UNIVERSITY OF CALIFORNIA, SAN DIEGO

Visual Understanding of Complex Human Behavior via Attribute Dynamics

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering
(Signal and Image Processing)

by

Weixin Li

Committee in charge:

Professor Nuno Vasconcelos, Chair
Professor Serge J. Belongie
Professor Kenneth Kreutz-Delgado
Professor Gert R. G. Lanckriet
Professor Lawrence K. Saul

2016

The dissertation of Weixin Li is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2016

DEDICATION

To my parents:

Jianfeng Li and Weixian Zhan

EPIGRAPH

Good mathematicians see analogies between theorems or theories; the very best ones see analogies between analogies.

— Stefan Banach

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Finishing the doctoral work is not an easy quest on my own. Undoubtedly, the outcome of my graduate study during these years, including this thesis, would not be made possible without the help, support, motivation and contributions from many people, to whom I would like to deliver my sincere acknowledgment.

As always, the most impactful person during this course I would like to express my heartfelt gratitude to is my research advisor, Professor Nuno Vasconcelos. It has been a privilege to finish my graduate work under his supervision. His patient guidance, visionary inspiration, and spectacular insights into a large variety of topics in computational vision and machine learning have always come to me in a timely manner, of which I can never overestimate the value. I also feel deeply honored to have an exceptional doctoral committee to serve as my advisory board. To these world-renowned intellectual scholars: Professor Lawrence K. Saul, Professor Gert R. G. Lanckriet, Professor Serge J. Belongie, and Professor Kenneth Kreutz-Delgado, I wish to deeply appreciate their enlightening discussion, active involvement, and invaluable advice that have driven me all through the way.

My research work has been supported and recognized in the financial form by several sources. I am sincerely grateful to the China Scholarship Council[1], which, on behalf of my country, China, conferred the award for outstanding students abroad to me. My acknowledgment also goes to the Electrical and Computer Engineering department at University of California San Diego, which kindly provided me with a fellowship to initialize my odyssey at North America;

[1] The Chinese Ministry of Education's non-profit organization that provides Student financial aid to Chinese citizens and foreigners to study abroad or to study in China, respectively.

the National Science Foundation (NSF), which has consistently propelled my research work via grant CCF-0830535 and IIS-1208522 (primary investigator: Professor Nuno Vasconcelos); the Neural Information Processing Systems (NIPS) Foundation and the Institute of Electrical and Electronics Engineers (IEEE) Computer Society technical committee on Pattern Analysis and Machine Intelligence (PAMI), for their generous travel awards for me to attend their premier academic conferences.

During my years at the Statistical and Visual Computing Laboratory (SVCL), where most of my work was accomplished, there are numerous people I have worked with and/or got assistance from: Antoni Chan, Hamed Masnadi-Shirazi, Sunhyoung Han, Vijay Mahadevan, Nikhil Rasiwasia, Kritika Muralidharan, Mulloy Morrow, Ehsan Saberian, Jose Costa Pereira and Mandar Dixit, and these of the new generation, Can Xu, Song Lu, Si Chen, Zhaowei Cai, Bo Liu, Yingwei Li, Pedro Morgado. I would like to thank them all and wish them the best luck in their career and/or path to finish graduate study. Assistance and support from these colleagues also count quite much in my work, and I feel it fortunate to have them around to facilitate all my achievements. I started working with Dr. Antoni Chan, who guided me into the area of dynamic modeling, and drew some preliminary inspiration from him. I authored the first two papers with Dr. Vijay Mahadevan, and watched his practice closely on how a work germinates from a raw idea and grows into mature publication. I also collaborated with Dr. Jose Costa Pereira on the hardware maintenance and many other issues in the lab to keep the infrastructure actively responding to the need of people inside and outside SVCL. Yingwei Li has helped me quite a lot during my last year in the lab with experiments that brought many ideas into concrete reality, which constitutes a crucial portion of this thesis work.

I have to thank all my friends in San Diego, in California, in United States, and around the world, too, for their generous hands that made my post-graduate life far easier. It has been entertaining to share experience in many perspectives with my roommates: Menglai Han, Yu Xiang, and Jianlin Zheng. I also had a very warm memory with my close friend, Yangbin Gao, of the days when we dined (especially at Crab Hut) and hung out around together, and when he spent all night through in my living room playing the endless exotic video game *Dark Souls*. Similar stories were shared with other friends too: Shaohe Wang, Bo Yang, Linchun Chen, Jun Zhou, *etc*. There are another special group of people, dubbed *San Diego Red Army*, I interacted almost on a weekly basis, with whom I played soccer games on every Saturday. I appreciate their assistance very much for the two-figure goals they kindly allowed me to score. Although separated at two opposite places of the globe for most of the time, the connections between me and my intimate friends in China have never gone faded: Wen Li, Kechun Liang, Qianyu Liu, Jiawei Lv, to name a few. The list can definitely go far longer, but due to lack of space, for these friends I cannot explicitly enumerate their names, I also extend my appreciation and best wishes to you all.

Lastly, and most importantly, I would like to thank my family. I owe my parents, Jianfeng Li and Weixian Zhan, a debt I can never pay back, including the life itself, the love all the way, the unconditional support in every possible form, and the willingness to let me embark on this long quest so far away from home to pursue the dream, to identify my ego, and to explore every possibility in my life. This thesis is also a tribute to my late grandmother, Xiuying Lin, who was my close custodian during childhood and has given me all her love since then. I really wish she could have the chance to witness the finish of this thesis.

The text of Chapter II is, in part, based on the material as it appears in the

following publications: The binary dynamic system was originally proposed in W.-X. LI and N. Vasconcelos, “Recognizing Activities by Attribute Dynamics,” *Advances in Neural Information Processing Systems* (NIPS), 2012. The mixture of binary dynamic systems was originally proposed in W.-X. LI, Y. Li and N. Vasconcelos, “Mixtures of Binary Dynamic Systems,” under review at *Journal of Machine Learning Research* (JMLR). The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter III is, in part, based on the material as it appears in the following publications: The variational scheme for BDS using LJ bound and Fisher vector were originally proposed in W.-X. LI and N. Vasconcelos, “Complex Activity Recognition via Attribute Dynamics,” to appear at *International Journal of Computer Vision* (IJCV). The variational scheme for BDS using JJ bound and the associated expectation-maximization algorithm were originally proposed in W.-X. LI, Y. Li and N. Vasconcelos, “Efficient Variational Inference, Learning and Probabilistic Kernels for Binary Dynamic Systems,” under review at *Neural Information Processing Systems* (NIPS), 2016. The variational inference scheme for the mixture of binary dynamic systems was originally proposed in W.-X. LI, Y. Li and N. Vasconcelos, “Mixtures of Binary Dynamic Systems,” under review at *Journal of Machine Learning Research* (JMLR). The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter IV is, in part, based on the material as it appears in the following publications: The sub-optimal learning scheme for BDS was originally proposed in W.-X. LI and N. Vasconcelos, “Recognizing Activities by Attribute Dynamics,” *Advances in Neural Information Processing Systems* (NIPS), 2012. The variational expectation-maximization algorithm for parameter estimation of BDS using JJ bound was originally proposed in W.-X. LI, Y. Li and N. Vasconcelos, "Efficient Variational Inference, Learning and Probabilistic Kernels for Binary Dynamic Systems," under review at *Neural Information Processing Systems* (NIPS), 2016. The variational expectation-maximization algorithm for parameter estimation of the mixture of binary dynamic systems was originally proposed in W.-X. LI, Y. Li and N. Vasconcelos, "Mixtures of Binary Dynamic Systems," under review at *Journal of Machine Learning Research* (JMLR). The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter V is, in part, based on the material as it appears in the following publications: The bag-of-model encoding scheme with zeroth and first order statistics were originally proposed in W.-X. LI and N. Vasconcelos, "Complex Activity Recognition via Attribute Dynamics," to appear at *International Journal of Computer Vision* (IJCV). The probabilistic kernels for BDS was originally proposed in W.-X. LI, Y. Li and N. Vasconcelos, "Efficient Variational Inference, Learning and Probabilistic Kernels for Binary Dynamic Systems," under review at *Neural Information Processing Systems* (NIPS), 2016. The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter VI is, in part, based on the material as it appears in the following publications: W.-X. LI and N. Vasconcelos, "Complex Activity Recognition via Attribute Dynamics," to appear at *International Journal of Computer Vision* (IJCV), and W.-X. LI, Y. Li and N. Vasconcelos, "Efficient Variational Inference, Learning and Probabilistic Kernels for Binary Dynamic Systems," under review at *Neural Information Processing Systems* (NIPS), 2016. The dissertation author was a primary researcher and an author of the cited material.

VITA

2008	B.S. in Automatic Control *cum laude*, Tsinghua University, Beijing, China
2011	M.S. in Electrical Engineering, University of California San Diego, United States
2016	Ph.D. in Electrical Engineering, University of California San Diego, United States

PUBLICATIONS

Wei-Xin LI, Yingwei Li and Nuno Vasconcelos, "Mixtures of Binary Dynamic Systems," under review at *Journal of Machine Learning Research* (JMLR)

Wei-Xin LI, Yingwei Li and Nuno Vasconcelos, "Efficient Variational Inference, Learning and Probabilistic Kernels for Binary Dynamic Systems," under review at *Neural Information Processing Systems* (NIPS), 2016

Wei-Xin LI and Nuno Vasconcelos, "Complex Activity Recognition via Attribute Dynamics," to appear at *International Journal of Computer Vision* (IJCV), 2016

Yingwei Li, Wei-Xin LI, Vijay Mahadevan and Nuno Vasconcelos, "VLAD^3: Encoding Dynamics of Deep Features for Action Recognition," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2016

Wei-Xin LI and Nuno Vasconcelos, "Multiple Instance Learning for Soft Bags via Top Instances," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2015

Wei-Xin LI, Vijay Mahadevan and Nuno Vasconcelos, "Anomaly Detection and Localization in Crowded Scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (TPAMI), Vol. 36, No. 1, 2014

Wei-Xin LI, Qian Yu, Ajay Divakaran and Nuno Vasconcelos, "Dynamic Pooling for Complex Event Recognition," *Proceedings of IEEE International Conference on Computer Vision* (ICCV), 2013

Wei-Xin LI, Qian Yu, Harpreet Sawhney and Nuno Vasconcelos, "Recognizing Activities via Bag of Words for Attribute Dynamics," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2013

Wei-Xin LI and Nuno Vasconcelos, "Recognizing Activities by Attribute Dynamics," *Advances in Neural Information Processing Systems* (NIPS), 2012

Vijay Mahadevan, Wei-Xin LI, Viral Bhalodia and Nuno Vasconcelos, “Anomaly Detection in Crowded Scenes,” *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2010

ABSTRACT OF THE DISSERTATION

Visual Understanding of Complex Human Behavior via Attribute Dynamics

by

Weixin Li

Doctor of Philosophy in Electrical Engineering
(Signal and Image Processing)

University of California, San Diego, 2016

Professor Nuno Vasconcelos, Chair

Visual understanding of human behavior in video sequences is one of the fundamental topics in computational vision. Being a sequential signal by nature, most critical insights of human activity can only be perceived via modeling the temporal structure. Despite an intuitive proposition, this task is non-trivial to accomplish. One of the most significant obstacles comes from the enormous variability and distinct properties of temporal structure at different levels of the human motion hierarchy, which spans a wide range of collectiveness, time and space, semantic granularity, and so forth. This has posed a rigorous challenge

for a solution that is supposed to be capable of simultaneously capturing the instantaneous movements, encoding the mid-level evolution patterns, coping with to long-term non-stationarity or content drifts, and being invariant to intra-class variation and other visual noise.

While most of the previous works in the literature focus on addressing some aspects of this problem, we aim to develop a unified framework to handle them all for complex human activity analysis. Specifically, we propose to model the temporal structure of human behavior on a robust, stable yet general representation platform that encodes some semantically meaningful concepts (or attributes). This platform bridges the gap between low-level visual feature and the high-level logical reasoning, bringing in benefits such as better generalization, knowledge transfer, and so forth. While attributes take care of abstracting semantic information from *short-term motion* in low-level visual signal, the dynamic model focuses on charactering the *mid-range evolution patterns* in this space. To cope with *long-term non-stationarity* and *intra-class variation* for complex events, we derive two encoding schemes that capture the zeroth and first order statistics of the attribute dynamics in video snippets, instead of precisely characterizing the whole sequence, which is prone to over-fitting due to the sparse nature of complex event instantiation.

The proposed framework is implemented via several novel models, together with the corresponding technical tools for statistical inference, parameter estimation, similarity measure, encoding statistics at the model manifold, and so on. In particular, a dynamic model is proposed to capture the evolution pattern in sequential binary data, denoted the *binary dynamic system* (BDS), which consists of a binary principal component analysis for modeling appearance and Gauss-Markov process to encode dynamics. A mixture model is further derived from

BDS to characterize multiple types of dynamics in a large data corpus. Based on variational methods, an accurate and efficient approximate inference scheme is developed for the state posterior to handle the intrinsic intractability; and a variational expectation-maximization algorithm is also derived for parameter estimation. Through these tools, measurements that quantify the similarity or dissimilarity of two binary sequences are devised from the perspective of control theory, information geometry, and kernel methods. Besides, approaches to encode the statistics of sequential binary data in the manifold of statistical models are proposed, resulting in the bag-of-words for attribute dynamics (BoWAD) and vector of locally aggregated descriptor for attribute dynamics (VLADAD).

Empirical study on challenging tasks of complex human activity analysis justifies the effectiveness of the proposed framework. Our solution not only produces the state-of-the-art results for event detection, but also enables recounting that provides the visual evidence anchored over time in the video for the prediction, and facilitates tasks like semantic video segmentation, content based video summarization, and so forth.

Chapter I

Introduction

I.A Visual Understanding of Human Motion

I.A.1 Background

Computational vision (*a.k.a.* machine vision, computer vision) is a subject of scientific research and engineering that studies the acquisition, extraction, processing, analysis, and interpretation of visual signals (*e.g.*, infra-red, visible lights) recorded from the real world in order to produce specific information to facilitate the understanding of the signal sources [102, 44].

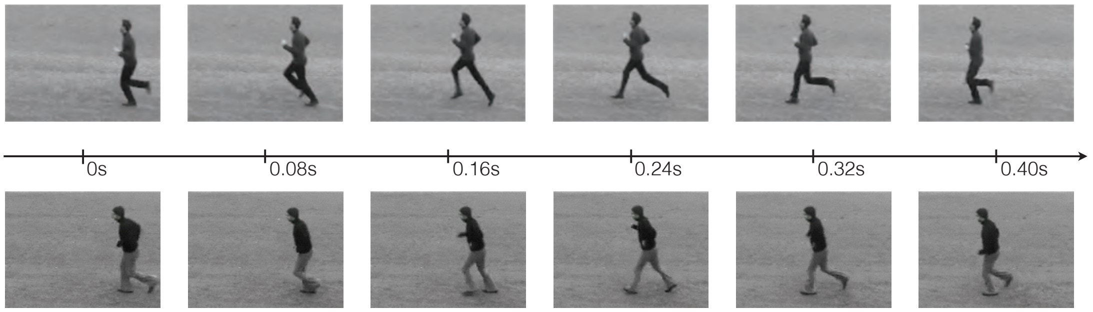
Among many subfields of computational vision, visual understanding of human behavior has been one of the most fundamental topics dating back to the early age of the research subject, when it was specifically developed as the visually sensing component for robotics, or a computational model to interpret biological visual systems [75, 76, 102]. The goal of visual understanding of human behavior is to extract information from video sequences to answer questions such as the identity of the subject(s) (who), the categories of events in the past, now and in the future (what), the time and place of the event (when and where), and the fine-grained patterns of the event (how) [3]. Facilitated by the processing power of modern computing machines, and spurred by the demand of managing tremendous amount of visual data generated by ubiquitous mobile recording devices during the Internet era, the application of visual understanding of human behavior has reached a far broader horizon with practical applications to machine-human interaction [37], augmented or virtual reality [1, 27], automated media data management [172, 81], intelligent surveillance [25, 96], *etc*.

While the early focus is set on recognizing some simple gestures and primitive motion [34, 19, 15, 121, 14, 142, 51], recently the major attention has been turn to more challenging and realistic tasks, where more complex human activities are considered in unconstraint enviroments [93, 130, 111, 85, 114, 58, 81]. This not only enables a substantially larger range of applicability of behavior understanding, but also poses several technical challenges.

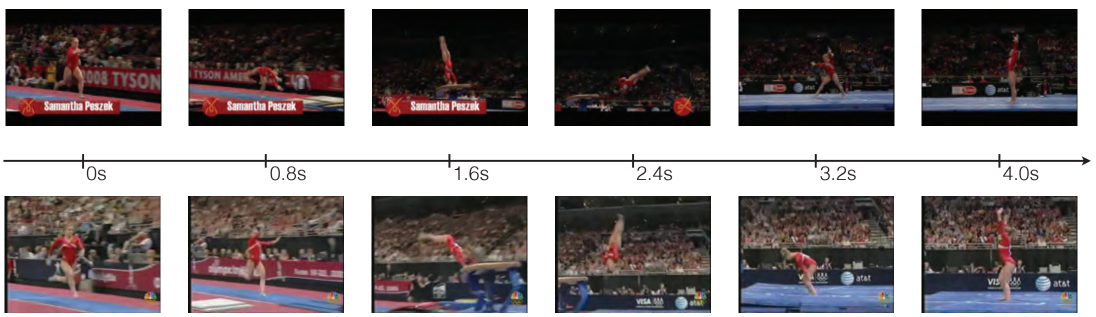
I.A.2 Challenge of Modeling Temporal Structure

Being a proposition of both scientific and practical values, understanding human motion in video via computational machinery, however, is non-trivial to fulfill in the technical point of view. In the big picture, human behavior is a broad topic that can be represented in a hierarchical structure spanning a large scale of time and space, collectiveness, semantic granularity, and so forth. Analysis of this complex concept via visual signal incurs difficulties from several sources.

The temporal structure modeling is one of the most prominent challenges. A video sequence is not a random collection of images. The temporal order of video frames conveys intrinsic information of the event critical for interpreting the story. Human behavior at different temporal scales, however, exhibits divergent properties, as illustrated in Fig. I.1. These diverse temporal properties require distinct strategies to characterize. Recent studies have shown that, instantaneous or primitive types of motion, *e.g.*, running, jumping, can be effectively captured by 1) low-level image features computed within local spatiotemporal visual support, and 2) the statistics of these features aggregated over a few video frames [92, 142, 140, 165, 117], which has its roots in the classical research on biological vision and motion perception [77]. Movements of longer duration with mid-range temporal structure, such as sports activity "long jump," typically requires the characterization of the temporal distribution of the sub-module actions that compose the activity [111, 155, 48]. In the even more complex case of high-level events, which can last for hours, the visual content are so sophisticated



(a) Examples of short-term instantaneous primitive motion "running". These types of movements can be captured by statistics aggregated over low-level image features.



(b) Examples of mid-term continuos smoothing activity "gymnastics vault". These types of behavior are best characterized with sequential description of short-term actions (*e.g.*, "running"-"jumping"-"touching pad"-"somersault"-"landing").

(c) Examples of long-term complex event "wedding ceremony." Intra-class variation is so significant that learning holistic temporal structure most likely leads to instance-specific depictions that hardly apply to other examples from the same event class.

Figure I.1: Divergent properties of human behavior at different temporal scales. Key frames of two video instances at each granularity are exemplified.

that local evidences are commonly used in an orderless fashion to justify the event recognition for better generalization [154, 98, 88]. In the extreme of the temporal scale, where a typical application is surveillance video analysis, continuous visual content are streaming in endlessly. For this type of problem, both the short-term evolution patterns and the long-term non-stationarity due to content drift are two critical aspects of the data to account for. Challenges due to the variability in temporal structure of human activity are further compounded by the sparseness of training examples as the temporal scale increases [114]. Overall, the interplay between the stationarity and non-stationarity of human motion results from several complex sources, including sociological, psychological, physical, biological factors [149, 20, 9, 59, 104]. As such, modeling temporal structure of human behavior is a complex proposition that requires a principled way of capturing these divergent properties at different level of granularity, to achieve the best balance among representativeness, selectivity, and invariance to noise such as intra-class variation.

Many approaches in computational vision for modeling human motion mostly focus on only one of these critical factors, making them somehow biased to a specific case of motion. The popular *bag-of-visual-words* (BoVW) has been widely adopted for human action recognition [142, 93, 166]. This paradigm posits that a visual entity (*e.g.*, an image, a video sequence) can be represented by an *orderless* corpus of lower-level visual features aggregated from it. While very robust to noise, BoVW is not flexible enough to encode critical temporal information in many scenarios, even after enhancement with rigid pooling cells over time [93, 99, 90]. Similarly, the visual data representation via semantically meaningful concepts, of arising interest recently, also ignores the temporal information in human action, though it provides a more general intermediate platform that bridges the gap between low level features and high level semantic reasoning [89, 125, 115, 101, 72]. On the other hand, another popular proposal for human motion analysis exclusively focus on modeling the evolution pattern, in appreciation of the significance of temporal structure for human motion. While motivated by insights, most of works in this direction aim to solve the problem with one single model, or operate on the unstable, low-level, task-specific, or computationally expensive representations, which cannot generalize to more challenging scenarios, *e.g.*, open-source videos [82, 28, 95].

I.B A Unified Temporal Structure Hierarchy for Human Behavior

To motivate and justify our technical solution, we start by introducing the unified temporal structure hierarchy for human behavior. In the big picture, we propose that, according to the stationarity of visual content, any human behavior can be categorized into one of the three layers in the hierarchy of Fig. I.2.

At the very low-level of the hierarchy resides the primitive motion, *e.g.*, "running," "jumping," "waving hands." These types of instantaneous movements are 1) the fundamental constituent elements of more complex actions [3]; and 2) mostly constraint by physical motion laws of human bodies, *e.g.*, Newtonian mechanics, thus the space of possible configurations is bounded [157]. In this light, learning the representation for these movements by exhaustive instantiation of the whole example space is feasible given today's data resources and computational technology. In practice, this is frequently implemented with data-driven strategies such as descriptors computed by statistics of low-level image features in local spatiotemporal support with salient motion followed by unsupervised motion prototype clustering [92, 165] or recently popular neural networks that learn low-level action templates from tremendous amount of video data [73, 81, 148]. It has been shown that, these schemes can confidently and precisely model behavior at this level, achieving spectacular results in action recognition [117, 118].

More complex behavior is observed at the middle layer of the hierarchy. One such activity is typically comprised of a sequence of local primitive movements in a particular pattern, which results from the underlying procedure controlled or driven by social convention (*e.g.*, a couple exchanging rings at a wedding ceremony), legal regulations (*e.g.*, crowd crossing roads at a street intersection), domain knowledge or instructions (*e.g.*, sport activity "high-jump"), *etc* [20, 9]. Due to this constraint, homogeneity holds reasonably well for these activities of the same category despite some possibility of variation. For example, while athletes may perform the sport activity "triple jump" in slightly different styles (*e.g.*, various duration of running, in-air movements), they always follow the sequence of "running-skipping-jumping-landing," as determined by the expert instruction. Another critical observation at this level is that the homogeneity is only preserved on top of an *appropriate* basis of representation for local constituting movements. Unreliable features (*e.g.*, low-level image optical flow) can lead to activity representation significantly vulnerable to noise and difficult to generalize [82, 28, 95], which nullifies the uniformity among instances from a category.

Finally, a long-term sophisticated story anchors at the top layer of the hierarchy. In most cases, such events are subject to very loose constraints, if any, and exhibit substantial intra-class variation or flexibility in plot, since the latent factors (*e.g.*, human psychological processes) governing behind are highly

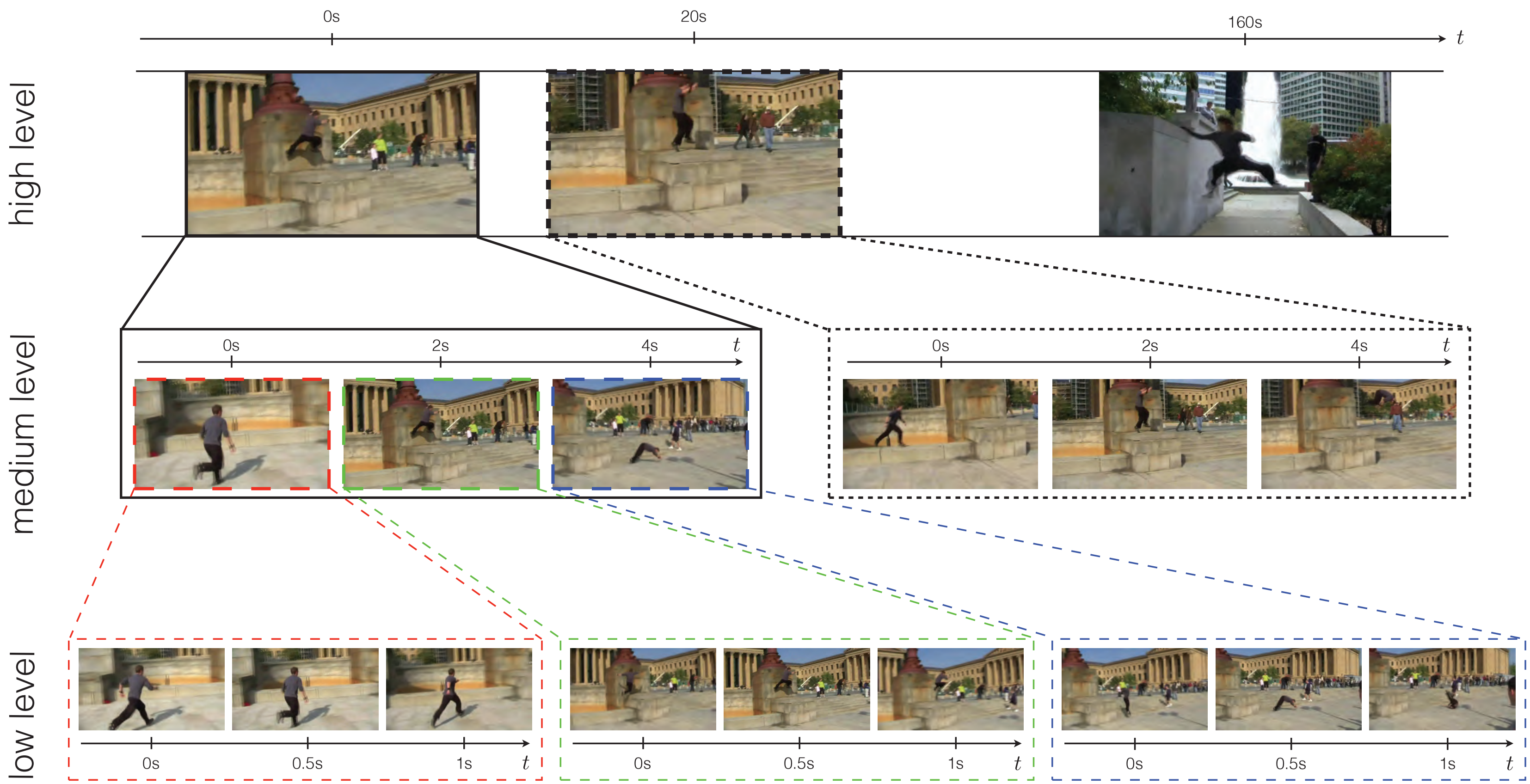


Figure I.2: Hierarchy of complex human behavior "parkour." Key frames are illustrated for 1) the high-level event of the whole complex video sequence comprised of various shots of parkour activities; 2) medium-level activities of sequential movements (*e.g.*, "running"-"kicking-and-turning"-"rolling-after-landing" in the solid box); and 3) low-level instantaneous motion (*e.g.*, "running" in the red dashed box, "kicking-and-turning" in the green dashed box, and "rolling-after-landing" in the blue dashed box) . Time axis indicates the temporal scale and key frame anchor points (not to scale).

unpredictable and diversified [149]. This inhomogeneity is further compounded by the fact that a video sequence is not necessarily an objective visual recording of chronological events, but a product of video post-processing such as montage sequences for artistic representation in filmmaking [129]. Both of these unique properties pose a challenge for temporal structure modeling at this level that do not exist at previous two. As such, the underlying factors that generate the video event must also be perceived at a very high level of abstraction, where it is usual that temporal stationarity may not hold and most clues are loosely connected over time. In this case, robustness plays a far more important role in event characterization, rather than the precise instance-level selectivity.

Overall, the hierarchy provides a unified interpretation of temporal structure at different levels. In the bottom-up direction, both intra-class diversity and variation increase as the space of feasible behavior configurations expands exponentially, while possible instantiation becomes more sparse at the same time. This motivates our strategy for complex human behavior modeling comprised of technical solutions of three distinct flavors, which correspond to various balance points for the trade-off between selectivity and invariance at different layers of the hierarchy. For the low-level, we rely on the BoVW benchmark for instantaneous action representation since it can be learnt with the non-parametric method given moderate amount of training data, as in our case. At the mid-level, we propose to combine the semantic attribute representation, which preserves the temporal homogeneity, and the dynamic model, which regularizes the temporal structure characterization. This provides a solution that models the temporal structure with flexibility on top of a reliable basis that can generalizes well. For inhomogeneous complex events at the high-level, due to the sparse examples for training and loosely correlated local event evidences scattered over time, we resort to the corpora encoding frameworks that capture the distribution of multiple local sub-events in the statistical manifold of models of attribute dynamics. Despite losing the capability of depicting the holistic chronological story, these frameworks can still characterize the evolution patterns of local events at the mid-level that are sufficient to identify most high-level event categories, while exhibits better robustness to intra-class variation than those modeling the global temporal structure, as will be seen in the experiment. This is further shown to enable a recounting scheme that can provide visual content evidences to justify high-level event recognition.

I.C Contributions of the Thesis

In this thesis, we address the problem of modeling temporal structure for visual human behavior understanding across several scales via a statistical perspective. We specifically focus on the use of dynamic systems for encoding insightful properties of complex human behavior at the mid-level. This results in, from the theoretical viewpoint, a hierarchical representation of human behavior that characterizes temporal structure at distinct levels; and, from the technical viewpoint, a new set of statistical tools for modeling, analyzing, and encoding discrete time-series. The main contributions of the thesis are summarized as follows.

