévy process with Lévy measure $\mu(dp, d\omega) = \delta_1(dp)B(d\omega)$.*

When B is continuous, X is a Poisson process with $X = \sum_{i=1}^{N} \delta_{\omega_i}$ where $N \sim Poi(B(\Omega))$ and ω_i are independent realized from the distribution $B/B(\Omega)$. When B is discrete with the form of $B = \sum_i p_i \delta_{\omega_i}$, then $X = \sum_i b_i \delta_{\omega_i}$ where b_i are independent Bernoulli variables with the probability $b_i = 1$ equal to p_i.

A Bernoulli process can be viewed as a special case of a Poisson process with the atom (singleton) weights at interval $[0,1]$, regardless of whether the base measure B is discontinuous. An intuitive interpretation is to consider Ω as a space of features and X as an object characterized by the features it possesses. The random measure B then renders the probability that X possesses each particular feature ω, as shown in Figure 5.2.

The important property of the beta process and the Bernoulli process is that they are a pair of conjugate stochastic processes. Let $B \sim BP(c, B_0)$, and let $X_i|B \sim BeP(B)$ for $i = 1, ..., n$ be n independent Bernoulli process samples from B. Let $X_{1...n}$ denote the set of observations $\{X_1, ..., X_n\}$. The posterior distribution of B given $X_{1...n}$ is still a beta process with updated parameters:

$$B|X_{1...n} \sim BP(c+n, \frac{c}{c+n}B_0 + \frac{1}{c+n}\sum_{i=1}^{n} X_i) \tag{5.4}$$

This result is derived in (118).

5.4 Hierarchical Beta Process

In various domains, there are a number of groups of data which are yielded from related, but distinct generative processes. Although we can analyze each group independently, it neglects critical information shared among groups. On the contrary, combining groups in a single exchangeable dataset may lead to biased estimates and obscure characteristics distinguishing particular groups. Hierarchical Bayesian models provide an elegant compromise (73, 120). Estimates based on posterior dependencies between parameters are "shrunk" together, so that groups share the strength of each other while retaining distinctive features.

The hierarchical beta process allows us to model a type of stochastic phenomena in which features are shared among multiple subjects from a number of groups. To lay down its theoretical ground, we discussed the statistical and computational properties of several more basic stochastic processes in previous sections.

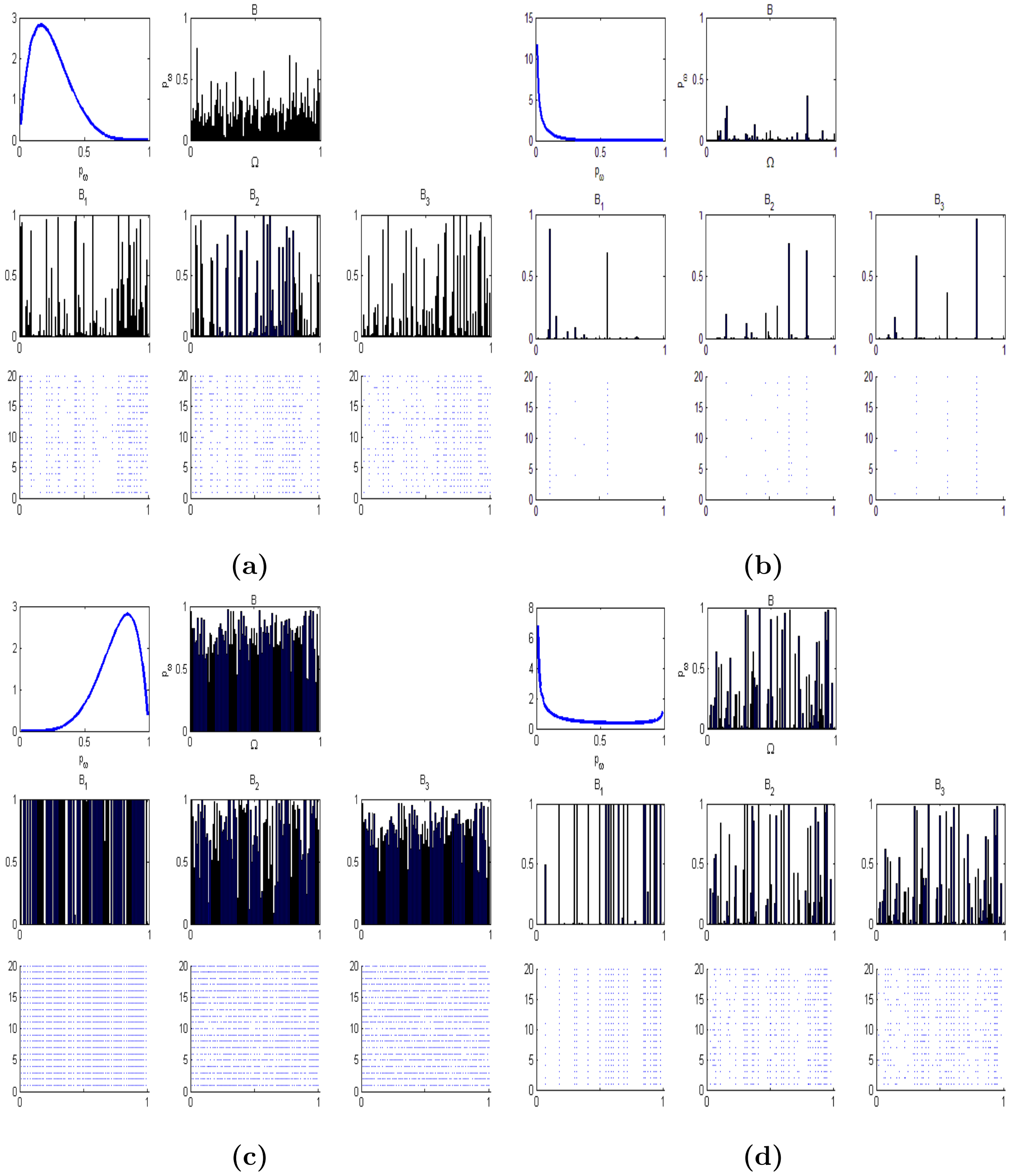


Figure 5.4: Realizations from a hierarchal beta process with n=3 and $n_1 = n_2 = n_3 = 20$. We vary the concentration parameters: c_1, c_2 and c_3, and the base measure parameters: a_0 and b_0. In (a), the parameters are $c_1 = c_2 = c_3 = 1$ and $a_0 = 2$, $b_0 = 6$. In (b), the parameters are $c_1 = c_2 = c_3 = 1$ and $a_0 = 2$, $b_0 = 0.6$. In (c), the parameters are $c_1 = 0.02$, $c_2 = 3$ and $c_3 = 400$, and $a_0 = 6$, $b_0 = 2$. In (d), the parameters are $c_1 = 0.02$, $c_2 = 3$ and $c_3 = 400$, and $a_0 = 0.2$, $b_0 = 0.6$.

If we view Ω as a space of all the possible "features", X can be taken as an object characterized by the features it possesses such that the random measure B encodes the probability that X possesses each particular features. To construct a hierarchical beta process, a beta process prior $BP(c_0, B_0)$ is put on B, in which B becomes a realization of the baseline measure B_0. The advantage of this structure is that we can share statistical strength among multiple object groups. Suppose an object i within the group j is represented by a binary vector $X_{i,j}$ of which each component indicates whether the particular feature ω is possessed by the object with a probability p_ω^j specific to group j. These probabilities are atom weights of a discrete measure G_j over the space Ω, and G_j itself is a realization of beta process $BP(c_j, B)$. In summary, the model with graphical representation shown in Figure 5.3is specified as following:

$$B \sim BP(c_0, B_0) \tag{5.5}$$

$$G_j \sim BP(c_j, B) \qquad \forall j \leq n \tag{5.6}$$

$$X_{i,j} \sim BeP(A_j) \qquad \forall i \leq n_j \tag{5.7}$$

According to Eqn. 5.5, we illustrate four sets of hierarchical beta process each of which contains three groups of beta-Bernoulli processes from a common beta process with specified parameters in Figure 5.4. This illustration highlights the effect of the concentration parameters for c_{1-3} and mass parameters for (a_0, b_0).

5.5 Hierarchical Dynamic Model

The modeling approach of the expertise-specific groups' eye movements for the dermatological images is diagrammed in Figure 5.5.

In Figure 5.5 (a), the hierarchy represents the heterogeneous structure produced by individuals with different expertise levels examining medical image view-

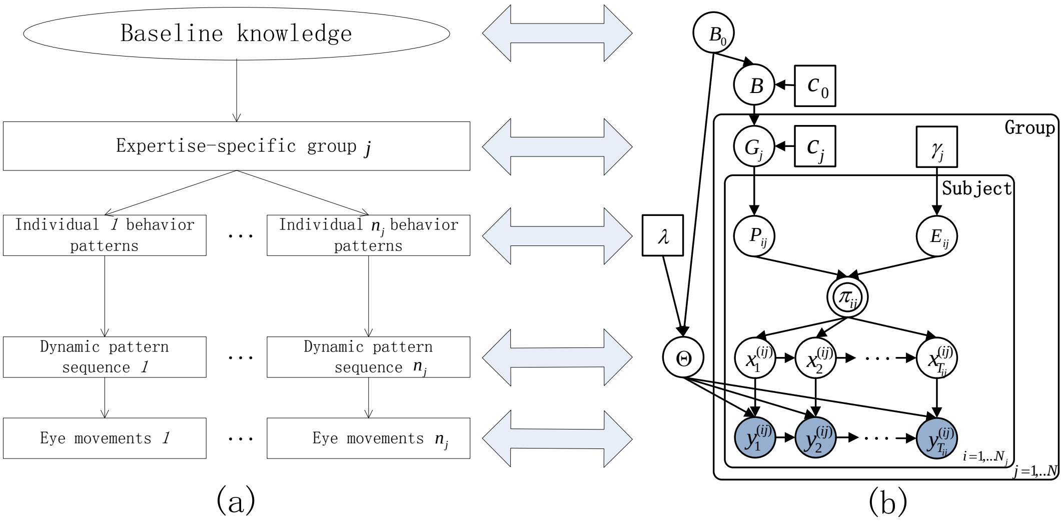


Figure 5.5: Graphical model representation of the hierarchical dynamic model. B_0 denotes a fixed continuous complete random measure as a global baseline on the space of all possible eye movement patterns Θ. B denotes a beta process to measure the eye movement patterns shared among N groups. G_j denotes a beta process to represent the eye movement patterns shared among N_j subjects of group j. The transition distribution π_{ij} of subject i in group j is deterministic. It is determined by both the pattern indicator variable P_{ij} of which each component p_{ijk} is Bernoulli-distributed given the probability g_{jk} and the transition variable E_{ij} which is Gamma-distributed given γ_j. $z_{t_{ij}}^{(ij)}$ and $x_{t_{ij}}^{(ij)}$ denotes the hidden state variable and the observation variable of the hidden Markov model. Θ denotes the emission distributions as eye movement patterns. The total number of eye movement patterns exhibited is denoted by the dimensionality of Θ: K which is depend on the beta process B_0.

ing strategies as well as to provide the flexibility of learning new patterns as new eye movement data are observed. A group of subjects with the same expertise level share a set of behavior patterns based on their knowledge. In accordance with these common behavior patterns, each group member's time-evolving behaviors also display their individualized temporal patterns in terms of unique subsets of behaviors and/or their unique sequential combinations. At the lowest level, each behavior is measured based on observed eye movements. Figure 5.5 (b) shows the graphical representation of the hierarchical dynamic model corresponding to (a)'s structure. B_0 is the global base measure on the space of all possible behaviors Θ. The common behavior pattern of the group defined as $\{(\theta_k, E_k)\}$ is characterized by the shared behaviors among p group members and the probabilities that it possesses each particular behavior is encoded by B_0. A group member p performs individualized behavior pattern defined as $\{(\theta_k, S_{pk})\}$ which is a Bernoulli process realization of the group common pattern $\{(\theta_k, E_k)\}$. The transition matrix π_p follows a Dirichlet distribution specified by the non-zero entries of S_p.

Since fixation and saccadic data are deployed in a sequential manner we use a hidden Markov model (HMM) to characterize their temporal dynamic nature. Since eye movements are inherently not smooth and highly correlated, the strong Markovian assumption of HMMs is inappropriate. We therefore employ autoregressive HMMs to relax the Markovian assumption by modeling eye movement data as a noisy linear combination of some finite set of past observations plus additive white noise.

5.5.1 Dynamical likelihoods

Auto-HMMs has been proposed to be a simpler but often effective way to describe dynamical systems (121). Let $y_t^{(ij)}$ denote the eye movement data of the i^{th} subject at time step t in the j^{th} group. We associate each time-step's observation with one fixation and its successive saccade as one observation unit. Let $x_t^{(ij)}$ denote the corresponding latent dynamic mode. We have

$$x_t^{(ij)} \sim \pi_{x_{t-1}^{(ij)}} \tag{5.8}$$

$$y_t^{(ij)} = A_{x_t^{(ij)}} \tilde{y}_t^{(ij)} + e_t(x_t^{(ij)}) \tag{5.9}$$

where $e_t^{(ij)}(k) \sim N(0, \Sigma_k)$ which is an additive white noise, $A_k = [A_{1,k}, ..., A_{r,k}]$ as the set of lag matrices, and $\tilde{y}_t^{(ij)} = [y_{t-1}^{(ij)}, ..., y_{t-r}^{(ij)}]$. In our case, we specify $r = 1$.We thus define $\theta_k = (A_k, \Sigma_k)$ as one eye movement pattern.