I.C.1 A Hierarchical Representation of Temporal Structure for Human Behavior

To cope with the highly divergent characteristics of human behavior at different levels, and leverage the power of dynamic modeling in capturing these insights, we propose a hierarchical representation of human behavior according to the nature of temporal structure. In this hierarchy, we posit that, while short-term instantaneous movements are modeled by statistics of low-level features, mid-range activities are represented in the space of semantically meaningful concepts or attributes, whose evolution is depicted by smooth dynamic processes (denoted *attribute dynamics*). Higher level events, however, frequently exhibits substantial non-stationarity. Together with the sparseness of training examples, temporal structure at this level is encoded with robust framework such as the zeroth and first order statistics of mid-level dynamics, resulting in *bag-of-words for attribute dynamics* (BoWAD) representation and *locally aggregated descriptors for attribute dynamics* (VLADAD). Combined with proper choices of models at different levels, we show that state-of-the-art results in complex activity or event recognition and recounting can be achieved.

I.C.2 Statistical Models of Dynamics for Sequential Binary Data

We present a novel statistical model that captures the evolution patterns behind sequences of multi-dimensional binary observations (referred as binary sequences, sequential binary signals, for the rest of the thesis), denoted the *binary dynamic system* (BDS). BDS consists of two major modules: 1) the binary principal component analysis (PCA) for observation, and 2) the Gauss-Markov process for dynamics. This formulation generalizes the conventional linear dynamic system to the binary signal. A mixture model, denoted the *mixture of binary dynamic systems* (mix-BDS), is derived to enhance representation power of BDS for large corpus where multiple distinct types of patterns are present. A simplified version of mix-BDS, *bag-of-words for attribute dynamics* (BoWAD) is also introduced to model dynamics of binary data for large-scale problems.

I.C.3 A Statistical Toolkit for Reasoning, Learning and Encoding of Sequential Data with Dynamic Systems

We also develop technical solutions via principled paradigms to address challenges such as statistical inference, parameter estimation, similarity measure, and discriminative data encoding for the proposed dynamic models. Specifically, a variational inference scheme is devised, via rigorous lower-bounds of log sigmoid nonlinearity, to compute the posterior of hidden states in the BDS. This is shown to provide a tight approximation to the exact result that is intractable, outperforming the state of the art in both accuracy and efficiency. In the similar way, a variational expectation-maximization algorithm is also proposed for parameter estimation of BDS and its mixture model. Further more, to facilitate the use of the proposed model in discriminative tasks, similarity or dissimilarity measures between sequential binary data are derived from three distinct perspectives, including information geometry, dynamic system theory, and kernel methods, which generalize previous techniques for real-valued data domain. These are shown to produce competitive results on complex activity or event recognition tasks.

I.C.4 Applications to Complex Human Activity Recognition

Using the proposed dynamics modeling framework, technical toolkits, and hierarchical interpretation of human behavior, we accomplish state-of-the-art performance on several popular tasks of human behavior analysis. We propose that, while short-term primitive human motion can be captured by statistics of low-level image features, characterization of finer-grained temporal structure is critical for describing mid- and high-level activities. To this end, mid-range activities should be characterized on an intermediate layer of visual concepts, instead of low-level representations suffers from noisy, unstable, or task-specific computationally expensive observations. This is implemented by modeling dynamics on the mid-level semantically meaningful attribute space for complex human activity understanding. Further more, we show that modeling dynamics for mid-level behavior while encoding higher-level events with robust schemes achieves the best balance between selectivity to target categories and invariance to the inherent huge intra-class variations of the problem. Empirical study on benchmark complex event detection datasets shows that, our strategies not only produce competitive recognition results, but also enable the finer-grained recounting outputs that provide semantically meaningful visual content anchored in video as the evidence to justify the event prediction.

I.D Organization of the Thesis

The rest of the thesis is organized as follows. We start by introduction of technical tools for dynamics modeling. In Chapter II, the technical formulation of the dynamic models are presented, including the review of the linear dynamic systems, the proposed binary dynamic system, and its mixture version. We derive the statistical inference schemes for these models in Chapter III. These consist of the review of the variational inference framework for models with hidden variables, the lower bounds adopted to approach the intractable log sigmoid nonlinearity, and the efficient routines to compute the evidence lower bound, mean and covariance of the variational distributions based on the popular Kalman smoothing filter from the control theory literature. Chapter IV details the parameter estimation for the proposed models. These are implemented either from the perspective of dynamic texture learning, resulting in a sub-optimal routine; or via the maximum likelihood estimation principle, resulting in the variational expectation-maximization algorithm. Different encoding schemes for sequential data are introduced in Chapter V, where three types of representation architectures are derived to capture dynamics in sequences for discriminative tasks, using results from document analysis, information geometry, and kernel methods. We address the problem of complex activity recognition and recounting in Chapter VI. We present the motivation and insights into the temporal structure modeling of complex events, and derive solution based on our technical tools and analysis of the problem. This leads to a unified activity representation that efficiently and effectively captures evolution patterns of human motion at different temporal scales, producing state-of-the-art results on event recognition and recounting. Finally, the thesis is concluded in Chapter VII, where some possibilities of future works are also discussed.

Chapter II

Statistical Models of Dynamic Systems

One of the most popular strategies to capture the temporal structure of sequential data in literature is implemented via the dynamic Bayesian network (DBN) [133], which characterizes the probabilistic dependency among multiple factors at each temporal instant and over time. A common scheme of DBN is formulated as the *state-space model* (SSM) [107], which posits sequence of multi-dimensional data as noisy observations mapped from a state process in a hidden lower-dimensional space. Originating from the control theory for description of physical systems [45], SSMs have been shown to be flexible in modeling dynamic processes in many other applications across numerous fields of science and engineering [62, 40, 12]. Within the large family of SSMs, one of the most popular architectures is the linear dynamic system (LDS), which assumes linear Gaussianity for both hidden states and observed signals. Despite its limitation of linear assumption, LDS has not only achieved substantial successes, but also inspired other variants that can handle more complex scenarios [41, 107, 169]. One notable enhancement of LDS is the generalized linear dynamic system (GLDS) that combines the exponential-family distributions with the Gauss-Markov process to handle a large variety of types of data [49], including the binary dynamic system (BDS), which is of specific interest to human motion analysis in the attribute space [101].

II.A Definitions and Notations

In this section, notations, definitions and brief results are presented to facilitate the understanding of the technical presentation throughout this thesis. Table II.1 summarizes the notations and definitions.

Table II.1: Notations and definitions.

notation	definition
$\boldsymbol{x}$ (boldface)	a vector.
$\boldsymbol{x}_{1:\tau}$	a vector sequence: $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_\tau\}$.
$\boldsymbol{A}$ (capital)	a matrix
x_i, $A_{i,j}$	the i-th element of $\boldsymbol{x}$, the element at (i,j) of matrix $\boldsymbol{A}$.
A (capital)	a scalar constant, or random variable.
$\boldsymbol{A}^\intercal$, $\boldsymbol{x}^\intercal$	transpose of $\boldsymbol{A}$, $\boldsymbol{x}$.
$\mathrm{tr}(\boldsymbol{A})$	trace of square matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$.
$A_{[r,s]}$, $\boldsymbol{x}_{[i]}$	block t, s of a matrix $\boldsymbol{A}$, and block i of a vector $\boldsymbol{x}$.
$A_{r,:}$, $A_{:,c}$	row r of matrix $\boldsymbol{A}$, column c of matrix $\boldsymbol{A}$.
$\boldsymbol{A}^\dagger$	pseudoinverse of $\boldsymbol{A}$.
$\mathcal{S}^d$	the set of $d \times d$ symmetric matrices: $\{\boldsymbol{A} \vert \boldsymbol{A} \in \mathbb{R}^{d\times d}, \boldsymbol{A} = \boldsymbol{A}^\intercal\}$.
$\mathcal{S}^d_{++}$	the set of $d \times d$ positive-definite matrices: $\{\boldsymbol{A} \vert \boldsymbol{A} \in \mathcal{S}^d, \boldsymbol{A} \succ \boldsymbol{0}\}$.
$p(\boldsymbol{x};\theta)$, $p_\theta(\boldsymbol{x})$, or p_θ	the probability density (or mass) function (PDF or PMF) of a random vector $\boldsymbol{x}$, with parameter θ.
$\langle f(\boldsymbol{x}) \rangle_{p(\boldsymbol{x};\theta)}$	expectation of function $f(\boldsymbol{x})$ with respect to $\boldsymbol{x}$: $\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x};\theta)}[f(\boldsymbol{x})]$.
$\mathrm{KL}(p_{\theta_1} \vert\vert p_{\theta_2})$	the *Kullback-Leibler* (KL) *divergence* [86] between distributions p_{θ_1} and p_{θ_2}: $\langle \ln p_{\theta_1}(\boldsymbol{x}) \rangle_{p_{\theta_1}} - \langle \ln p_{\theta_2}(\boldsymbol{x}) \rangle_{p_{\theta_1}}$.
$\vert\vert\boldsymbol{x} - \boldsymbol{y}\vert\vert^2_{\boldsymbol{\Sigma}}$	the (squared) Mahalanobis distance: $(\boldsymbol{x}-\boldsymbol{y})^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{y})$.
$\mathcal{N}(\boldsymbol{\mu}, \Sigma)$	a Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$.
$\mathrm{H}[q(X)]$	the entropy of X distributed as $q(X)$: $-\int q(x) \ln q(x) dx$.
$\mathcal{G}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$	the PDF of $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$: $(2\pi)^{-d/2}\vert\Sigma\vert^{-1}\exp\{-\frac{1}{2}\vert\vert\boldsymbol{x}-\boldsymbol{\mu}\vert\vert^2_\Sigma\}$, $\boldsymbol{x}, \boldsymbol{\mu} \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$.
$Y \vert X$	random variable Y conditional on random variable X.

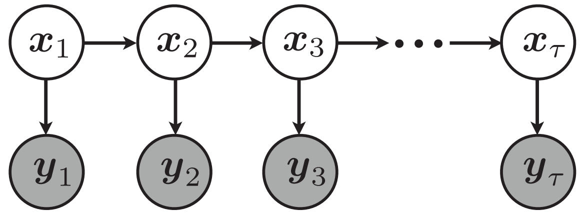


Figure II.1: Graphical model for the linear dynamic system or binary dynamic system.

It can be shown [86] that, when $p_{\theta_1} = \mathcal{G}(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma_1)$ and $p_{\theta_2} = \mathcal{G}(\boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma_2)$,

$$\mathrm{KL}(p_{\theta_1} || p_{\theta_2}) = \frac{1}{2}\left[\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2_{\Sigma_2} - \ln\left|\Sigma_2^{-1}\Sigma_1\right| - d\right]. \tag{II.1}$$

The entropy of two random variables X and Z can be factorized according to

$$\begin{aligned} \mathrm{H}[q(\mathrm{X},\mathrm{Z})] &= -\int_{x,z} q(x,z) \ln q(x,z) dx dz \\ &= -\int_{x,z} q(x|z)q(z) \ln q(x|z)q(z) dx dz \\ &= -\int_z q(z)\left[\int_x q(x|\mathrm{Z}=z) \ln q(x|\mathrm{Z}=z) dx + \ln q(\mathrm{Z}=z)\right] dz \\ &= \int_z q(z)\mathrm{H}[q(\mathrm{X}|\mathrm{Z}=z)] dz + \mathrm{H}[q(\mathrm{Z})]. \end{aligned} \tag{II.2}$$

II.B Linear Dynamic Systems

Video sequences are frequently modeled as samples of a *linear dynamic system* (LDS)

$$\begin{cases} \boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{v}_t, & \text{(II.3a)} \\ \boldsymbol{y}_t = \boldsymbol{C}\boldsymbol{x}_t + \boldsymbol{w}_t + \boldsymbol{u}, & \text{(II.3b)} \end{cases}$$

where $\boldsymbol{x}_t \in \mathbb{R}^L$ and $\boldsymbol{y}_t \in \mathbb{R}^D$ (of mean $\boldsymbol{u}$) are a hidden *state* and *observation* variable at time t, respectively; $\boldsymbol{A} \in \mathbb{R}^{L \times L}$ a state transition matrix that encodes dynamics; $\boldsymbol{C} \in \mathbb{R}^{D \times L}$ an observation matrix that maps state to observations; and $\boldsymbol{x}_1 = \boldsymbol{\mu} + \boldsymbol{v}_0$ an initial condition. Both states and observations have additive Gaussian noise $\boldsymbol{v}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{S})$, $\boldsymbol{v}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q})$ and $\boldsymbol{w}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{R})$ $(t \geqslant 1, t \in \mathbb{Z})$.

LDS parameters can be learned by maximum likelihood (ML), using the expectation-maximization (EM) algorithm [146]. A simpler approximate learning procedure was, however, introduced by [39]. This is known as the dynamic texture (DT) and decouples the learning of observation and state variables by interpreting the LDS as the combination of a principal component analysis (PCA) and a Gauss-Markov process. Under this interpretation, the columns of C are principal components of the observed video data and the hidden state $\boldsymbol{x}$ is a vector of PCA coefficients. The observation parameters are first learned through a PCA of the video frames, and the state parameters are then learned by least squares. This simple approximate learning algorithm tends to perform very well, and is popular in computer vision.

II.C Binary Dynamic Systems

Motivated by the linear dynamic system (LDS), the *binary dynamic system* (BDS), specified by parameter $\boldsymbol{\theta} = \{\boldsymbol{S}, \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{Q}, \boldsymbol{u}\}$, models a sequence of *binary* vectors $\boldsymbol{y}_{1:\tau} \in \{0,1\}^{D \times \tau}$ by

$$\begin{cases} \boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{v}_t, & \text{(II.4a)} \\ \boldsymbol{y}_t | \boldsymbol{x}_t \sim \mathrm{Bern}(\sigma(\boldsymbol{C}\boldsymbol{x}_t + \boldsymbol{u})), & \text{(II.4b)} \end{cases}$$

where $\sigma(\theta) = (1 + e^{-\theta})^{-1}$ is the sigmoid function ($\sigma(\boldsymbol{\theta}) \equiv [\sigma(\theta_1), \cdots, \sigma(\theta_K)]^\intercal$); $\text{Bern}(\boldsymbol{\pi})$ the multivariate Bernoulli distribution, *i.e.*, $\boldsymbol{y} \sim \text{Bern}(\boldsymbol{\pi})$ such that $p(\boldsymbol{y}) = \prod_d \pi_d^{y_d}(1 - \pi_d)^{(1-y_d)}$; $\boldsymbol{x}_t \in \mathbb{R}^L$ and $\boldsymbol{u} \in \mathbb{R}^D$ are the hidden state variable and observation bias, respectively; $\boldsymbol{A} \in \mathbb{R}^{L\times L}$ is the state transition matrix; and $\boldsymbol{C} \in \mathbb{R}^{D\times L}$ the observation matrix; the initial condition is given by $\boldsymbol{x}_1 = \boldsymbol{\mu} + \boldsymbol{v}_0 \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$; and the state noise process is $\boldsymbol{v}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q})$. For brevity, we denote $\tilde{\boldsymbol{C}} = [\boldsymbol{C}, \boldsymbol{u}]$ and $\tilde{\boldsymbol{x}}_t = [\boldsymbol{x}_t^\intercal, 1]^\intercal$. Alternatively, the observation model of (II.4b) can be regarded as a binary principal component analysis (PCA) of [139] with C as the principal components and $\boldsymbol{x}_t$ being the coefficients, which evolve according to the Markov-Gaussian process of (II.4a). This interpretation has motivated a very efficient learning scheme consisting of a binary PCA and a least-square estimation [97], which serves as a good initialization for other more principled learning algorithms. The graphical model of the BDS is illustrated in Fig. II.1.

Given the above definition of BDS, the distributions of the initial state, conditional states, and conditional observations are

$$\begin{aligned}
p(\boldsymbol{x}_1) &= \mathcal{G}(\boldsymbol{x}_1; \boldsymbol{\mu}_0, \boldsymbol{S}), && \text{(II.5)}\\
p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) &= \mathcal{G}(\boldsymbol{x}_{t+1}; \boldsymbol{A}\boldsymbol{x}_t, \boldsymbol{Q}), && \text{(II.6)}\\
p(\boldsymbol{y}_t|\boldsymbol{x}_t) &= \prod_{d=1}^{D} \sigma(\omega_{d,t})^{y_{dt}} \sigma(-\omega_{d,t})^{(1-y_{dt})}, && \text{(II.7)}\\
\omega_{d,t} &= \boldsymbol{C}_{d,:}\boldsymbol{x}_t + u_d. && \text{(II.8)}
\end{aligned}$$

The joint distribution of observation $\boldsymbol{y}_{1:\tau}$ and state $\boldsymbol{x}_{1:\tau}$ is

$$p(\boldsymbol{x}_{1:\tau}, \boldsymbol{y}_{1:\tau}; \boldsymbol{\theta}) = p(\boldsymbol{y}_{1:\tau}|\boldsymbol{x}_{1:\tau})p(\boldsymbol{x}_{1:\tau}) = p(\boldsymbol{x}_1)\prod_{t=1}^{\tau-1} p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)\prod_{t=1}^{\tau} p(\boldsymbol{y}_t|\boldsymbol{x}_t). \quad \text{(II.9)}$$

Note that, although both the BDS shares the same graphical model as the conventional LDS, the latter has a Gaussian state as the conjugate prior to its Gaussian observation, which leads to exact and efficient inference by the Kalman smoothing filter [146, 131], while Gaussian states and binary observations are entangled in the BDS case. This complex form incurs a challenge that makes exact inference of the state posterior intractable for the BDS. We will show that, however, the problem can be well addressed by approximation schemes, *e.g.*, *variational inference* [79] . This also inspires a parameter estimation framework that generalizes the conventional *expectation-maximization* (EM) algorithm [35] for BDS learning in Section V.A.

II.D Mixture Models for Binary Dynamic Systems

While a BDS can only encode one type of binary sequences, the *mixture of binary dynamic systems* (mix-BDS) accounts for multiple evolution patterns in binary vector sequence corpora. Under the mix-BDS, a binary vector sequence is sampled from one of K BDS components. Specifically, given a prior probability $\boldsymbol{\alpha} = \{\alpha_1, \cdots, \alpha_K\}$ ($\alpha_k \geqslant 0,\ \sum_k \alpha_k = 1$) of K components, a component indicator variable $\boldsymbol{z}$ is first sampled from a categorical distribution parametrized by $\boldsymbol{\alpha}$ as

$$\boldsymbol{z} \ \sim \ \mathrm{Cat}(K, \boldsymbol{\alpha}). \tag{II.10}$$

Then the binary vector sequence $\boldsymbol{y}_{1:\tau} \in \{0,1\}^{D\times\tau}$ is drawn from the z-th BDS component of the mixture model according to

$$\begin{cases} \boldsymbol{x}_{t+1} | \boldsymbol{x}_t, \boldsymbol{z} = z \quad \sim \quad \mathcal{N}(\boldsymbol{A}\boldsymbol{x}_t, \boldsymbol{Q}_z), & \text{(II.11a)} \\ \boldsymbol{y}_t | \boldsymbol{x}_t, \boldsymbol{z} = z \quad \sim \quad \mathrm{Bern}(\sigma(\boldsymbol{C}_z \boldsymbol{x}_t + \boldsymbol{u}_z)), & \text{(II.11b)} \end{cases}$$

where the z-th BDS component is parameterized by $\boldsymbol{\theta}_z = \{\boldsymbol{S}_z, \boldsymbol{\mu}_z, \boldsymbol{A}_z, \boldsymbol{C}_z, \boldsymbol{Q}_z, \boldsymbol{u}_z\}$ as defined in (II.4).

Under the definition of mix-BDS, the probability of a binary vector sequence $\boldsymbol{y}_{1:\tau}$ is

$$p(\boldsymbol{y}_{1:\tau}) = \sum_{z=1}^{K} p(\boldsymbol{z}=z) p(\boldsymbol{y}_{1:\tau}|\boldsymbol{z}=z) = \sum_{z=1}^{K} \alpha_z p(\boldsymbol{y}_{1:\tau}|\boldsymbol{z}=z), \tag{II.12}$$

where $p(\boldsymbol{y}_{1:\tau}|\boldsymbol{z}=z)$ is the probability of $\boldsymbol{y}_{1:\tau}$ under the z-th BDS component. Similar to a single BDS, the conditional probabilities of the initial state, intermediate states, and observations for the z-th BDS component are

$$\begin{aligned} p(\boldsymbol{x}_1|z) &= \mathcal{G}(\boldsymbol{x}_1; \boldsymbol{\mu}_{0,z}, \boldsymbol{S}_z), & \text{(II.13)} \\ p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, z) &= \mathcal{G}(\boldsymbol{x}_{t+1}; \boldsymbol{A}_z \boldsymbol{x}_t, \boldsymbol{Q}_z), & \text{(II.14)} \\ p(\boldsymbol{y}_t|\boldsymbol{x}_t, z) &= \prod_{d=1}^{D} \sigma(\omega_{d,t,z})^{y_{dt}} \sigma(-\omega_{d,t,z})^{(1-y_{dt})}, & \text{(II.15)} \\ \omega_{d,t,z} &= \boldsymbol{C}_{z,d,:} \boldsymbol{x}_t + u_{d,z}, & \text{(II.16)} \end{aligned}$$

where notations follow those of (II.5) to (II.8) respectively. Following the convention in the literature of mixture models [35], an unit assignment vector $\boldsymbol{z} \in \{0,1\}^K$ is used for brevity such that $z_j = 1$ if and only if $\boldsymbol{z} = j$ in (II.10). Using (II.13) to (II.16), the joint probability of the complete data is

$$p(\boldsymbol{x}_{1:\tau}, \boldsymbol{y}_{1:\tau}, z; \boldsymbol{\theta}) = p(z) p(\boldsymbol{x}_{1:\tau}|z) p(\boldsymbol{y}_{1:\tau}|\boldsymbol{x}_{1:\tau}, z) = p(z) \prod_{j=1}^{K} \Big[p(\boldsymbol{x}_{1:\tau}|j) p(\boldsymbol{y}_{1:\tau}|\boldsymbol{x}_{1:\tau}, j) \Big]^{z_j}, \tag{II.17}$$

where the mixture model is specified by $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \{\boldsymbol{\theta}_k\}_1^K\}$ with its graphical model illustrated in Fig. II.2. Although the graph is moralized and triangulated, and its junction tree resembles that of Fig. II.1 with z added to each clip [112, 23], exact

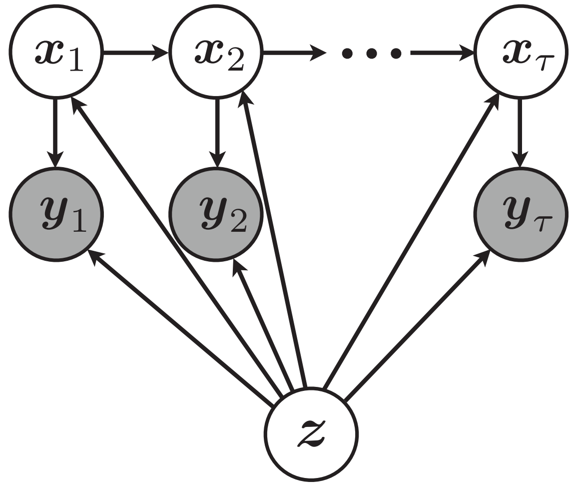


Figure II.2: Graphical model for the mixture of binary dynamic systems.

inference for the mixture model is intractable due to the difficulty in that of its BDS component. Nevertheless, using the paradigm in Section III.C, strategies for posterior approximation and parameter estimation of the mixture model can be derived as presented in Section III.D and Section V.A, respectively.

II.E Acknowledgment

The text of Chapter II is, in part, based on the material as it appears in the following publications: The binary dynamic system was originally proposed in W.-X. LI and N. Vasconcelos, "Recognizing Activities by Attribute Dynamics," *Advances in Neural Information Processing Systems* (NIPS), 2012. The mixture of binary dynamic systems was originally proposed in W.-X. LI, Y. Li and N. Vasconcelos, "Modeling, Clustering, and Segmenting Binary Sequences with Mixtures of Binary Dynamic Systems," under review at *Journal of Machine Learning Research* (JMLR). The dissertation author was a primary researcher and an author of the cited material.

Chapter III

Inference for Dynamic Systems

In this chapter, we first review the variational inference framework, and then the exact inference of the linear dynamic system via the Kalman smoothing filter, before presenting the scheme to approximate the posterior of hidden states in the binary dynamic system. The algorithms for inference of the mixture model is derived in the end.

III.A Variational Inference

Assume that a probabilistic model $p(Y;\theta)$ of parameter θ contains an observed variable Y and a hidden variable X. Let $q(X)$ be a member from a family of tractable distributions $\mathcal{D}_q$. Variational inference [79] approximates the posterior $p(X|Y;\theta)$ with $q^*(X) \in \mathcal{D}_q$ that is closest to $p(X|Y;\theta)$ in the KL sense, such that

$$q^*(X) = \arg\min_{q \in \mathcal{D}_q} \mathrm{KL}(q(X)||p(X|Y;\theta)). \tag{III.1}$$

Intuitively, if the posterior $p(X|Y;\theta)$ is tractable, *i.e.*, $p(X|Y;\theta) \in \mathcal{D}_q$, the variational inference is *exact* as $q^*(X) = p(X|Y;\theta)$ in (III.1) since 1) KL divergence is always non-negative, and 2) it vanishes if and only if p and q are identical [33]. It is worth noting that, using (III.1) as the metric to minimize the dissimilarity to the ground true posterior $p(X|Y;\theta)$, a unimodal approximate distribution $q(X)$ will most likely fit into only one mode of $p(X|Y;\theta)$ [12]. Thus care should be taken when the multi-modality of $p(X|Y;\theta)$ is crucial in the problem of interest.

III.A.1 Variational Inference for One Hidden Variable

In the case of intractable posteriors, direct solution to problem (III.1) is challenging in general. Alternatively, consider following decomposition of the log-evidence

$$\ln p(Y;\theta) = \mathscr{L}(q;\theta) + \mathrm{KL}(q(X)||p(X|Y;\theta)) \geqslant \mathscr{L}(q;\theta), \tag{III.2}$$

where

$$\mathscr{L}(q;\theta) = \int_X q(X) \ln \frac{p(Y,X;\theta)}{q(X)} dX = \langle \ln p(X,Y;\theta) \rangle_q + \mathrm{H}_q(X) \tag{III.3}$$

is an evidence lower bound (ELBO) of $\log p(Y;\theta)$ due to the non-negativeness of the KL divergence; and the last equality of (III.2) holds if and only if $q(X) = p(X|Y;\theta)$. Note that, since the log-evidence $\log p(Y;\theta)$ is fixed for the given model and observation, maximization of the lower bound $\mathscr{L}(q;\theta)$ with respect to q also minimizes $\mathrm{KL}(q(X)||p(X|Y;\theta))$:

$$q^*(X) = \arg\max_{q \in \mathcal{D}_q} \mathscr{L}(q;\theta) = \arg\min_{q \in \mathcal{D}_q} \mathrm{KL}(q(X)||p(X|Y;\theta)), \tag{III.4}$$

which is often adopted to determine $q^*(x)$ in practice. It could happen that evaluation of $\mathscr{L}(q,\theta)$ is impractical due to the complex form of $p(Y,X;\theta)$, *e.g.*, binary dynamic systems. In such case, we resort to optimizing another tractable lower bound $\tilde{\mathscr{L}}$ such that $\mathscr{L}(q;\theta) \geqslant \tilde{\mathscr{L}}(q;\theta)$. Another critical observation on the lower bound $\mathscr{L}(q;\theta)$ is that, given the observed data, it can also be regarded as a function of both the model $p(Y,X;\theta)$ and the variational distribution q. This has been shown to play a fundamental role in the generalized expectation-maximization algorithm [108], which is adopted in this work for parameter estimation in Section V.A.

III.A.2 Chain-rule of Variational Inference for Multiple Hidden Variables

If the model of interest contains more than one type of hidden variable, there are typically two general strategies to handle this case: *sequential* and *block approaches* [79]. Since the challenge in inference for binary dynamic systems comes from the irregularity of distribution rather than the scale, we followed the first strategy in this work. For this, we present an operational scheme, denoted the *chain rule of variational inference*, to compute the *joint* variational distribution of multiple hidden variables. Essentially, the scheme reduces the evaluation of the joint posterior distribution into a series of local estimation sub-problems, and solve them one by one. In each sub-problem, only one hidden variable is handled; and analytic approximation is applied only when it is needed, depending on the form of the distribution being considered, to guarantee a tight induced global lower bound.

We start by considering a model with an observed variable Y and two hidden variables X and Z [1]. Using the same notations in Section III.A.1, the lower bound of (III.3) becomes

$$\mathscr{L}(q;\theta) = \langle \ln p(X,Y,Z;\theta)\rangle_{q_{X,Z}} + \mathrm{H}_q(X,Z) \tag{III.5}$$

$$= \int_z q(z)\left[\int_x q(x|Z=z)\ln p(x,Y,z;\theta)dx + \mathrm{H}_q(X|Z=z)\right]dz + \mathrm{H}_q(Z) \tag{III.6}$$

$$= \int_z q(z)\mathscr{L}(q_{X|z};\theta,z)dz + \mathrm{H}_q(Z), \tag{III.7}$$

[1]For brevity, we only use two hidden variables for illustration. The idea, however, can be easily adapted to groups of variables, *e.g.*, both X and Z can contain multiple members such that $\boldsymbol{X} = \{X_i\}$ and $\boldsymbol{Z} = \{Z_i\}$.

where

$$\mathscr{L}(q_{X|z};\theta,z) = \int_x q(x|Z=z)\ln p(x,Y,z;\theta)dx + \mathrm{H}_q(X|Z=z) \tag{III.8}$$

is a lower bound of $\ln p(Y,z;\theta)$, which is a function of $q(X|Z=z)$; and (III.6) results from (III.5) due to (II.2). Maximization of (III.7) gives

$$\begin{aligned}
\max_{q_{X,Z}} \mathscr{L}(q;\theta) &= \max_{q_{X|Z},q_Z} \left[\int_z q(z)\mathscr{L}(q_{X|z};\theta,z)dz + \mathrm{H}_q(Z)\right] && \text{(III.9)}\\
&= \max_{q_Z} \left\{\int_z q(z)\left[\max_{q_{X|Z=z}} \mathscr{L}(q_{X|z};\theta,z)\right]dz + \mathrm{H}_q(Z)\right\} && \text{(III.10)}\\
&= \max_{q_Z} \left[\int_z q(z)\ \ln p^*(Y,z;\theta)dz + \mathrm{H}_q(Z)\right], && \text{(III.11)}
\end{aligned}$$

where

$$\ln p^*(Y,z;\theta) = \max_{q_{X|Z=z}} \mathscr{L}(q_{X|z};\theta,z); \tag{III.12}$$

and (III.10) holds since the coefficients for the expectation $\langle\cdot\rangle_{q_Z}$, which is a convex combination, are non-negative. Similar to the problem of (III.4), if $\mathscr{L}(q_{X|z};\theta,z)$ is intractable, another manageable lower bound $\tilde{\mathscr{L}}(q_{X|z};\theta,z)$ is used instead. This is where the approximation is applied for $p_{X|Z}$, which only changes the form of $p(X|Z)$ but not necessarily that of $p(Z)$. Intuitively, (III.10) factorizes the original variational inference of (III.9) into two sub-problems: 1) the *nested problem* of (III.12), which can be solved via single-variable variational inference as the problem of (III.4); and 2) the *root problem* of (III.11), which also can be solved the same way as that of (III.4) once the conditional variational distribution in the nested problem of (III.12) is determined. In the same fashion, schemes for models with more than two types of hidden variables can be derived too, which compute the conditional variational posteriors in the innermost-to-outermost direction. The

Algorithm 1: Chain Rule of Variational Inference

Input: a probabilistic model $p(Y, X_{1:n_X}; \theta)$, an observation y, a set of tractable distributions $\mathcal{D}_q$;

$i \leftarrow 1$;
$\ln p^*(y, x_{i+1:n_X}, x_i; \theta) \leftarrow \ln p(y, x_{i+1:n_X}, x_i; \theta)$;
for $i := 1$ *to* n_X **do**
choose a tractable (with respect to x_i) lower bound $\ln \tilde{p}^*(y, x_i, x_{i+1:n_X}; \theta)$ of the log-evidence $\ln p^*(y, x_i, x_{i+1:n_X}; \theta)$ such that

$$\ln p^*(y, x_i, x_{i+1:n_X}; \theta) \geqslant \ln \tilde{p}^*(y, x_i, x_{i+1:n_X}; \theta);$$

compute $q^*_{X_i|X_{i+1:n_X}}$ that optimizes $\tilde{\mathscr{L}}(q(x); \theta, x_{i+1:n_X})$ by solving

$$\ln p^*(y, x_{i+1:n_X}; \theta) \leftarrow \max_{q(x) \in \mathcal{D}_{q_{X_i}}} \tilde{\mathscr{L}}(q(x); \theta, x_{i+1:n_X}),$$

where
$\tilde{\mathscr{L}}(q(x); \theta, x_{i+1:n_X}) = \int q(x) \ln \tilde{p}^*(y, x, x_{i+1:n_X}; \theta) dx + \mathrm{H}_q(X_i)$;
$i \leftarrow i + 1$;
end

Output: $q(X_{1:n_X}) = \prod_i q^*(X_i | X_{i+1:n_X})$

procedure is summarized in Algorithm 1. To facilitate evaluation in practice, the order to evaluate hidden variables can be derived by exploiting the topological structure of the original graphical model, *e.g.*, using the factorization properties in Bayesian networks [133]. Note that, no assumption of independence is made on any steps in the derivation above. Instead, the correlation in the original model, which could be crucial, is preserved and encoded via the conditional variational distribution in each sub-problem. On the other hand, the cost of the chain rule is that, the complexity may quickly become prohibitive as the scale of the problem increases, making it impractical or impossible to implement.

If full independence is assumed among each hidden variables of the variational distribution in Algorithm 1 such that

$$q(\{X_i\}) = \prod_i q(X_i), \tag{III.13}$$

the procedure becomes another closely related and popular technique called *factorial approximation* [12], or *mean field approximation* [163], which is inspired by the mean field theory from the statistical mechanics literature [116]. The representation is designed to efficiently depict the behavior of an enormous stochastic models with a large number of random nodes that interact with each other. In this case, the intractability typically stems from the combinatorial configurations and the complex entanglement. To cope with these challenges, the model is fully factorized into a field of *independent* variables; and the inter-node interaction is approximated by an averaged effect or estimated mean (which justifies its name). This leads to a manageable inference that can be solved iteratively through gradient descent with convergence to local optimum [163]. While this scheme can handle problems at scale [120, 178, 137, 65, 170, 24, 135], performance in other scenarios can be seriously affected as the oversimplified assumption of full independence fails to capture some critical correlation [80, 61, 12]. For this reason, factorial approximation is less desirable than the chain rule for inference in mixtures of binary dynamic systems, where the dependence between mixture cluster assignments and state sequences plays a crucial role. Nevertheless, both methods are not exclusive to each other. Actually, they can work in a hybrid framework to complement each other for much more powerful representation with both flexibility and tractability, *e.g.*, using partial factorization to reduce global complexity through removal of weak inter-group correlation, while applying the finer modeling scheme to local substructure [136, 8, 171].

III.B Inference for Linear Dynamic Systems

Before presenting the inference for the binary dynamic system and its mixture version, we first brief review the inference for the linear dynamic system. As we will see later, the inference of BDS can be solved efficiently with similar message passing routine.

III.B.1 Solution to Inference of Linear Dynamic Systems

Consider the LDS of (II.3) with parameters $\boldsymbol{\theta}_{LDS} = \{\boldsymbol{S}, \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{u}\}$, an observation sequence $\boldsymbol{y}_{1:\tau}$ ($\boldsymbol{y}_t \in \mathbb{R}^D$), and the variational distribution $q(\boldsymbol{x})$ of (III.36). The ELBO of (III.3) for the LDS is

$$\mathscr{L}(q;\boldsymbol{\theta},\boldsymbol{y}) = \langle \ln p(\boldsymbol{x}_1) \rangle_q + \sum_{t=1}^{\tau-1} \langle \ln p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) \rangle_q + \sum_{t=1}^{\tau} \langle \ln p(\boldsymbol{y}_t|\boldsymbol{x}_t) \rangle_q + \mathrm{H}_q(X). \tag{III.14}$$

It can be shown that (see Appendix III.F.2), the optimal q^* that maximizes (III.14) is a Gaussian of the form

$$q(\boldsymbol{x}_{1:\tau}) = \mathcal{G}(\boldsymbol{x}_{1:\tau}; \boldsymbol{m}, \boldsymbol{\Phi}),\ \boldsymbol{m} \in \mathbb{R}^{L\tau \times 1},\ \boldsymbol{\Phi} \in \mathcal{S}_{++}^{L\tau}, \tag{III.15}$$

where $\boldsymbol{m}_{[i]} \in \mathbb{R}^L$ and $\boldsymbol{\Phi}_{[i,j]} \in \mathbb{R}^{L \times L}$ are the mean of $\boldsymbol{x}_i$ and covariance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, respectively,

$$\boldsymbol{m}_{[i]} = \langle \boldsymbol{x}_i \rangle_q,\ \ \boldsymbol{\Phi}_{[i,j]} = \left\langle (\boldsymbol{x}_i - \boldsymbol{m}_{[i]})(\boldsymbol{x}_j - \boldsymbol{m}_{[j]})^{\mathsf{T}} \right\rangle_q.$$

Defining $\tilde{\boldsymbol{y}}_t = \boldsymbol{y}_t - \boldsymbol{u}$,

$$\mathscr{L}(q;\boldsymbol{\theta},\boldsymbol{y}) = \langle \ln \mathcal{G}(\tilde{\boldsymbol{y}}_t; \boldsymbol{C}\boldsymbol{x}_t, \boldsymbol{R}) \rangle_{q(\boldsymbol{x}_t)} = \langle \ln \mathcal{G}(\boldsymbol{x}_t; \tilde{\boldsymbol{y}}_t, \boldsymbol{R}) \rangle_{\mathcal{G}(\boldsymbol{x}_t; \boldsymbol{C}\boldsymbol{m}_{[t]}, \boldsymbol{C}\boldsymbol{\Phi}_{[t,t]}\boldsymbol{C}^{\mathsf{T}})}, \tag{III.16}$$

and, from (II.3b),

$$\langle \ln p(\boldsymbol{y}_t|\boldsymbol{x}_t)\rangle_q \;\; \propto \;\; -\frac{1}{2}\Big[||\tilde{\boldsymbol{y}}_t - \boldsymbol{C}\boldsymbol{m}_{[t]}||^2_{\boldsymbol{R}} + \mathrm{tr}(\boldsymbol{R}^{-1}\boldsymbol{C}\boldsymbol{\Phi}_{[t,t]}\boldsymbol{C}^{\mathsf{T}})\Big].$$

It follows that

$$\begin{aligned}\mathscr{L}(q;\boldsymbol{\theta},\boldsymbol{y}) \propto -\frac{1}{2}\Bigg\{ & ||\boldsymbol{\mu} - \boldsymbol{m}_{[1]}||^2_{\boldsymbol{S}} + \mathrm{tr}(\boldsymbol{S}^{-1}\boldsymbol{\Phi}_{[1,1]}) \\ & + \sum_{t=1}^{\tau-1} \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t+1} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \boldsymbol{A}^{\mathsf{T}}\boldsymbol{Q}^{-1}\boldsymbol{A} & -\boldsymbol{A}^{\mathsf{T}}\boldsymbol{Q}^{-1} \\ -\boldsymbol{Q}^{-1}\boldsymbol{A} & \boldsymbol{Q}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t+1} \end{bmatrix} \\ & + \sum_{t=1}^{\tau} \mathrm{tr}(\boldsymbol{R}^{-1}\boldsymbol{C}\boldsymbol{\Phi}_{[t,t]}\boldsymbol{C}^{\mathsf{T}}) + \sum_{t=1}^{\tau} ||\tilde{\boldsymbol{y}}_t - \boldsymbol{m}_{[t]}||^2_{\boldsymbol{R}} \Bigg\} + \frac{1}{2}\ln|\boldsymbol{\Phi}|. \end{aligned} \tag{III.17}$$

The optimization of (III.17) with respect to the variational distribution q can be factorized into two optimization problems

$$\{\boldsymbol{m}^*, \boldsymbol{\Phi}^*\} = \underset{\{\boldsymbol{m},\boldsymbol{\Phi}\}\in\mathbb{R}^{L\tau}\times\mathcal{S}^{L\tau}_{++}}{\arg\max} \mathscr{L}(q;\boldsymbol{\theta},\boldsymbol{y}) = \Big\{ \underset{\boldsymbol{m}\in\mathbb{R}^{L\tau}}{\arg\max} \mathscr{L}(q;\boldsymbol{\theta},\boldsymbol{y}), \underset{\boldsymbol{\Phi}\in\mathcal{S}^{L\tau}_{++}}{\arg\max} \mathscr{L}(q;\boldsymbol{\theta},\boldsymbol{y}) \Big\}.$$

Consolidating the terms containing $\boldsymbol{\Phi}$,

$$\begin{gathered}\boldsymbol{\Phi}^* = \underset{\boldsymbol{\Phi}}{\arg\max}\; \ln|\boldsymbol{\Phi}| - \mathrm{tr}(\boldsymbol{W}_{\mathrm{LDS}}\boldsymbol{\Phi}), \\ s.t.\; \boldsymbol{\Phi} \in \mathcal{S}^{L\tau}_{++}, \end{gathered} \tag{III.18}$$

where

$$
W_{\mathrm{LDS}[i,j]} = \begin{cases} A^\intercal Q^{-1} A + S^{-1} + C^\intercal R^{-1} C, & i = j = 1, \\ A^\intercal Q^{-1} A + Q^{-1} + C^\intercal R^{-1} C, & 1 < i = j < \tau, \\ Q^{-1} + C^\intercal R^{-1} C, & i = j = \tau, \\ -Q^{-1} A, & i = j + 1, \\ -A^\intercal Q^{-1}, & i = j - 1, \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{III.19}
$$

It can be shown that (see Appendix III.F.3), the solution to (III.18) is

$$
\mathbf{\Phi}^* = {W_{\mathrm{LDS}}}^{-1}. \tag{III.20}
$$

Similarly, we have

$$
m^* = W_{\mathrm{LDS}}^{-1} \beta, \;\; \beta = \begin{bmatrix} \beta_{[1]} \\ \vdots \\ \beta_{[\tau]} \end{bmatrix}, \;\; \beta_{[t]} = \begin{cases} S^{-1} \mu + C^\intercal {R_1}^{-1} \tilde{u}_1, & t = 1, \\ C^\intercal {R_t}^{-1} \tilde{u}_t, & 1 < t \leqslant \tau. \end{cases} \tag{III.21}
$$

On the other hand, since all random variables x and y (as well as all marginal or conditional distributions) of the LDS are Gaussian, the variational inference is *exact* in the case of LDS, and

$$
q^*(x) = p(x|y;\theta_{LDS}) = \mathcal{G}(x_{1:\tau}; m, \mathbf{\Phi}).
$$

In the following section, we briefly review the Kalman smoothing filter [146, 131], which efficiently computes the solution of (III.20) and (III.21).

III.B.2 Kalman Smoothing Filter

The key step in the variational inference of Section III.B.1 is to determine

$$\begin{aligned} \boldsymbol{m}_{[t]} &= \langle \boldsymbol{x}_t \rangle_q , \\ \boldsymbol{\Phi}_{[t,t]} &= \left\langle (\boldsymbol{x}_t - \boldsymbol{m}_{[t]})(\boldsymbol{x}_t - \boldsymbol{m}_{[t]})^\top \right\rangle_q , \\ \boldsymbol{\Phi}_{[t,t+1]} &= \left\langle (\boldsymbol{x}_t - \boldsymbol{m}_{[t]})(\boldsymbol{x}_{t+1} - \boldsymbol{m}_{[t+1]})^\top \right\rangle_q . \end{aligned}$$

In this appendix, we derive an efficient method for this computation, which draws on the solution of the identical variational inference problem for the LDS of (II.3).

Defining expectations conditioned on the observed sequence from time $t = 1$ to $t = r$ as

$$\hat{\boldsymbol{x}}_t^r = \langle \boldsymbol{x}_t \rangle_{p(\boldsymbol{x}_t|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_r)} , \tag{III.22}$$

$$\boldsymbol{V}_{t,k}^r = \left\langle (\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t^r)(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k^r)^\top \right\rangle_{p(\boldsymbol{x}_t,\boldsymbol{x}_k|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_r)} , \tag{III.23}$$

the estimates are calculated via the *forward* and *backward* recursions:

- In the **forward recursion**, for $t = 1, \cdots, \tau$, compute

$$\boldsymbol{V}_{t,t}^{t-1} = \boldsymbol{A}\boldsymbol{V}_{t-1,t-1}^{t-1}\boldsymbol{A}^\top + \boldsymbol{Q}, \tag{III.24}$$

$$\boldsymbol{K}_t = \boldsymbol{V}_t^{t-1}\boldsymbol{C}^\top(\boldsymbol{C}\boldsymbol{V}_{t,t}^{t-1}\boldsymbol{C}^\top + \boldsymbol{R}_t)^{-1}, \tag{III.25}$$

$$\boldsymbol{V}_{t,t}^{t} = \boldsymbol{V}_{t,t}^{t-1} - \boldsymbol{K}_t\boldsymbol{C}\boldsymbol{V}_{t,t}^{t-1}, \tag{III.26}$$

$$\hat{\boldsymbol{x}}_t^{t-1} = \boldsymbol{A}\hat{\boldsymbol{x}}_{t-1}^{t-1}, \tag{III.27}$$

$$\hat{\boldsymbol{x}}_t^{t} = \hat{\boldsymbol{x}}_t^{t-1} + \boldsymbol{K}_t(\tilde{\boldsymbol{y}}_t - \boldsymbol{C}\hat{\boldsymbol{x}}_t^{t-1}), \tag{III.28}$$

with initial conditions $\hat{\boldsymbol{x}}_1^0 = \boldsymbol{\mu}$ and $\boldsymbol{V}_{1,1}^0 = \boldsymbol{S}$.

- In the **backward recursion**, for $t = \tau, \cdots, 1$,

$$\boldsymbol{J}_{t-1} = \boldsymbol{V}_{t-1,t-1}^{t-1} \boldsymbol{A}^\intercal (\boldsymbol{V}_{t,t}^{t-1})^{-1}, \tag{III.29}$$

$$\hat{\boldsymbol{x}}_{t-1}^{\tau} = \hat{\boldsymbol{x}}_{t-1}^{t-1} + \boldsymbol{J}_{t-1}(\hat{\boldsymbol{x}}_t^{\tau} - \boldsymbol{A}\hat{\boldsymbol{x}}_{t-1}^{t-1}), \tag{III.30}$$

$$\boldsymbol{V}_{t-1,t-1}^{\tau} = \boldsymbol{V}_{t-1,t-1}^{t-1} + \boldsymbol{J}_{t-1}(\boldsymbol{V}_{t,t}^{\tau} - \boldsymbol{V}_{t,t}^{t-1})\boldsymbol{J}_{t-1}{}^\intercal, \tag{III.31}$$

and for $t = \tau, \cdots, 2$,

$$\boldsymbol{V}_{t-1,t-2}^{\tau} = \boldsymbol{V}_{t-1,t-1}^{t-1} \boldsymbol{J}_{t-2}{}^\intercal + \boldsymbol{J}_{t-1}(\boldsymbol{V}_{t,t-1}^{\tau} - \boldsymbol{A}\boldsymbol{V}_{t-1,t-1}^{t-1})\boldsymbol{J}_{t-2}{}^\intercal \tag{III.32}$$

with initial condition $\boldsymbol{V}_{\tau,\tau-1}^{\tau} = (\boldsymbol{I} - \boldsymbol{K}_\tau \boldsymbol{C})\boldsymbol{A}\boldsymbol{V}_{\tau-1,\tau-1}^{\tau-1}$.

The final result for the inference of LDS is

$$q^*(\boldsymbol{x}_t) = \mathcal{G}(\boldsymbol{x}_t; \boldsymbol{m}_{[t]}, \boldsymbol{\Phi}_{[t,t]}) = \mathcal{G}(\boldsymbol{x}_t; \hat{\boldsymbol{x}}_t^{\tau}, \hat{\boldsymbol{V}}_{t,t}^{\tau}); \tag{III.33}$$

and

$$\begin{aligned} q^*(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) &= \mathcal{G}\left(\begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t+1} \end{bmatrix}; \begin{bmatrix} \boldsymbol{m}_{[t]} \\ \boldsymbol{m}_{[t+1]} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Phi}_{[t,t]} & \boldsymbol{\Phi}_{[t,t+1]} \\ \boldsymbol{\Phi}_{[t+1,t]} & \boldsymbol{\Phi}_{[t+1,t+1]} \end{bmatrix} \right) \\ &= \mathcal{G}\left(\begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t+1} \end{bmatrix}; \begin{bmatrix} \hat{\boldsymbol{x}}_t^{\tau} \\ \hat{\boldsymbol{x}}_{t+1}^{\tau} \end{bmatrix}, \begin{bmatrix} \hat{\boldsymbol{V}}_{t,t}^{\tau} & \hat{\boldsymbol{V}}_{t,t+1}^{\tau} \\ \hat{\boldsymbol{V}}_{t+1,t}^{\tau} & \hat{\boldsymbol{V}}_{t+1,t+1}^{\tau} \end{bmatrix} \right). \end{aligned} \tag{III.34}$$

The overall complexity is $O(L^\kappa \tau)$, where $\kappa \in [2.38, 3]$ is the constant coefficient in the complexity of $n \times n$ matrix product $O(n^\kappa)$ [32], since the routine only involves matrix manipulation for $L \times L$ matrices, and the number of operations is linear in the length of the sequence (τ).

III.C Inference of Hidden States in Binary Dynamic Systems

To apply the variational method of Section III.A to inferring hidden state in the BDS, we consider its complete-data log-evidence by taking the logarithm of (II.9) (up to constants independent of all variables and the parameter):

$$\begin{aligned}\ln p(\boldsymbol{x}_{1:\tau},\boldsymbol{y}_{1:\tau};\boldsymbol{\theta}) = \\ &-\frac{1}{2}\ln|\boldsymbol{S}| - \Big(\frac{\tau-1}{2}\Big)\ln|\boldsymbol{Q}| - \frac{1}{2}\|\boldsymbol{x}_1-\boldsymbol{\mu}\|^2_{S} - \frac{1}{2}\sum_{t=1}^{\tau-1}\|\boldsymbol{x}_{t+1}-\boldsymbol{A}\boldsymbol{x}_t\|^2_{Q} \\ &+\sum_{t,d}\Big[y_{dt}\ln\sigma(\boldsymbol{C}_{d,:}\boldsymbol{x}_t+u_d)+(1-y_{dt})\ln\sigma(-\boldsymbol{C}_{d,:}\boldsymbol{x}_t-u_d)\Big]+\text{const.}\end{aligned} \tag{III.35}$$

The irregular form of the sigmoid non-linearity makes the posterior $p(\boldsymbol{x}_{1:\tau}|\boldsymbol{y}_{1:\tau})$ intractable (note that $p(\boldsymbol{x}_{1:\tau}|\boldsymbol{y}_{1:\tau}) \propto p(\boldsymbol{x}_{1:\tau},\boldsymbol{y}_{1:\tau})$). It can be shown that, however, the log-evidence of (III.35) is a *concave* function in $\boldsymbol{x}_{1:\tau}$, thus the ground true posterior $p(\boldsymbol{x}_{1:\tau}|\boldsymbol{y}_{1:\tau})$ is *unimodal* (see Appendix III.F.1 for discussion), which justifies the appropriateness of variational methods in approximating $p(\boldsymbol{x}_{1:\tau}|\boldsymbol{y}_{1:\tau};\boldsymbol{\theta})$. To address the technical difficulty of the expectation of $\ln\sigma(\cdot)$ in (III.3), two lower-bounds are considered. These lead to two algorithms for inference and learning of different complexities. For brevity, we denote the mean of the variational distribution as $\boldsymbol{m}\in\mathbb{R}^{L\tau\times 1}$ (and $\tilde{\boldsymbol{m}}_{[t]} = [\boldsymbol{m}_{[t]}{}^{\intercal},1]^{\intercal}$.), the covariance as $\boldsymbol{\Phi}\in\mathcal{S}^{L\tau}_{++}$, the second order moment as $\boldsymbol{P}\in\mathcal{S}^{L\tau}_{++}$.

III.C.1 Variational Inference with ELBO_{SJ}

Consider a multivariate Gaussian distribution of *full* covariance for $q(\boldsymbol{x})$,

$$q(\boldsymbol{x}_{1:\tau}) = \mathcal{G}(\boldsymbol{x}_{1:\tau}; \boldsymbol{m}, \boldsymbol{\Phi}),\ \boldsymbol{m} \in \mathbb{R}^{L\tau\times 1}, \boldsymbol{\Phi} \in \mathcal{S}_{++}^{L\tau}, \tag{III.36}$$

where $\boldsymbol{m}_{[t]} \in \mathbb{R}^{L}$ and $\boldsymbol{\Phi}_{[r,s]} \in \mathbb{R}^{L\times L}$ are the mean of $\boldsymbol{x}_t$ and covariance between $\boldsymbol{x}_r$ and $\boldsymbol{x}_s$, respectively,

$$\boldsymbol{m}_{[t]} = \langle \boldsymbol{x}_t \rangle_q,\ \ \boldsymbol{\Phi}_{[r,s]} = \Big\langle (\boldsymbol{x}_r - \boldsymbol{m}_{[r]})(\boldsymbol{x}_s - \boldsymbol{m}_{[s]})^\intercal \Big\rangle_q .$$

Since ω (a linear projection of $\boldsymbol{x}$) is Gaussian, $\langle \ln \sigma(\omega) \rangle_q$ is bounded by

$$\langle \ln \sigma(\omega) \rangle_q \geqslant \ln \sigma(\langle \omega \rangle_q) - \frac{1}{8}\mathrm{var}(\omega), \tag{III.37}$$

which results from setting $\xi = 1/2$ in (A.10) of [138]. This leads to a new lower bound $\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q)$ of (III.35)

$$\begin{aligned}\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q) = -\frac{1}{2}\Big\{ & ||\boldsymbol{\mu} - \boldsymbol{m}_{[1]}||_{S}^{2} + \mathrm{tr}(\boldsymbol{S}^{-1}\boldsymbol{\Phi}_{[1,1]}) + \frac{1}{4}\sum_t \mathrm{tr}(\boldsymbol{C}\boldsymbol{\Phi}_{[t,t]}\boldsymbol{C}^\intercal) \\ & + \sum_{t=1}^{\tau-1} \mathrm{tr}\Big(\boldsymbol{Q}^{-1}(\hat{\boldsymbol{P}}_{t+1,t+1} - \hat{\boldsymbol{P}}_{t+1,t}\boldsymbol{A}^\intercal - \boldsymbol{A}\hat{\boldsymbol{P}}_{t,t+1} + \boldsymbol{A}\hat{\boldsymbol{P}}_{t,t}\boldsymbol{A}^\intercal)\Big)\Big\} \\ & + \sum_{t,d}\Big[y_{d,t}\ln\sigma(\hat{\omega}_{d,t}) + (1 - y_{d,t})\ln\sigma(-\hat{\omega}_{d,t})\Big] + \frac{1}{2}\ln|\boldsymbol{\Phi}| + \mathrm{const},\end{aligned} \tag{III.38}$$

where $\hat{\boldsymbol{P}}_{r,s} = \langle \boldsymbol{x}_r\boldsymbol{x}_s^\intercal \rangle_{q_i(\boldsymbol{x}|j)} = \boldsymbol{\Phi}_{[r,s]} + \boldsymbol{m}_{[r]}\boldsymbol{m}_{[s]}{}^\intercal$ and $\hat{\omega}_{d,t} = \langle \omega_{d,t} \rangle_q = \boldsymbol{C}_{d,:}\boldsymbol{m}_{[t]} + u_d$.

The variational distribution $q^*(\boldsymbol{x})$ is the solution of

$$\{\boldsymbol{m}^*, \boldsymbol{\Phi}^*\} = \underset{\{\boldsymbol{m},\boldsymbol{\Phi}\}\in\mathbb{R}^{L\tau}\times\mathcal{S}_{++}^{L\tau}}{\arg\max}\ \hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q). \tag{III.39}$$

This is a *convex optimization* problem, since all terms of $\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q)$, depend on either $\boldsymbol{\Phi}$ or $\boldsymbol{m}$ separately (not on both), have the convex domain $(\boldsymbol{m}, \boldsymbol{\Phi}) \in \mathbb{R}^{L\tau} \times \mathcal{S}_{++}^{L\tau}$ and are concave - either a) linear functions, b) quadratic functions of negative definite coefficient matrices, c) negative log-sum-exp functions, or d) log determinant of $\boldsymbol{\Phi}$. Furthermore, (III.39) can be factorized into

$$\{\boldsymbol{m}^*, \boldsymbol{\Phi}^*\} = \underset{\{\boldsymbol{m},\boldsymbol{\Phi}\}\in\mathbb{R}^{L\tau}\times\mathcal{S}_{++}^{L\tau}}{\arg\max}\ \hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q) = \left\{ \underset{\boldsymbol{m}\in\mathbb{R}^{L\tau}}{\arg\max} \hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q), \underset{\boldsymbol{\Phi}\in\mathcal{S}_{++}^{L\tau}}{\arg\max} \hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q) \right\}.$$

Consolidating the terms containing $\boldsymbol{\Phi}$,

$$\begin{gathered} \boldsymbol{\Phi}^* = \underset{\boldsymbol{\Phi}}{\arg\max}\ \ln|\boldsymbol{\Phi}| - \operatorname{tr}(\boldsymbol{W}_{SJ}\boldsymbol{\Phi}), \\ s.t.\ \boldsymbol{\Phi} \in \mathcal{S}_{++}^{L\tau}, \end{gathered} \tag{III.40}$$

where $\boldsymbol{W}_{SJ} \in \mathcal{S}_{++}^{L\tau}$ is a positive-definite matrix such that

$$\boldsymbol{W}_{SJ[i,j]} = \begin{cases} \boldsymbol{A}^\intercal \boldsymbol{Q}^{-1}\boldsymbol{A} + \boldsymbol{S}^{-1} + \frac{1}{4}\boldsymbol{C}^\intercal\boldsymbol{C}, & i = j = 1, \\ \boldsymbol{A}^\intercal \boldsymbol{Q}^{-1}\boldsymbol{A} + \boldsymbol{Q}^{-1} + \frac{1}{4}\boldsymbol{C}^\intercal\boldsymbol{C}, & 1 < i = j < \tau, \\ \boldsymbol{Q}^{-1} + \frac{1}{4}\boldsymbol{C}^\intercal\boldsymbol{C}, & i = j = \tau, \\ -\boldsymbol{Q}^{-1}\boldsymbol{A}, & i = j + 1, \\ -\boldsymbol{A}^\intercal\boldsymbol{Q}^{-1}, & i = j - 1, \\ \boldsymbol{0}, & \text{otherwise}. \end{cases}$$

It can be shown that (see Appendix III.F.3), this has optimal solution

$$\boldsymbol{\Phi}^{*} = \boldsymbol{W}_{SJ}{}^{-1}. \tag{III.41}$$

While (III.41) is conceptually straightforward, the inversion of the matrix $\boldsymbol{W}_{SJ}^{-1}$ can be too expensive for long video sequences (large τ).

In most cases, nevertheless, only 1) the value of the ELBO , 2) the mean $\boldsymbol{m}$, and 3) state covariances at each time step and between two adjacent time steps, $\boldsymbol{\Phi}_{[t,t]}$ and $\boldsymbol{\Phi}_{[t,t+1]}$, are needed, *e.g.*, in model parameter estimation. Alternatively, note that the structure of (III.41) resembles that of the LDS of (III.19), thus the popular *Kalman smoothing filter* [131] can be adopted to compute the ELBO, parameters $\boldsymbol{\Phi}^{*}{}_{[t,t]}$ and $\boldsymbol{\Phi}^{*}{}_{[t,t+1]}$, using the same routine in Section III.B.2 with proper substitution of parameters.