5.5.2 Hierarchical prior

The hierarchical beta-Bernoulli processes proposed by Thibaux et al. (119) is a suitable tool to describe the situation where multiple groups of subjects are defined by countably infinite shared features following the Levy measure. We utilize this process in the following specification based on our problem scenario.

Let B_0 denote a fixed continuous random base measure on a measurable space $\Theta = \{\theta_k\}$ which represents a library of all the potential eye movements patterns. To characterize patterns shared among multiple groups, let B denote a discrete realization of a beta process given the prior $BP(c_0, B_0)$. Let G_j be a discrete random measure on Θ drawn from B following the beta process which represents a measure on the eye movement patterns shared among multiple subjects within the group j. Let P_{ij} denote a Bernoulli measure given the beta process G_j. P_{ij} is a binary vector of Bernoulli random variables representing whether a particular eye movement pattern exhibited in the eye movement sequence of subject i within group j. This hierarchical construction can be formulated as follow:

$$B|B_0 \sim BP(c_0, B_0) \tag{5.10}$$

$$G_j|B \sim BP(c_j, B) \qquad j = 1, ..., N \tag{5.11}$$

$$P_{ij}|G_j \sim BeP(G_j) \qquad i = 1, ..., N_j \tag{5.12}$$

where $B = \sum_k b_k \delta_{\theta_k}$ with $\{\theta_k\}$ drawn from the library Θ and coupled with their weights b_k. b_k is beta-distributed given b_0 and c_0. Furthermore, $G_j = \sum_k g_{jk} \delta_{\theta_{jk}}$ shows that G_j is associated with both $\{\theta_{jk}\}$ which is a subset of countable number of eye movement patterns drawn from $\{\theta_k\}$ and their corresponding probability masses $\{g_{jk}\}$ given group j. $\{g_{jk}\}$ is also beta-distributed given b_k and c_j. The combination of these two variables characterizes how the eye movement patterns shared among subjects within expertise-specific group j. Thus P_{ij} as a Bernoulli process realization from the random measure G_j is denoted as:

$$P_{ij} = \sum_k p_{ijk} \delta_{\theta_{jk}} \tag{5.13}$$

where p_{ijk} as a binary variable denotes whether subject i within group j exhibits eye movement pattern k given probability mass g_{jk}. Based on the above formulation, for $k = 1...K_j$ patterns $\{(\theta_{jk}, g_{jk})\}$ characterize how a set of common eye movement patterns likely shared among group j and $\{(\theta_{jk}, p_{ijk})\}$ represent subject i's personal subset of eye movement patterns given group j.

The transition distribution $\pi_{ij} = \{\pi_{x_t^{(ij)}}\}$ of the auto-HMMs at the bottom level governs the transitions between the i^{th} subject's personal subset of eye movement patterns θ_{jk} of group j. It is determined by the element-wise multiplication between the eye movement subset $\{p_{ijk}\}$ of subject i in group j and the gamma-distributed variables $\{e_{ijk}\}$:

$$e_{ijk}|\gamma_j \sim Gamma(\gamma_j, 1) \tag{5.14}$$

$$\pi_{ij} \propto E_{ij} \bigotimes P_{ij} \tag{5.15}$$

where $E_{ij} = [e_{ij1}, ...e_{ijK_j}]$. P_{ij} determines the effective dimensionality of π_{ij}, which is inferred from observations.

5.6 Posterior Inference with Gibbs Sampler

We use the Gibbs sampler to do the posterior inference. In one iteration of the sampler, each latent variable is visited and assigned a value by drawing from the distribution of that variable conditional on the assignments to all other latent variables as well as the observation. In particular, based on the sampling algorithm proposed in (119), we developed a Gibbs sampling solution to the hierarchical beta processes part of the model.

We adopt normal-inverse-Wishart distribution to provide an appropriate conjugate matrix prior to pattern space Θ. The conjugate prior on the set of design matrix A and the noise covariance Σ is the matrix normal-inverse-Wishart prior. This distribution places a conditionally matrix normal prior on A given Σ:

$$p(A|\Sigma, M, K) = \frac{|K|^{\frac{d}{2}}}{|2\pi\Sigma|} exp\{-\frac{1}{2} tr((A-M)^T \Sigma^{-1} (A-M)K)\} \tag{5.16}$$

and an inverse-Wishart prior on Σ

$$\Sigma \sim \mathcal{W}(\nu, \Delta) \tag{5.17}$$

Consider a set of observations $D = \{X, Y\}$, the posterior distribution of $\{A, \Sigma\}$ can be decomposed as the product of posterior A as $\mathcal{MN}(A; S_{yx}S_{xx}^{-1}, \Sigma, S_{xx})$

with $S_{xx} = XX^T + K$, $S_{yx} = YX^T + MK$, and $S_{yy} = YY^T + MKM^T$ and the marginal posterior of Σ as $\mathcal{W}(\nu + N, \Delta + S_{y|x})$ where $S_{y|x} = S_{yy} - S_{yx}S_{xx}^{-1}S_{yx}^T$.

When sampling the pattern indicator matrix P_j of group j, we need to address two situations separately. For a pattern which has non-zero probability because of either its priori or having already been instantiated by at least one subject, we compute its posterior as follows.

Let $\{\omega\}$ denote the atoms (eye movement patterns) that have been observed at least once. We define the variables to perform inference: $b_0 = B_0(\{\omega\})$, $b = B(\{\omega\}) = \sum_k b_k\delta_\omega$, $g_j = G_j(\{\omega\}) = \sum_k g_{jk}\delta_\omega$, and $p_{ij} = P_{ij}(\{\omega\}) = \sum_k p_{ijk}\delta_\omega$. According to Equation 5.10 - 5.12, these variables from their respective processes have the following distributions:

$$B(\omega) \sim Beta(c_0 B_0(\omega), c_0(1 - B_0(\omega))) \tag{5.18}$$

$$G_j(\omega) \sim Beta(c_j B(\omega), c_j(1 - B(\omega))) \tag{5.19}$$

$$P_{ij}(\omega) \sim Ber(G_j) \tag{5.20}$$

We marginalize out G using conjugacy. Let $m_j = \sum_{i=1}^{n_j} p_{ij}$, and use $\Gamma(x+1) = x\Gamma(x)$, the posterior distribution of b given P_j:

$$p(b|b_0, P) \propto p(b|b_0)\frac{\Gamma(m_j + c_j b)\Gamma(n_j - m_j + c_j(1-b))}{\Gamma(c_j b)\Gamma(c_j(1-b))} \tag{5.21}$$

This posterior is log-concave, which we can use adaptive rejection sampling method (122) to approximate in our Gibbs sampler. We can sample g_j from its conditional posterior distribution by conjugacy:

$$p(g_j|b, P) \propto Beta(c_j b + m_j, c_j(1-b) + n_j - m_j) \tag{5.22}$$

Given the i^{th} subject's eye movements data sequence $y_{1:T_{ij}}^{(ij)}$ in the group j, transition variable E_{ij} and within-group-j shared pattern set $\theta_{1:K_j}$, the current sampling pattern indicator p_{ijk} of pattern k exhibited by subjects i in group j follows this posterior distribution:

$$\begin{aligned}&p(p_{ijk}|P^{(-ijk)}, y_{1:T_{ij}^{(ij)}}, \theta_{1:K_j}^{(-ijk)}, E_{ij}, B_0) \propto \\ &p(p_{ijk}|P^{(-ijk)}, B_0)p(y_{1:T_{ij}^{(ij)}}|P_{ij}, E_{ij}, \theta_{1:K_j}^{(-ijk)})\end{aligned} \tag{5.23}$$

where $P^{(-ijk)}$ denotes the set of all P_{ij} except p_{ijk}. In particular, for the instantiated patterns

$$\begin{aligned}&p(p_{ijk}|P^{(-ijk)}, B_0) = \\ &\int p(p_{ijk}|G_j) \int p(G_j|B, P)p(B|B_0, P)dBdG_j\end{aligned} \tag{5.24}$$

Both $p(G_j|B, P)$ and $p(B|B_0, P)$ can be sampled as in Equation 5.22 and Equation 5.21, respectively.

For the yet-instantiated patterns of group j, since they can be directly sampled from the conjugate prior distribution of Θ, we only need to infer the distribution of their number the prior distribution of which is Poisson-distributed $K \sim Poi(\frac{c_0\lambda}{c_0+k-1})$. Given that all other patterns from all other groups are zero:

$$\begin{aligned}&p(k_{ij}|P_{ij}, y_{1:T_{ij}^{(ij)}}, \theta_{1:K_j}^{(-ijk)}, E_{ij}, \lambda) \propto \\ &p(p_{ijk}|P^{(-ijk)}, \lambda)p(y_{1:T_{ij}^{(ij)}}|P_{ij}, E_{ij}, \theta_{1:K_j}^{(-ijk)})\end{aligned} \tag{5.25}$$

Given transition distributions π_{ij}, shared patterns $\{\theta_k\}$, and observations $y_{1:T_{ij}}$, within forward-backward massage passing algorithm, we sample the forward message to update the hidden state sequence $x_{1:T_{ij}}^{(ij)}$:

$$\begin{aligned}&p(x_{t_{ij}}|x_{(t_{ij}-1)}, y_{1:T_{ij}}^{(ij)}, \pi_{ij}, \{\theta_k\}) \propto \\ &\pi_{x_{(t_{ij}-1)}^{(ij)}} N(y_{t_{ij}}^{(ij)}; A_{x_{t_{ij}}^{(ij)}}\tilde{y}_{t_{ij}}^{(ij)}, \Sigma_{x_{t_{ij}}^{(ij)}})m_{t+1,t}(x_{t_{ij}}^{(ij)})\end{aligned} \tag{5.26}$$

5.7 Synthetic Experiment

We generated six 4 dimensional time series from an auto-regressive HMM $y_t^{(i)} = a_{z_t^{(i)}}y_{t-1}^{(i)} + e_t^{(i)}(x_t^{(i)})$ with $a_k \in \{-0.9, -0.6, -0.3, 0, 0.3, 0.6, 0.9\}$, and noise covariance Σ_k drawn from an $IW(6, 0.2I_4)$ prior. We use these simulated time series to test the model. The shared patterns were sampled from a beta process using $a = 2, b = 6$, as described in Section 2.7. This setting allows us to simulate observation sequences with their true states as shown in Figure 5.6 (a). In particular, different subset of a_k are randomly selected to generate each time series.

We used six HMMs tied together with a shared set of transition and dynamic parameters, and infinite Gaussian mixture models to compare the performance between these models on the generated time series. Each model was initialized with the sufficient statistics of the simulated data. The results shown in Figure 5.6 indicate our model performs better at distinguishing these dynamical matrices. The Hamming distance error in Figure 5.6 (b) indicates our model fits the observation better. One possible cause is that the iGMM makes a strong assumption that the eye movement data are independent which is hardly true. On the contrary, our model only assumes that the eye movement patterns are exchangeable in order. What's more, the iGMM and HMM approaches assume that each time series exhibits the same structure. Our model allows each time series to share a subset of common structures. Additionally, our model and the HMM take sequential information of eye movements into account. Example segmentations representing the median Hamming distance error are shown in Figure 5.6 (c) and (d). In particular, the key point is illustrated that our model emphasizes the sharing of behavior patterns instead of assuming all time series are behaving on the same dynamics.

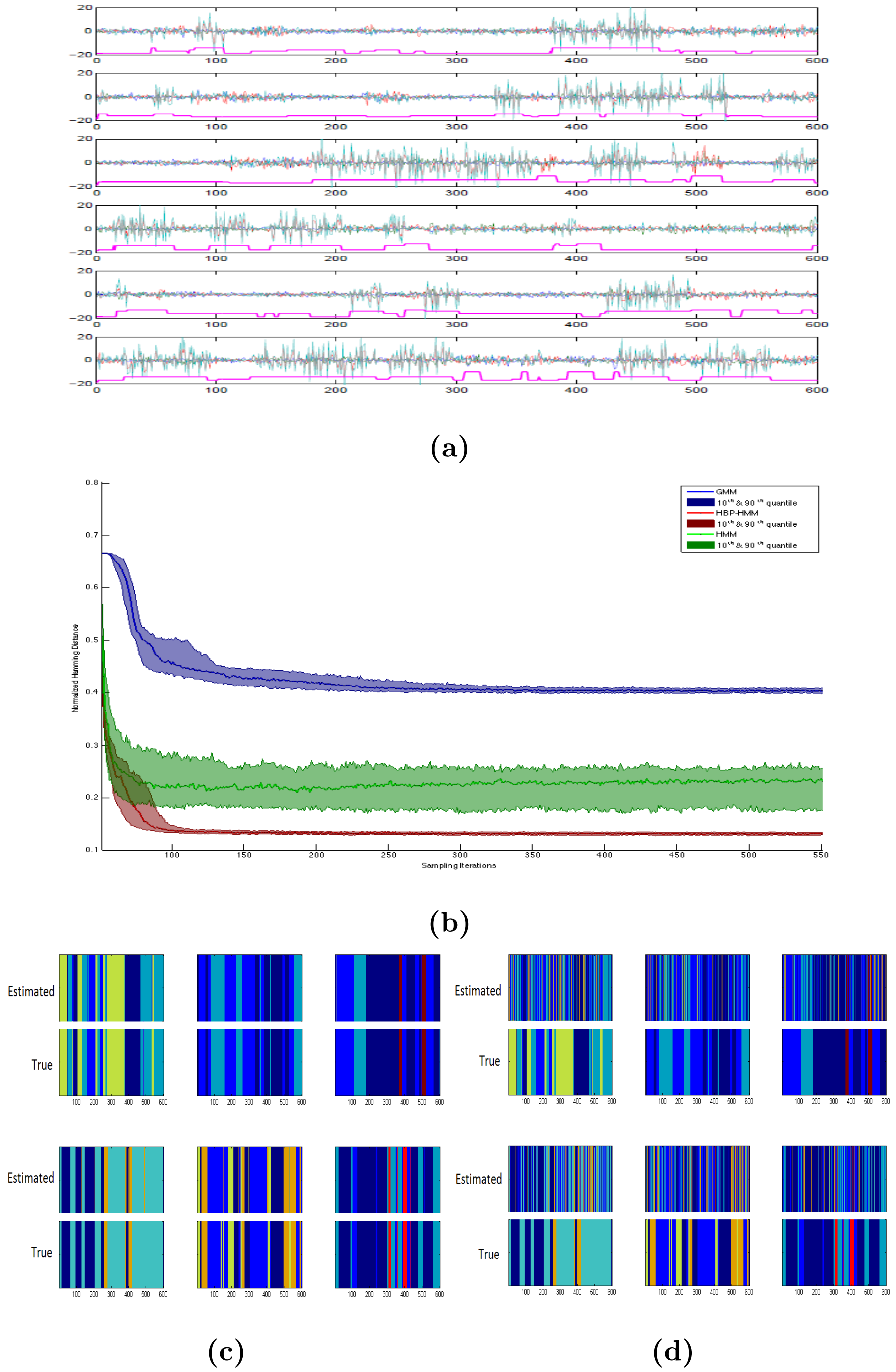


Figure 5.6: (a) Simulated sequences from 6 AR-generated time series and their true state sequences (cyan). (b) The medians and 10^{th} and 90^{th} quantiles of Hamming distances between the true and estimated mode sequence on 100 trials. (c)-(d) Typical segmentations for the six time series at 550 sampling iterations for the two models. The top and bottom panels display the estimated and true sequences.