The optimal variational mean parameter $\boldsymbol{m}^{*}$ has no closed form solution, due to the log-sigmoid terms of (III.38). We rely on a numerical procedure for determining the stationary point of $\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q^{*})$ for $\boldsymbol{m}$. Since the problem is convex, this suffices to guarantee a global optimum. Specifically, the variational mean $\boldsymbol{m}$ is the solution of

$$\begin{aligned}
\boldsymbol{m}^{*} &= \arg\max_{\boldsymbol{m}} \hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q) && \text{(III.42)}\\
&= \arg\max_{\boldsymbol{m}} \Bigg\{ \boldsymbol{\mu}^{\intercal}\boldsymbol{S}^{-1}\boldsymbol{m}_{[1]} - \frac{1}{2}\boldsymbol{m}_{[1]}^{\intercal}\boldsymbol{S}^{-1}\boldsymbol{m}_{[1]} && \text{(III.43)}\\
&\qquad - \frac{1}{2}\sum_{t=1}^{\tau-1} \begin{bmatrix} \boldsymbol{m}_{[t]} \\ \boldsymbol{m}_{[t+1]} \end{bmatrix}^{\intercal} \begin{bmatrix} \boldsymbol{A}^{\intercal}\boldsymbol{Q}^{-1}\boldsymbol{A} & -\boldsymbol{A}^{\intercal}\boldsymbol{Q}^{-1} \\ -\boldsymbol{Q}^{-1}\boldsymbol{A} & \boldsymbol{Q}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{m}_{[t]} \\ \boldsymbol{m}_{[t+1]} \end{bmatrix}\\
&\qquad + \sum_{t,k} \Big[y_{d,t} \ln \sigma(\hat{\omega}_{d,t}) + (1 - y_{d,t}) \ln \sigma(-\hat{\omega}_{d,t}) \Big] \Bigg\}.
\end{aligned}$$

This can be rewritten as

$$\boldsymbol{m}^* = \arg\max_{\boldsymbol{m}} \Big\{ -\boldsymbol{m}^\intercal \tilde{\boldsymbol{W}} \boldsymbol{m} + \boldsymbol{b}_1^\intercal \boldsymbol{m}_{[1]} - \sum_{t,k} \Big[y_{d,t} \ln(1 + \exp(-\boldsymbol{C}_{d,:}\boldsymbol{m}_{[t]} - u_d)) \\ + (1 - y_{d,t}) \ln(1 + \exp(\boldsymbol{C}_{d,:}\boldsymbol{m}_{[t]} + u_d)) \Big] \Big\}, \tag{III.44}$$

where

$$\tilde{\boldsymbol{W}}_{[i,j]} = \begin{cases} \boldsymbol{A}^\intercal \boldsymbol{Q}^{-1} \boldsymbol{A} + \boldsymbol{S}^{-1}, & i = j = 1, \\ \boldsymbol{A}^\intercal \boldsymbol{Q}^{-1} \boldsymbol{A} + \boldsymbol{Q}^{-1}, & 1 < i = j < \tau, \\ \boldsymbol{Q}^{-1}, & i = j = \tau, \\ -\boldsymbol{Q}^{-1}\boldsymbol{A}, & i = j + 1, \\ -\boldsymbol{A}^\intercal \boldsymbol{Q}^{-1}, & i = j - 1, \\ \boldsymbol{0}, & \text{otherwise}, \end{cases} \tag{III.45}$$

$\hat{\omega}_{d,t} = \boldsymbol{C}_{d,:}\boldsymbol{m}_{[t]} + u_d$, and $\boldsymbol{b}_1 = 2\boldsymbol{S}^{-1}\boldsymbol{\mu}$. Since $\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q)$ is a concave function of $\boldsymbol{m} \in \mathbb{R}^{\tau L}$, gradient-based methods can be applied to search for the stationary point where global optimum is guaranteed.

The gradient of $\hat{\mathscr{L}}(\boldsymbol{\theta}, q)$ is

$$\frac{\partial}{\partial \boldsymbol{m}} \hat{\mathscr{L}}(\boldsymbol{\theta}, q) = -\tilde{\boldsymbol{W}}\boldsymbol{m} + \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{C}^\intercal & & \\ & \ddots & \\ & & \boldsymbol{C}^\intercal \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_\tau \end{bmatrix}, \tag{III.46}$$

where

$$\boldsymbol{\beta}_t = [\sigma(\hat{\omega}_{d,t}) - y_{1t}, \ \cdots, \ \sigma(\hat{\omega}_{Dt}) - y_{Dt}]^\intercal;$$

The second-order partial derivatives of $\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q)$ is

$$\frac{\partial^2}{\partial \boldsymbol{m}^2} \hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q) = -\tilde{\boldsymbol{W}} - \begin{bmatrix} \boldsymbol{C}^\mathsf{T} \Xi_1 \boldsymbol{C} & & \\ & \ddots & \\ & & \boldsymbol{C}^\mathsf{T} \Xi_\tau \boldsymbol{C} \end{bmatrix}, \tag{III.47}$$

where $\Xi_t = \mathrm{diag}(\sigma(\hat{\omega}_{1,t})\sigma(-\hat{\omega}_{1,t}), \; \cdots, \; \sigma(\hat{\omega}_{D,t})\sigma(-\hat{\omega}_{D,t}),)$. Given the concavity and smoothness of $\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q)$, many popular numerical optimization algorithms can be utilized to search for its optimum, *e.g.*, gradient descent, Newton-Raphson method, Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, *etc*.

III.C.2 Variational Inference with ELBO$_{JJ}$

Noting that (II.4b) can also be interpreted as the Bayesian logistic regression, we adopt the lower bound $\tilde{\sigma}(x;\xi)$ of $\sigma(\cdot)$ in [66] such that

$$\sigma(x) \geqslant \tilde{\sigma}(x;\xi) = \sigma(\xi) \exp\Big\{ -\lambda(\xi)(x^2 - \xi^2) + \frac{x-\xi}{2} \Big\}, \;\; \lambda(\xi) = \frac{1}{2\xi}\Big[\sigma(\xi) - \frac{1}{2}\Big], \tag{III.48}$$

where $\xi > 0$ is the parameter that controls the shape of $\tilde{\sigma}(x;\xi)$. This has been shown to achieve good performance in Bayesian logistic regression [66, 67]. Combining (III.48) with (III.35) and substituting them into (III.3) leads to the variational lower bound (up to constants independent of $q(\boldsymbol{x})$ and $\boldsymbol{\xi}$)

$$\begin{aligned}\tilde{\mathscr{L}}_{JJ}(q,\boldsymbol{\xi};\boldsymbol{\theta}) =\Big\langle & -\frac{1}{2}(\boldsymbol{x}_1-\boldsymbol{\mu})^\intercal \boldsymbol{S}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}) - \frac{1}{2}\sum_{t=1}^{\tau}(\boldsymbol{C}\boldsymbol{x}_t-\tilde{\boldsymbol{u}}_t)^\intercal \tilde{\boldsymbol{R}}_t^{-1}(\boldsymbol{C}\boldsymbol{x}_t-\tilde{\boldsymbol{u}}_t)\\ & -\frac{1}{2}\sum_{t=1}^{\tau-1}\begin{bmatrix}\boldsymbol{x}_t\\ \boldsymbol{x}_{t+1}\end{bmatrix}^\intercal \begin{bmatrix}\boldsymbol{A}^\intercal\boldsymbol{Q}^{-1}\boldsymbol{A} & -\boldsymbol{A}^\intercal\boldsymbol{Q}^{-1}\\ -\boldsymbol{Q}^{-1}\boldsymbol{A} & \boldsymbol{Q}^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_t\\ \boldsymbol{x}_{t+1}\end{bmatrix}\Big\rangle_{q(\boldsymbol{x})}\\ & +\sum_{t,d}\zeta(\xi_{d,t}) + \mathrm{H}_q(X) + \text{const},\end{aligned} \tag{III.49}$$

where $\boldsymbol{\xi}\in\mathbb{R}_{++}^{D\times\tau}$ is the variational parameter,

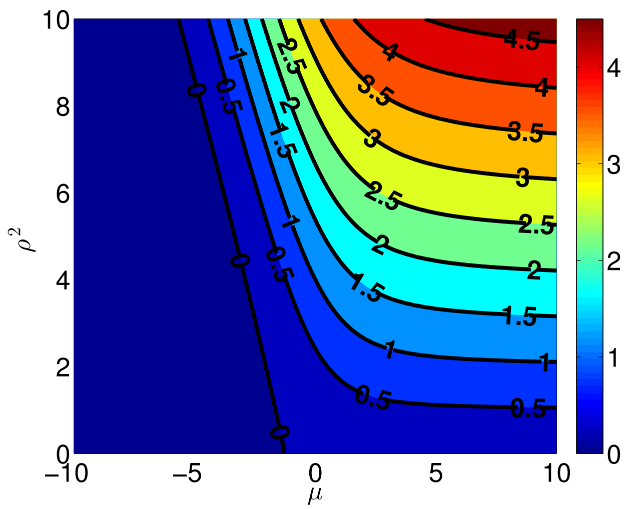
$$\tilde{\boldsymbol{R}}_t^{-1} = 2\text{diag}\{\lambda(\xi_{1,t}),\cdots,\lambda(\xi_{D,t})\}\succ \boldsymbol{0},\quad \tilde{\boldsymbol{u}}_t = \frac{1}{4}\Big[\frac{2y_{1,t}-1}{\lambda(\xi_{1,t})},\cdots,\frac{2y_{D,t}-1}{\lambda(\xi_{D,t})}\Big]^\intercal - \boldsymbol{u}, \tag{III.50}$$

and

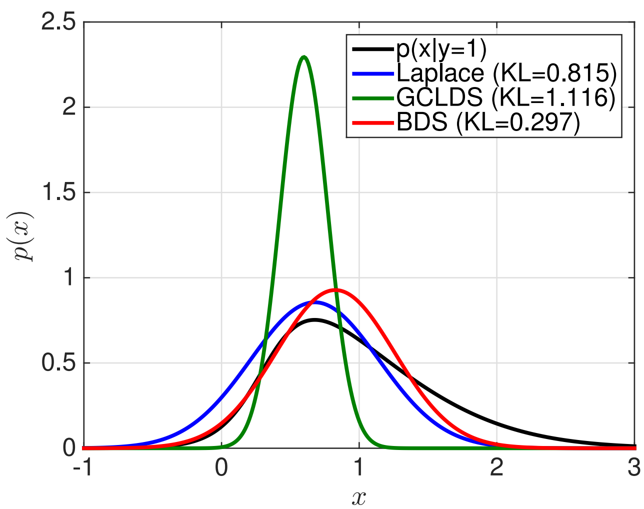
$$\zeta(\xi) = \ln\sigma(\xi) + \lambda(\xi)\xi^2 - \frac{1}{2}\xi + \frac{1}{16\lambda(\xi)}. \tag{III.51}$$

The lower bound of (III.49) is a function of the variational distribution q and the variational parameter $\boldsymbol{\xi}$. Since both are entangled in a complex way, we resort to coordinate descent to search the optimum. This inspires an optimization scheme similar to the EM algorithm, which alternates between maximizing $\tilde{\mathscr{L}}_{JJ}(q,\boldsymbol{\xi};\boldsymbol{\theta})$ over q while fixing $\boldsymbol{\xi}$ and vice versa. The whole procedure of the EM-style algorithm for BDS variational inference is summarized in Algorithm 2.

The rigorous bound of (III.49) leads to the significant improvement of our method in accuracy over previous state-of-the-art GCLDS [49]. To see this significance, the gap between two bounds and results of approximate inference for a 1D example are illustrated in Fig. III.1.



(a) Contour of the difference between two bounds.



(b) Approximate distributions for posterior $p(x|y=1)$.

Figure III.1: Comparison of variational bounds and approximate distributions. Top: contour of the difference $b_1(\mu,\rho^2) - b_2(\mu,\rho^2)$ between two lower bounds of $\langle \ln \sigma(x) \rangle_{p(x)}, x \sim \mathcal{N}(\mu,\rho^2)$ for different (μ,ρ^2); $b_1(\mu,\rho^2) = \ln \sigma(\sqrt{\mu^2+\rho^2}) + (\mu - \sqrt{\mu^2+\rho^2})/2$ is our bound for BDS, $b_2(\mu,\rho^2) = \mu + \ln \sigma(-\mu - \rho^2/2)$ the bound for GCLDS. Bottom: approximate inference results in 1D case for $p(x|y=1)$ with prior $x \sim \mathcal{N}(0,1)$ and conditional probability $p(y|x) = \text{Bern}(\sigma(6x-2))$; KL divergence to $p(x|y=1)$ (black) is shown in parentheses for Laplace approximation (blue), GCLDS (green), and our method (red). Best viewed in color.

E-step

In this step, (III.49) is optimized over q given $\boldsymbol{\xi}$ fixed. After dropping terms that do not depend on q in (III.49), we have

$$q^* = \underset{q(x)\in\mathcal{D}_q}{\arg\max} \Big\langle -\frac{1}{2}(\boldsymbol{x}_1-\boldsymbol{\mu})^\intercal \boldsymbol{S}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}) - \frac{1}{2}\sum_{t=1}^{\tau}(\boldsymbol{C}\boldsymbol{x}_t-\tilde{\boldsymbol{u}}_t)^\intercal \tilde{\boldsymbol{R}}_t^{-1}(\boldsymbol{C}\boldsymbol{x}_t-\tilde{\boldsymbol{u}}_t) \\ -\frac{1}{2}\sum_{t=1}^{\tau-1}\begin{bmatrix}\boldsymbol{x}_t\\ \boldsymbol{x}_{t+1}\end{bmatrix}^\intercal \begin{bmatrix}\boldsymbol{A}^\intercal\boldsymbol{Q}^{-1}\boldsymbol{A} & -\boldsymbol{A}^\intercal\boldsymbol{Q}^{-1}\\ -\boldsymbol{Q}^{-1}\boldsymbol{A} & \boldsymbol{Q}^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_t\\ \boldsymbol{x}_{t+1}\end{bmatrix}\Big\rangle_{q(x)} + \mathrm{H}_q(X). \tag{III.52}$$

Since all terms subject to the expectation are *quadratic* or linear functions in $\boldsymbol{x}$, it can be shown that (see Appendix III.F.2 for details), the solution to (III.52) is a Gaussian distribution

$$q^*(\boldsymbol{x}_{1:\tau}) = \mathcal{G}(\boldsymbol{x}_{1:\tau}; \boldsymbol{m}, \boldsymbol{\Phi}),\ \ \boldsymbol{m}\in\mathbb{R}^{L\tau\times 1},\ \boldsymbol{\Phi}\in\mathcal{S}_{++}^{L\tau}, \tag{III.53}$$

where

$$\boldsymbol{\Phi} = \boldsymbol{W}_{JJ}^{-1} \tag{III.54}$$

with $\boldsymbol{W}_{JJ} \in \mathcal{S}_{++}^{L\tau}$ defined by

$$\boldsymbol{W}_{JJ[r,s]} = \begin{cases} \boldsymbol{A}^\intercal \boldsymbol{Q}^{-1}\boldsymbol{A} + \boldsymbol{S}^{-1} + \boldsymbol{C}^\intercal \tilde{\boldsymbol{R}}_1^{-1}\boldsymbol{C}, & r = s = 1, \\ \boldsymbol{A}^\intercal \boldsymbol{Q}^{-1}\boldsymbol{A} + \boldsymbol{Q}^{-1} + \boldsymbol{C}^\intercal \tilde{\boldsymbol{R}}_r^{-1}\boldsymbol{C}, & 1 < r = s < \tau, \\ \boldsymbol{Q}^{-1} + \boldsymbol{C}^\intercal \tilde{\boldsymbol{R}}_\tau^{-1}\boldsymbol{C}, & r = s = \tau, \\ -\boldsymbol{Q}^{-1}\boldsymbol{A}, & r = s + 1, \\ -\boldsymbol{A}^\intercal \boldsymbol{Q}^{-1}, & r = s - 1, \\ \boldsymbol{0}, & \text{otherwise}; \end{cases} \tag{III.55}$$

and

$$\boldsymbol{m} = \boldsymbol{W}_{JJ}^{-1}\boldsymbol{\beta}, \;\; \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_{[1]} \\ \vdots \\ \boldsymbol{\beta}_{[\tau]} \end{bmatrix}, \;\; \boldsymbol{\beta}_{[t]} = \begin{cases} \boldsymbol{S}^{-1}\boldsymbol{\mu} + \boldsymbol{C}^\intercal \tilde{\boldsymbol{R}}_1^{-1}\tilde{\boldsymbol{u}}_1, & t = 1, \\ \boldsymbol{C}^\intercal \tilde{\boldsymbol{R}}_t^{-1}\tilde{\boldsymbol{u}}_t, & 1 < t \leqslant \tau. \end{cases} \tag{III.56}$$

Although both (III.54) and (III.56) are conceptually straightforward to compute, inversion of $\boldsymbol{W}_{JJ}$ can be computationally expensive for a very long sequence (*e.g.*, τ is large), at a complexity around $O(L^\kappa \tau^\kappa)$.

In many scenarios, however, only 1) the value of $\tilde{\mathscr{L}}_{JJ}(q, \boldsymbol{\xi}; \boldsymbol{\theta})$ in (III.49) (the lower-bound of the data log-evidence), 2) the mean $\boldsymbol{m}$, and 3) the covariance of states at each time step and between two adjacent time steps, $\boldsymbol{\Phi}_{[t,t]}$ and $\boldsymbol{\Phi}_{[t,t+1]}$, are needed, *e.g.*, in model parameter estimation. Thus, these critical results can be efficiently computed by an efficient solution that is derived from the popular *Kalman smoothing filter* (KSF) with a complexity of $O(L^\kappa \tau)$ [131], with similar routine in Section III.B.2 with proper parameter substitution. For convenience, we define the Kalman filtering as a mapping from $\boldsymbol{y}^{(i)}$ to

$\{\tilde{\mathscr{L}}_{JJ}(q^*,\boldsymbol{\xi};\boldsymbol{\theta}),\boldsymbol{m}^{(i,j)},\{\boldsymbol{\Phi}^{(i,j)}_{[t,t]}\},\{\boldsymbol{\Phi}^{(i,j)}_{[t,t+1]}\}\}$ given $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ as parameters:

$$\{\tilde{\mathscr{L}}_{JJ}(q^*,\boldsymbol{\xi};\boldsymbol{\theta}),\boldsymbol{m}^{(i,j)},\{\boldsymbol{\Phi}^{(i,j)}_{[t,t]}\},\{\boldsymbol{\Phi}^{(i,j)}_{[t,t+1]}\}\} = \mathrm{KSF}(\boldsymbol{y}^{(i)};\boldsymbol{\xi},\boldsymbol{\theta}). \tag{III.57}$$

M-step

In this step, (III.49) is optimized over $\boldsymbol{\xi}$ given $q(\boldsymbol{x};\boldsymbol{m},\boldsymbol{\Phi})$ fixed. After dropping terms that do not depend on $\boldsymbol{\xi}$ in (III.49), we have

$$\boldsymbol{\xi}^* = \underset{\boldsymbol{\xi}\in\mathbb{R}^{D\times\tau}_{++}}{\arg\max} \sum_{t,d} \ln\sigma(\xi_{d,t}) + \lambda(\xi_{d,t})(\xi^2_{d,t} - \left\langle \omega^2_{d,t} \right\rangle_q) - \frac{\xi_{d,t}}{2}, \tag{III.58}$$

where $\left\langle \omega^2_{dt} \right\rangle_q = (\tilde{\boldsymbol{C}}_{j,d,:}\boldsymbol{m}_{[t]} + u_d)^2 + \tilde{\boldsymbol{C}}_{j,d,:}\boldsymbol{\Phi}_{[t,t]}\tilde{\boldsymbol{C}}_{j,d,:}{}^{\mathsf{T}}$. (III.58) can be solved by optimization over each $\xi_{d,t}$ individually, which yields the solution

$$\xi^*_{d,t} = \left\langle \omega^2_{dt} \right\rangle_q^{\frac{1}{2}} = \left[\left(\tilde{\boldsymbol{C}}_{j,d,:}\boldsymbol{m}_{[t]} + u_d \right)^2 + \tilde{\boldsymbol{C}}_{j,d,:}\boldsymbol{\Phi}_{[t,t]}\tilde{\boldsymbol{C}}_{j,d,:}{}^{\mathsf{T}} \right]^{\frac{1}{2}}. \tag{III.59}$$

See Appendix III.F.4 for derivations.

Since each M-step requires $O(L^2D\tau)$ operations, the total complexity of our inference algorithm is $O((DL^2 + L^\kappa)\tau)$ [2]. Note that, while it is possible to plug (III.59) into (III.49) to derive a gradient descent algorithm for optimizing the ELBO, this will nullify the elegant Markovian structure of (III.49) and result in a complex non-linear objective function, whose expensive gradient needs to be evaluated frequently, as in the case of GCLDS. In contrast, our EM-like inference routine only requires very efficient closed-form update rules, achieving a tremendous boost in speed over GCLDS.

[2] More precisely, there is another factor n_{EM} in the complexity, *i.e.*, $O((DL^2 + L^\kappa)\tau n_{EM})$, where n_{EM} is the average number of EM iterations. Nevertheless, it is still fair to consider n_{EM} as a constant since our EM algorithm for inference always converges after several iterations regardless of the value of D, L, and τ.

Algorithm 2: Variational Inference of BDS (VarInf$_{\text{BDS}}$) with ELBO$_{JJ}$ via Coordinate Descent

Input: a binary vector sequence $\boldsymbol{y}_{1:\tau}$, a BDS parameter $\boldsymbol{\theta}$, initial variational parameter $\boldsymbol{\xi}$;

$n \leftarrow 0$, $\boldsymbol{\xi}^{(0)} \leftarrow \boldsymbol{\xi}$;

repeat

(*VE-step*): update the variational distribution $q(\boldsymbol{x}; \boldsymbol{m}, \boldsymbol{\Phi})$ by

$$\{\tilde{\mathscr{L}}, \boldsymbol{m}^{(n+1)}, \{\boldsymbol{\Phi}^{(n)}_{[t,t]}\}, \{\boldsymbol{\Phi}^{(n)}_{[t,t+1]}\}\} \leftarrow \text{KSF}(\boldsymbol{y}_{1:\tau}; \boldsymbol{\xi}^{(n)}, \boldsymbol{\theta}),$$

where KSF$(\cdot; \cdot, \cdot)$ is the Kalman smoothing filter of (III.57).

(*VM-step*):

for $d := 1$ *to* D **do**

for $t := 1$ *to* τ **do**

update $\xi_{d,t}$ according to

$$\xi^{(n+1)}_{d,t} \leftarrow \left[\left(\tilde{\boldsymbol{C}}_{j,d,:}\boldsymbol{m}^{(n+1)}_{[t]} + u_d\right)^2 + \tilde{\boldsymbol{C}}_{j,d,:}\boldsymbol{\Phi}^{(n+1)}_{[t,t]}\tilde{\boldsymbol{C}}_{j,d,:}{}^{\mathsf{T}}\right]^{\frac{1}{2}};$$

end

end

$n \leftarrow n + 1$;

until *convergence*;

Output: $\tilde{\mathscr{L}}$, $\boldsymbol{m}^{(n)}$, $\{\boldsymbol{\Phi}^{(n)}_{[t,t]}\}$, $\{\boldsymbol{\Phi}^{(n)}_{[t,t+1]}\}$, $\boldsymbol{\xi}^{(n)}$.

III.D Inference for Mixture of Binary Dynamic Systems

To capture the full dependence between the indicator and hidden state sequence, a mixture model is assumed for the variational distribution $q(\boldsymbol{x}_{1:\tau}, \boldsymbol{z})$ as

$$q(\boldsymbol{x}_{1:\tau}, \boldsymbol{z}) = q(\boldsymbol{x}_{1:\tau}|\boldsymbol{z})q(\boldsymbol{z}) = \prod_{j=1}^{K} \Big[q(\boldsymbol{x}_{1:\tau}|z_j = 1)q(z_j = 1) \Big]^{z_j}, \tag{III.60}$$

where

$$q(\boldsymbol{z} = j) = q(z_j = 1) = q(j) = \gamma_j, \tag{III.61}$$

$$q(\boldsymbol{x}_{1:\tau}|z_j = 1) = q(\boldsymbol{x}_{1:\tau}|j) \in \mathcal{D}_{q(\boldsymbol{x}|\boldsymbol{z})}, \tag{III.62}$$

with $q(\boldsymbol{z};\boldsymbol{\gamma}) \in \mathcal{D}_{q(\boldsymbol{z})} = \{\{q(\boldsymbol{z} = j) = \gamma_j\}_{j=1}^{K} | \sum_j \gamma_j = 1, \gamma_j \geqslant 0\}$ and $\mathcal{D}_{q(\boldsymbol{x}|\boldsymbol{z})} = \{q(\boldsymbol{x})|q(\boldsymbol{x}) \geqslant 0, \int q(\boldsymbol{x})d\boldsymbol{x} = 1\}$. Note that, in the variational model above, both the indicator $\boldsymbol{z}$ and the state sequence $\boldsymbol{x}_{1:\tau}$ are subject to *free-form* distributions over their support: $\boldsymbol{z}$ is sampled from an arbitrary categorical distribution $\mathrm{Cat}(K, \boldsymbol{\gamma})$ over integer set $\{1, \cdots, K\}$; and $\boldsymbol{x}_{1:\tau}$, conditional on $\boldsymbol{z}$, is sampled from an arbitrary distribution over $\mathbb{R}^L$.

Given the mixture model parameter $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \{\boldsymbol{S}_j, \boldsymbol{\mu}_j, \boldsymbol{A}_j, \boldsymbol{C}_j, \boldsymbol{Q}_j, \boldsymbol{u}_j\}_{j=1}^{K}\}$, consider the following lower-bound $\mathscr{L}(q;\boldsymbol{\theta})$ for an observed sequence $\boldsymbol{y}$, by applying the chain rule of variational inference in Section III.A.2 with $q(\boldsymbol{x}_{1:\tau}, \boldsymbol{z})$ of (III.60)

$$\mathscr{L}(q;\boldsymbol{\theta}) = \textstyle\sum_j q(j)\mathscr{L}(q_{\boldsymbol{x}|z};\boldsymbol{\theta}, j) + \mathrm{H}_q(Z), \tag{III.63}$$

where

$$\mathscr{L}(q_{\boldsymbol{x}|z};\boldsymbol{\theta},j) = \int_{\boldsymbol{x}} q(\boldsymbol{x}|j)\ln p(\boldsymbol{x},\boldsymbol{y},j;\boldsymbol{\theta})d\boldsymbol{x} + \mathrm{H}_q(X|Z=j). \tag{III.64}$$

Plugging the complete-data log-evidence of (II.17) for the mixture model (up to scalar constants)

$$\begin{aligned} &\ln p(\boldsymbol{x}_{1:\tau},\boldsymbol{y}_{1:\tau},j;\boldsymbol{\theta}) = \\ &\quad \ln\alpha_j - \frac{1}{2}\ln\left|\boldsymbol{S}_j\right| - \left(\frac{\tau-1}{2}\right)\ln\left|\boldsymbol{Q}_j\right| - \frac{1}{2}\left\|\boldsymbol{x}_1 - \boldsymbol{\mu}_{0,j}\right\|^2_{\boldsymbol{S}_j} - \frac{1}{2}\sum_{t=1}^{\tau-1}\left\|\boldsymbol{x}_{t+1} - \boldsymbol{A}_j\boldsymbol{x}_t\right\|^2_{\boldsymbol{Q}_j} \\ &\qquad + \sum_{t,d}\Big[y_{dt}\ln\sigma(\boldsymbol{C}_{j,d,:}\boldsymbol{x}_t + u_{j,d}) + (1-y_{dt})\ln\sigma(-\boldsymbol{C}_{j,d,:}\boldsymbol{x}_t - u_{j,d})\Big] + \text{const} \end{aligned} \tag{III.65}$$

into (III.64), and following (III.9) to (III.11), yield two sets of optimization problems to determine the optimal variational distribution $q^*(\boldsymbol{x}_{1:\tau},\boldsymbol{z}) = q^*(\boldsymbol{x}_{1:\tau}|\boldsymbol{z})q^*(\boldsymbol{z})$.

The first set consists of K nested problems (as discussed in Section III.A.2)

$$q^*(\boldsymbol{x}_{1:\tau}|j) \;=\; \underset{q\in\mathcal{D}_{q(x|z)}}{\arg\max}\ \mathscr{L}(q;\boldsymbol{\theta},j), \quad j=1,\cdots,K. \tag{III.66}$$

This is the inference of Section III.C, thus it can be solved with the identical algorithm there, which gives the result

$$q^*(\boldsymbol{x}_{1:\tau}|j) = \underset{q\in\mathcal{D}_{q(x|z)}}{\arg\max}\tilde{\mathscr{L}}(q;\boldsymbol{\theta},j) = \mathcal{G}(\boldsymbol{x}_{1:\tau};\boldsymbol{m}_{[j]},\boldsymbol{\Phi}_j), \tag{III.67}$$

and

$$\ln p^*(\boldsymbol{y}_{1:\tau},j;\boldsymbol{\theta}) = \tilde{\mathscr{L}}(q^*(\boldsymbol{x}_{1:\tau}|j);\boldsymbol{\theta},j) = \ln\alpha_j + \ln p^*(\boldsymbol{y}_{1:\tau}|j;\boldsymbol{\theta}_j), \tag{III.68}$$

where $\ln p^*(\boldsymbol{y}_{1:\tau}|j;\boldsymbol{\theta}_j)$ is the lower bound to the conditional log-evidence $\ln p(\boldsymbol{y}_{1:\tau}|j;\boldsymbol{\theta}_j)$, which is identical to $\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta},q)$ of (III.38) in Section III.C.1, or $\tilde{\mathscr{L}}_{JJ}(q^*,\boldsymbol{\xi}^*;\boldsymbol{\theta}_j)$ of (III.49) in Section III.C.2.

The second problem is the root problem of (III.11)

$$\max_{q(\boldsymbol{z};\gamma)\in\mathcal{D}_{q(\boldsymbol{z})}} \sum\nolimits_j \gamma_j \ln p^*(\boldsymbol{y}_{1:\tau},j;\boldsymbol{\theta}) - \sum\nolimits_j \gamma_j \ln \gamma_j, \tag{III.69}$$

by using the result of (III.68) from the nested problems of (III.66). It can be shown that, solution to (III.69) is given by (see Appendix III.F.5 for details)

$$\gamma_j^* = \frac{p^*(\boldsymbol{y}_{1:\tau},j;\boldsymbol{\theta})}{\sum_{k=1}^K p^*(\boldsymbol{y}_{1:\tau},k;\boldsymbol{\theta})} = \frac{\alpha_j p^*(\boldsymbol{y}_{1:\tau}|j;\boldsymbol{\theta}_j)}{\sum_{k=1}^K \alpha_k p^*(\boldsymbol{y}_{1:\tau}|k;\boldsymbol{\theta}_j)}. \tag{III.70}$$

It is worth noting that, (III.70) resembles the form of posterior of cluster assignment in a mixture model (*e.g.*, a regular Gaussian mixture). The difference is that (III.70) uses a lower-bound of the cluster-conditional data evidence $p^*(\boldsymbol{y}_{1:\tau}|j;\boldsymbol{\theta})$ instead of the ground truth $p(\boldsymbol{y}_{1:\tau}|j;\boldsymbol{\theta})$ (which is intractable in our case). If the inference of cluster-conditional data evidence is exact in (III.66), the posterior of the cluster assignment estimated in (III.70) is also exact.

III.E Acknowledgement

The text of Chapter III is, in part, based on the material as it appears in the following publications: The variational scheme for BDS using LJ bound and Fisher vector were originally proposed in W.-X. LI and N. Vasconcelos, "Complex Activity Recognition via Attribute Dynamics," to appear at *International Journal of Computer Vision* (IJCV). The variational scheme for BDS using JJ bound and the associated expectation-maximization algorithm were originally proposed

in W.-X. LI, Y. Li and N. Vasconcelos, "Efficient Variational Inference, Learning and Probabilistic Kernels for Binary Dynamic Systems," under review at *Neural Information Processing Systems* (NIPS), 2016. The variational inference scheme for the mixture of binary dynamic systems was originally proposed in W.-X. LI, Y. Li and N. Vasconcelos, "Modeling, Clustering, and Segmenting Binary Sequences with Mixtures of Binary Dynamic Systems," under review at *Journal of Machine Learning Research* (JMLR). The dissertation author was a primary researcher and an author of the cited material.

III.F Appendix

III.F.1 Unimodality of the State Posterior of the BDS

By rewriting the complete data log-evidence of (III.35) as a function of $\boldsymbol{x}$, the log-posterior is of the form (up to constants independent of $\boldsymbol{x}$)

$$\begin{aligned}
&\ln p(\boldsymbol{x}_{1:\tau}|\boldsymbol{y}_{1:\tau};\boldsymbol{\theta}) = \\
&\quad -\frac{1}{2}(\boldsymbol{x}_1-\boldsymbol{\mu})^\intercal \boldsymbol{S}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}) - \frac{1}{2}\sum_{t=1}^{\tau-1}(\boldsymbol{x}_{t+1}-\boldsymbol{A}\boldsymbol{x}_t)^\intercal \boldsymbol{Q}^{-1}(\boldsymbol{x}_{t+1}-\boldsymbol{A}\boldsymbol{x}_t) \\
&\quad -\sum_{t,d}\Big\{y_{dt}\ln[1+\exp(-\tilde{\boldsymbol{C}}_{j,d,:}\boldsymbol{x}_t - u_d)] + (1-y_{dt})\ln[1+\exp(\tilde{\boldsymbol{C}}_{j,d,:}\boldsymbol{x}_t + u_d)]\Big\} \\
&\quad + \text{const.}
\end{aligned}$$

Note that, (III.71) is *strictly concave* in $\boldsymbol{x} \in \mathbb{R}^{L\tau}$ since all terms are (with $\boldsymbol{x}$ as arguments subject to linear transformations) either 1) quadratic functions of negative definite coefficient matrices; or 2) negative log-sum-exp functions [16].

To see this, the second-order partial derivative of $\ln p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta})$ is

$$\frac{\partial^2}{\partial \boldsymbol{x}^2} \ln p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}) = -\boldsymbol{W}^\circ - \begin{bmatrix} \boldsymbol{C}^\intercal \mathbf{Y}_1 \boldsymbol{C} & & \\ & \ddots & \\ & & \boldsymbol{C}^\intercal \mathbf{Y}_\tau \boldsymbol{C} \end{bmatrix}, \tag{III.71}$$

where $\boldsymbol{W}^\circ \in \mathcal{S}_{++}^{L\tau}$ is defined by

$$\boldsymbol{W}^\circ{}_{[r,s]} = \begin{cases} \boldsymbol{A}^\intercal \boldsymbol{Q}^{-1} \boldsymbol{A} + \boldsymbol{S}^{-1}, & r = s = 1, \\ \boldsymbol{A}^\intercal \boldsymbol{Q}^{-1} \boldsymbol{A} + \boldsymbol{Q}^{-1}, & 1 < r = s < \tau, \\ \boldsymbol{Q}^{-1}, & r = s = \tau, \\ -\boldsymbol{Q}^{-1}\boldsymbol{A}, & r = s + 1, \\ -\boldsymbol{A}^\intercal \boldsymbol{Q}^{-1}, & r = s - 1, \\ \boldsymbol{0}, & \text{otherwise;} \end{cases}$$

and

$$\mathbf{Y}_t = \operatorname{diag}(\sigma(\omega_{1,t})\sigma(-\omega_{1,t}),\ \cdots,\ \sigma(\omega_{D,t})\sigma(-\omega_{D,t})) \in \mathcal{S}_{++}^{D}.$$

The Hessian of $\ln p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta})$ is negative-definite because 1) $\boldsymbol{W}^\circ \in \mathcal{S}_{++}^{L\tau}$ is positive-definite; and 2) the second matrix in (III.71) is positive-semidefinite. Hence, $\ln p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta})$ is strictly concave.

On the other hand, $p(\boldsymbol{\infty}|\boldsymbol{y};\boldsymbol{\theta}) = 0$ since 1) $p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta})$ is smooth in $\boldsymbol{x} \in \mathbb{R}^{L\tau}$, and 2) $\int p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}) d\boldsymbol{x} = 1$. Thus there exists a *closed and bounded* set $\mathcal{X} \subset \mathbb{R}^{L\tau}$ such that $p(\boldsymbol{x}_1|\boldsymbol{y};\boldsymbol{\theta}) > p(\boldsymbol{x}_2|\boldsymbol{y};\boldsymbol{\theta}), \forall \boldsymbol{x}_1 \in \partial\mathcal{X}, \boldsymbol{x}_2 \in \mathbb{R}^{L\tau} \setminus \mathcal{X}$. By extreme value theorem, $p(\boldsymbol{x}_1|\boldsymbol{y};\boldsymbol{\theta})$ (and $\ln p(\boldsymbol{x}_1|\boldsymbol{y};\boldsymbol{\theta})$) must achieve a maximum at $\boldsymbol{x}^* \in \mathcal{X}$.

Altogether, we have

$$x^* = \arg\max_{x \in \mathbb{R}^{L\tau}} p(x|y;\theta) = \arg\max_{x \in \mathbb{R}^{L\tau}} \ln p(x|y;\theta) \neq \infty. \tag{III.72}$$

It follows that there is a global maximum at $x^* \neq \infty$ for the concave function of $\ln p(x_{1:\tau}|y_{1:\tau};\theta)$. Therefore, $p(x_{1:\tau}|y_{1:\tau};\theta)$ is a unimodal distribution peaking at x^*.

III.F.2 Optimal Variational Distribution for Dynamic Systems

The optimization problem of (III.14) or (III.52) is of the general form

$$\begin{aligned} &\max_q F[q] \\ &s.t.\ F[q] = \int_x q(x)[g(x) - \ln q(x)]dx, \\ &\quad q(x) \in \mathcal{D}_q, \end{aligned} \tag{III.73}$$

where $F[q]$ is a functional of q; $\mathcal{D}_q = \{q(x)|q(x) \geqslant 0, \int q(x)dx = 1, x \in \mathbb{R}^n\}$ is the set of all PDFs defined on $\mathbb{R}^n$; and

$$g(x) = -\frac{1}{2}x^\intercal W x + b^\intercal x + c, \ \ W \in \mathcal{S}^n_{++}, b \in \mathbb{R}^n, c \in \mathbb{R} \tag{III.74}$$

is a *strictly concave* quadratic function in $x \in \mathbb{R}^n$. Note that, problem of (III.73) is a convex problem as 1)the objective function $F[q]$ is concave in q, and 2) the domain $\mathcal{D}_q$ is a convex set. Using the method of Lagrange multipliers, the constraint problem of (III.73) can be converted to an unconstraint one

$$\max_{\{q(x),\nu(x),\lambda\}} \int_x q(x)[g(x) - \ln q(x)]dx + \int_x \nu(x)q(x)dx + \lambda(\int_x q(x)dx - 1), \tag{III.75}$$

where $\nu(\boldsymbol{x}) \geqslant 0, \forall \boldsymbol{x} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ are the multipliers. By calculus of variations [134] and Karush-Kuhn-Tucker (KKT) conditions [16], the sufficient and necessary conditions for the optimal point $\{q^*(\boldsymbol{x}), \nu^*(\boldsymbol{x}), \lambda^*\}$ are

$$g(\boldsymbol{x}) - \ln q^*(\boldsymbol{x}) - 1 + \nu^*(\boldsymbol{x}) + \lambda^* = 0,\ \forall \boldsymbol{x} \in \mathbb{R}^n, \qquad \text{(stationarity)} \tag{III.76}$$

$$\int_{\boldsymbol{x}} q^*(\boldsymbol{x}) d\boldsymbol{x} = 1, \qquad \text{(primal feasibility)} \tag{III.77}$$

$$q^*(\boldsymbol{x}) \geqslant 0,\ \forall \boldsymbol{x} \in \mathbb{R}^n, \qquad \text{(primal feasibility)} \tag{III.78}$$

$$\nu^*(\boldsymbol{x}) \geqslant 0,\ \forall \boldsymbol{x} \in \mathbb{R}^n, \qquad \text{(dual feasibility)} \tag{III.79}$$

$$\nu^*(\boldsymbol{x}) q^*(\boldsymbol{x}) = 0,\ \forall \boldsymbol{x} \in \mathbb{R}^n. \quad \text{(complementary slackness)} \tag{III.80}$$

From (III.76), it follows that

$$q^*(\boldsymbol{x}) > 0,\ \ \forall \boldsymbol{x} \in \mathbb{R}^n. \tag{III.81}$$

Combining (III.81), (III.79) and (III.80) leads to

$$\nu^*(\boldsymbol{x}) = 0,\ \ \forall \boldsymbol{x} \in \mathbb{R}^n. \tag{III.82}$$

Substituting (III.82) into (III.76), and noting that $q^*(\boldsymbol{x})$ is a PDF by definition, we have

$$q^*(\boldsymbol{x}) \propto \exp(g(\boldsymbol{x})) = \exp(-\frac{1}{2}\boldsymbol{x}^\intercal W \boldsymbol{x} + \boldsymbol{b}^\intercal \boldsymbol{x}). \tag{III.83}$$

It is now clear that $q^*(\boldsymbol{x})$ is a Gaussian distribution of the form

$$q^*(\boldsymbol{x}) = \mathcal{G}(\boldsymbol{x}; \boldsymbol{W}^{-1}\boldsymbol{b}, \boldsymbol{W}^{-1}). \tag{III.84}$$

Reorganizing terms of (III.52) into the form of (III.73) gives the result of (III.53) by (III.84).

III.F.3 Solution to Covariance of the Variational Distribution

We study an optimization problem of general form for brevity, before deriving the solution to problems in the maintext.

The general optimization problem is

$$\max_{\boldsymbol{X} \in \mathcal{S}_{++}} b \ln |\boldsymbol{X}| - \mathrm{tr}(\boldsymbol{A}\boldsymbol{X}), \quad s.t.\ \boldsymbol{A} \in \mathcal{S}_{++},\ b > 0. \tag{III.85}$$

Since 1) both $b \ln |\boldsymbol{X}|$ and $-\mathrm{tr}(\boldsymbol{A}\boldsymbol{X})$ are smooth and concave functions in $\boldsymbol{X}$, and 2) the domain $\mathcal{S}_{++}$ is a convex set, the maximum of problem (III.85) is achieved at either 1) its stationary point(s) (if there is any), or 2) the boundary of its domain (could be at infinity) [16].

The derivative of the objective function in the problem of (III.85) is

$$\frac{\partial}{\partial \boldsymbol{X}} \left\{ b \ln |\boldsymbol{X}| - \mathrm{tr}(\boldsymbol{A}\boldsymbol{X}) \right\} = b\boldsymbol{X}^{-\intercal} - \boldsymbol{A}. \tag{III.86}$$

Setting (III.86) to zero leads to

$$\boldsymbol{X}^* = b\boldsymbol{A}^{-1} \in \mathcal{S}_{++}, \tag{III.87}$$

which achieves the global maximum for the problem of (III.85).

The solution to problem (III.18) and (III.41) can be derived by setting $b = 1$ and $\boldsymbol{A}$ to $\boldsymbol{W}_{\text{LDS}}$ or $\boldsymbol{W}_{SJ}$, respectively.

III.F.4 Update Rules in the M-step for Variational Inference

The general form of the objective function in (III.58) is (with $a \in \mathbb{R}$ as the parameter)

$$f(\xi) = \ln \sigma(\xi) + \frac{1}{2\xi}(\sigma(\xi) - \frac{1}{2})(\xi^2 - a^2) - \frac{\xi}{2}, \;\; \xi > 0. \tag{III.88}$$

The first-order derivative of (III.88) is

$$f'(\xi) = \frac{1}{2}\Big(\frac{a^2}{\xi^2} - 1\Big)\Big[\sigma(\xi) - \frac{1}{2} - \xi\sigma(\xi)\sigma(-\xi)\Big], \;\; \xi > 0. \tag{III.89}$$

Since

$$\sigma(\xi) - \frac{1}{2} - \xi\sigma(\xi)\sigma(-\xi) = \frac{1 - e^{-2\xi} - 2\xi e^{-\xi}}{2(1 + e^{-\xi})} = \frac{e^{-\xi}(e^{\xi} - e^{-\xi} - 2\xi)}{2(1 + e^{-\xi})} > 0, \;\; \forall \xi > 0, \tag{III.90}$$

we have

$$f'(\xi) > 0, \;\; \forall \xi \in (0, |a|), \tag{III.91}$$

$$f'(\xi) = 0, \;\; \xi = |a|, \tag{III.92}$$

$$f'(\xi) < 0, \;\; \forall \xi \in (|a|, +\infty), \tag{III.93}$$

and

$$\xi^* = \arg\max_{\xi} f(\xi) = \sqrt{a^2} = |a|. \tag{III.94}$$

Setting $a^2 = \left\langle \omega_{d,t}^2 \right\rangle_q$ in (III.88) gives the result of (III.59) by (III.94).

III.F.5 Inference of the Cluster Assignments in the Mixture Model

Problem (III.69) is of the form

$$\max_{\gamma} \sum_j \gamma_j \big(\ln \beta_j - \ln \gamma_j\big), \qquad \text{(III.95)}$$

$$s.t.\ \ \boldsymbol{\gamma} \succeq \mathbf{0},\ \mathbf{1}^{\mathsf{T}}\boldsymbol{\gamma} = 1,$$

where $\mathbf{1} = [1, \cdots, 1]^{\mathsf{T}} \in \mathbb{R}^K$ and $\beta_j > 0$ are the parameter. Note that, problem (III.95) is a *convex* problem because 1) the objective function of problem (III.95) is a concave function in $\boldsymbol{\gamma}$ since

$$\frac{\partial^2}{\partial \boldsymbol{\gamma}^2} \sum_j \gamma_j \big(\ln \beta_j - \ln \gamma_j\big) = -\mathrm{diag}(\frac{1}{\gamma_1}, \cdots, \frac{1}{\gamma_K}) \in \mathcal{S}_{--}, \forall \boldsymbol{\gamma} \succ \mathbf{0}$$

where $\mathcal{S}_{--}$ is the set of negative-definite matrices, and 2) its domain is a convex set (more precisely, a standard $(K-1)$-simplex).

By introducing Lagrange multipliers $\lambda \in \mathbb{R}$ and $\boldsymbol{\nu} \succeq \mathbf{0}$, the constraint problem of(III.95) is converted to an unconstraint one:

$$\max_{\gamma, \lambda, \nu} \ \sum_j \gamma_j (\ln \beta_j - \ln \gamma_j) + \boldsymbol{\nu}^{\mathsf{T}}\boldsymbol{\gamma} + \lambda(\mathbf{1}^{\mathsf{T}}\boldsymbol{\gamma} - 1). \qquad \text{(III.96)}$$

According to Karush-Kuhn-Tucker conditions, at the optimal point $\{\boldsymbol{\gamma}^*, \lambda^*, \boldsymbol{\nu}^*\}$, we have

$$\forall j,\ \ln \beta_j - \ln \gamma_j^* - 1 + \nu_j^* + \lambda^* = 0, \qquad \text{(III.97)}$$

$$\sum_j \gamma_j^* = 1, \qquad \text{(III.98)}$$

$$\forall j,\ {\nu_j}^* \gamma_j^* = 0. \qquad \text{(III.99)}$$

It is obvious that $\boldsymbol{\gamma}^* \succ \mathbf{0}$, thus

$$\boldsymbol{\nu}^* = \mathbf{0}. \tag{III.100}$$

Combining (III.97), (III.98) and (III.100) leads to solution

$$\gamma_j^* = \frac{\beta_j}{\sum_k \beta_k}. \tag{III.101}$$

Finally, substituting $\beta_j = \tilde{p}^*(\boldsymbol{y}_{1:\tau}, j; \boldsymbol{\theta})$ of (III.68) into (III.95) gives the result of (III.70).

Chapter IV

Parameter Estimation for Dynamic Systems

IV.A Parameter Estimation via Suboptimal Procedures

IV.A.1 Binary Principal Component Analysis

Binary PCA [139] is a dimensionality reduction technique for binary data, which belongs to the generalized exponential family PCA [31]. It fits a linear model to binary observations, by embedding the natural parameters of Bernoulli distributions in a low-dimensional subspace. Let Y denote a $K \times \tau$ binary matrix ($y_{kt} \in \{0,1\}$, *e.g.*, the indicator of occurrence of attribute k at time t) where each column is a vector of K binary observations sampled from a multivariate Bernoulli distribution

$$Y_{kt} \sim B(y_{kt};\pi_{kt}) = \pi_{kt}^{y_{kt}}(1-\pi_{kt})^{1-y_{kt}} = \sigma(\theta_{kt})^{y_{kt}}\sigma(-\theta_{kt})^{1-y_{kt}} \tag{IV.1}$$

of natural parameters $\theta_{kt} = \log(\frac{\pi_{kt}}{1-\pi_{kt}})$. Binary PCA finds a L-dimensional ($L \ll K$) embedding of the natural parameters, by maximizing the log-likelihood of the binary matrix Y

$$\mathcal{L} = \ln p(\{y_{kt}\};\mathbf{\Theta}) = \sum_{k,t}\Big[y_{kt}\ln\sigma(\Theta_{kt}) + (1-y_{kt})\ln\sigma(-\Theta_{kt})\Big] \tag{IV.2}$$

under the constraint

$$\mathbf{\Theta} = \boldsymbol{C}\boldsymbol{X} + \boldsymbol{u}\mathbf{1}^{\mathsf{T}}, \tag{IV.3}$$

where $\boldsymbol{C} \in \mathbb{R}^{K\times L}$, $\boldsymbol{X} \in \mathbb{R}^{L\times\tau}$, $\boldsymbol{u} \in \mathbb{R}^{K}$ and $\mathbf{1} \in \mathbb{R}^{\tau}$ is the vector of all ones. Each column of $\boldsymbol{C}$ is a basis vector of a latent subspace and the t-th column of $\boldsymbol{X}$ contains the coordinates of the t-th binary vector in this basis (up to a translation by $\boldsymbol{u}$).

Algorithm 3: Sub-optimal Algorithm for Learning BDS

Input : a set of n sequences of attribute score vectors $\{\boldsymbol{y}_{1:\tau_i}^{(i)}\}_{i=1}^{n}$, state space dimension L.

Binary PCA [139]:

$$\{\boldsymbol{C},\boldsymbol{X},\boldsymbol{u}\} = \text{B-PCA}(\{\boldsymbol{y}_{1:\tau_i}^{(i)}\}_{i=1}^{n}, L);$$

Assemble state sequences ($\boldsymbol{X}_{t_1:t_2} \equiv [\boldsymbol{x}_{t_1}, \cdots, \boldsymbol{x}_{t_2}]$):

$$\hat{\boldsymbol{X}}_2^{\tau} = [\boldsymbol{X}_{2:\tau_1}^{(1)}, \cdots, \boldsymbol{X}_{2:\tau_n}^{(n)}], \quad \hat{\boldsymbol{X}}_1^{\tau-1} = [\boldsymbol{X}_{1:\tau_1-1}^{(1)}, \cdots, \boldsymbol{X}_{1:\tau_n-1}^{(n)}];$$

Estimate state palrameters:

$$\boldsymbol{A} = \hat{\boldsymbol{X}}_2^{\tau}(\hat{\boldsymbol{X}}_1^{\tau-1})^{\dagger},\ \boldsymbol{V} = \hat{\boldsymbol{X}}_2^{\tau} - \boldsymbol{A}\hat{\boldsymbol{X}}_1^{\tau-1},\ \boldsymbol{Q} = \frac{1}{\sum_i(\tau_i - 1)}\boldsymbol{V}(\boldsymbol{V})^{\intercal},$$

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_1^{(i)}, \quad \boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_1^{(i)} - \boldsymbol{\mu})(\boldsymbol{x}_1^{(i)} - \boldsymbol{\mu})^{\intercal}.$$

Output: $\boldsymbol{\Omega} = \{\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{Q}, \boldsymbol{u}, \boldsymbol{\mu}, \boldsymbol{S}\}$

IV.A.2 Learning Binary Dynamic Systems via Sub-optimal Algorithm

The discussion above suggests a generalization of the DT learning procedure to the BDS. The binary PCA basis is learned first, by maximizing the expected log-likelihood of (IV.2) subject to the constraint of (IV.3). Since the Bernoulli is a member of exponential family, (IV.2) is concave in $\boldsymbol{\Theta}$, but not in $\boldsymbol{C}, \boldsymbol{X}$ and $\boldsymbol{u}$ jointly. The ML parameters can be found with the procedure of [139], which iterates between the optimization with respect to one of the variables $\boldsymbol{C}, \boldsymbol{X}$ and $\boldsymbol{u}$ as the other two are held constant. Each iteration is a convex sub-problem that can be solved efficiently with a fixed-point auxiliary function [139].

Once the optimal embedding $\boldsymbol{C}^*, \boldsymbol{X}^*$ and $\boldsymbol{u}^*$ of the attribute sequence is recovered, the remaining parameters are estimated by solving a least-squares problem for $\boldsymbol{A}$ and $\boldsymbol{Q}$, and using ML estimates for the Gaussian parameters of the initial condition ($\boldsymbol{\mu}$ and $\boldsymbol{S}$). Since this is identical to the least squares procedure of [39], we omit the details. The learning procedure, including the least squares equations, is summarized in Algorithm 3. Since the optimal solution maximizes the most natural measure of similarity (KL divergence) between probability distributions, this extension is conceptually equivalent to the procedure used to learn the LDS, which finds the subspace that best fits the observations in the Euclidean sense, the natural similarity measure for Gaussian data. This is unlike previous extensions of the LDS, *e.g.*, kernel dynamic systems (KDS) that rely on a non-linear kernel PCA (KPCA) [141] of the observation space but still assume an Euclidean measure (Gaussian noise) [22, 28].

IV.B Parameter Estimation for Mixtures of Binary Dynamic Systems via Maximum Likelihood Estimation

In this section, we review the *maximum likelihood estimation* (MLE) for models with hidden variables yet intractable posteriors. Since the BDS contains hidden variables yet the posterior is intractable, we rely on the scheme of *variational expectation maximization* (VEM) [108, 79], which generalizes the conventional *expectation-maximization* (EM) [35] algorithm by optimizing a lower bound of the log-likelihood via coordinate descent. Then, in Section IV.C and Section IV.D, we present algorithms to learn the mixture of binary dynamic systems via the VEM algorithm.

Consider the same model $p(Y;\theta)$ with hidden variable X in Section III.A again. In MLE, given training data $\mathcal{T}_y$, the parameter is estimated by maximizing the log-likelihood:

$$\theta_{MLE} = \arg\max_{\theta} \mathcal{L}(\theta;\mathcal{T}_y) = \arg\max_{\theta} p(\mathcal{T}_y;\theta). \tag{IV.4}$$

Since evaluation of the log-likelihood for the model $p(Y;\theta)$ is difficult, its variational lower-bound $\mathscr{L}(q,\theta)$ of (III.2) is maximized instead:

$$\mathcal{L}(\theta;\mathcal{T}_y) \geqslant \mathscr{L}(q,\theta) = \int_{\mathcal{T}_x} q(\mathcal{T}_x) \ln \frac{p(\mathcal{T}_y,\mathcal{T}_x;\theta)}{q(\mathcal{T}_x)} d\mathcal{T}_x = \left\langle \ln p(\mathcal{T}_x,\mathcal{T}_y;\theta) \right\rangle_q + \mathrm{H}[q(\mathcal{T}_x)], \tag{IV.5}$$

where $\mathcal{T}_x$ is the (unobserved) training data of hidden variable X. Note that, here $\mathscr{L}(q,\theta)$ is a function of both the model parameter θ and variational distribution q. This suggests a coordinate descent algorithm that alternates between optimizing the variational distribution (*expectation* step) and the optimal parameter (*maximization* step):

E(xpectation)-step: Optimize $\mathscr{L}(q,\theta)$ over q with θ fixed such that

$$q^* = \arg\max_{q} \mathscr{L}(q,\theta) = \arg\max_{q} \left\langle \ln p(\mathcal{T}_x,\mathcal{T}_y;\theta) \right\rangle_q + \mathrm{H}[q(\mathcal{T}_x)]. \tag{IV.6}$$

M(aximization)-step: Optimize $\mathscr{L}(q,\theta)$ over θ with q fixed such that

$$\theta^* = \arg\max_{\theta} \mathscr{L}(q,\theta) = \arg\max_{\theta} \left\langle \ln p(\mathcal{T}_x,\mathcal{T}_y;\theta) \right\rangle_q. \tag{IV.7}$$

The algorithm is summarized in Algorithm 4. The E-step determines the optimal

Algorithm 4: Maximize $\mathscr{L}(q,\theta)$ via variational EM

Input: initial value of $\theta^{(0)}$;

$n \leftarrow 0$;

repeat

(VE-step): increase $\mathscr{L}(\theta^{(n)},q^{(n)})$ by solving

$$q^{(n+1)}(x) = \text{arg max}_{q\in\mathcal{D}_q}\mathscr{L}(\theta^{(n)},q); \tag{IV.8}$$

(VM-step): increase $\mathscr{L}(\theta^{(n)},q^{(n+1)})$ by solving

$$\theta^{(n+1)} = \text{arg max}_{\theta\in\mathcal{D}_\theta}\mathscr{L}(q,\theta^{(n+1)}); \tag{IV.9}$$

$n \leftarrow n+1$;

until *convergence*;

Output: $\theta^{(n)}$

variational distribution given the current estimate of the parameter, and then computes the expectation in (IV.5), which justifies its name; while the M-step optimizes the expected complete data log-likelihood given the current estimate of the variational distribution. If the optimal variational distribution in (IV.8) is identical to the ground true posterior of $p(\mathcal{T}_x|\mathcal{T}_y;\theta)$, Algorithm 4 reduces to the conventional EM algorithm [35]. Despite the limitation that the algorithm cannot guarantee convergence to a (local) maximum of the data log-evidence if the posterior is approximately inferred [54], it has been shown to achieve satisfactory performance in practice when given a tight lower bound, as is the case of mixtures of binary dynamic systems (see Section VI.E for details).

IV.C Variational EM for Mixtures of Binary Dynamic Systems using ELBO$_{SJ}$

Given an *independent and identically distributed* (*i.i.d.*) N-example training set $\mathcal{T}_y = \{\boldsymbol{y}^{(i)}\}_{i=1}^N$, the parameter $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \{\boldsymbol{S}_j, \boldsymbol{\mu}_j, \boldsymbol{A}_j, \boldsymbol{C}_j, \boldsymbol{Q}_j, \boldsymbol{u}_j\}_{j=1}^K\}$ of a K-component mixture of binary dynamic systems is estimated via the MLE framework of Section IV.B (which reduces to a single BDS learning when $K = 1$). Using the same assumption of (III.60), the variational distribution $q(\mathcal{T}_x, \mathcal{T}_z)$ of the hidden states $\mathcal{T}_x = \{\boldsymbol{x}^{(i)}\}_{i=1}^N$ and assignment vectors $\mathcal{T}_z = \{\boldsymbol{z}^{(i)}\}_{i=1}^N$ is

$$q(\mathcal{T}_x, \mathcal{T}_z) = \prod_{i=1}^N q_i(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) = \prod_{i=1}^N \prod_{j=1}^K \left[q_i(\boldsymbol{x}^{(i)}|j) q_i(j) \right]^{z_j^{(i)}}, \tag{IV.10}$$

where $\boldsymbol{z}^{(i)} \sim \mathrm{Cat}(K, \boldsymbol{\gamma}^{(i)})$ and $q_i(\boldsymbol{x}|j) \in \mathcal{D}_{q(\boldsymbol{x}|\boldsymbol{z})} = \{q(\boldsymbol{x}) | q(\boldsymbol{x}) \geqslant 0, \int q(\boldsymbol{x}) d\boldsymbol{x} = 1\}$. According to (II.17), the complete-data log-evidence is (up to scalar constants)

$$\begin{aligned}
&\ln p(\mathcal{T}_x, \mathcal{T}_z, \mathcal{T}_y; \boldsymbol{\theta}) = \\
&\quad \sum_{i,j} z_j^{(i)} \ln \alpha_j - \frac{1}{2} \sum_{i,j} z_j^{(i)} \ln \left| \boldsymbol{S}_j \right| - \frac{1}{2} \sum_{i,j} z_j^{(i)} (\tau_i - 1) \ln \left| \boldsymbol{Q}_j \right| \\
&\quad - \frac{1}{2} \sum_{i,j} z_j^{(i)} \mathrm{tr} \left[\boldsymbol{S}_j^{-1} \left(\boldsymbol{P}_{1,1}^{(i)} - \boldsymbol{x}_1^{(i)} \boldsymbol{\mu}_j^\intercal - \boldsymbol{\mu}_j \boldsymbol{x}_1^{(i)\intercal} + \boldsymbol{\mu}_j \boldsymbol{\mu}_j^\intercal \right) \right] \\
&\quad - \frac{1}{2} \sum_{i} z_j^{(i)} \sum_{t=1}^{\tau_i - 1} \mathrm{tr} \left[\boldsymbol{Q}_j^{-1} \left(\boldsymbol{P}_{t+1,t+1}^{(i)} - \boldsymbol{P}_{t+1,t}^{(i)} \boldsymbol{A}_j^\intercal - \boldsymbol{A}_j \boldsymbol{P}_{t+1,t}^{(i)\intercal} + \boldsymbol{A}_j \boldsymbol{P}_{t,t}^{(i)} \boldsymbol{A}_j^\intercal \right) \right] \\
&\quad + \sum_{i,j} z_j^{(i)} \sum_{t=1}^{\tau_i} \sum_{d} \left[y_{dt}^{(i)} \ln \sigma(\omega_{d,t}^{(i)}) + (1 - y_{dt}^{(i)}) \ln \sigma(-\omega_{d,t}^{(i)}) \right] + \mathrm{const},
\end{aligned} \tag{IV.11}$$

where $\boldsymbol{P}_{r,s}^{(i)} = \boldsymbol{x}_r^{(i)} \boldsymbol{x}_s^{(i)\intercal}$ and $\omega_{d,t}^{(i)} = \boldsymbol{C}_{j,d,:} \boldsymbol{x}_t^{(i)} + u_{j,d}$.