6

Discovery of Eye Movement Patterns

Our model converges to generate 387 eye movement patterns based on eleven attending subjects diagnosing on fifty dermatological images. These results allow us to analyze and describe the dermatological images based on a novel perspective of experts' perceptual strategies.

6.1 Eye Movement Pattern Estimation

In Figure 6.1, our model generated 87933 fixation-saccade units to simulate the 15 professionals including both attending and resident groups. We then compare the distributions of the simulation and the observations. Note that the results also validate the discovered eye movement patterns, because such simulation needs to generate a set of realizations of eye movement patterns explicitly from the hierarchical prior, simulate multiple possible sequences of these patterns, and

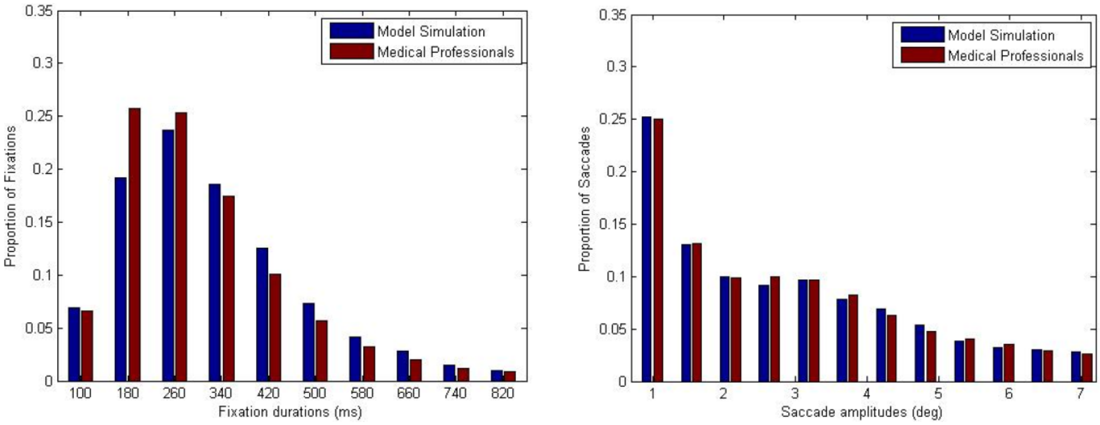


Figure 6.1: Quantitative performance evaluations. (a) The histogram of the observed fixation duration distributions of the 15 professionals including both attending and resident group (red) and our model's simulations over 50 images (blue). (b) The histograms of the saccade amplitude distributions of the 15 professionals (red) and our model's simulations over 50 images (blue).

then we are able to draw fixation-saccade samples from them.

Figure 6.2 illustrates the eleven dermatologists diagnosing a case of a skin manifestation of endocarditis by showing one set of observed eye movement sequences and the model's discovered eye movement patterns shared by the dermatologists which correspond to descriptively meaningful perceptual units. In the medical image, there are multiple skin lesions spreading over the thumbnail and tip, the two parts of index finger, and the middle finger. A primary abnormality is on the thumb tip. The eye movement sequences in Figure 6.2 indicate that dermatologists examine the image in a highly patterned manner by fixating on the primary abnormality heavily and switching their visual attention actively between and within the primary and secondary abnormalities. Our model decomposes each eye movement sequence into several subsets of its segments. Each subset is characterized by one estimated latent state and a Gaussian emission distribution which summarizes the similar temporal-spatial properties shared among multiple sequences. The way that the patterns are shared among the subjects is also indicated by their matrix in Figure 6.2. For example the first subject's eye movements evolve over time with the first eight out of nine patterns, and the eleventh subject has seven patterns except pattern 5 and pattern 9. In other words, most but not all patterns are shared by all physicians, as one would expect when modeling human behaviors where there almost certainly exist some variation and some individual differences. Again, our model is able to capture both the shared (stereotypical) behaviors and the individualized (idiosyncratic) ones. Transition probability matrices indicated these patterns are persistent with high self-transition probabilities which measure the likelihood of a given pattern transiting into itself in our dynamic model. In Figure 6.3, we demonstrate the same inference process on another image.

Some similar patterns also emerged in the resident group but are lacking in the novice group as shown in Figure 6.4 (c). This suggests that experts, equipped with domain knowledge organized in finer gradations of functional categories, can discriminate the significance of their findings in a particular context. In contrast, in Figure 6.4 (c) the novices failed to do so, although they perceive the same abnormalities too. Compare Figure 6.4 (b), Figure 6.4 (b) and Figure 6.4 (b), the difference between the transition probability matrices of the medical professional groups and the novice group suggests professionals' eye movement patterns are more persistent than the novices'.

Figure 6.4 and 6.5 show the discovered eye movement patterns from three expertise-specific groups on two dermatological images. These two images are among the most difficult cases to make a correct diagnosis, as estimated by dermatologists, and some of the patterns exhibited on them are critical to inform

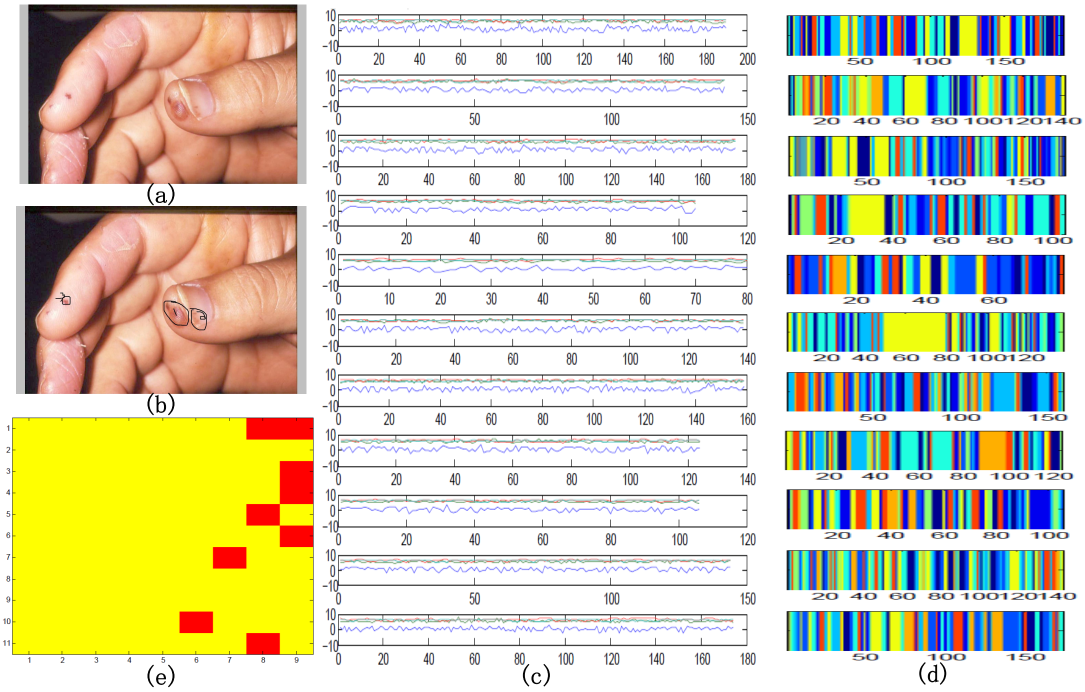


Figure 6.2: The model performance running on the eye movement data of 11 subjects viewing one case. (a) shows the original medical image. (b) is primary and secondary abnormalities were explicitly marked and numbered by an experienced dermatologist. (c) shows eleven time series, each observation of which is composed of 4 components: log values of fixation location (xy coordinate), fixation duration and saccade amplitude. (d) shows the HMM-derived eye movement pattern sequences for the corresponding 11 time series with 4 chains of 55000 sampling iterations. The color coding corresponds to the segments of the each specific eye movement pattern. (e) shows the shared eye movement pattern matrix of which the row number indicates the subjects and the column number indicates the shared patterns. For example, yellow color at the first row represents the time series of subject 1 exhibits pattern 1-7 but lacks pattern 8 and 9.

some properties of the images.

Taking the first illustrated image in Figure 6.4 for example, there are multiple skin lesions spreading over the thumb nail and tip, the two parts of index finger and the middle finger. A primary abnormality is on the thumb tip. The eye movement sequences indicate that attending dermatologists fixated on the primary abnormality heavily and switched their visual attention actively between and within the primary and secondary findings. The same patterns are also exhibited in the resident dermatologist group. The reason for lacking other patterns is probably because the number of participants at this expertise-specific group is limited in the dataset (only four participants). In contrast, the novice group exhibits significantly different eye movement patterns compared to the other groups. According to the novices' patterns, we can see shorter saccades so as to leave long fixation durations at the center of the image as seen in Pattern 1 and 9 of Figure 6.4 (c) and do not exhibit the eye movements switching between primary and secondary abnormalities as dermatologists' Pattern 2, 3, and 5 in Figure 6.4 (a). What is more, the more random transition matrices in Figure 6.4 (c) indicate that novice group's patterns are not persistent, which suggests that novices' focus of attention is more random when viewing the image. We reason that these relatively unstable viewing behavior reflect that fact that the novice cannot perceive the important diagnostic relationships among the multiple abnormalities and fail to prioritize them. All the pattern differences between expertise-specific groups holds for the other images studied here.

Some shared patterns emerged in the attending and the resident groups but are lacking in the novice group as shown in Figure 6.4 (c). This suggests that experts, equipped with domain knowledge organized in finer gradations of functional categories (36), can discriminate the significance of their findings in a particular context. In contrast, in Figure 6.4 (c) the novices failed to do so, although

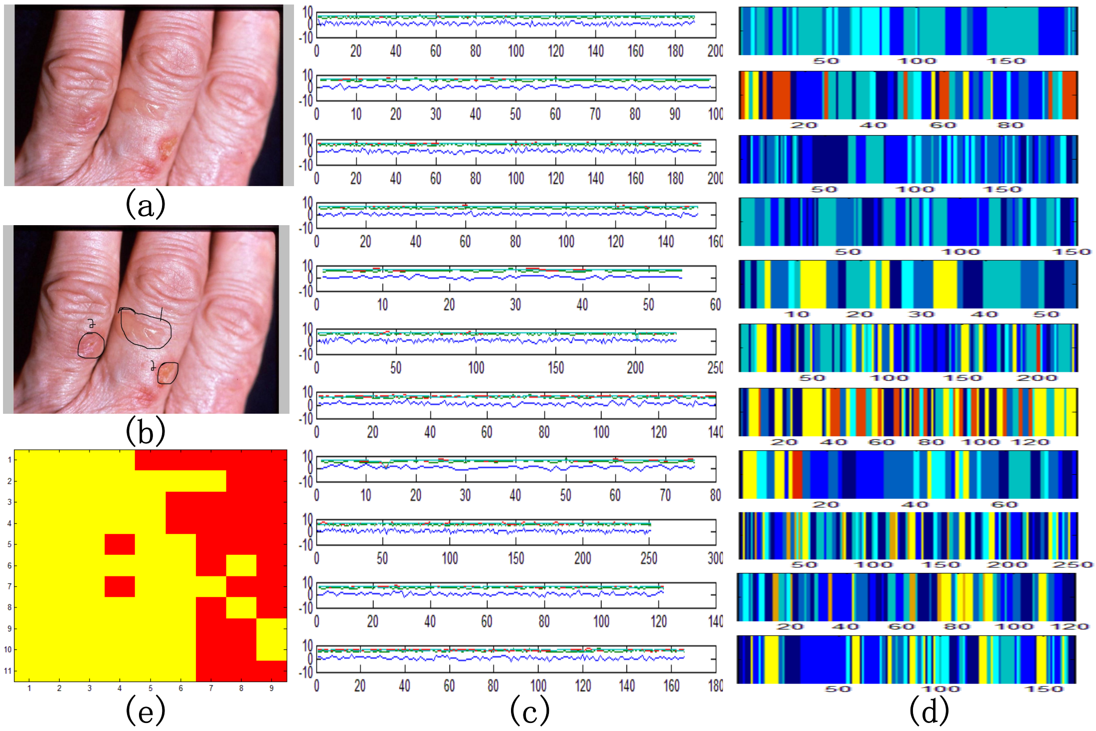
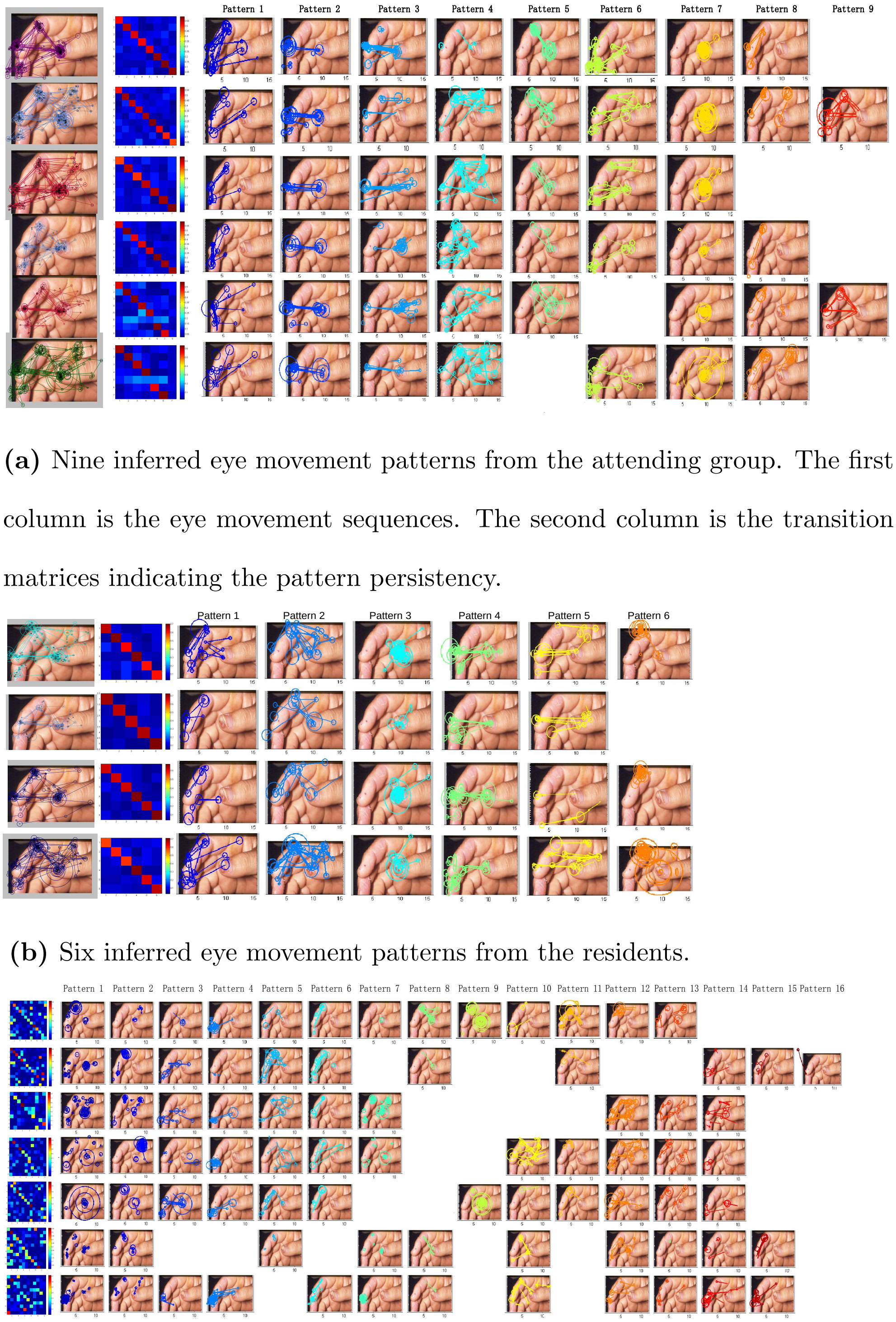


Figure 6.3: The model performance running on the eye movement data of 11 subjects viewing one case. (a) shows the original medical image. (b) is primary and secondary abnormalities were explicitly marked and numbered by an experienced dermatologist. (c) shows eleven time series, each observation of which is composed of 4 components: log values of fixation location (xy coordinate), fixation duration and saccade amplitude. (d) shows the HMM-derived eye movement pattern sequences for the corresponding 11 time series with 4 chains of 55000 sampling iterations. The color coding corresponds to the segments of the each specific eye movement pattern. (e) shows the shared eye movement pattern matrix of which the row number indicates the subjects and the column number indicates the shared patterns. For example, yellow color at the first row represents the time series of subject 1 exhibits pattern 1-7 but lacks pattern 8 and 9.



(a) Nine inferred eye movement patterns from the attending group. The first column is the eye movement sequences. The second column is the transition matrices indicating the pattern persistency.

(b) Six inferred eye movement patterns from the residents.

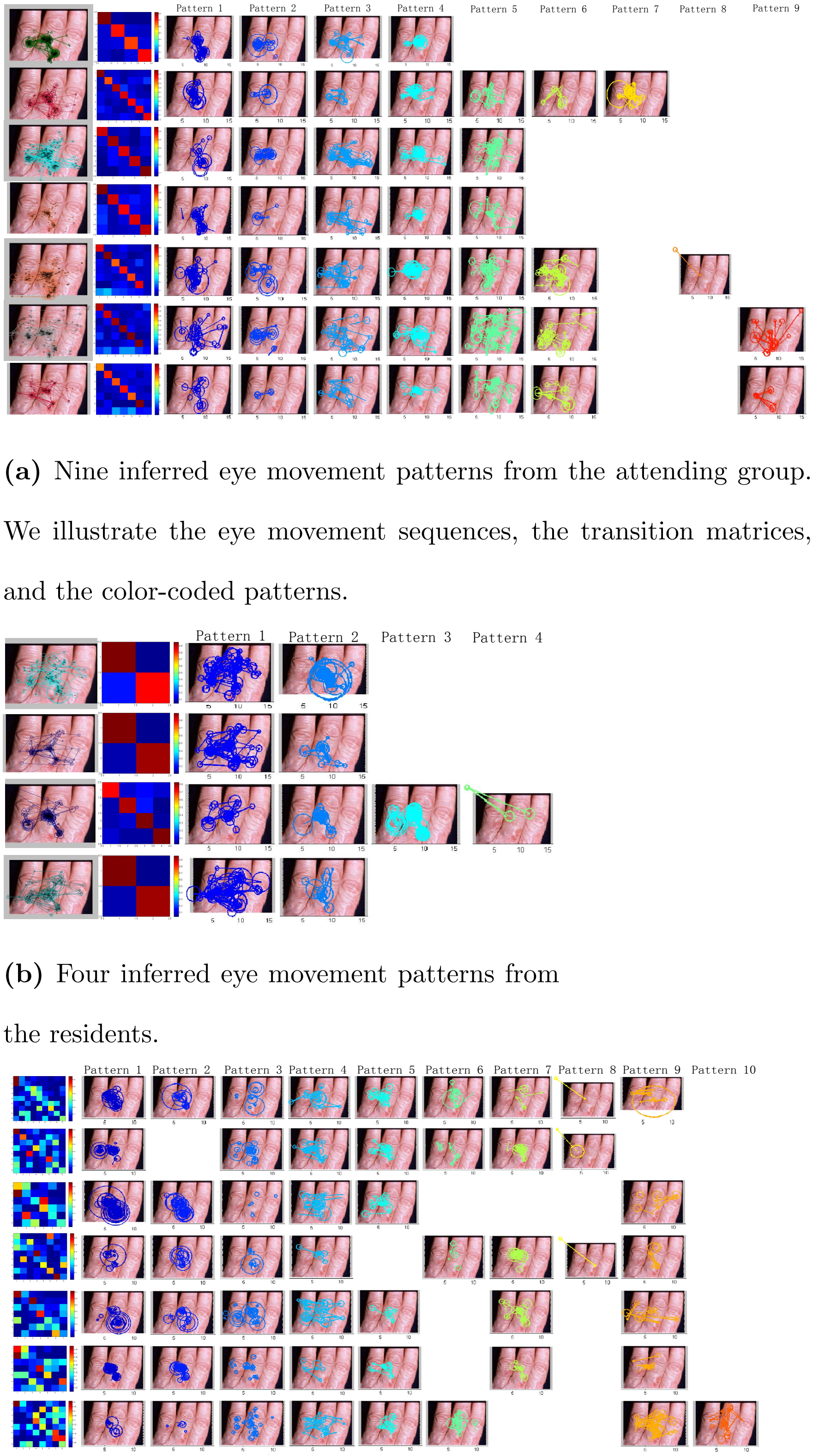
(c) Sixteen inferred eye movement patterns from the novices.

Figure 6.4: The eye movement patterns of the three expertise-specific groups signify the different perceptual behaviors.

their eye movement patterns indicate that they notice the same abnormalities too. When comparing the transition probability matrices between the expertise-specific groups in the second column of Figure 6.4 (a-c) and Figure 6.5 (a-c), it becomes clear that professionals' eye movement patterns are more persistent than the novices'.

6.2 Eye Movement Pattern Interpretation

To further analyze the meanings of the discovered eye movement patterns, we mapped thought units (see section 3.2) to patterns discovered in the eye movement data in order to see whether they correspond consistently during the diagnostic process. Pattern occurrence and thought unit alignment resulted in assignment of each fixation in a complete eye movement sequence to a specific pattern and to a thought unit such as PRI or LOC (or None). Although thought units are often spread out across eye movement patterns, some trends can be discerned. Initial integration of eye movement patterns with thought units was accomplished by calculating the counts of their time-aligned correspondence in Figure 6.6. Analysis on the left column diagram of Figure 6.6 shows, for example, that primary morphology (PRI) is closely related to the combination of two specific patterns: Pattern 2 is characterized by fixations switching between the primary and the different secondary abnormalities; and Pattern 7 by long fixations only on the primary abnormality. It is worth to point out that identification of the primary morphology is an early key diagnostic step which helps the physician to place the lesion in the correct category. Pattern 7 has relationship to location (LOC) which appears to correspond to the primary morphology location. Pattern 4 consists of eye movement sequence segments which are characterized by shorter fixation durations and longer saccades. This scanning behavior corresponds to



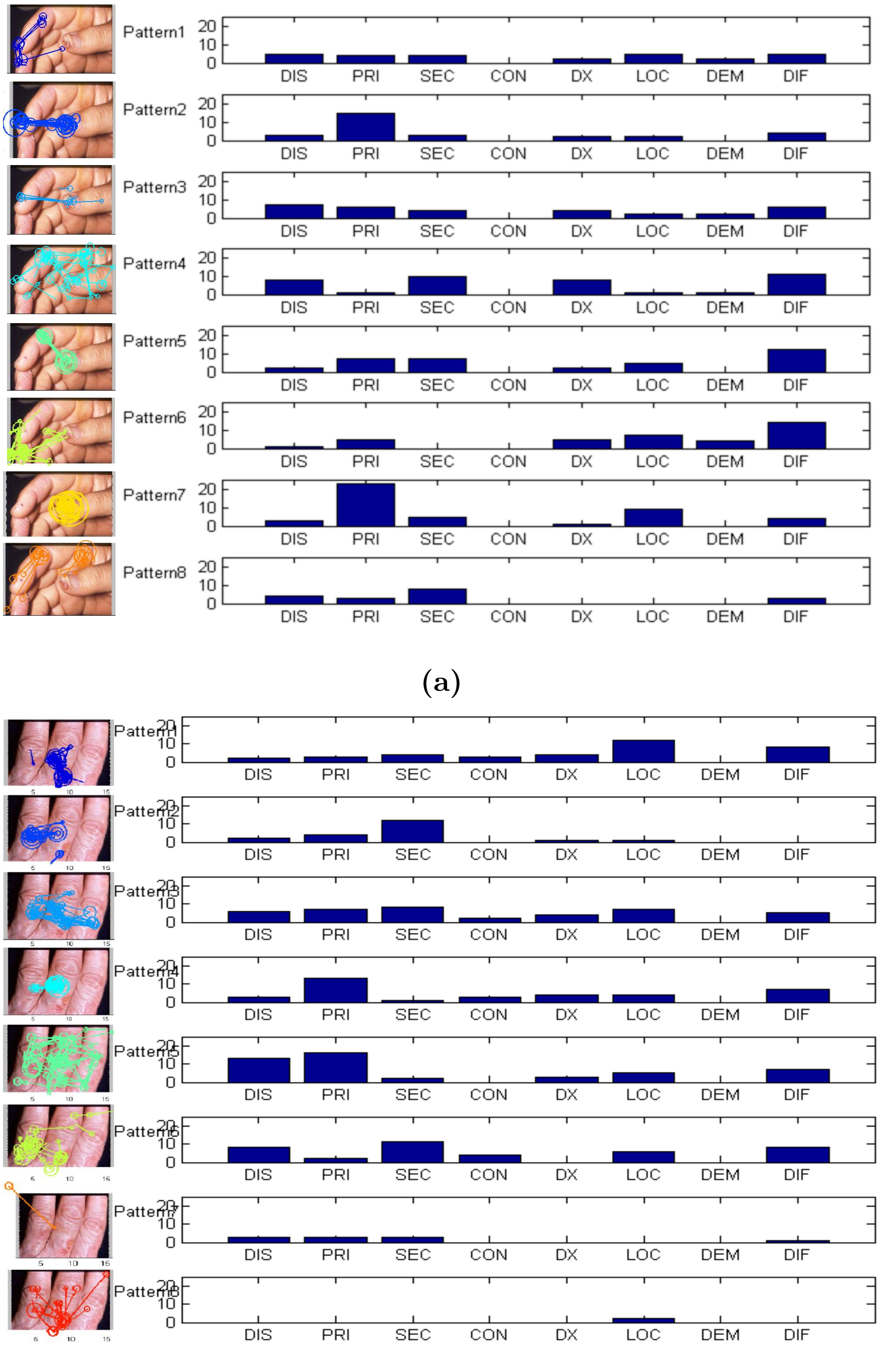
(a) Nine inferred eye movement patterns from the attending group. We illustrate the eye movement sequences, the transition matrices, and the color-coded patterns.