Substituting (IV.11) into (III.63), using the ELBO in (III.C.1), and following the same derivation in Section III.D, yields the objective function to optimize in

VEM of Algorithm 4:

$$\begin{aligned}
&\tilde{\mathscr{L}}_{SJ}(q(\mathcal{T}_x,\mathcal{T}_z),\boldsymbol{\theta}) = \\
&\quad \sum_{i,j}\gamma_j^{(i)}\ln\alpha_j - \frac{1}{2}\sum_{i,j}\gamma_j^{(i)}\ln\left|\boldsymbol{S}_j\right| - \frac{1}{2}\sum_{i,j}\gamma_j^{(i)}(\tau_i-1)\ln\left|\boldsymbol{Q}_j\right| \\
&\quad - \frac{1}{2}\sum_{i,j}\gamma_j^{(i)}\mathrm{tr}\left[\boldsymbol{S}_j^{-1}(\hat{\boldsymbol{P}}_{1,1|j}^{(i)} - \boldsymbol{m}_{[1]}^{(i,j)}\boldsymbol{\mu}_j^{\intercal} - \boldsymbol{\mu}_j\boldsymbol{m}_{[1]}^{(i,j)\intercal} + \boldsymbol{\mu}_j\boldsymbol{\mu}_j^{\intercal})\right] \\
&\quad - \frac{1}{2}\sum_{i,j}\gamma_j^{(i)}\sum_{t=1}^{\tau_i-1}\mathrm{tr}\left[\boldsymbol{Q}_j^{-1}(\hat{\boldsymbol{P}}_{t+1,t+1|j}^{(i)} - \hat{\boldsymbol{P}}_{t+1,t|j}^{(i)}\boldsymbol{A}_j^{\intercal} - \boldsymbol{A}_j\hat{\boldsymbol{P}}_{t+1,t|j}^{(i)\intercal} + \boldsymbol{A}_j\hat{\boldsymbol{P}}_{t,t|j}^{(i)}\boldsymbol{A}_j^{\intercal})\right] \\
&\quad + \sum_{i,j,d}\gamma_j^{(i)}\sum_{t=1}^{\tau_i}\left[y_{dt}^{(i)}\ln\sigma(\tilde{\boldsymbol{C}}_{j,d,:}\tilde{\boldsymbol{m}}_{[t]}^{(i,j)}) + (1-y_{dt}^{(i)})\ln\sigma(-\tilde{\boldsymbol{C}}_{j,d,:}\tilde{\boldsymbol{m}}_{[t]}^{(i,j)})\right] \\
&\quad - \frac{1}{8}\sum_{i,j}\gamma_j^{(i)}\sum_{t=1}^{\tau_i}\mathrm{tr}(\boldsymbol{C}_j\boldsymbol{\Phi}_{[t,t]}^{(i,j)}\boldsymbol{C}_j^{\intercal}) + \sum_{i,j}\gamma_j^{(i)}\mathrm{H}[(q_i(X|j)] - \sum_{i,j}\gamma_j^{(i)}\ln\gamma_j^{(i)},
\end{aligned} \tag{IV.12}$$

where $\hat{\boldsymbol{P}}_{r,s|j}^{(i)} = \langle \boldsymbol{x}_r\boldsymbol{x}_s^{\intercal}\rangle_{q_i(\boldsymbol{x}|j)} = \boldsymbol{\Phi}_{[r,s]}^{(i,j)} + \boldsymbol{m}_{[r]}^{(i,j)}\boldsymbol{m}_{[s]}^{(i,j)\intercal}$, $\tilde{\boldsymbol{C}}_j = [\boldsymbol{C}_j,\boldsymbol{u}_j]$, and $\tilde{\boldsymbol{m}}_{[t]}^{(i,j)} = [\boldsymbol{m}_{[t]}^{(i,j)\intercal},1]^{\intercal}$.

IV.C.1 E-step

In the E-step, given the current estimate of the model parameter $\boldsymbol{\theta} = \{\boldsymbol{\alpha},\{\boldsymbol{S}_j,\boldsymbol{\mu}_j,\boldsymbol{A}_j,\boldsymbol{C}_j,\boldsymbol{Q}_j,\boldsymbol{u}_j\}_{j=1}^{K}\}$, the variational distribution q is updated by maximizing (IV.12) over q:

$$q^* = \arg\max_q \tilde{\mathscr{L}}_{SJ}(q(\mathcal{T}_x,\mathcal{T}_z),\boldsymbol{\theta}). \tag{IV.13}$$

This is exactly the variational inference for the mixture model. Specifically, due to *i.i.d.* training data, the algorithm of Section III.D is repeated for each training example using parameter $\boldsymbol{\theta}$. During the i-th pass for $\boldsymbol{y}^{(i)}$, 1) the inference of Section III.C.1 is first repeated for $\boldsymbol{y}^{(i)}$ under each of the K BDS components; and then 2) the expected responsibilities of K component to $\boldsymbol{y}^{(i)}$ are computed by

(III.69) using the component-conditional log-evidence of (III.68).

IV.C.2 M-step

In the M-step, given the current variational distribution q estimated in the E-step, the model parameter $\boldsymbol{\theta}$ is updated by maximizing (IV.12) over $\boldsymbol{\theta}$:

$$\begin{aligned}\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \; & \sum_j \hat{N}_j \Big\{ \ln \alpha_j - \frac{1}{2}(\hat{\tau}_j - 1)\ln\left|\boldsymbol{Q}_j\right| - \frac{1}{2}\ln\left|\boldsymbol{S}_j\right| \Big\} \\ & - \frac{1}{2}\sum_j \operatorname{tr}\Big[\boldsymbol{S}_j^{-1}(\boldsymbol{\eta}_j - \boldsymbol{\chi}_j\boldsymbol{\mu}_j^{\intercal} - \boldsymbol{\mu}_j\boldsymbol{\chi}_j^{\intercal} + \hat{N}_j\boldsymbol{\mu}_j\boldsymbol{\mu}_j^{\intercal})\Big] \\ & - \frac{1}{2}\sum_j \operatorname{tr}\Big[\boldsymbol{Q}_j^{-1}(\boldsymbol{\varphi}_j - \boldsymbol{\Psi}_j\boldsymbol{A}_j^{\intercal} - \boldsymbol{A}_j\boldsymbol{\Psi}_j^{\intercal} + \boldsymbol{A}_j\boldsymbol{\phi}_j\boldsymbol{A}_j^{\intercal})\Big] \\ & + \sum_{i,j,d}\gamma_j^{(i)}\sum_{t=1}^{\tau_i}\Big[y_{dt}^{(i)}\ln\sigma(\tilde{\boldsymbol{C}}_{j,d,:}\tilde{\boldsymbol{m}}_{[t]}^{(i,j)}) + (1 - y_{dt}^{(i)})\ln\sigma(-\tilde{\boldsymbol{C}}_{j,d,:}\tilde{\boldsymbol{m}}_{[t]}^{(i,j)})\Big] \\ & - \frac{1}{8}\sum_j \operatorname{tr}(\boldsymbol{C}_j\boldsymbol{\Gamma}_j\boldsymbol{C}_j^{\intercal}),\end{aligned} \tag{IV.14}$$

where the aggregate statistics are

$$\begin{aligned} \hat{N}_j &= \textstyle\sum_i \gamma_j^{(i)}, & \boldsymbol{\varphi}_j &= \textstyle\sum_i \gamma_j^{(i)} \sum_{t=2}^{\tau_i} \hat{\boldsymbol{P}}_{t,t|j}^{(i)}, \\ \hat{\tau}_j &= \textstyle\sum_i \gamma_j^{(i)}\tau_i/\hat{N}_j, & \boldsymbol{\phi}_j &= \textstyle\sum_i \gamma_j^{(i)} \sum_{t=1}^{\tau_i-1} \hat{\boldsymbol{P}}_{t,t|j}^{(i)}, \\ \boldsymbol{\eta}_j &= \textstyle\sum_i \gamma_j^{(i)} \hat{\boldsymbol{P}}_{1,1|j}^{(i)}, & \boldsymbol{\Psi}_j &= \textstyle\sum_i \gamma_j^{(i)} \sum_{t=1}^{\tau_i-1} \hat{\boldsymbol{P}}_{t+1,t|j}^{(i)}, \\ \boldsymbol{\chi}_j &= \textstyle\sum_i \gamma_j^{(i)} \boldsymbol{m}_{[1]}^{(i,j)}, & \boldsymbol{\Gamma}_j &= \textstyle\sum_i \gamma_j^{(i)} \sum_{t=1}^{\tau_i} \hat{\boldsymbol{P}}_{t,t|j}^{(i)}. \end{aligned} \tag{IV.15}$$

The solution to (IV.19) leads to following explicit update rules of $\boldsymbol{\theta}^*$ for each component j:

$$\begin{aligned} \boldsymbol{\mu}_j^* &= \tfrac{1}{\hat{N}_j}\boldsymbol{\chi}_j, & \boldsymbol{S}_j^* &= \tfrac{1}{\hat{N}_j}\boldsymbol{\eta}_j - \boldsymbol{\mu}_j^*\boldsymbol{\mu}_j^{*\intercal}, & \alpha_j^* &= \hat{N}_j/N, \\ \boldsymbol{A}_j^* &= \boldsymbol{\Psi}_j\boldsymbol{\phi}_j^{-1}, & \boldsymbol{Q}_j^* &= \tfrac{1}{(\hat{\tau}_j-1)\hat{N}_j}(\boldsymbol{\varphi}_j - \boldsymbol{A}_j^*\boldsymbol{\Psi}_j^{\intercal}), \end{aligned} \tag{IV.16}$$

while $\tilde{\boldsymbol{C}}_j$ is updated by numerical solutions. See Appendix IV.F for details.

IV.C.3 Initialization

Initialization plays a critical role in the parameter estimation of models with hidden variables.

For $\boldsymbol{\theta}$, we consider two ways of initializing model parameters: 1) the suboptimal learning scheme of [97], which consists of a binary PCA and a least squares problem; and 2) the spectral learning of [18]. We notice from empirical results that, while the former defines an initial model inferior to that of the latter, its final convergent result slightly outperforms that of the latter.

For learning the mixture model, we adopt the strategies of [23] and learn mixture models by component splitting.

IV.D Variational EM for Mixtures of Binary Dynamic Systems using ELBO$_{JJ}$

If we substitute (IV.11) into (III.63), use the ELBO in Section III.C.2 and follow the same derivation in Section III.D, the objective function to optimize in

VEM of Algorithm 4 is:

$$
\begin{aligned}
&\mathscr{L}_{JJ}(q(\mathcal{T}_x,\mathcal{T}_z),\{\boldsymbol{\xi}^{(i,j)}\},\boldsymbol{\theta}) = \\
&\quad \sum_{i,j}\gamma_j^{(i)}\ln\alpha_j - \frac{1}{2}\sum_{i,j}\gamma_j^{(i)}\ln|\boldsymbol{S}_j| - \frac{1}{2}\sum_{i,j}\gamma_j^{(i)}(\tau_i-1)\ln\left|\boldsymbol{Q}_j\right| \\
&\quad -\frac{1}{2}\sum_{i,j}\gamma_j^{(i)}\mathrm{tr}\Big[\boldsymbol{S}_j^{-1}\big(\hat{\boldsymbol{P}}_{1,1|j}^{(i)} - \hat{\mathbf{s}}_{1|j}^{(i)}\boldsymbol{\mu}_j^{\intercal} - \boldsymbol{\mu}_j\hat{\mathbf{s}}_{1|j}^{(i)\intercal} + \boldsymbol{\mu}_j\boldsymbol{\mu}_j^{\intercal}\big)\Big] \\
&\quad -\frac{1}{2}\sum_{i,j}\gamma_j^{(i)}\sum_{t=1}^{\tau_i-1}\mathrm{tr}\Big[\boldsymbol{Q}_j^{-1}\big(\hat{\boldsymbol{P}}_{t+1,t+1|j}^{(i)} - \hat{\boldsymbol{P}}_{t+1,t|j}^{(i)}\boldsymbol{A}_j^{\intercal} - \boldsymbol{A}_j\hat{\boldsymbol{P}}_{t+1,t|j}^{(i)\intercal} + \boldsymbol{A}_j\hat{\boldsymbol{P}}_{t,t|j}^{(i)}\boldsymbol{A}_j^{\intercal}\big)\Big] \\
&\quad -\frac{1}{2}\sum_{i,j}\gamma_j^{(i)}\sum_{t=1}^{\tau_i}\mathrm{tr}\Big[\tilde{\boldsymbol{R}}_{t,j}^{(i)-1}\big(\tilde{\boldsymbol{u}}_{t,j}^{(i)}\tilde{\boldsymbol{u}}_{t,j}^{(i)\intercal} - \tilde{\boldsymbol{u}}_{t,j}^{(i)}\hat{\mathbf{s}}_{t|j}^{(i)\intercal}\boldsymbol{C}_j^{\intercal} - \boldsymbol{C}_j\hat{\mathbf{s}}_{t|j}^{(i)}\tilde{\boldsymbol{u}}_{t,j}^{(i)\intercal} + \boldsymbol{C}_j\hat{\boldsymbol{P}}_{t,t|j}^{(i)}\boldsymbol{C}_j^{\intercal}\big)\Big] \\
&\quad +\sum_{i,j}\gamma_j^{(i)}\sum_{t,d}\zeta(\xi_{d,t}^{(i,j)}) + \sum_{i,j}\gamma_j^{(i)}\mathrm{H}[(q_i(\mathrm{X}|j)] - \sum_{i,j}\gamma_j^{(i)}\ln\gamma_j^{(i)}, \qquad \text{(IV.17)}
\end{aligned}
$$

where $\hat{\mathbf{s}}_{r|j}^{(i)} = \langle \boldsymbol{x}_r\rangle_{q_i(\boldsymbol{x}|j)}$, $\hat{\boldsymbol{P}}_{r,s|j}^{(i)} = \langle \boldsymbol{x}_r\boldsymbol{x}_s^{\intercal}\rangle_{q_i(\boldsymbol{x}|j)} = \mathrm{cov}(\boldsymbol{x}_r,\boldsymbol{x}_s)_{q_i(\boldsymbol{x}|j)} + \hat{\mathbf{s}}_{r|j}^{(i)}\hat{\mathbf{s}}_{s|j}^{(i)\intercal}$, and $\tilde{\boldsymbol{R}}_{t,j}^{(i)-1}$, $\tilde{\boldsymbol{u}}_{t,j}^{(i)}$ and $\zeta(\xi)$ are defined by (III.50) and (III.51) using $\boldsymbol{\xi}^{(i,j)}$ and $\boldsymbol{y}^{(i)}$, respectively. The whole VEM algorithm for maximizing (IV.17) is depicted in Algorithm 5, with E-step, M-step and initialization discussed below.

IV.D.1 E-step

In the E-step, given the current estimate of the model parameter $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \{\boldsymbol{S}_j,\boldsymbol{\mu}_j,\boldsymbol{A}_j,\boldsymbol{C}_j,\boldsymbol{Q}_j,\boldsymbol{u}_j\}_{j=1}^K\}$, the variational distribution q and parameter $\boldsymbol{\xi}$ are updated by maximizing (IV.17) jointly over q and $\boldsymbol{\xi}$:

$$
\{q^*,\boldsymbol{\xi}^*\} = \underset{\{q,\boldsymbol{\xi}\}}{\arg\max}\ \mathscr{L}_{JJ}(q(\mathcal{T}_x,\mathcal{T}_z),\{\boldsymbol{\xi}^{(i,j)}\},\boldsymbol{\theta}). \qquad \text{(IV.18)}
$$

This is exactly the variational inference for the mixture model. Specifically, due to *i.i.d.* training data, the algorithm of Section III.D is repeated for each training

Algorithm 5: Variational EM for Mixtures of Binary Dynamic Systems with ELBO_{JJ}

Input: training corpora $\mathcal{T}_y = \{\boldsymbol{y}^{(i)}\}_{i=1}^{N}$, initial model parameter $\boldsymbol{\theta}^{(0)}$, initial variational parameter $\{\boldsymbol{\xi}^{(i,j)}\}$, number of components K;

$n \leftarrow 0$;
repeat
 (VE-step):
 $\boldsymbol{\varphi}_j, \boldsymbol{\phi}_j, \hat{N}_j, \tilde{\tau}_j, \boldsymbol{\Gamma}_j, \boldsymbol{\Psi}_j, \boldsymbol{\eta}_j, \boldsymbol{\Lambda}_{j,d}, \boldsymbol{v}_{j,d}^{\intercal}, \boldsymbol{\chi}_j \leftarrow 0$;
 for $i := 1$ *to* N **do**
  $\tilde{\gamma}^{(i)} \leftarrow 0$;
  for $j := 1$ *to* K **do**
   compute the variational distribution for $\boldsymbol{y}^{(i)}$ under j-th BDS component:

$$\{\ln p_{ij}^{*}, \boldsymbol{m}^{(i,j)}, \{\boldsymbol{\Phi}_{[t,t]}^{(i,j)}\}, \{\boldsymbol{\Phi}_{[t,t+1]}^{(i,j)}\}, \boldsymbol{\xi}^{(i,j)}\} \leftarrow \text{VarInf}_{\text{BDS}}(\boldsymbol{y}^{(i)}, \boldsymbol{\theta}_j^{(n)}, \boldsymbol{\xi}^{(i,j)});$$

   $\tilde{\gamma}^{(i)} \leftarrow \tilde{\gamma}^{(i)} + \alpha_j^{(n)} p_{ij}^{*}$;
  end
  for $j := 1$ *to* K **do**
   update $\gamma_j^{(i)}$ and statistics according to ($\hat{\boldsymbol{P}}_{r,s}^{(i,j)} = \boldsymbol{\Phi}_{[r,s]}^{(i,j)} + \boldsymbol{m}_{[r]}^{(i,j)} \boldsymbol{m}_{[s]}^{(i,j)\intercal}$)

$$\begin{array}{lll} \gamma_j^{(i)} \leftarrow \alpha_j^{(n)} p_{ij}^{*} / \tilde{\gamma}^{(i)}; & \boldsymbol{\varphi}_j \leftarrow \boldsymbol{\varphi}_j + \gamma_j^{(i)} \sum_{t=2}^{\tau_i} \hat{\boldsymbol{P}}_{t,t}^{(i,j)}; & \boldsymbol{\chi}_j \leftarrow \boldsymbol{\chi}_j + \gamma_j^{(i)} \boldsymbol{m}_{[1]}^{(i,j)}; \\ \hat{N}_j \leftarrow \hat{N}_j + \gamma_j^{(i)}; & \boldsymbol{\phi}_j \leftarrow \boldsymbol{\phi}_j + \gamma_j^{(i)} \sum_{t=1}^{\tau_i - 1} \hat{\boldsymbol{P}}_{t,t}^{(i,j)}; & \boldsymbol{\eta}_j \leftarrow \boldsymbol{\eta}_j + \gamma_j^{(i)} \hat{\boldsymbol{P}}_{1,1}^{(i,j)}; \\ \tilde{\tau}_j \leftarrow \tilde{\tau}_j + \gamma_j^{(i)} \tau_i; & \boldsymbol{\Psi}_j \leftarrow \boldsymbol{\Psi}_j + \gamma_j^{(i)} \sum_{t=1}^{\tau_i - 1} \hat{\boldsymbol{P}}_{t+1,t}^{(i,j)}; & \boldsymbol{\Gamma}_j \leftarrow \boldsymbol{\Gamma}_j + \gamma_j^{(i)} \sum_{t=1}^{\tau_i} \boldsymbol{\rho}_{t,j}^{(i)} [\boldsymbol{m}_{[t]}^{(i,j)\intercal}, 1]; \end{array}$$

   for $d := 1$ *to* D **do**

$$\boldsymbol{v}_{j,d}^{\intercal} \leftarrow \boldsymbol{v}_{j,d}^{\intercal} + \gamma_j^{(i)} \sum_{t=1}^{\tau_i} (2 y_{d,t}^{(i)} - 1) \, [\boldsymbol{m}_{[t]}^{(i,j)\intercal}, 1];$$
$$\boldsymbol{\Lambda}_{j,d} \leftarrow \boldsymbol{\Lambda}_{j,d} + \gamma_j^{(i)} \sum_{t=1}^{\tau_i} \lambda(\xi_{d,t}^{(i,j)}) \hat{\hat{\boldsymbol{P}}}_{t,t}^{(i,j)};$$

   end
  end
 end
 (VM-step):
 for $j := 1$ *to* K **do**
  update parameter $\boldsymbol{\theta}_j$ and α_j according to

$$\begin{array}{ll} \alpha_j^{(n+1)} \leftarrow \hat{N}_j / N; & \boldsymbol{\mu}_j^{(n+1)} \leftarrow \frac{1}{\hat{N}_j} \boldsymbol{\chi}_j; \\ \boldsymbol{S}_j^{(n+1)} \leftarrow \frac{1}{\hat{N}_j} \boldsymbol{\eta}_j - \boldsymbol{\mu}_j^{(n+1)} \boldsymbol{\mu}_j^{(n+1)\intercal}; & \boldsymbol{A}_j^{(n+1)} \leftarrow \boldsymbol{\Psi}_j \boldsymbol{\phi}_j^{-1}; \\ \boldsymbol{Q}_j^{(n+1)} \leftarrow \frac{1}{\tilde{\tau}_j - \hat{N}_j} (\boldsymbol{\varphi}_j - \boldsymbol{A}_j^{(n+1)} \boldsymbol{\Psi}_j^{\intercal}); & [\boldsymbol{C}_{j,d,:}{}^{(n+1)}, u_{j,d}^{(n+1)}] \leftarrow \frac{1}{4} \boldsymbol{v}_{j,d}^{\intercal} \boldsymbol{\Lambda}_{j,d}^{-1}; \end{array}$$

 end
 $n \leftarrow n + 1$;
until *convergence*;

Output : $\boldsymbol{\theta}^{(n)}$

example using parameter $\boldsymbol{\theta}$. During the i-th pass for $\boldsymbol{y}^{(i)}$, 1) Algorithm 2 is first repeated for $\boldsymbol{y}^{(i)}$ under each of the K BDS components, denoted as the *inner* EM; and then 2) the expected responsibilities of K component to $\boldsymbol{y}^{(i)}$ are computed by (III.69) using the component-conditional log-evidence of (III.68) estimated in the inner EM.

IV.D.2 M-step

In the M-step, given the current variational distribution q and parameter $\boldsymbol{\xi}$ estimated in the E-step, the model parameter $\boldsymbol{\theta}$ is updated by maximizing (IV.17) over $\boldsymbol{\theta}$:

$$
\begin{aligned}
\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \ & \sum_j \hat{N}_j \Big\{ \ln \alpha_j - \frac{1}{2}(\hat{\tau}_j - 1)\ln \Big|\boldsymbol{Q}_j\Big| - \frac{1}{2}\ln\big|\boldsymbol{S}_j\big| \Big\} \\
& - \frac{1}{2}\sum_j \operatorname{tr}\Big[\boldsymbol{S}_j^{-1}(\boldsymbol{\eta}_j - \boldsymbol{\chi}_j\boldsymbol{\mu}_j^{\intercal} - \boldsymbol{\mu}_j\boldsymbol{\chi}_j^{\intercal} + \hat{N}_j\boldsymbol{\mu}_j\boldsymbol{\mu}_j^{\intercal})\Big] \\
& - \frac{1}{2}\sum_j \operatorname{tr}\Big[\boldsymbol{Q}_j^{-1}(\boldsymbol{\varphi}_j - \boldsymbol{\Psi}_j\boldsymbol{A}_j^{\intercal} - \boldsymbol{A}_j\boldsymbol{\Psi}_j^{\intercal} + \boldsymbol{A}_j\boldsymbol{\phi}_j\boldsymbol{A}_j^{\intercal})\Big] \\
& - \frac{1}{2}\sum_j \Big\{ \sum_i \gamma_j^{(i)} \sum_{t=1}^{\tau_i} \operatorname{tr}\Big[\tilde{\boldsymbol{R}}_{t,j}^{(i)-1}\Big(\tilde{\boldsymbol{C}}_j\hat{\tilde{\boldsymbol{P}}}_{t,t|j}^{(i)}\tilde{\boldsymbol{C}}_j^{\intercal} - 2\boldsymbol{\Gamma}_j\tilde{\boldsymbol{C}}_j^{\intercal}\Big)\Big]\Big\},
\end{aligned}
\tag{IV.19}
$$

where the aggregate statistics are

$$
\begin{aligned}
&\hat{N}_j = \textstyle\sum_i \gamma_j^{(i)}, && \boldsymbol{\varphi}_j = \textstyle\sum_i \gamma_j^{(i)} \sum_{t=2}^{\tau_i} \hat{\boldsymbol{P}}_{t,t|j}^{(i)}, \\
&\hat{\tau}_j = \textstyle\sum_i \gamma_j^{(i)} \tau_i / \hat{N}_j, && \boldsymbol{\phi}_j = \textstyle\sum_i \gamma_j^{(i)} \sum_{t=1}^{\tau_i - 1} \hat{\boldsymbol{P}}_{t,t|j}^{(i)}, \\
&\boldsymbol{\eta}_j = \textstyle\sum_i \gamma_j^{(i)} \hat{\boldsymbol{P}}_{1,1|j}^{(i)}, && \boldsymbol{\Psi}_j = \textstyle\sum_i \gamma_j^{(i)} \sum_{t=1}^{\tau_i - 1} \hat{\boldsymbol{P}}_{t+1,t|j}^{(i)}, \\
&\boldsymbol{\chi}_j = \textstyle\sum_i \gamma_j^{(i)} \boldsymbol{m}_{[1]}^{(i,j)}, && \boldsymbol{\Gamma}_j = \textstyle\sum_i \gamma_j^{(i)} \sum_{t=1}^{\tau_i} \boldsymbol{\rho}_{t,j}^{(i)} \tilde{\boldsymbol{m}}_{[t]}^{(i,j)\intercal}, \\
&\boldsymbol{v}_{j,d}^{\intercal} = \textstyle\sum_i \gamma_j^{(i)} \sum_{t=1}^{\tau_i} (2y_{d,t}^{(i)} - 1)\tilde{\boldsymbol{m}}_{[t]}^{(i,j)\intercal}, && \boldsymbol{\Lambda}_{j,d} = \textstyle\sum_i \gamma_j^{(i)} \sum_{t=1}^{\tau_i} \lambda(\xi_{d,t}^{(i,j)}) \hat{\tilde{\boldsymbol{P}}}_{t,t|j}^{(i)},
\end{aligned}
\tag{IV.20}
$$

with $\tilde{\boldsymbol{C}}_j = \begin{bmatrix} \boldsymbol{C}_j, & \boldsymbol{u}_j \end{bmatrix}$ and

$$\boldsymbol{\rho}_{t,j}^{(i)} = \frac{1}{4}\Big[\frac{2y_{1,t}^{(i)}-1}{\lambda(\xi_{1,t}^{(i,j)})}, \cdots, \frac{2y_{D,t}^{(i)}-1}{\lambda(\xi_{D,t}^{(i,j)})}\Big]^\intercal, \quad \hat{\hat{\boldsymbol{P}}}_{t,t|j}^{(i)} = \begin{bmatrix} \hat{\boldsymbol{P}}_{t,t|j}^{(i)} & \boldsymbol{m}_{[t]}^{(i,j)} \\ \boldsymbol{m}_{[t]}^{(i,j)\intercal} & 1 \end{bmatrix}. \tag{IV.21}$$

The solution to (IV.19) leads to following explicit update rules of $\boldsymbol{\theta}^*$ for each component j:

$$\begin{aligned} &\boldsymbol{\mu}_j^* = \tfrac{1}{\hat{N}_j}\boldsymbol{\chi}_j, \quad \boldsymbol{S}_j^* = \tfrac{1}{\hat{N}_j}\boldsymbol{\eta}_j - \boldsymbol{\mu}_j^*\boldsymbol{\mu}_j^{*\intercal}, \quad \tilde{\boldsymbol{C}}_{j,d,:}^* = \tfrac{1}{4}\boldsymbol{v}_{j,d}^\intercal \boldsymbol{\Lambda}_{j,d}^{-1}, \\ &\boldsymbol{A}_j^* = \boldsymbol{\Psi}_j \boldsymbol{\phi}_j^{-1}, \quad \boldsymbol{Q}_j^* = \tfrac{1}{(\hat{\tau}_j-1)\hat{N}_j}(\boldsymbol{\varphi}_j - \boldsymbol{A}_j^*\boldsymbol{\Psi}_j^\intercal), \quad \alpha_j^* = \hat{N}_j/N. \end{aligned} \tag{IV.22}$$

Our further analysis shows that, these update rules achieve *global optimality* at each M-step, despite that the problem of (IV.19) is *non-convex* and the update rules are derived from its stationary point. Note that, this observation is significant in the sense that, although (IV.22) resemble update rules of the LDS and other variants, which are widely used in the community [146, 50, 131, 24], they are solely derived from *stationary points* of likelihood functions in the parameter space, yet (even local) optimality is seldom confirmed in the literature. See Appendix IV.F for complete derivation and proof.

IV.D.3 Initialization

For ξ, the initial value is set by $\xi_{d,t}^{(i,j)} = 5$ in Algorithm 5. In practice, we found that the result is not sensitive to this initial value, and the inner EM procedure converges in less than 10 iterations in almost all cases.

Initialization for $\boldsymbol{\theta}$ and the mixture model is the same as in Section IV.C.3.

IV.E Acknowledgement

The text of Chapter IV is, in part, based on the material as it appears in the following publications: The sub-optimal learning scheme for BDS was originally proposed in W.-X. LI and N. Vasconcelos, "Recognizing Activities by Attribute Dynamics." *Advances in Neural Information Processing Systems* (NIPS), 2012. The variational expectation-maximization algorithm for parameter estimation of BDS using JJ bound was originally proposed in W.-X. LI, Y. Li and N. Vasconcelos, "Efficient Variational Inference, Learning and Probabilistic Kernels for Binary Dynamic Systems." under review at *Neural Information Processing Systems* (NIPS), 2016. The variational expectation-maximization algorithm for parameter estimation of the mixture of binary dynamic systems was originally proposed in W.-X. LI, Y. Li and N. Vasconcelos, "Modeling, Clustering, and Segmenting Binary Sequences with Mixtures of Binary Dynamic Systems." under review at *Journal of Machine Learning Research* (JMLR). The dissertation author was a primary researcher and an author of the cited material.