(b) Four inferred eye movement patterns from the residents.

(c) Ten inferred eye movement patterns from the novices.

Figure 6.5: The eye movement patterns of the three expertise-specific groups.

the thought units, including distribution (DIS), secondary morphology (SEC), diagnosis (DX) and differential diagnosis (DIF). For example, the scanning pattern coupled with thought unit DX is possibly related to confirmation of secondary findings to support or rule out diagnostic hypotheses.



(a)

(b)

Figure 6.6: Analysis of the correspondence between eye movement patterns and thought units for the two example images. For each pattern we plotted the counts of fixations which are labeled as the corresponding thought units. The pattern numbering is consistent with previous figures.

7

Dermatological Image Understanding

7.1 Image Understanding

There has been significant progress in automatic algorithms for image understanding (123, 124, 125, 126, 127, 128, 129, 130, 131), as shown in Figure 2.2 (a). However, when the cues in images are not sufficient to generate a good interpretation automatically, active learning methods are necessary to incorporate human perceptual capability into this process (132, 133, 134, 135, 136), as shown in Figure 2.2 (b). The idea of active learning is that automatic machine learning algorithms can achieve greater accuracy with fewer training labels by querying the user to provide support for the uncertain elements (137).

Besides locating and identifying the objects of interest in an image by bounding boxes or image segments with semantic labels (123, 124), recent image understanding studies also aim at exploring the underlying scene structure by estimating a qualitative 3D layout of the scene to recover the spatial relationships among multiple objects in the original 3D space (125, 126, 127, 128). These geometric approaches approximate the 3D space by planar surfaces or volumes from monocular images and some of them extend the idea to combine global consistency constraints (129). Dynamic 3D scene reconstruction is another focus. Various computation approaches such as the Markov random field and generative non-parametric graphical models are developed to robustly infer the 3D layout of roads, the locations of buildings, as well as dynamic traffic in the scene (130, 131).

There is significant success with the above automatic algorithms. However, human interaction becomes critical when key information such as strong edges and lines cannot be detected easily (135, 138). To borrow human perceptual power, active learning was proposed and benefited a broad range of computer vision applications (132, 133, 134, 135, 136). Essentially, the advantage of active learning methods relies on combining the human capability of image understanding with rich information from images using machine learning approaches. This is particularly important when understanding images requires domain expertise and rich background knowledge. Some active learning studies attempted to maximize the knowledge gain from users while valuing their effort (132). Others strived to simplify human interaction by fully utilizing the automatic algorithms and providing intuitive scribbles (134, 135).

Image understanding in knowledge-rich, visual domains, such as image inspection, is challenging, since complex perceptual and conceptual processing are engaged to transform image pixels into meaningful contents at semantic level (36). Active learning methods via manually marking and annotating become not only labor intensive for experts but also ineffective because of the variability and noise of experts' performance (41, 135). To address this problem, we propose to combine perceptual expertise as an effortless yet valuable cognitive resource into active learning methods of image understanding. This requires a means to extract and represent experts' perceptual expertise in a form that is ready to be applied in active learning schemes, as shown in Figure 2.2 (c).

When expanding the analysis to multiple images, we discover several basic yet distinctive types of patterns shared across multiple images which we term as *signature patterns* with respect to the patterns' fixation duration and saccade amplitude.

7.2 Signature Pattern Recognition

We define a type of signature patterns by three criteria. First, its self-transition probability, which is indicated by the transition matrix, is no less than the median 0.65, so the signature patterns are stably retained by attending group. Second, it manifests clear diagnostic regions, for example pattern 7 in Figure 6.2 corresponds to a long fixation duration on the primary abnormality. Third, the temporal-spatial properties of signature pattern exemplars within each type are similar but distinctive from other types, which is elaborated in Figure 7.1. The other discovered patterns are not identified as signature patterns because they lack one or more of the three criteria. In the illustrated case in Figure 6.2, there are three patterns recognized as the signature patterns. Pattern 2 and Pattern 5 are characterized by fixations switching back and forth between the primary and the different secondary abnormalities with long saccade amplitudes and relatively short fixation durations. These patterns suggest that subjects compare and associate the two types of abnormalities. Pattern 7 is characterized by a series of long-duration fixations only on the primary abnormality with extremely short saccades. This pattern suggests that subjects fixate on the primary abnormality to make a diagnosis.

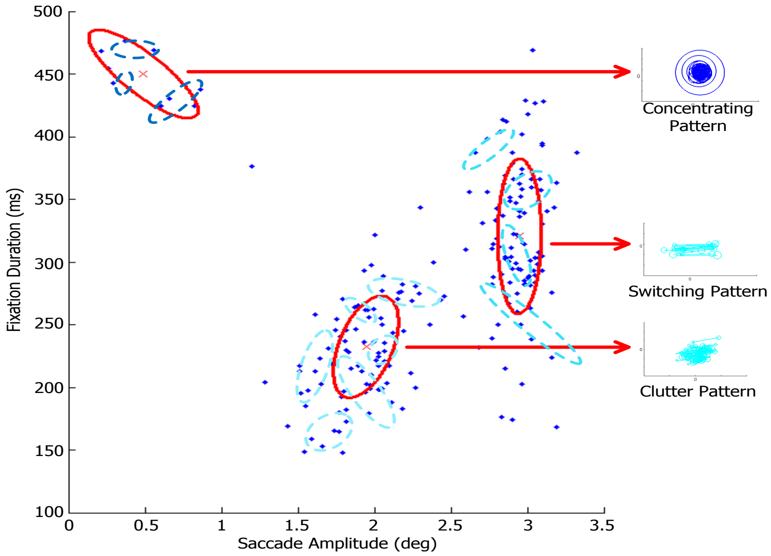


Figure 7.1: Distinctive temporal-spatial properties of 217 fixation-saccade pairs from 12 exemplars forms the three types of signature patterns. Each blue dot represents one eye movement unit from a signature pattern exemplar. The exemplars are indicated by dash-line Gaussian emission distributions estimated from our model. Both eye movement units and their corresponding exemplars are projected from a four-dimension space (including x-y coordinate, fixation duration and saccade amplitude) onto this space. The signature patterns are characterized by a three-component Gaussian mixture. The one on the upper left represents *Concentrating Pattern*, the one on the right captures *Switching Pattern*, and the one on the lower middle represents *Clutter Pattern*. For each type, we project the units back into x-y coordinate space centered on the origin and visualize them on the right side of the main diagram.

Based on the eye movement patterns generated from our model over fifty images, we are able to specify three types of signature patterns. The first type is

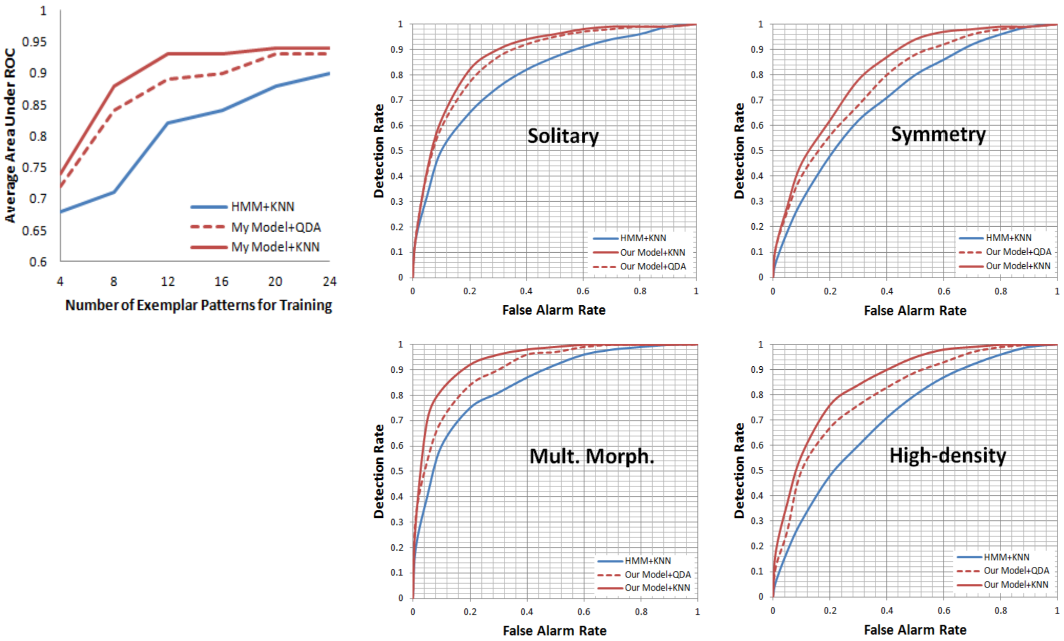


Figure 7.2: ROC curves summarizing categorization performance for the four perceptual categories. Left: Area under average ROC curves for different numbers of exemplar patterns. Right: We compare our model using two different classification techniques with canonical Hidden Markov Models.

named *Concentrating Pattern* which is characterized by a series of long-duration fixations and short-amplitude saccades usually fixating on primary abnormalities. The second is the *Switching Pattern* characterized by a series of relatively short-duration fixations and long-amplitude saccades usually switching back and forth between two abnormalities. And the third is *Clutter Pattern* characterized by a series of shorter fixations and relatively long saccades usually scanning within localized abnormal regions. To quantify the temporal-spatial properties of the three types of signature patterns, we illustrate some of their exemplars in Figure 7.1.

The estimation of the signature patterns based on their exemplar features can be solved using different classification techniques. Since Gaussian mixture is one intuitively appropriate tool to describe the distributions of these signature patterns according to Figure 7.1, we first adopt quadratic discrimination analysis (QDA) by assuming a simple parametric model for the densities of the temporal-spatial properties of the eye movement units. A training set includes 217 eye movement units of 12 exemplar patterns from 10 images, which are shown in Figure 7.1. We test the validity of the classifier through comparing the image categorization performance based on QDA with K nearest neighbors (K-NN) and experts' performance.

7.3 Perceptual Category Specification

Based on our consulting dermatologist's suggestion, we propose four broad perceptual categories in terms of lesion distribution and configuration. We further determine the associations between the combinations of the exhibitions of these three types of signature patterns and the four specified categories:

- If the set of eye movement patterns exhibited on an image only includes *Concentrating Patterns*, the image is categorized as *Localized* which means that the image contains a solitary lesion as primary abnormality.

- If the set of eye movement patterns exhibited on an image solely includes *Switching Patterns*, the image is categorized as *Symmetrical* which means that the lesions in the image are symmetrically distributed.

- If the set of eye movement patterns exhibited on an image includes both *Concentrating Patterns* and *Switching Patterns*, the image is categorized as *Multiple Morphologies* which means that the lesions in the image belong to

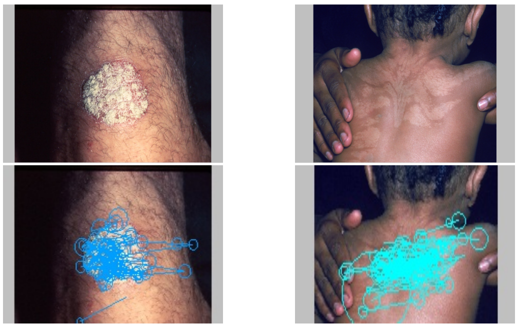


Figure 7.3: Two false positive cases. The left panel shows an image labeled as *Localized* lesion with *Clutter Pattern* recognized on it. The right panel shows a case labeled as *symmetric* lesion with *Clutter Pattern* recognized on it. Image used with permission from Logical Images, Inc

different dermatological morphologies and usually one lesion is identified as primary abnormalities and the other are secondary ones.

- If the set of eye movement patterns exhibited on an image includes *Clutter Patterns*, the image is categorized as *High-Density Lesions* which means that the image contains multiple lesions distributed in either scattered or clustered manner.