IV.F Appendix

The parameter estimation is implemented via the variational EM algorithm (Algorithm 5). The details of E-step are discussed in Appendix III.D. In the M-step, we need to solve the optimization problem of (IV.19) to update

$\boldsymbol{\Theta} = \{\alpha_j, \boldsymbol{S}_j, \boldsymbol{\mu}_j, \boldsymbol{A}_j, \boldsymbol{C}_j, \boldsymbol{Q}_j, \boldsymbol{u}_j\}$ such that

$$\begin{aligned}
\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \; & \sum_j \hat{N}_j \Big\{ \ln \alpha_j - \frac{1}{2}\big[(\hat{\tau}_j - 1)\ln\Big|\boldsymbol{Q}_j\Big| - \frac{1}{2}\ln|\boldsymbol{S}_j|\big]\Big\} \\
& - \frac{1}{2}\sum_j \mathrm{tr}\Big[\boldsymbol{S}_j^{-1}(\boldsymbol{\eta}_j - \boldsymbol{\chi}_j\boldsymbol{\mu}_j^\intercal - \boldsymbol{\mu}_j\boldsymbol{\chi}_j^\intercal + \hat{N}_j\boldsymbol{\mu}_j\boldsymbol{\mu}_j^\intercal)\Big] \\
& - \frac{1}{2}\sum_j \mathrm{tr}\Big[\boldsymbol{Q}_j^{-1}(\boldsymbol{\varphi}_j - \boldsymbol{\Psi}_j\boldsymbol{A}_j^\intercal - \boldsymbol{A}_j\boldsymbol{\Psi}_j^\intercal + \boldsymbol{A}_j\boldsymbol{\phi}_j\boldsymbol{A}_j^\intercal)\Big] \\
& + \sum_j g^{(j)}(\boldsymbol{C}_j),
\end{aligned} \tag{IV.23}$$

where $g^{(j)}(\boldsymbol{C}_j)$ is the lower bound for the observation model in ELBOs, *i.e.*, for EBLO_{SJ} of (IV.12)

$$\begin{aligned}
g_1^{(j)}(\boldsymbol{C}_j) = & \sum_{i,d} \gamma_j^{(i)} \sum_{t=1}^{\tau_i} \Big[y_{kt}^{(i)} \ln\sigma(\tilde{\boldsymbol{C}}_{j,d,:}\tilde{\boldsymbol{m}}_{[t]}^{(i,j)}) + (1 - y_{kt}^{(i)})\ln\sigma(-\tilde{\boldsymbol{C}}_{j,d,:}\tilde{\boldsymbol{m}}_{[t]}^{(i,j)})\Big] \\
& - \frac{1}{8}\mathrm{tr}\Big(\tilde{\boldsymbol{C}}_j \begin{pmatrix} \boldsymbol{\Gamma}_j & 0 \\ 0 & 0 \end{pmatrix} \tilde{\boldsymbol{C}}_j^\intercal\Big),
\end{aligned} \tag{IV.24}$$

for ELBO_{JJ} of (IV.17)

$$g_2^{(j)}(\boldsymbol{C}_j) = -\frac{1}{2}\mathrm{tr}\Big[\Big(\sum_i \gamma_j^{(i)} \sum_{t=1}^{\tau_i} \Big(\tilde{\boldsymbol{R}}_{t,j}^{(i)}\Big)^{-1} \tilde{\boldsymbol{C}}_j \hat{\tilde{\boldsymbol{P}}}_{t,t|j}^{(i)} \tilde{\boldsymbol{C}}_j^\intercal\Big) - 2\boldsymbol{\Gamma}_j\tilde{\boldsymbol{C}}_j^\intercal\Big], \tag{IV.25}$$

and notations are defined the same as in Section IV.C.2 and Section IV.D.2. Let $f(\boldsymbol{\theta})$ be the objective function of the problem (IV.23). A major challenge here is that, $f(\boldsymbol{\theta})$ is *neither concave not convex* in $\boldsymbol{\theta}$, thus setting the gradient to zero with negative-definite Hessian only guarantees local optimum at best. We show that, however, the unique stationary point of $f(\boldsymbol{\theta})$ actually achieves *global optimum*. To this end, we first derive an algorithm to identify the sub-optimal point $\boldsymbol{\theta}^\circ$ that maximizes $f(\boldsymbol{\theta})$ with respect to each of its parameters individually; and then we prove that, $f(\boldsymbol{\theta})$ achieves *global optimum* at $\boldsymbol{\theta}^\circ$, *i.e.*, $\boldsymbol{\theta}^* = \boldsymbol{\theta}^\circ$.

IV.F.1 Optimization

Before presenting the results, we study optimization problems of general forms for brevity, which are used to derive solutions to problems in the rest of this chapter. Two typical forms of optimization problems are discussed in this part. They are all shown to be convex problems and closed form solutions are derived.

Problem 1

The first problem is

$$\max_{\boldsymbol{X}\in\mathcal{S}_{++}} \; -b\ln|\boldsymbol{X}| - \mathrm{tr}(\boldsymbol{A}\boldsymbol{X}^{-1}), \quad s.t.\;\; \boldsymbol{A}\in\mathcal{S}_{++},\, b>0. \tag{IV.26}$$

Let $\boldsymbol{Y} = \boldsymbol{X}^{-1} \in \mathcal{S}_{++}$, we have

$$(\boldsymbol{X}^*)^{-1} = \boldsymbol{Y}^* = \arg\max_{\boldsymbol{Y}\in\mathcal{S}_{++}} \; b\ln|\boldsymbol{Y}| - \mathrm{tr}(\boldsymbol{A}\boldsymbol{Y}), \quad s.t.\;\; \boldsymbol{A}\in\mathcal{S}_{++},\, b>0, \tag{IV.27}$$

since $\boldsymbol{X} \to \boldsymbol{Y}$ is a bijection between $\mathcal{S}_{++}$ and $\mathcal{S}_{++}$. Note that problem (IV.27) is identical to problem (III.85), thus the solution to problem (IV.26) is

$$\boldsymbol{X}^* = \frac{1}{b}\boldsymbol{A}. \tag{IV.28}$$

Problem 2

The second problem is

$$\max_{\boldsymbol{X}} \; -\mathrm{tr}\big[\boldsymbol{D}(\boldsymbol{X}\boldsymbol{C}\boldsymbol{X}^\intercal - 2\boldsymbol{B}\boldsymbol{X}^\intercal)\big], \quad s.t. \;\; \boldsymbol{X},\boldsymbol{B} \in \mathbb{R}^{n\times m},\ \boldsymbol{C} \in \mathcal{S}^m_{++}, \boldsymbol{D} \in \mathcal{S}^n_{++}. \tag{IV.29}$$

Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{C}^{\frac{1}{2}} \in \mathbb{R}^{n\times m}$, we have

$$\begin{aligned} \boldsymbol{X}^*\boldsymbol{C}^{\frac{1}{2}} = \boldsymbol{Y}^* = \arg\max_{\boldsymbol{Y}} \; & -\mathrm{tr}(\boldsymbol{D}\boldsymbol{Y}\boldsymbol{Y}^\intercal) + 2\mathrm{tr}(\boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{B}^\intercal\boldsymbol{D}\boldsymbol{Y}), \\ & s.t. \;\; \boldsymbol{Y},\boldsymbol{B} \in \mathbb{R}^{n\times m},\ \boldsymbol{C} \in \mathcal{S}^m_{++}, \boldsymbol{D} \in \mathcal{S}^n_{++}, \end{aligned} \tag{IV.30}$$

since $\boldsymbol{X} \to \boldsymbol{Y}$ is a *bijection* between $\mathbb{R}^{n\times m}$ and $\mathbb{R}^{n\times m}$. Note that, the objective function of problem (IV.30) is *strictly concave* as it consists of 1) a quadratic term in $\boldsymbol{Y}$ with negative-definite matrix as the coefficient, and 2) a linear term in $\boldsymbol{Y}$; and the domain is a convex set $\mathbb{R}^{n\times m}$. Thus problem of (IV.30) is a convex problem whose maximum is achieved at either 1) its stationary point(s) (if there is any), or 2) the boundary of its domain (possibly at infinity, *i.e.*, only the supremum is available).

The derivative of the objective function of problem (IV.30) is

$$\frac{\partial}{\partial \boldsymbol{Y}}\big\{-\mathrm{tr}(\boldsymbol{D}\boldsymbol{Y}\boldsymbol{Y}^\intercal) + 2\mathrm{tr}(\boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{B}^\intercal\boldsymbol{D}\boldsymbol{Y})\big\} = -2\boldsymbol{D}\boldsymbol{Y} + 2\boldsymbol{D}\boldsymbol{B}\boldsymbol{C}^{-\frac{1}{2}}. \tag{IV.31}$$

Setting $(IV.31)$ to zero leads to

$$\boldsymbol{Y}^* = \boldsymbol{B}\boldsymbol{C}^{-\frac{1}{2}}, \tag{IV.32}$$

and

$$\boldsymbol{X}^* = \boldsymbol{B}\boldsymbol{C}^{-1}. \tag{IV.33}$$

IV.F.2 Finding the Stationary Point

The sub-optimal point $\boldsymbol{\theta}^\circ$ of $f(\boldsymbol{\theta})$ can be computed by optimizing $f(\boldsymbol{\theta})$ with respect to each of its parameters individually, in the order of $\boldsymbol{\alpha}$, $\{\boldsymbol{\mu}_j\}$, $\{\boldsymbol{S}_j\}$, $\{\boldsymbol{A}_j\}$, $\{\boldsymbol{Q}_j\}$, $\{\boldsymbol{C}_j\}$ and $\{\boldsymbol{u}_j\}$.

Component Proportion $\{\alpha_j\}$

Optimizing $f(\boldsymbol{\theta})$ over $\boldsymbol{\alpha}$ requires solving the problem of

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \textstyle\sum_j \hat{N}_j \ln \alpha_j \\ &s.t.\ \ \forall j,\ \alpha_j \geqslant 0, \\ &\textstyle\sum_j \alpha_j = 1. \end{aligned} \tag{IV.34}$$

Note that, problem (IV.34) is a convex problem since 1) the objective function of (IV.34) is a concave function in $\boldsymbol{\alpha}$, and 2) its domain is a convex set (more precisely, a standard $(K-1)$-simplex).

Using Lagrange multipliers $\lambda \in \mathbb{R}$ and $\boldsymbol{\nu} \succeq \mathbf{0}$, problem of (IV.34) is converted to an unconstraint one:

$$\max_{\boldsymbol{\alpha},\lambda,\boldsymbol{\nu}} \textstyle\sum_j \hat{N}_j \ln \alpha_j + \sum_j \nu_j \alpha_j + \lambda(\sum_j \alpha_j - 1). \tag{IV.35}$$

By Karush-Kuhn-Tucker conditions, the optimal point $\{\boldsymbol{\alpha}^\circ, \lambda^\circ, \boldsymbol{\nu}^\circ\}$ shall satisfies

$$\frac{\hat{N}_j}{\alpha_j^\circ} + \nu_j^\circ + \lambda^\circ = 0, \ \forall j, \tag{IV.36}$$

$$\textstyle\sum_j \alpha_j^\circ = 1, \tag{IV.37}$$

$$\nu_j{}^\circ \alpha_j^\circ = 0, \ \forall j, \tag{IV.38}$$

$$\boldsymbol{\alpha}^\circ \succeq \mathbf{0}. \tag{IV.39}$$

Obviously, $\boldsymbol{\alpha}^\circ \succ \mathbf{0}$, thus

$$\boldsymbol{\nu}^\circ = \mathbf{0}. \tag{IV.40}$$

Combining (IV.36), (IV.37) and (IV.40) leads to solution

$$\alpha_j^\circ = \frac{\hat{N}_j}{\sum_k \hat{N}_k} = \frac{\hat{N}_j}{N}. \tag{IV.41}$$

Initial State Mean $\boldsymbol{\mu}_j$

Optimizing $f(\boldsymbol{\theta})$ over $\boldsymbol{\mu}_j$ requires solving the problem of

$$\max_{\boldsymbol{\mu}_j} \ -\operatorname{tr}\left[\boldsymbol{S}_j^{-1}(\hat{N}_j \boldsymbol{\mu}_j \boldsymbol{\mu}_j^\intercal - 2\boldsymbol{\chi}_j \boldsymbol{\mu}_j^\intercal)\right]. \tag{IV.42}$$

This is of the form of the general problem (IV.29), thus the solution is

$$\boldsymbol{\mu}_j^\circ = \frac{1}{\hat{N}_j} \boldsymbol{\chi}_j. \tag{IV.43}$$

Initial State Covariance Matrix S_j

Optimizing $f(\boldsymbol{\theta})$ over $\boldsymbol{S}_j$ requires solving the problem of

$$\max_{\boldsymbol{S}_j \succ \mathbf{0}} \; -\hat{N}_j \ln \left|\boldsymbol{S}_j\right| - \operatorname{tr}\left[(\boldsymbol{\eta}_j - 2\boldsymbol{\chi}_j \boldsymbol{\mu}_j^\intercal + \hat{N}_j \boldsymbol{\mu}_j \boldsymbol{\mu}_j^\intercal)\boldsymbol{S}_j^{-1}\right]. \tag{IV.44}$$

This is of the form of the general problem (IV.26), thus the solution is

$$\boldsymbol{S}_j^\circ = \frac{1}{\hat{N}_j}(\boldsymbol{\eta}_j - 2\boldsymbol{\chi}_j \boldsymbol{\mu}_j^\intercal + \hat{N}_j \boldsymbol{\mu}_j \boldsymbol{\mu}_j^\intercal) = \frac{1}{\hat{N}_j}\boldsymbol{\eta}_j - \boldsymbol{\mu}_j^\circ \boldsymbol{\mu}_j^{\circ\intercal} \tag{IV.45}$$

using the result of (IV.43).

State Transition Matrix A_j

Optimizing $f(\boldsymbol{\theta})$ over $\boldsymbol{A}_j$ requires solving the problem of

$$\max_{\boldsymbol{A}_j} \; -\operatorname{tr}\left[\boldsymbol{Q}_j^{-1}(\boldsymbol{A}_j \boldsymbol{\phi}_j \boldsymbol{A}_j^\intercal - 2\boldsymbol{\Psi}_j \boldsymbol{A}_j^\intercal)\right]. \tag{IV.46}$$

This is of the form of the general problem (IV.29), thus the solution is

$$\boldsymbol{A}_j^\circ = \boldsymbol{\Psi}_j \boldsymbol{\phi}_j^{-1}. \tag{IV.47}$$

State Noise Matrix Q_j

Optimizing $f(\boldsymbol{\theta})$ over $\boldsymbol{Q}_j$ requires solving the problem of

$$\max_{\boldsymbol{Q}_j \succ \mathbf{0}} \; -\hat{N}_j(\hat{\tau}_j - 1)\ln\left|\boldsymbol{Q}_j\right| - \operatorname{tr}\left[(\boldsymbol{\varphi}_j - 2\boldsymbol{\Psi}_j \boldsymbol{A}_j^\intercal + \boldsymbol{A}_j \boldsymbol{\phi}_j \boldsymbol{A}_j^\intercal)\boldsymbol{Q}_j^{-1}\right]. \tag{IV.48}$$

This is of the form of the general problem (IV.26), thus the solution is

$$\boldsymbol{Q}_j^{\circ} = \frac{1}{(\hat{\tau}_j - 1)\hat{N}_j}(\boldsymbol{\varphi}_j - 2\boldsymbol{\Psi}_j\boldsymbol{A}_j^{\intercal} + \boldsymbol{A}_j\boldsymbol{\phi}_j\boldsymbol{A}_j^{\intercal}) = \frac{1}{(\hat{\tau}_j - 1)\hat{N}_j}\left(\boldsymbol{\varphi}_j - \boldsymbol{A}_j^{\circ}\boldsymbol{\Psi}_j^{\intercal}\right) \tag{IV.49}$$

using the result of (IV.47).

Observation Matrix $\boldsymbol{C}_j$ and Mean Vector $\boldsymbol{u}_j$ for ELBO 1

Updating $\tilde{\boldsymbol{C}}_j$ in Section IV.C.2 requires solving the optimization problem of

$$\tilde{\boldsymbol{C}}_j^{*} = \arg\max_{\tilde{\boldsymbol{C}}_j} \; g_1^{(j)}(\tilde{\boldsymbol{C}}_j), \tag{IV.50}$$

$$g_1^{(j)}(\tilde{\boldsymbol{C}}_j) = \sum_{i,d}\gamma_j^{(i)}\sum_{t=1}^{\tau_i}\left[y_{kt}^{(i)}\ln\sigma(\tilde{\boldsymbol{C}}_{j,d,:}\tilde{\boldsymbol{m}}_{[t]}^{(i,j)}) + (1 - y_{kt}^{(i)})\ln\sigma(-\tilde{\boldsymbol{C}}_{j,d,:}\tilde{\boldsymbol{m}}_{[t]}^{(i,j)})\right] - \frac{1}{8}\mathrm{tr}\left(\tilde{\boldsymbol{C}}_j\begin{pmatrix}\boldsymbol{\Gamma}_j & 0\\ 0 & 0\end{pmatrix}\tilde{\boldsymbol{C}}_j^{\intercal}\right). \tag{IV.51}$$

Note that, problem (IV.50) is a convex optimization problem since 1) its objective function is the sum of a quadratic term with semi-negative definite matrix as the coefficient ($\mathrm{diag}(\boldsymbol{\Gamma}_j, 0) \succeq \boldsymbol{0}$), and a conical combination (with $\gamma_j^{(i)} \geqslant 0$ as coefficients) of negative log-sum-exp of $\tilde{\boldsymbol{C}}_j$, and 2) its domain $\mathbb{R}^{D\times(L+1)}$ is a convex set.

Defining the vector form of $\tilde{\boldsymbol{C}}_j$ as

$$\tilde{\boldsymbol{c}}_j = [\tilde{\boldsymbol{C}}_{j,1,:}, \cdots, \tilde{\boldsymbol{C}}_{j,D,:}]^{\intercal}, \tag{IV.52}$$

the derivative of (IV.51) is

$$\frac{\partial}{\partial \tilde{\boldsymbol{c}}_j} g_1^{(j)}(\tilde{\boldsymbol{c}}_j) = -\frac{1}{4}\mathrm{diag}(\boldsymbol{\Gamma}_j, \cdots, \boldsymbol{\Gamma}_j)\tilde{\boldsymbol{c}}_j - \frac{1}{4}\sum_{i,t} \gamma_j^{(i)} \begin{bmatrix} (\sigma(\tilde{\boldsymbol{C}}_{j,1,:}\tilde{\boldsymbol{m}}_{[t]}^{(i,j)}) - y_{1t}^{(i)})\tilde{\boldsymbol{m}}_{[t]}^{(i,j)} \\ \vdots \\ (\sigma(\tilde{\boldsymbol{C}}_{j,D,:}\boldsymbol{b}_{t,i}) - y_{Dt}^{(i)})\tilde{\boldsymbol{m}}_{[t]}^{(i,j)} \end{bmatrix}; \tag{IV.53}$$

and the second-order derivative of $g_1^{(j)}(\tilde{\boldsymbol{c}}_j)$ is

$$\begin{aligned} \frac{\partial^2}{\partial \tilde{\boldsymbol{c}}_j^2} g_1^{(j)}(\tilde{\boldsymbol{c}}_j) = & -\frac{1}{4}\mathrm{diag}(\boldsymbol{\Gamma}_j, \cdots, \boldsymbol{\Gamma}_j) \\ & -\frac{1}{4}\sum_i \gamma_j^{(i)} \sum_{t=1}^{\tau_i} \begin{bmatrix} \beta_1 \tilde{\boldsymbol{m}}_{[t]}^{(i,j)} \tilde{\boldsymbol{m}}_{[t]}^{(i,j)\top} & & \\ & \ddots & \\ & & \beta_D \tilde{\boldsymbol{m}}_{[t]}^{(i,j)} \tilde{\boldsymbol{m}}_{[t]}^{(i,j)\top} \end{bmatrix}, \end{aligned} \tag{IV.54}$$

where

$$\beta_k = \sigma(\tilde{\boldsymbol{C}}_{j,d,:}\tilde{\boldsymbol{m}}_{[t]}^{(i,j)})\sigma(-\tilde{\boldsymbol{C}}_{j,d,:}\tilde{\boldsymbol{m}}_{[t]}^{(i,j)}).$$

Numerical solvers (*e.g.*, gradient ascent, Newton-Raphson method, BFGS algorithm) can be used to search for the unique stationery point, *i.e.*, the global optimal point.

Observation Matrix C_j and Mean Vector u_j for ELBO 2

Defining $\boldsymbol{X} = \tilde{\boldsymbol{C}}_j^\mathsf{T}$ for convenience, optimizing $f(\boldsymbol{\theta})$ over $\tilde{\boldsymbol{C}}_j$ in Section IV.D.2 requires solving the problem of

$$\max_{\boldsymbol{X}} \; g_2^{(j)}(\boldsymbol{X}) \tag{IV.55}$$

$$s.t. \;\; g_2^{(j)}(\boldsymbol{X}) = -\sum_i \gamma_j^{(i)} \left\{ \sum_{t=1}^{\tau_i} \mathrm{tr}\Big[\boldsymbol{D}_{t,i}\boldsymbol{X}\boldsymbol{B}_{t,i}\boldsymbol{X}^\mathsf{T} - 2\boldsymbol{E}_{t,i}\boldsymbol{B}_{t,i}\boldsymbol{X}^\mathsf{T} \Big] \right\}, \tag{IV.56}$$

where

$$\boldsymbol{B}_{t,i} = \left(\tilde{\boldsymbol{R}}_{t,j}^{(i)}\right)^{-1}, \;\; \boldsymbol{D}_{t,i} = \hat{\boldsymbol{P}}_{t,t|j}^{(i)}, \;\; \boldsymbol{E}_{t,i} = \tilde{\boldsymbol{m}}_{[t]}^{(i,j)} \boldsymbol{\rho}_{t,j}^{(i)\mathsf{T}}.$$

using the statistics of (IV.20). The first-order derivative of (IV.56) is

$$\frac{\partial}{\partial \boldsymbol{X}} g_2^{(j)}(\boldsymbol{X}) = 2\sum_i \gamma_j^{(i)} \left\{ \sum_{t=1}^{\tau_i} \Big[\boldsymbol{E}_{t,i}\boldsymbol{B}_{t,i} - \boldsymbol{D}_{t,i}\boldsymbol{X}\boldsymbol{B}_{t,i}) \Big] \right\}; \tag{IV.57}$$

or, in the vectorized form,

$$\nabla_{\mathrm{vec}(\boldsymbol{X})}\, g_2^{(j)}(\boldsymbol{X}) = -2\left[\sum_i \gamma_j^{(i)} \sum_{t=1}^{\tau_i} \left(\boldsymbol{B}_{t,i} \otimes \boldsymbol{D}_{t,i} \right) \right] \mathrm{vec}(\boldsymbol{X}) + 2\sum_i \gamma_j^{(i)} \sum_{t=1}^{\tau_i} \mathrm{vec}(\boldsymbol{E}_{t,i}\boldsymbol{B}_{t,i}), \tag{IV.58}$$

where $\mathrm{vec}(\boldsymbol{A})$ is the vectorization of $\boldsymbol{A}$ by concatenating the columns of $\boldsymbol{A}$, and $\boldsymbol{A} \otimes \boldsymbol{B}$ is the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$. Using (III.50), the Hessian of (IV.56) in the vectorized form is

$$\begin{aligned} \boldsymbol{H} &= \nabla^2_{\mathrm{vec}(\boldsymbol{X})}\, g_2^{(j)}(\boldsymbol{X}) \\ &= -2\sum_i \gamma_j^{(i)} \sum_{t=1}^{\tau_i} \boldsymbol{B}_{t,i} \otimes \boldsymbol{D}_{t,i} \\ &= -4\sum_i \gamma_j^{(i)} \sum_{t=1}^{\tau_i} \mathrm{diag}(\lambda(\xi_{1,t}^{(i,j)})\boldsymbol{D}_{t,i}, \cdots, \lambda(\xi_{D,t}^{(i,j)})\boldsymbol{D}_{t,i}). \end{aligned} \tag{IV.59}$$

This leads to the vectorized form of (IV.56)

$$g_2^{(j)}(\boldsymbol{X}) = \frac{1}{2}\mathrm{vec}(\boldsymbol{X})^\intercal \cdot \boldsymbol{H} \cdot \mathrm{vec}(\boldsymbol{X}) + \boldsymbol{b}_{\boldsymbol{X}}^\intercal \cdot \mathrm{vec}(\boldsymbol{X}), \tag{IV.60}$$

where

$$\begin{aligned} \boldsymbol{b}_{\boldsymbol{X}} &= \sum_i \gamma_i \sum_{t=1}^{\tau_i} \mathrm{vec}(\boldsymbol{B}_{t,i}\boldsymbol{E}_{t,i}) \\ &= \sum_i \gamma_j^{(i)} \sum_{t=1}^{\tau_i} \left[(2y_{1,t}^{(i)} - 1)\tilde{\boldsymbol{m}}_{[t]}^{(i,j)\intercal}, \cdots, (2y_{D,t}^{(i)} - 1)\tilde{\boldsymbol{m}}_{[t]}^{(i,j)\intercal}\right]^\intercal. \end{aligned} \tag{IV.61}$$

Note that, $g_2^{(j)}(\boldsymbol{X})$ is quadratic in $\boldsymbol{X}$ with a negative-definite Hessian $\boldsymbol{H}$, and its domain $\mathbb{R}^{(L+1)\times D}$ is a convex set; therefore problem (IV.55) is a convex optimization problem. It follows that, (IV.56) is a strictly concave function in $\boldsymbol{X}$, and the optimal point $\boldsymbol{X}^\circ$ of (IV.56) or (IV.60) is computed in the closed form by

$$\begin{aligned} \mathrm{vec}(\boldsymbol{X}^\circ) &= -\boldsymbol{H}^{-1}\boldsymbol{b}_{\boldsymbol{X}} \\ &= \frac{1}{4}\begin{bmatrix} \left\{\sum_{i,t}\left[\gamma_j^{(i)}\lambda(\xi_{1,t}^{(i,j)})\boldsymbol{D}_{t,i}\right]\right\}^{-1}\left\{\sum_{i,t}\left(\gamma_j^{(i)}(2y_{1,t}^{(i)}-1)\tilde{\boldsymbol{m}}_{[t]}^{(i,j)}\right)\right\} \\ \vdots \\ \left\{\sum_{i,t}\left[\gamma_j^{(i)}\lambda(\xi_{D,t}^{(i,j)})\boldsymbol{D}_{t,i}\right]\right\}^{-1}\left\{\sum_{i,t}\left(\gamma_j^{(i)}(2y_{D,t}^{(i)}-1)\tilde{\boldsymbol{m}}_{[t]}^{(i,j)}\right)\right\} \end{bmatrix}. \end{aligned} \tag{IV.62}$$

Thus,

$$\tilde{\boldsymbol{C}}_{j,d,:}^\circ = \frac{1}{4}\left\{\sum_{i,t}\gamma_j^{(i)}(2y_{1,t}^{(i)}-1)\tilde{\boldsymbol{m}}_{[t]}^{(i,j)\intercal}\right\}\left\{\sum_{i,t}\left[\gamma_j^{(i)}\lambda(\xi_{1,t}^{(i,j)})\hat{\boldsymbol{P}}_{t,t|j}^{(i)}\right]\right\}^{-1}. \tag{IV.63}$$

IV.F.3 Global Optimality of the M-step

Although the objective function $f(\boldsymbol{\theta})$ of (IV.23) is generally *non-concave*, the sub-optimal point $\boldsymbol{\theta}^\circ$ that is determined in Section IV.F.2 via stationary point conditions, nevertheless, achieves *global optimum* for $f(\boldsymbol{\theta})$ by the following theorem.

Theorem 1 (Global Optimality) *For the objective function $f(\boldsymbol{\theta})$ of problem (IV.23), and the parameter $\boldsymbol{\theta}^\circ$ determined in Section IV.F.2,*

$$f(\boldsymbol{\theta}^\circ) \geqslant f(\boldsymbol{\theta}), \quad \forall\, \boldsymbol{\theta} \in \mathcal{T}_{\boldsymbol{\theta}}, \tag{IV.64}$$

where $\mathcal{T}_{\boldsymbol{\theta}}$ is the feasible set of problem (IV.23).

Proof (*Proof by contradiction*) Assume that there exists another point $\boldsymbol{\theta}' \neq \boldsymbol{\theta}^\circ$ such that $f(\boldsymbol{\theta}') > f(\boldsymbol{\theta}^\circ)$. Consider the following procedure.

1. Define $\boldsymbol{\theta}'_1 \equiv \{\tilde{\boldsymbol{\alpha}}, \{\boldsymbol{\mu}'_j\}, \{\boldsymbol{S}'_j\}, \{\boldsymbol{A}'_j\}, \{\boldsymbol{Q}'_j\}, \{\boldsymbol{C}'_j\}, \{\boldsymbol{u}'_j\}\}$, where

$$\tilde{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \{\boldsymbol{\mu}'_j\}, \{\boldsymbol{S}'_j\}, \{\boldsymbol{A}'_j\}, \{\boldsymbol{Q}'_j\}, \{\boldsymbol{C}'_j\}, \{\boldsymbol{u}'_j\}). \tag{IV.65}$$

   From the solution of (IV.34),

$$\tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^\circ. \tag{IV.66}$$

   Thus $\boldsymbol{\theta}'_1 = \{\boldsymbol{\alpha}^\circ, \{\boldsymbol{\mu}'_j\}, \{\boldsymbol{S}'_j\}, \{\boldsymbol{A}'_j\}, \{\boldsymbol{Q}'_j\}, \{\boldsymbol{C}'_j\}, \{\boldsymbol{u}'_j\}\}$, and

$$f(\boldsymbol{\theta}'_1) \geqslant f(\boldsymbol{\theta}'). \tag{IV.67}$$

2. Define $\boldsymbol{\theta}'_2 \equiv \{\boldsymbol{\alpha}^\circ, \{\tilde{\boldsymbol{\mu}}_j\}, \{\boldsymbol{S}_j\}, \{\boldsymbol{A}'_j\}, \{\boldsymbol{Q}'_j\}, \{\boldsymbol{C}'_j\}, \{\boldsymbol{u}'_j\}\}$, where

$$\{\tilde{\boldsymbol{\mu}}_j\} = \arg\max_{\{\boldsymbol{\mu}_j\}} f(\boldsymbol{\alpha}^\circ, \{\boldsymbol{\mu}_j\}, \{\boldsymbol{S}'_j\}, \{\boldsymbol{A}'_j\}, \{\boldsymbol{Q}'_j\}, \{\boldsymbol{C}'_j\}, \{\boldsymbol{u}'_j\}). \tag{IV.68}$$

   From the solution of (IV.42),

$$\tilde{\boldsymbol{\mu}}_j = \boldsymbol{\mu}_j^\circ. \tag{IV.69}$$

   Thus $\boldsymbol{\theta}'_2 = \{\boldsymbol{\alpha}^\circ, \{\boldsymbol{\mu}_j^\circ\}, \{\boldsymbol{S}'_j\}, \{\boldsymbol{A}'_j\}, \{\boldsymbol{Q}'_j\}, \{\boldsymbol{C}'_j\}, \{\boldsymbol{u}'_j\}\}$, and

$$f(\boldsymbol{\theta}'_2) \geqslant f(\boldsymbol{\theta}'_1). \tag{IV.70}$$

3. So forth.

In this way, a sequence of parameters $\boldsymbol{\theta}'_1, \cdots, \boldsymbol{\theta}'_6$ can be produced such that

$$f(\boldsymbol{\theta}'_i) \geqslant f(\boldsymbol{\theta}'_{i-1}) \tag{IV.71}$$

by repeating the above procedure in the order of $\boldsymbol{\alpha}$, $\{\boldsymbol{\mu}_j\}$, $\{\boldsymbol{S}_j\}$, $\{\boldsymbol{A}_j\}$, $\{\boldsymbol{Q}_j\}$, $\{\tilde{\boldsymbol{C}}_j\}$, and, at each step, using the parameter of the last step $\boldsymbol{\theta}'_{i-1}$ ($\boldsymbol{\theta}'_0 = \boldsymbol{\theta}'$) as the initial point to optimize over the i-th parameter while fixing others . Note that, the solution to each of these problems in Section IV.F.2 is *unique* and *deteministic*. Thus it follows that

$$\boldsymbol{\theta}'_6 = \boldsymbol{\theta}^\circ, \tag{IV.72}$$

and

$$f(\boldsymbol{\theta}^\circ) = f(\boldsymbol{\theta}'_6) \geqslant f(\boldsymbol{\theta}'_0) = f(\boldsymbol{\theta}') > f(\boldsymbol{\theta}^\circ). \tag{IV.73}$$

The contradiction of $f(\boldsymbol{\theta}^{\circ}) > f(\boldsymbol{\theta}^{\circ})$ in (IV.73) negates the initial proposition on the existence of $\boldsymbol{\theta}'$. Therefore,

$$f(\boldsymbol{\theta}^{\circ}) \geqslant f(\boldsymbol{\theta}), \quad \forall\, \boldsymbol{\theta} \in \mathcal{T}_{\boldsymbol{\theta}}. \tag{IV.74}$$

■

The theorem justifies the *global optimality* of the update rules of (IV.22) in the M-step of Algorithm 5. Similar conclusions can also be made in the same way for other popular Gaussian state-space models, *e.g.*, [146, 50, 131, 24], where little result has been reported on this crucial property before.