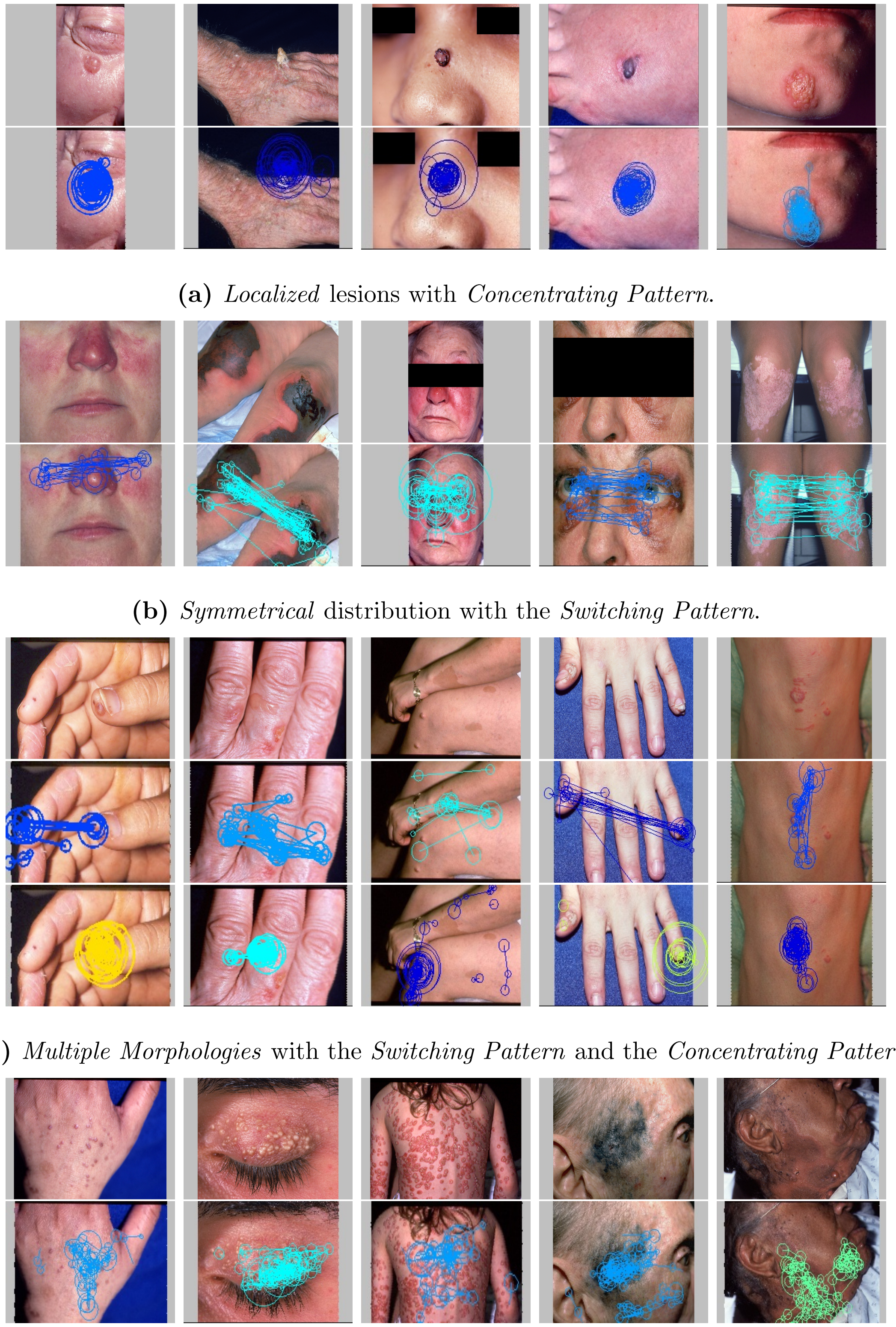
According to the signature patterns recognized on the images, we can catalogue the images into the four categories as shown in Figure 9 (a)-(d).

To measure the performance of our image categorization approach, we conduct an experiment following the same procedure by recruiting another ten dermatologists and using a different set of forty dermatological images as stimuli. These images are also randomly selected from a dermatological image database. Our three consulting dermatologists achieve consensus to categorize the forty images into the four perceptual categories. We use 232 estimated eye movement patterns on these images and the ones from the previous experiment as a testing set. In Figure 7.2 (a), we examine categorization performance given training sets containing between 4 and 24 exemplars. We assume each eye movement sequence exhibits the same set of patterns in order to implement the canonical HMMs. We see that our model lead to significant improvements in categorization performance, particularly when few training exemplars are available. The highest accuracy is achieved on detection of the *Multiple Morphologies* category. This may be caused by the requirement of detections of the two different signature patterns to determine the varied distributions and significance of the lesions. The difference between *Multiple Morphologies* images and *Symmetrical* images is that the eye movement patterns exhibited on the latter do not contain *Concentrating Pattern*. This is because the symmetrical visual-spatial structures imply that lesions are equivalent important without single primary one for the subjects to concentrate their focus on as shown in Figure 6.4 (b). Since the specifications of signature patterns are heuristic, we may be able to improve the categorization performance by identifying extra meaningful and distinctive eye movement patterns, and these extra patterns may also lead to image categorization at a finer detailed level.

The difference between *Multiple Morphologies* images and *Symmetrical* images is that the eye movement patterns exhibited on the latter do not contain *Concentrating Pattern*. This is because the symmetrical visual-spatial structures imply that lesions are equivalent important without single primary one for the subjects to concentrate their focus on as shown in Figure 6.4 (a) and (b). Since the specifications of signature patterns are determined heuristically, we may be able to improve the categorization performance by identifying additional meaningful and distinctive eye movement patterns, and these extra patterns may also lead to image categorization at a finer detailed level.

Since the dermatological images are collected for future diagnosis, and training purposes, the dermatologists took them in a particular way. They tend to center primary abnormalities and preserve as much related contextual information as possible, such as patients' demographic information, body parts, lesion size and so on. Nonetheless, these high-resolution images have complex backgrounds, and large appearance variations for luminance and camera angles. These factors cause some false alarms, as shown in Figure 7.3. In particular, photographic scales of some lesions in the images tend to influence our model's performance. For instance, the localized lesions are at large scale in some images, leading to cluttered eye movement patterns rather than concentrating ones as shown in Figure 7.3 (a). In another case shown in Figure 7.3 (b) there is an angle between the camera and the patient's back, so the symmetric shape lesions are skewed in the image. When dermatologists are examining this image, they tend to focus on the half of the lesion that are closer. This leads to a *Clutter Pattern* instead of a *Symmetric Pattern*. Since both the number of fixations and their durations are indicative of the depth of information processing associated with the particular image regions, the exhibition of *Concentrating Pattern* usually corresponds to a localized primary abnormality as shown in Figure 6.4 (a) and (c). The saccade amplitudes of *Switching Pattern* and *Clutter Pattern* inform both the image visual-spatial structures (symmetry) as in Figure 6.4 (b) and distributions of multiple abnormalities (primary abnormality versus secondary abnormality) as in Figure 6.4 (c).

Note that the different viewing times of dermatologists yield length-varying eye movement sequences. Since each sequence is modeled with one HMM separately, the emission distributions of which group multiple fixation-saccadic units into one pattern exhibited repeatedly. Thus longer sequence means that its corresponding longer HMM draws more pattern samples from the prior distribution, so besides containing more repeated common patterns, it likely has some unique patterns.



(a) *Localized* lesions with *Concentrating Pattern.*

(b) *Symmetrical* distribution with the *Switching Pattern.*

(c) *Multiple Morphologies* with the *Switching Pattern* and the *Concentrating Pattern.*

(d) *High-density Lesions* with the *Clutter Pattern.*

Figure 6.4: Example images with the signature patterns are illustrated.

8

Conclusions

8.1 Overview of the Computational Framework

The fundamental computational problem of this thesis is how to infer and represent perceptual skill based on impoverished human behavioral data, and how to use manifested perceptual skill to advance image understanding in domain of expertise. I then proposed a solution to combine hierarchical stochastic processes with dynamic models based on the principles of non-parametric Bayesian inference. This computational framework has four major components, which we illustrated on studying the perceptual and conceptual processing of three expertise-specific groups of subjects.

The first component is to assume a *hypothesis space* of candidate extensions for the structure to be learned. In a continuous measurable feature space of behaviors, such as the eye movements, a natural hypothesis space consisted of all potential statistical regularities in that space. These hypotheses can be thought of as a possible "micro-rule" for generalizing the eye movements across multiple stimuli, or alternatively, as a candidate "feature" that could distinguish expertise-specific visual behaviors.

The second component in the framework is to formulate and incorporate *prior knowledge* to each element of the hypothesis space. The prior encodes our beliefs about which subset of hypotheses are the most likely candidates for new observed eye movement data in general, independent of any observed data. In some sense, the hypothesis space itself is a derivation of the prior. In some extreme cases, excluding logically possible hypotheses from our hypothesis space is equivalent to including them but assigning them a prior probability of zero. Over a hypothesis space of eye movement patterns, our prior distribution might embody the knowledge that some patterns tend to be very common within a group of subjects with the same expertise level, and thus give preference to eye movement patterns with approximately that property. Over a hierarchy of expertise-specific groups, our prior assumed that these eye movement patterns map onto related yet distinctive groups of subjects' eye movements.

The third component is a *generative model* of the observed eye movement sequences, which allows us to evaluate hypotheses base on their likelihood of giving rise to the eye movements we observed. The most basic of the Markov processes is the hidden Markov model (HMM), which assumes that the data can be modeled as conditionally independent given an underlying discrete-valued Markov sequence. Motivated by our particular application of human eye movements, we relax its assumption from conditionally independent observations to conditionally linear dynamics using autoregressive process.

Finally, the actual generalization of eye movement patterns is determined by the *posterior probability* and the principle of *hypothesis averaging*. The posterior probability of each subset of patterns is equal to the product of its prior probability and dynamic likelihood. This gives the rational degree of belief in each pattern subset as a function of both our prior knowledge about more or less reasonable pattern candidate extensions and the statistical information carried by the observed eye movement sequences. When the posterior is spread out broadly over many possible eye movement patterns, the predictive procedure is essentially an average of all of these possible "micro-rules". When the posterior is concentrated on a single subset of eye movement patterns, the weighted average will only focus on the extension of that one best hypothesis.

Although we can analyze the four components separately, to really understand the framework and its implication, it is necessary to see how these components interact as a function of the prior knowledge which the subjects bring into their task and their behaviors which are observed. I explored these issues over the course of dermatologists' examining medical images and reached several conclusions.

8.2 Summary of the Major Contributions

Human intelligence is valuable in terms of providing directions of building smarter machines. How to extract and represent such intelligence and incorporate it into computational approaches to provide effective computing solutions for digital image retrieval is challenging.

One of the obstacles is that the image-feature-based visual similarity does not necessarily correspond to subjective perceptual similarity. For instance, a query image is used in traditional content-based image retrieval as an example to illustrate a user's information need. A critical retrieval step is to evaluate the significance of the visual cues in the query image so as to effectively interpret it and extrapolate the user's intentions. However, in knowledge-rich domains such as image-based medical diagnosis it is impossible to evaluate visual cues of medical images without the inputs of domain experts' perceptual and conceptual reasoning processes.

Through our studies we can summarize subjects' perceptual skill in an objective and automatic manner. Understanding of the relationships between eye movement patterns, image properties, and semantic concepts will be beneficial from at least several viewpoints. First, since eye movement data are deployed in a sequential manner, exploration of their temporal characteristics will provide us deeper understanding of the influence of expertise on perceptual processing. Second, compared to the observed eye movement data, the extracted patterns have semantic features and can serve as a more robust and consistent human capability measure. Third, instead of annotation, our learned perceptual skill can be embedded into an active learning scheme as a more efficient approach to integrate human knowledge into image understanding.

Solely behavioral variables from task manipulations, such as response time or accuracy, are insufficient to determine whether a particular cognitive process is engaged or whether a particular cognitive architecture theory is correct. Since visual attention, as a selective dynamic cognitive process, is dominated by knowledge, interest, and expectations of the scene (1, 23), it is possible to acquire insight into some aspects of subjects' interests or cognitive strategies by analyzing their eye movement sequences while they are pursuing certain tasks in domains of expertise where perceptual skills are paramount. One key step to manifest perceptual skill and uncover underlying cognitive processes is to discover expertise-specific perceptual viewing behaviors and differentiate the stereotypical and idiosyncratic behavioral patterns that characterize a group of subjects at the same training level. Addressing this problem requires segmenting an eye movement sequence into a set of time intervals that have a useful interpretation, as well as summarizing the commonality of eye movement patterns shared within and between expertise-specific groups. Furthermore, these meaningful patterns enable us to uncover time-evolving properties of subjects' perceptual reasoning processes and to understand images at a domain-knowledge level.

Our approach identified expertise-specific eye movement patterns that exist over time. Dermatology images and experts are an appropriate test-bed, but we can also apply our approach to other problem domains. We elaborate 50 images and delivered an extensive discussion on two illustrated cases. As our future work, we will use the discovered meaningful patterns to parse corresponding image features, which correspond to deep perceptual skills (as opposed to detailed surface features only), and that, accordingly, have potential to fill the semantic gap described at the paper's beginning.

We successfully discover certain aspects of experts' domain-specific knowledge by summarizing stereotypical and idiosyncratic behavioral patterns from their eye movements while examining medical images. The domain-specific knowledge unveils the meaning and significance of the visual cues as well as the relations among functionally integral visual cues without segmentation or processing of individual objects or regions. This will benefit the traditional pixel-based statistical methods for image understanding by evaluating perceptual meanings and relations of the image features which spatially correspond to the eye movement patterns. This combination of expert knowledge and image features will help us to generalize our approach to images for which there is no experts' eye movements recorded.

9

Future Work

Although Gibbs sampler provides theoretical guarantees of accuracy, mixing rates on large datasets can often be slow, and difficult to characterize in general. Variational methods provide an alternative by a fast deterministic approximation to posteriors with an optimization criterion that can be easily utilized to assess convergence.

Another keen interest is online learning. Gaze-contingent applications require inferences to be made sequentially as eye movement data arrive. The batch processing algorithms may also be impractical for long time series. The standard issue of a progressively impoverished particle representation introduces challenges that are interesting to explore in the future.

We obtain certain aspects of experts' domain-specific knowledge by summarizing their perceptual skills from their eye movements while diagnosing images. The domain-specific knowledge unveils the meaning and significance of the visual cues as well as the relations among functionally integral visual cues without segmentation or processing of individual objects or regions. This will benefit the traditional pixel-based statistical methods for image understanding by evaluating perceptual meanings and relations of the image features which spatially correspond to the eye movement patterns. This combination of expert knowledge and image features allows us to generalize our approach to images on which there is no record of experts' eye movements.

Many potential new applications would require estimation of the subjects' eye movement patterns in real time based on the observed eye movements. Here, we use sequential Monte Carlo methods to filter the current latent eye movement pattern based on the learned library of eye movement patterns from experienced dermatologists.

In the future, we will also design and develop a prototype of an adaptive image retrieval system. This system will retrieve a collection of medical images based on the inferred informational needs from users' eye movements. Our model will be integrated into this system as an component to evaluate users' informational interests through the real-time estimation of meaningful eye movement patterns.

Besides, we attempt to project the meaningful eye movement patterns from their spatial-temporal space into image feature space. This will facilitate us to identify valuable diagnostic image features and generalize eye movement patterns over the different dermatological images. Evaluation of a subject's expertise level is another future application in medical training. We also intend to evaluate diagnostic processes given a subject's visual interaction with test images through calculating the model's posterior probability. Compared to simply calculating diagnosis error rates to evaluate expertise level, our approach can unveil which diagnostic reasoning steps lead to wrong diagnosis and the possible cognitive factors such as misconception, miscategorization and misperception, and form the basis of support systems.

References

[1] John M. Henderson and George L. Malcolm. **Searching in the Dark Cognitive Relevance Drives Attention in Real-world Scenes**. *Psychonomic Bulletin and Review*, **16**(5):850–856, 2009. 1, 9, 126

[2] M. Smuc, E. Mayr, and F. Windhager. **The Game Lies in the Eye of the Beholder: The Influence of Expertise on Watching Soccer**. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 1631–1636, Austin, TX, 2010. Lawrence Erlbaum Associates. 1, 13

[3] Merim Bilalic, Robert Langner, Michael Erb, and Wolfgang Grodd. **Mechanisms and neural basis of object and pattern recognition: a study with chess experts**. *J. Exp Psychol Gen*, **139**:728–742, 2010. 1, 13

[4] Eugene Levin, Alexander Zarnowski, Cheryl A. Cohen, and Robert Liimakka. **Human Centric Approach to Inhomogenious Geospatial Data Fusion and Actualization**. In *Proc. of ASPRS*, pages 1–5, 2010. 1, 13

[5] Jason S. McCarley, Arthur F. Kramer, Christopher D. Wickens, Eric D. Vidoni, and Walter R. Boot. **Visual Skills in Airport Security Screening**. *Psychological Science*, **15**:302–306, 2004. 1, 13

[6] Elizabeth Krupinski, Allision Tillack, Lynne Richter, Jeffrey Henderson, Achyut Bhatacharyya, Katherine Scott, Anna Graham, Michael Descour, John Davis, and Ronald Weinstein. **Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience**. *Journal of Human Pathology*, **37**(12):1543–1556, 2006. 1, 13, 14

[7] David Manning, Susan Ethell, Tim Donovan, and Trevor Crawford. **How do radiologists do it? The influence of experience and training on searching for chest nodules**. *Journal of Radiography*, **12**(2):134–142, 2006. 1, 13

[8] Philip G. Zimbardo and Richard J. Gerrig. *Psychology and Life*. HarperCollins, New York, 1996. 3, 7

[9] Thomas P. Habif, James L. Campbell Jr., M. Shane Chapman, James GH. Dinulos, and Kathryn A. Zug. *Skin Disease Diagnosis and Treatment*. Elsevier Mosby, 2005. 4

[10] Rui Li, Jeff Pelz, Pengcheng Shi, Cecilia Ovesdotter Alm, and Anne R. Haake. **Learning Eye Movement Patterns for Characterization of Perceptual Expertise**. In *ETRA*, pages 393–396, 2012. 5

[11] Rui Li, Jeff Pelz, Pengcheng Shi, and Anne R. Haake. **Learning Image-Derived Eye Movement Patterns to Characterize Perceptual Expertise**. In *CogSci*, pages 190–195, 2012. 5, 69

[12] Rui Li, Pengcheng Shi, and Anne R. Haake. **Image Understanding from Experts' Eyes by Modeling Perceptual Skills of Diagnostic Reasoning Processes**. In *IEEE Conf. on Computer Vison and Pattern Recognition*, 2013. 5

[13] Rui Li, Evelyn Rozanski, and Anne R. Haake. **Framework of a Real-Time Adaptive Hypermedia System**. In *workshop of SIGIR*, pages 1–5, 2009. 6

[14] Rui Li, Sai K. Mulpuru, Cara F. Calvelli, Jeff B. Pelz, Pengcheng Shi, and Anne R. Haake. **A Human-Centered Content-Based Image Retrieval System**. In *Poster of AMIA*, 2010. 6

[15] Rui Li, Evelyn Rozanski, and Anne R. Haake. **A User Model to Predict Web Page Viewing Behavior**. In *Adjuct Proc. of UMAP*, pages 19–21, 2009. 6

[16] Rui Li, Preethi Vaidyanathan, Sai K. Mulpuru, Cara F. Calvelli, Jeff B. Pelz, Pengcheng Shi, and Anne R.

Haake. **Human-Centric Approaches to Image Understanding and Retrieval**. In *West New York Image Processing Workshop*, 2010. 6

[17] Wilson McCoy, Cecilia Ovesdotter Alm, Cara Calvelli, Rui Li, Jeff B. Pelz, Pengcheng Shi, and Anne Haake. **Annotation Schemes to Encode Domain Knowledge in Medical Narratives**. In *Language Annotation Workshop VI of ACL*, page accepted, 2012. 6, 79

[18] Corey Engelman, Rui Li, Jeff Pelz, Pengcheng Shi, and Anne Haake. **Exploring Interaction Modes for Image Retrieval**. In *Proc. of NGCA*, pages 1–5, 2011. 6

[19] Christof Koch and Shimon Ullman. **Shifts in selective visual attention: towards the underlying neural circuitry**. *Human Neurobiology*, **4**. 10, 12

[20] Antonio Torralba, Aude Oliva, Monica S. Castelhano, and John M. Henderson. **Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search**. *Psychological Review*, **113**(4):766–786, 2006. 10

[21] Wei Wang, Cheng Chen, Yizhou Wang, Tingting Jiang, Fang Fang, and Yuan Yao. **Simulating Human Saccadic Scanpaths on Natural Images**. In *CVPR*, pages 441–448, 2011. 10

[22] Marianne DeAngelus and Jeff B. Pelz. **Top-down control of eye movements: Yarbus revisited**. *Visual Cognition*, **17**:790–811, 2009. 10

[23] Monica S. Castelhano, Michael L. Mack, and John M. Henderson. **Viewing task influences eye movement control during active scene perception**. *Journal of Vision*, **9**(3):1–15. 10, 11, 126

[24] von Helmholtz. *Handbuch der physiologischen optik*. Leipzig: Voss, 1867. 11

[25] Laurent Itti, Christof Koch, and Ernst Niebur. **A Model of Saliency-Based Visual Attention for Rapid Scene Analysis**. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, **20**(11):1254–1259. 11, 12

[26] Keith Rayner. **Eye Movements in Reading and Information Processing: 20 Yeas of Research**. *Psychological Bulletin*, **124**(3):372–422. 11

[27] Sebastian Pannasch and Boris M. Velichkovsky. **Distractor effect and saccade amplitudes: further evidence on different modes of processing in free exploration of visual images**. *Visual Cognition*, **17**(6). 11

[28] Pieter J. A. Unema, Sebastian Pannasch, Markus Joos, and Boris M. Velichkovsky. **Time course of information processing during scene perception: the relationship between saccade amplitued and fixation duration**. *Visual Cognition*, **12**(3):473–494. 11

[29] Anne M. Treisman and Garry Gelade. **A feature-integration theory of attention**. *Cognitive Psychology*, **12**. 12

[30] Laurent Itti and Christof Koch. **Computational Modelling of Visual Attention**. *Nature Reviews, Neuroscience*, **2**(4):194–203. 12

[31] Jixu Chen and Qiang Ji. **Probabilistic Gaze Estimation Without Active Personal Calibration**. In *Proc. of CVPR*, pages 609–616, 2011. 12

[32] S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Gurin-Dugu. **Modeling Spatio-Temporal Saliency to Predict Gaze Direction for Short Videos**. *International Journal of Computer Vision*, pages 231–243, 2009. 12

[33] Aude Oliva, Antonio Torralba, Monica S. Catelhano, and John M. Henderson. **Top-Down Control of Visual Attention in Object Detection**. In *Proc. ICIP*, pages 253–256. ACM, April 2003. 12

[34] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. **SUN: A Bayesian Framework for Saliency using Natural Statistics**. *Journal of Vision*, **8**(32):1–20, 2008. 12

[35] Yuan-Chi Tseng and Andrew Howes. **The Adaptation of Visual Search Strategy to Expected Information Gain**. In *Proc. CHI*, pages 1075–1084. ACM Press. 12

[36] R. Hoffman and M. S. Fiore. **Perceptual (Re)learning : A Leverage Point for Human-Centered Computing**. *J. Intelligent Systems*, **22**(3):79–83, 2007. 13, 105, 113

[37] Thomas J. Palmeri, Alan C-N. Wong, and Isabel Gauthier. **Computational approaches to the development of perceptual expertise**. *TRENDS in Cognitive Sciences*, **8**(8):378–386, 2004. 13

[38] Mikael Hans Sodergren, Felipe Orihuela-Espina, James Clarke, Ara Darzi, and Guang-Zhong Yang. **A Hidden Markov Model Model-based Analylsis Framework Using Eye Tracking Data to Characterise Re-Orientation Strategies in Minimally Invasive Surgery**. *Journal of Cognitive Processes*, pages 275–3283, 2010. 13, 18

[39] Lucia Ballerini, Xiang Li, Robert B. Fisher, and Jonathan Rees. **A Query-by-Example Content-Based Image Retrieval System of Non-Melanoma Skin Lesions**. In *MCRCDS workshop of MICCAI*, pages 312–319, New York, 2009. Springer Press. 14

[40] James W.Woods, Charles A. Sneiderman, Kamran Hameed, Michael J. Ackerman, and Charlie Hatton. **Using UMLS Metathesaurus Concepts to Describe Medical Images Dermatology Vocabulary**. *J. Computers in Biology and Medicine*, **36**:89–100, 2006. 14

[41] S. Gordon, S. Lotenberg, J. Jeronimo, and H. Greenspan. **Evaluation of Uterine Cervis Segmentations using Ground Truth from Multiple Experts**. *J. Computerized Medical Imaging and Graphics*, **33**(3):205–216, 2009. 14, 113

[42] Antti Ajanki, David R. Hardoon, Samuel Kaski, Kai Puolamki, and John Shaw-Taylor. **Can eyes reveal interest? Implicit queries from gaze patterns**. *User Model User-Adapted Interaction*, **19**(4):307–339, 2009. 15

[43] Tomasz D. Loboda, Peter Brusilovsky, and Jorg Brunstein. **Inferring Word Relevance from Eye-movements of Readers**. In *Proc. IUI*, pages 175–184. ACM Press, 2011. 15

[44] Marcelo G. Armentano and Analia A. Amandi. **Recognition of Uer Intentions for Interface Agents with Variable Order Markov Models**. In *Proc. of UMAP*, pages 173–184, 2009. 15

[45] Kai Puolamaki, Antti Ajanki, and Samuel Kaski. **Learning to Learn Implicit Qeries from Gaze Patterns**. In *Proceedings of International Conference on Machine Learning*, pages 760–767, 2008. 15

[46] Georg Buscher, Andreas Dengel, and Ludger van Elst. **Query Expansion Using Gaze-Based Feedback on the Subdocument Level**. In *Proc. SIGIR*, pages 387–394. ACM Press, 2008. 15

[47] Alejandro Jaimes, Jeff Pelz, Tim Grabowski, Jason Babcock, and Shih-Fu Chang. **Using Human Observers' Eye Movements in Automatic Image Classifiers**. In *Proc. of SPIE*, pages 373–384, 2001. 15

[48] Jeremy Goecks and Jude Shavlik. **Learning Users' Interests by Unobtrusively Observing Their Normal Behavior**. In *Proc. IUI 2000*, pages 129–132. ACM Press, 2000. 15

[49] Yoshinori Hijikata. **Implicit User Profiling for On Demand Relevance Feedback**. In *Proc. IUI*, pages 198–205. ACM Press, 2004. 15

[50] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. **Accurately Interpreting Clickthrough Data as Implicit Feedback**. In *Proc. SIGIR 2005*, pages 154–161. ACM Press, 2005. 15

[51] Fabio Pianesi, Massimo Zancanaro, Bruno Lepri, and Alessandro Capelletti. **A Multimodal Annotated Corpus of Consensus Decison Making Meetings**. *Language Resources and Evaluation*, **41**(3), 2007. 15

[52] Stephan Raaijmakers, Khiet Truong, and Theresa Wilson. **Multimodal Subjectivity Analysis of Multiparty Conversation**. In *Proc. CEMNLP 2008*, pages 466–477. Association for Computational Linguistics, 2008. 15

[53] Theresa Wilson and Gregor Hofer. **Using Linguistic and Vocal Expressiveness in Social Role Recognition**. In *Proc. IUI 2011*, pages 419–422. ACM Press, 2011. 15

[54] Junya Morita, Kazuhisa Miwa, T Kitasaka, Kensaku Mori, Yasuhito Suenaga, Shingo Iwano, Mitsuru Ikeda, and T Ishigaki. **Interactions of Perceptual and Conceptual Processing: Expertise in Medical Image Diagnosis**. *J. Human-Computer Studies*, **66**(5):370–390, 2008. 16

[55] Vimla L. Patel, Jose E. Arocha, and David R. Kaufman. **A primer on aspects of cognition for medical informatics**. *J Am Med Inform Assoc.*, **8**:324–343, 2001. 16

[56] Laura Dempere-Marco, Xiaopeng Hu, and Guang-Zhong Yang. **A novel framework for the analysis of eye movements during visual search for knowledge gathering**. *Cognitive Computation*, **3**:206–222, 2011. 16

[57] Joseph H. Goldberg and Jonathan I. Helfman. **Scanpath Clustering and Aggregation**. In *ETRA*, pages 227–234, 2011. 17

[58] Andrew T. Duchowski, Jason Driver, Sheriff Jolaoso, Beverly N. Ramey, and Ami Robbins. **Scanpath Comparison Revisited**. In *ETRA*, pages 219–226, 2010. 17

[59] Anthony Santella and Doug DeCarlo. **Robust Clustering of Eye Movement Recordings for Quantification of Visual Interest**. In *ETRA*, pages 27–34, 2004. 17, 18

[60] Matt Feusner and Brian Lukoff. **Testing for Statistically Significant Differences between Groups of Scan Patterns**. In *ETRA*, pages 43–46, 2008. 17

[61] Julia M. West, Anne R. Haake, Evelyn P. Rozanski, and Keith S. Karn. **eyePatterns: Software for Identifying Patterns and Similarities Across Fixation Sequences**. In *ETRA*, pages 149–154, 2006. 17

[62] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. **It Depends on How You Look at It: Scanpath Comparison in Multiple Dimensions with MultiMatch, a Vector-based Approach**. *Behavior Reseasrch Methods*, 2012. 18, 19

[63] Laura Dempere-Marco, Xiao-Peng Hu, Stephen M. Ellis, David M. Hansell, and Guang-Zhong Yang. **Analysis of visual search patterns with EMD metric in normalized anatomical space**. *IEEE Transactions on Medical Imaging*, **25**(8):1011–1021. 18

[64] Raymond D. Rimey and Christopher M. Brown. **Controlling Eye Movements with Hidden Markov Models**. *Intl. J. of Computer Vision*, pages 47–65, 1991. 18

[65] Dario D. Salvucci. **Inferrring Intent in Eye-Based Interfaces: Tracing Eye Movements with Process Models**. In *Proc. of CHI*, pages 254–261, 1999. 18

[66] Matthew J. Beal, Zoubin Ghahramani, and Carl Edward Rasmussen. **The Infinite Hidden Markov Model**. In *Proc. of NIPS*, pages 577–584, 2002. 19

[67] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. **Hierarchical Dirichlet Processes**. *J. of the American Statistical Association*, **101(476)**:1566–1581, 2006. 19

[68] Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. **An HDP-HMM for Systems with State Persistence**. In *Proc. of ICML*, pages 312–319, 2008. 19

[69] Jurgen Van Gael, Yee Whye Teh, and Zoubin Ghahramani. **The Infinite Factorial Hidden Markov Model**. In *Proc. of NIPS*, pages 1697–1704, 2009. 19

[70] Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. **Sharing Features among Dynamical Systems with Beta Processes**. In *NIPS*, pages 549–557, 2009. 19

[71] B. de Finetti. *Funzione Caratteristica Di un Fenomeno Aleatorio*, chapter Meorie, Classe di Scienze Fisiche, Mathematiche e Naturali, pages 251–299. Academia Nazionale dei Lincei, 1931. 26

[72] David Heath and William Sudderth. **De Finetti's Theorem on Exchangeable Variables**. *The American Statistician*, pages 188–189, 1976. 26

[73] Jos M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. John Wiley, 2000. 26, 27, 29, 35, 36, 37, 38, 40, 42, 44, 89

[74] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis.* Chapman & Hall/CRC, 2004. 27, 29, 40, 42, 43, 44, 45, 63, 65, 67

[75] Lawrence D. Brown. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory.* John Wiley, 1978. 28, 30, 34, 38

[76] Ole E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory.* Institute of Mathematical Statistics, 1986. 28, 30, 34

[77] Martin J. Wainwright and Michael I. Jordan. **Graphical Models, Exponential Families, and Variational Inference**. *Machine Learning*, pages 1–305, 2004. 28, 30, 51, 62

[78] Bradley Efron. **The Geometry of Exponential Families**. *Annals of Statistics*, pages 362–376, 1978. 30

[79] Shun ichi Amari. **Information Geometry on Hierarchy of Probability Distributions**. *IEEE Transactions on Information Theory*, pages 1701–1711, 2001. 30, 34

[80] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* John Wiley, 1991. 30, 32

[81] Athanasios Papoulis. *Probability, Random Variable and Stochastic Processes.* McGraw Hill, 1991. 40, 44, 60, 81

[82] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* John Hopkins University Press, 1996. 48

[83] James W. Demmel. *Applied Numerical Linear Algebra.* SIAM, 1997. 48

[84] Mike West and Jeff Harrison. *Bayesian Forecasting and Dynamic Models.* Springer, 1997. 49

[85] Eugene Charniak. **Bayesian Networks without Tears**. *AI Magazine*, pages 50–63, 1991. 50

[86] Michael I. Jordan. **Graphical Models**. *Statistical Science*, pages 140–155, 2004. 50, 55

[87] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufman, 1988. 50, 51, 52, 60, 68

[88] Steffen L. Lauritzen. *Graphical Models.* Oxford University Press, 1996. 50

[89] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. *Understanding Belief Propagation and Its Generalizations*, chapter Exploring Artificial Intelligence in the New Millennium, pages 239–269. Morgan Kaufmann, 2003. 50

[90] Lawrence R. Rabiner. **A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition**. *Proceedings of the IEEE*, pages 257–286, 1989. 51, 54, 55

[91] David G. Forney. **The Viterbi Algorithm**. *Proceedings of the IEEE*, pages 268–278, 1973. 51, 57

[92] R. E. Kalman. **A New Approach to Linear Filtering and Prediction Problems**. *Journal of Badic Engineering*, pages 35–45, 1960. 51

[93] Robert G. Cowell, Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems.* Springer-Verlag, 1999. 52, 55

[94] David Heckerman. *Learning in Graphical Models*, chapter A Tutorial on Learning With Bayesian Networks, pages 301–354. MIT Press, 1999. 52, 68

[95] Brian D. O. Anderson and John B. Moore. *Optimal Filtering.* Prentice Hall, 1979. 54

[96] Arnaud Doucet, Nando de Freitas, Neil Gordon, and A. Smith. *Sequential Monte Carlo Methods in Practice.* Springer-Verlag, 2001. 54

[97] Thomas Kailath. **A View of Three Decades of Linear Filtering Theory**. *IEEE Transactions on Information Theory*, pages 146–181, 1974. 54, 55

[98] Sam Roweis and Zoubin Ghahramani. **A Unifying Review of Linear Gaussian Models**. *Neural Computation*, pages 305–345, 1999. 55

[99] Gregory F. Cooper. **The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks**. *Artificial Intelligence*, pages 393–405, 1990. 60, 62

[100] Steven M. Kay. *Fundamentals of Statistical Signal Processing.* Prentice Hall, 1993. 60

[101] Jose Luis Marroquin, Sanjoy K. Mitter, and Tomaso Poggio. **Probabilistic Solution of Ill-Posed Problems in Computational Vision**. *Journal of the American Statisticsal Association*, pages 76–89, 1987. 60, 61, 67, 68

[102] Brendan J. Frey and Nebjsa Jojic. **A Comparison of Algorithms for Inference and Learning in Probabilistic Graphical Models**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1392–1416, 2005. 60, 62, 68

[103] Richard Szeliski. **Bayesian Modeling of Uncertainty in Low-Level Vision**. *International Journal of Computer Vision*, pages 271–301, 1990. 60

[104] Solomon E. Shimony. **Finding MAPs for Belief Networks is NP-hard**. *Artificial Intelligence*, pages 399–410, 1994. 62

[105] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. **An Introduction to Variational Methods for Graphical Models**. *Machine Learning*, pages 183–233, 1999. 62

[106] D. J. C. MacKay. *Learning in Graphical Models*, chapter Introduction to Monte Carlo Methods, pages 175–204. MIT Press, 1999. 63, 67, 68

[107] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I. Jordan. **An Introduction to MCMC for Machine Learning**. *Machine Learning*, pages 5–43, 2003. 63, 64, 65, 67, 68

[108] Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1986. 64

[109] Stuart Geman and Donald Geman. **Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 721–741, 1984. 67, 68

[110] Alan E. Gelfand and Adrian F. M. Smith. **Sampling-Based Approaches to Calculating Marginal Densities**. *Journal of the American Statistical Association*, pages 398–409, 1990. 67, 68

[111] Jun S. Liu, Wing H. Wong, and Augustine Kong. **Covariance Structure and Covergence Rate of the Gibbs Sammpler with Various Scans**. *Journal of the Royal Statistical Society*, pages 157–169, 1995. 67, 68

[112] Gareth O. Roberts and Sujit K. Sahu. **Rate of Convergence of the Gibbs Sampler by Gaussian Approximation**. *Journal of the Royal Statistical Society*, pages 291–317, 1997. 68

[113] Wray L. Buntine. **Operations for Learning with Graphical Models**. *Journal of Artificial Intelligence Research*, pages 159–225, 1994. 68

[114] Hugh Beyer and Karen Holtzblatt. *Contextual design: defining customer-centered systems*. Morgan Kaufmann, 1997. 71

[115] P Boersma and D Weenink. **Praat: doing phonetics by computer (Version 5.1.05)**. `http://www.praat.org`. 80

[116] John Frank Charles Kingman. *Poisson Processes*. Oxford University Press, 1993. 86

[117] Michael I. Jordan. **Hierarchical Models, Nested Models and Completely Random Measures**. *Frontiers of Statistical Decison Making and Bayesian Analysis: In honors of James O. Bergers*, 2010. 86

[118] Nils Lid Hjort. **Nonparametric Bayes Estimators based on Beta Processes in Models for Life History Data**. *The Annals of Statistics*, pages 1259–1294, 1990. 86, 89

[119] Romain Thibaux and Michael I. Jordan. **Hierarchical Beta processes and the Indian Buffet Process**. *J. Machine Learning and Research*, **22**(3):25–31, 2007. 86, 94, 96

[120] Peter Mller and Fernando A. Quintana. **Nonparametric Bayesian Data Analysis**. *Statistical Science*, pages 96–110, 2004. 89

[121] Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. **Bayesian Nonparametric Methods for Learning Markov Switching Processes**. *IEEE Signal Processing Magazine*, pages 43–54, 2010. 94

[122] W. R. Gilks and P. Wild. **Adaptive Rejection Sampling for Gibbs Sampling**. *Journal of the Royal Statistical Society*, pages 337–348, 1992. 97

[123] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. **A Discriminatively Trained, Multiscale, Deformable Part Model**. In *CVPR*, pages 1–8, 2008. 112

[124] D. M. Gavrila and S. Munder. **Multi-cue pedestrian detection and tracking from a moving vehicle**. *IJCV*, **73**(1):41–59, 2007. 112

[125] Varsha Hedau, Derek Hoiem, and David Forsyth. **Recovering the Spatial Layout of Cluttered Rooms**. In *ICCV*, pages 1849–1856, 2009. 112, 113

[126] David C. Lee, Martial Hebert, and Takeo Kanade. **Geometric Reasoning for Single Image Structure Recovery**. In *CVPR*, pages 2136–2143, 2009. 112, 113

[127] Derek Hoiem, Alexei A. Efros, and Martial Hebert. **Recovering Surface Layout from an Image**. *IJCV*, **75**(1):151–172, 2007. 112, 113

[128] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. **Learning 3D scene structure from a single still image**. *PAMI*, **31**(5):824–840, 2009. 112, 113

[129] Abhinav Gupta, Alexei A. Efros, and Martial Hebert. **Blocks World Revisited : Image Understanding using Qualitative Geometry and Mechanics**. In *ECCV*, pages 482–496, 2010. 112, 113

[130] Andreas Geiger, Martin Lauer, and Raquel Urtasum. **A Generative Model for 3D Urban Scene Understanding from Movable Platforms**. In *CVPR*, pages 1945–1952, 2011. 112, 113

[131] David Crandall, Andrew Owens, Noah Snavely, and Dan Huttenlocher. **Discrete-Continuous Optimization for Large-Scale Structure from Motion**. In *CVPR*, pages 3001–3008, 2011. 112, 113

[132] Sudheendra Vijayanarasimhan, Prateek Jain, and Kristen Grauman. **Far-Sighted Actively Learning on a Budget for Image and Video Recognition**. In *CVPR*, pages 3035–3042, 2010. 112, 113

[133] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. **Active Learning with Gaussian Processes for Object Categorization**. In *ICCV*, pages 1–8, 2007. 112, 113

[134] Dhruv Batra, Adarsh Kowdle, and Devi Parikh. **iCoseg: Interactive Co-segmentation with Intelligent Scribble Guidance**. In *CVPR*, pages 3169–3176, 2010. 112, 113

[135] Adarsh Kowdle, Yao-Jen Chang, Andrew Gallagher, and Tsuhan Chen. **Active Learning for Piecewise Planar 3D Reconstruction**. In *CVPR*, pages 929–936, 2011. 112, 113

[136] Philippe Henri Gosselin and matthieu Cord. **Actively Learning Methods for Interactive Image Retrieval**. *IEEE Trans. on Image Processing*, **17**(7):1200–1211, 2008. 112, 113

[137] Burr Settles. **Active Learning Literature Survey**. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. 112

[138] Abhinav Gupta, Scott Satkin, Alexei A. Efros, and Martial Hebert. **From 3D Scene Geometry to Human Workspace**. In *CVPR*, pages 1961–1968, 2011. 113