Chapter V

Encoding Sequential Data with Dynamic Systems

While dynamic systems presented in Chapter II and the tools for inference and learning in Chapter III and Chapter IV provide a statistical framework for characterization of sequential data generation and probabilistic reasoning, discriminative tasks typically require features for these length-varying signals that exploit the generative properties. In this chapter, we present several methods of encoding binary sequential data for discriminative tasks via the dynamic systems we introduced. Although in this work we specifically focus on binary sequential signals (attribute dynamics), these ideas can be easily extended to other scenarios, including the special case of LDS [127, 100].

V.A Bag-of-Words for Attribute Dynamics

In this section, we introduce the *bag-of-words for attribute dynamics* (BoWAD) representation. Inspired by the bag-of-visual-words (BoVW) framework in image analysis, BoWAD essentially encodes the zeroth order statistics of sequential binary data using a vocabulary of BDS codewords. This consists of quantizing sequential signals recorded from a target into BDS, words of attribute dynamics (WADs), and representing the target with the histogram of occurrences of the codewords. For this purpose, we need to specify 1) how to learn the codewords by clustering training data; and 2) how the difference (or similarity) is quantified between a sequential signal and a codeword, between two codewords. One statistically plausible implementation is to learn a mixture of dynamic models using parameter estimation of Section IV.B, and to use the log-evidence as the similarity metric between a binary sequence and BDS codeword. Here we exploit another more computationally efficient alternative, which generalizes the principle of k-means to the binary sequences.

V.A.1 Clustering Samples in the Model Domain

Conventional clustering algorithms identify prototypes in the space of training examples (*e.g.*, in k-means, a cluster prototype is the centroid of the samples in the cluster), using a metric suited for that space (*e.g.*, Euclidean distance). Clustering a collection of binary sequences is not straightforward because 1) binary sequences can have different length; 2) the space of these sequences has non-Euclidean geometry; and 3) the search for optimal prototypes, under this geometry, may lead to intractable non-linear optimization. This is compounded by the fact that the dynamics of binary sequences are better summarized by a set of prototype BDSs than a set of prototype sequences.

The problem of learning a set of BDS prototypes is an instance of the problem of learning a *bag-of-models* (BoM). Given a training set $\mathcal{D} = \{\boldsymbol{z}_i\}_{i=1}^{N}$ ($\boldsymbol{z}_i \in \mathcal{Z}, \forall i$), the goal is to learn a dictionary of representative *models* $\{M_i(\boldsymbol{z})\}_{i=1}^{N_C}$ in a *model space* $\mathcal{M}$. The proposed solution is based on two mappings. The first

$$f_{\mathcal{M}} : \mathcal{Z} \supseteq \{\boldsymbol{z}_i\} \mapsto M \in \mathcal{M} \tag{V.1}$$

maps a set of examples $\{\boldsymbol{z}_i\} \subseteq \mathcal{D}$ into a model $M(\boldsymbol{z})$. The second,

$$\mathcal{M} \times \mathcal{M} \ni (M_1, M_2) \mapsto d_{\mathcal{M}}(M_1, M_2) \in \mathbb{R}_+ \tag{V.2}$$

measures the dissimilarity or distance between models.

The mapping of (V.1) is first used to produce a model $M(\boldsymbol{z}_i)$ per training example $\boldsymbol{z}_i$. Training samples are then clustered, at the model level, by alternating between two steps. In the *assignment step*, each $\boldsymbol{z}_i$ is assigned to the cluster whose model is closest to $M(\boldsymbol{z}_i)$, using the mapping of (V.2). In the *model refinement step*,

Algorithm 6: Bag-of-Models Clustering

Input : a set of samples $\mathcal{D} = \{\boldsymbol{z}_i\}_{i=1}^{N}$ ($\boldsymbol{z}_i \in \mathcal{Z}, \forall i$), number of clusters N_C, an initial set of models $\{M_i^{(0)}\}_{i=1}^{N_C}$.

set $t = 0$ *and* $S_i^{(0)} = \varnothing, i = 1, \cdots, N_C$;

repeat

$t = t + 1$;

Assignment-Step: $\forall i,\ S_i^{(t)} = \{\boldsymbol{z} \in \mathcal{D} \mid \forall j \neq i,\ \ d_{\mathcal{M}}(M(\boldsymbol{z}), M_i^{(t-1)}) \leqslant d_{\mathcal{M}}(M(\boldsymbol{z}), M_j^{(t-1)})\}$

Refinement-Step: $\forall i,\ M_i^{(t)} = M(\{S_i^{(t)}\})$

until $\forall i, S_i^{(t)} = S_i^{(t-1)}$;

Output: $\{M_i^{(t)}\}_{i=1}^{N_C}$ and $\{S_i^{(t)}\}_{i=1}^{N_C}$

the model associated with each cluster is relearned from the training samples assigned to it, via (V.1). This procedure is summarized in Algorithm 6 and denoted *bag-of-models clustering* (BMC).

BMC generalizes k-means, where $\boldsymbol{z}_i \in \mathbb{R}^d$ are feature vectors, $\mathcal{M}$ is the space of Gaussians of identity covariance

$$\mathcal{M} = \left\{ \mathcal{G}(\boldsymbol{z}; \boldsymbol{\mu}, I_d) \mid \boldsymbol{\mu} \in \mathbb{R}^d \right\}, \tag{V.3}$$

(V.1) selects the model

$$M(\{\boldsymbol{z}_i\}) = \mathcal{G}(\boldsymbol{z}; \hat{\boldsymbol{\mu}}, I), \tag{V.4}$$

where $\hat{\boldsymbol{\mu}}$ is the ML estimate of the mean

$$\hat{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}} p(\{\boldsymbol{z}_i\}; \boldsymbol{\mu}) = \frac{1}{|\{\boldsymbol{z}_i\}|} \sum\nolimits_i \boldsymbol{z}_i, \tag{V.5}$$

and (V.2) is the symmetric KL divergence derived from (II.1),

$$\mathrm{KL}(p_1 || p_2) + \mathrm{KL}(p_2 || p_1) = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2. \tag{V.6}$$

It should be noted that BMC differs from the *bag-of-systems* approach [128, 2] in two ways. First, it clusters *attribute sequences* rather than models. While, in the refinement step of Algorithm 6, models are re-learned from examples $\{\boldsymbol{z}_i\}$, the refinement step of [128, 2] only considers parameters of the models $M(\boldsymbol{z}_i)$ and not the examples $\boldsymbol{z}_i$ themselves. This usually entails loss of information. Second, Algorithm 6 *finds* the optimal representative for each cluster, according to the model fitting criterion of (V.1). In [128], the difficult geometry of the manifold defined by the LDS parameter tuple $(A, C) \in \mathbb{GL}(n) \times \mathbb{ST}(p, n)$, where $\mathbb{GL}(i)$ is the set of invertible matrices of size n and $\mathbb{ST}(p, n)$ the Stiefel manifold of $p \times n$ orthonormal matrices $(p \geqslant n)$, precludes a simple estimate of the optimal representative. Instead, this is approximated by model $M(\boldsymbol{z}_i)$ closest to the optimal representative. Although [2] introduce an approach to directly cluster LDS's in parameter space, its generalization to the BDS is unclear. We will show, in Section VI.E, that these differences can lead to significantly improved performance by Algorithm 6.

V.A.2 Dissimilarity Measure Between BDSs

Algorithm 6 requires a measure of distance Between BDSs. For this, we generalize a popular measure of distance between LDSs, the Binet-Cauchy kernel (BCK) of [161]. Given LDSs $\boldsymbol{\Omega}_a$ and $\boldsymbol{\Omega}_b$ driven by identical noise processes $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$ with observation sequences $\boldsymbol{y}^{(a)}$ and $\boldsymbol{y}^{(b)}$, the BCK is

$$K_{BC}(\boldsymbol{\Omega}_a, \boldsymbol{\Omega}_b) = \left\langle \sum_{t=0}^{\infty} e^{-\lambda t} (\boldsymbol{y}_t^{(a)})^\intercal \boldsymbol{W} \boldsymbol{y}_t^{(b)} \right\rangle_{p(\boldsymbol{v}, \boldsymbol{w})}, \tag{V.7}$$

where $\boldsymbol{W}$ is a semi-definite positive weight matrix and $\lambda \geqslant 0$ a temporal discounting factor. To extend (V.7) to BDSs $\boldsymbol{\Omega}_a$ and $\boldsymbol{\Omega}_b$, we note that $(\boldsymbol{y}_t^{(a)})^\intercal \boldsymbol{W} \boldsymbol{y}_t^{(b)}$

is the inner product of the Euclidean space of metric $d^2(\boldsymbol{y}_t^{(a)}, \boldsymbol{y}_t^{(b)}) = (\boldsymbol{y}_t^{(a)} - \boldsymbol{y}_t^{(b)})^\intercal \boldsymbol{W}(\boldsymbol{y}_t^{(a)} - \boldsymbol{y}_t^{(b)})$. For BDSs, whose observations y_t are Bernoulli distributed with parameters $\{\sigma(\boldsymbol{\theta}_t^{(a)})\}$, for $\boldsymbol{\Omega}_a$, and $\{\sigma(\boldsymbol{\theta}_t^{(b)})\}$, for $\boldsymbol{\Omega}_b$, this distance measure is naturally replaced by the symmetric KL divergence between Bernoulli distributions. This results in the *Binet-Cauchy KL divergence* (BC-KLD) [1]

$$\begin{aligned} D_{BC}(\boldsymbol{\Omega}_a, \boldsymbol{\Omega}_b) = \mathbb{E}_{\boldsymbol{v}}\Bigg[\sum_{t=0}^{\infty} e^{-\lambda t}\Big(&\mathrm{KL}(B(\sigma(\boldsymbol{\theta}_t^{(a)}))||B(\sigma(\boldsymbol{\theta}_t^{(b)}))) \\ &+ \mathrm{KL}(B(\sigma(\boldsymbol{\theta}_t^{(b)}))||B(\sigma(\boldsymbol{\theta}_t^{(a)})))\Big)\Bigg] \\ = \mathbb{E}_{\boldsymbol{v}}\Bigg[\textstyle\sum_{t=0}^{\infty} e^{-\lambda t}&\big(\sigma(\boldsymbol{\theta}_t^{(a)}) - \sigma(\boldsymbol{\theta}_t^{(b)})\big)^\intercal\big(\boldsymbol{\theta}_t^{(a)} - \boldsymbol{\theta}_t^{(b)}\big)\Bigg], \end{aligned} \qquad \text{(V.8)}$$

where $\boldsymbol{\theta}_t = \boldsymbol{C}\boldsymbol{x}_t + \boldsymbol{u}$ is the parameter of the multivariate Bernoulli distribution. The divergence at time t can be rewritten as

$$(\sigma(\boldsymbol{\theta}_t^{(a)}) - \sigma(\boldsymbol{\theta}_t^{(b)}))^\intercal(\boldsymbol{\theta}_t^{(a)} - \boldsymbol{\theta}_t^{(b)}) = (\boldsymbol{\theta}_t^{(a)} - \boldsymbol{\theta}_t^{(b)})^\intercal \hat{\boldsymbol{W}}_t(\boldsymbol{\theta}_t^{(a)} - \boldsymbol{\theta}_t^{(b)}), \qquad \text{(V.9)}$$

with $\hat{\boldsymbol{W}}_t$ a diagonal matrix whose k-th diagonal element is $\hat{W}_{t,k} = (\sigma(\theta_{t,k}^{(a)}) - \sigma(\theta_{t,k}^{(b)}))/(\theta_{t,k}^{(a)} - \theta_{t,k}^{(b)}) = \sigma'(\hat{\theta}_{t,k}^{(a,b)})$ (where, by the mean value theorem, $\hat{\theta}_{t,k}^{(a,b)}$ is some real value between $\hat{\theta}_{t,k}^{(a)}$ and $\hat{\theta}_{t,k}^{(b)}$). This reduces (V.9) to a form similar to (V.7), although with a time varying weight matrix $\boldsymbol{W}_t$. It is, nevertheless unclear whether (V.8) can be computed in closed-form. We rely on the approximation

$$D_{BC}(\boldsymbol{\Omega}_a, \boldsymbol{\Omega}_b) \approx \sum_{t=0}^{\infty} e^{-\lambda t}\left[\sigma(\bar{\boldsymbol{\theta}}_t^{(a)}) - \sigma(\bar{\boldsymbol{\theta}}_t^{(b)})\right]^\intercal \left[\bar{\boldsymbol{\theta}}_t^{(a)} - \bar{\boldsymbol{\theta}}_t^{(b)}\right], \qquad \text{(V.10)}$$

where $\bar{\boldsymbol{\theta}}$ is the mean of $\boldsymbol{\theta}$.

[1] Although the square root of the symmetric KL divergence is not a metric (since the triangle inequality does not hold), it has been shown effective for the design of probability distribution kernels, in the context of various applications [106, 159, 55, 21].

V.A.3 Learning a WAD Vocabulary

Given the BC-KLD distance between BDSs, it is possible to learn a WAD dictionary from a set of binary sequences $\mathcal{P} = \{\mathbf{\Pi}^{(i)}\}_{i=1}^{N}$, by applying Algorithm 6 as follows.

Refinement-Step: The mapping of (V.1) amounts to fitting a BDS to $\mathcal{P}' = \{\mathbf{\Pi}^{(i)}\} \subseteq \mathcal{P}$. This is done with Algorithm 3. The BDS learned per cluster jointly characterizes the appearance and dynamics of all attribute sequences in that cluster.

Assignment-Step: Each sample BDS is assigned to the closest centroid BDS, using (V.10).

To initialize the clustering algorithm, we follow the strategy of [23]. This has produced satisfactory results in all our experiments.

V.A.4 Quantization of BoAS with WAD Vocabulary

Given a WAD dictionary $\{\mathbf{\Omega}^{(i)}\}_{i=1}^{V}$, a set of binary sequences $\{\boldsymbol{y}_{1:\tau_i}^{(i)}\}_{i=1}^{N}$ is quantized by assigning the i-th attribute sequence to the k^*-th cluster according to

$$k^* = \arg\min_{j} d_{BC}\big(\mathbf{\Omega}(\boldsymbol{y}_{1:\tau_i}^{(i)}), \mathbf{\Omega}^{(j)}\big), \tag{V.11}$$

where $\mathbf{\Omega}(\boldsymbol{y}_{1:\tau_i}^{(i)})$ is the BDS learnt from $\boldsymbol{y}_{1:\tau_i}^{(i)}$ using (V.1). This produces a histogram of WAD counts, denoted *bag-of-words for attribute dynamics* (BoWAD), which can be used to classify video sequences of complex activities with the procedures commonly used for the BoVW [93, 166].

V.B Encoding Attribute Dynamics via Fisher Vector

In this section, we derive a scheme to encode the first-order statistics of a set of binary sequences using a BDS vocabulary, denoted as the vector of locally aggregated descriptors for attribute dynamics (VLADAD).

V.B.1 Bag-of-Models Interpretation of VLAD

The vector of locally aggregated descriptors (VLAD) [71] is an efficient representation of the first-order statistics of a data sample. It has been shown to outperform the BoVW histogram, which only captures zero-order statistics, in many image classification experiments. To extend the VLAD to the BoM, we start by interpreting it as an encoding of sample statistics with respect to a collection of local tensors of a model manifold.

Consider a Riemannian manifold $\mathcal{M}$ with geodesic distance $d_{\mathcal{M}}(M_1, M_2)$, such as (V.2), a set of reference models $\{M_i\}_{i=1}^{N_C}$, embedded in $\mathcal{M}$, and neighborhoods

$$\mathcal{R}_i = \{M \in \mathcal{M} | d_{\mathcal{M}}(M, M_i) \leqslant d_{\mathcal{M}}(M, M_j), j \neq i\},$$

where $\mathcal{R}_i$ is the neighborhood of M_i under $d_{\mathcal{M}}$. To encode a collection of examples $\mathcal{D} = \{z_i\}_{i=1}^{N}$ ($z_i \in \mathcal{Z}, \forall i$), these are first assigned to the regions $\mathcal{R}_i$

$$\mathcal{D}^i = \{z \in \mathcal{D} | f_{\mathcal{M}}(z) \in \mathcal{R}_i\} \tag{V.12}$$

using an assignment mapping $f_{\mathcal{M}}$, such as (V.1).

VLAD assumes examples $z \in \mathbb{R}^D$ and Gaussian models M_i, *i.e.*, a model manifold

$$\mathcal{M} = \{\mathcal{G}(\boldsymbol{z};\boldsymbol{\mu},\boldsymbol{\Sigma}) \mid \boldsymbol{\mu} \in \mathbb{R}^D, \boldsymbol{\Sigma} \in \mathcal{S}^D_{++}\}, \tag{V.13}$$

with geodesic distance approximated by the symmetric KL divergence

$$d_{\mathcal{M}}(M_1, M_2) = \mathrm{KL}(p_{M_1}||p_{M_2}) + \mathrm{KL}(p_{M_2}||p_{M_1}), \tag{V.14}$$

where $\mathrm{KL}(p_{M_1}||p_{M_2})$ is defined in (II.1). Most VLAD implementations assume that $\boldsymbol{\Sigma} = \boldsymbol{I}$, reducing (V.14) to the Euclidean metric $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$ ($\mathcal{D}^i$ assigned to the model of mean closest to the sample centroid). In this case, the assignment mapping maps an example $\boldsymbol{z}$ to a Gaussian of mean $\boldsymbol{\mu}$ and identity covariance, *i.e.*,

$$f_{\mathcal{M}}(\boldsymbol{z}) : \boldsymbol{z} \to \mathcal{G}(\boldsymbol{z};\boldsymbol{\mu},\boldsymbol{I}_D) \tag{V.15}$$

where $\boldsymbol{\mu} \in \{\boldsymbol{\mu}_i\}$ is the mean of one of the reference Gaussians.

As illustrated in Fig. V.1, the idea behind VLAD is to use the local tensor $\mathcal{G}_{M_i}$ defined by distance $d_{\mathcal{M}}(\cdot,\cdot)$ at M_i to encode the distribution of $\mathcal{D}^i$. A descriptor of $\mathcal{D}$ is then constructed by 1) aggregating the encoding of the examples in $\mathcal{D}^i$, for each region $\mathcal{R}_i$, and 2) concatenating the aggregate encodings from all regions. When $\mathcal{M}$ is a statistical manifold (of parameter $\boldsymbol{\theta}$), a commonly used metric tensor is the *Fisher kernel* [64]

$$K_M(z_1, z_2) = U_M^{\intercal}(z_1) I_M^{-1} U_M(z_2), \tag{V.16}$$

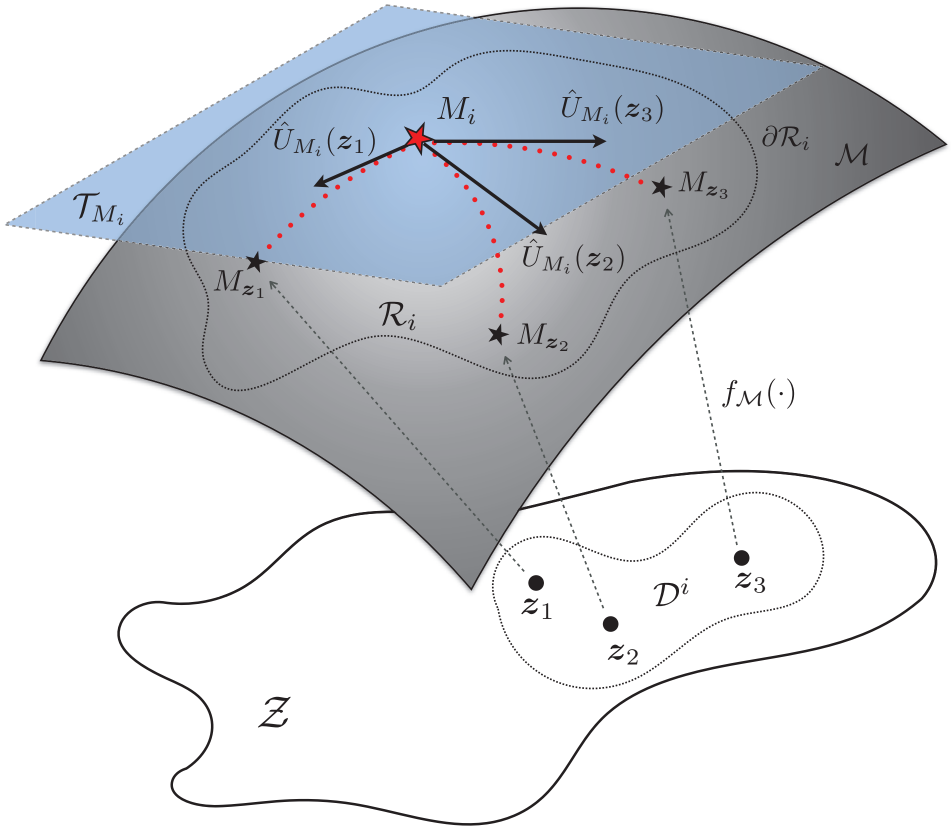


Figure V.1: VLAD encoding under the bag of models representation. The samples in $\mathcal{D}^i$ are first mapped into model manifold $\mathcal{M}$ by $f_{\mathcal{M}_i}(z)$, and then encoded by their statistics with respect to M_i (the red star in the figure), using the mapping $\hat{U}_{M_i}(z) = \boldsymbol{I}_{M_i}^{-1/2} U_{M_i}(z)$ defined by the local tensor $\mathcal{G}_{M_i}$, *i.e.*, the metric of the tangent space at M_i (the blue plane in the figure).

where

$$U_M(z) = \nabla_{\boldsymbol{\theta}} \log p_M(z;\boldsymbol{\theta}), \tag{V.17}$$

is the *Fisher score* and I_M is the *Fisher information metric*[2] at M. This tensor can be shown to approximate the KL-divergence in the neighborhood of M [4].

[2]In practice, the Fisher information metric I_M is often omitted, since the Fisher kernel is an Euclidean metric in the range space of the invertible linear transformation by $I_M^{\frac{1}{2}}$, of the tangent space of the manifold at M.

For the manifold of (V.13), the Fisher score is

$$U_M(z) = \begin{bmatrix} \nabla_{\mu} \log p_M(z;\mu,\Sigma) \\ \nabla_{\Sigma^{-1}} \log p_M(z;\mu,\Sigma) \end{bmatrix},$$

with

$$\nabla_{\mu} \log p_M(z) = \Sigma^{-1}(z-\mu), \tag{V.18}$$

$$\nabla_{\Sigma^{-1}} \log p_M(z) = \frac{1}{2}\Big[\Sigma - (z-\mu)(z-\mu)^{\intercal}\Big]. \tag{V.19}$$

After the aggregation over the sample $\mathcal{D}^i$, (V.18) encodes the relative position of the centroid of this sample w.r.t.
the region center μ_i (under the Mahalanobis metric defined by Σ_i^{-1}). Similarly, (V.19) encodes the relative shape of the sample w.r.t. that of the reference distribution, which is parametrized by Σ_i. Under the assumption that $\Sigma = I$, (V.18) reduces to $z - \mu$ and the second order statistics of (V.19) are usually omitted. This has some loss but reduces complexity [71].

V.B.2 Vector of Locally Aggregate Descriptors for Attribute Dynamics

The extension of the VLAD to the BDS requires evaluating the derivative of the expected log-likelihood of the sample with respect to the model parameters. This, however, is intractable, due to the intractability of the posterior distribution of BDS state given observations. To overcome this difficulty, we resort to approximate variational inference [79]. A similar strategy has recently been shown effective for image analysis [30].

The VLAD for attribute dynamics (VLADAD) approximates the Fisher score of the BDS by the derivatives of the ELBO with respect to the model parameters.

Using ELBO$_{SJ}$ in Section III.C.1

It can be shown that (see Appendix V.E.2), given an attribute sequence $\boldsymbol{y}$ and BDS $\boldsymbol{\theta} = \{\boldsymbol{S}^{-1}, \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{Q}^{-1}, \boldsymbol{C}, \boldsymbol{u}\}$[3], $\hat{\mathscr{L}}(\boldsymbol{\theta}, q^*)$ of (III.38) is of the form (using the same notations as in Section III.C.1)

$$\begin{aligned}\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q^*) = -\frac{1}{2}\Big\{&(\tau-1)\ln|\boldsymbol{Q}| + \ln|\boldsymbol{S}| + \mathrm{tr}\Big[\boldsymbol{S}^{-1}(\hat{\boldsymbol{P}}^*_{1,1} - \boldsymbol{\mu}\boldsymbol{m}^{*\intercal}_{[1]} - \boldsymbol{m}^*_{[1]}\boldsymbol{\mu}^\intercal + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal)\Big] \\ &+ \mathrm{tr}\Big[\boldsymbol{Q}^{-1}(\boldsymbol{\varphi} - \boldsymbol{\Psi}\boldsymbol{A}^\intercal - \boldsymbol{A}\boldsymbol{\Psi}^\intercal + \boldsymbol{A}\boldsymbol{\phi}\boldsymbol{A}^\intercal)\Big] + \frac{1}{4}\mathrm{tr}\Big[\boldsymbol{C}\big(\sum_t \boldsymbol{\Phi}^*_{[t,t]}\big)\boldsymbol{C}^\intercal\Big]\Big\} \\ &+ \sum_{t,k}\Big[y_{kt}\ln\sigma(\hat{\omega}^*_{kt}) + (1-y_{kt})\ln\sigma(-\hat{\omega}^*_{kt})\Big] + \text{const},\end{aligned} \tag{V.20}$$

which has derivatives

$$\frac{\partial}{\partial \boldsymbol{S}^{-1}}\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q^*) = \frac{1}{2}\Big(\boldsymbol{S} + \boldsymbol{\mu}\boldsymbol{m}^{*\intercal}_{[1]} + \boldsymbol{m}^*_{[1]}\boldsymbol{\mu}^\intercal - \hat{\boldsymbol{P}}^*_{1,1} - \boldsymbol{\mu}\boldsymbol{\mu}^\intercal\Big), \tag{V.21}$$

$$\frac{\partial}{\partial \boldsymbol{\mu}}\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q^*) = \boldsymbol{S}^{-1}(\boldsymbol{m}^*_{[1]} - \boldsymbol{\mu}), \tag{V.22}$$

$$\frac{\partial}{\partial \boldsymbol{A}}\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q^*) = \boldsymbol{Q}^{-1}(\boldsymbol{\Psi} - \boldsymbol{A}\boldsymbol{\phi}), \tag{V.23}$$

$$\frac{\partial}{\partial \boldsymbol{Q}^{-1}}\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q^*) = \frac{1}{2}\Big[\boldsymbol{\Psi}\boldsymbol{A}^\intercal + \boldsymbol{A}\boldsymbol{\Psi}^\intercal - \boldsymbol{A}\boldsymbol{\phi}\boldsymbol{A}^\intercal - \boldsymbol{\varphi} + (\tau-1)\boldsymbol{Q}\Big], \tag{V.24}$$

$$\frac{\partial}{\partial \tilde{\boldsymbol{C}}}\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q^*) = -\frac{1}{4}\left\{\tilde{\boldsymbol{C}}\tilde{\Upsilon} + \sum_{t=1}^{\tau}\begin{bmatrix}\sigma(\tilde{\boldsymbol{C}}_{1,:}\tilde{\boldsymbol{m}}^*_{[t]}) - y_{1t} \\ \vdots \\ \sigma(\tilde{\boldsymbol{C}}_{D,:}\tilde{\boldsymbol{m}}^*_{[t]}) - y_{Dt}\end{bmatrix}\tilde{\boldsymbol{m}}^{*\intercal}_{[t]}\right\}, \tag{V.25}$$

[3]For simplicity, we consider the precision matrices $\boldsymbol{S}^{-1}$ and $\boldsymbol{Q}^{-1}$ instead of the covariances $\boldsymbol{S}, \boldsymbol{Q}$ in the computation of Fisher scores.

where

$$\hat{\boldsymbol{P}}^*_{r,s} = \boldsymbol{\Phi}^*_{[r,s]} + \boldsymbol{m}^*_{[r]}\boldsymbol{m}^{*\intercal}_{[s]},\ \boldsymbol{\varphi} = \sum_{t=2}^{\tau}\hat{\boldsymbol{P}}^*_{t,t},\ \boldsymbol{\phi} = \sum_{t=1}^{\tau-1}\hat{\boldsymbol{P}}^*_{t,t},\ \boldsymbol{\Psi} = \sum_{t=2}^{\tau}\hat{\boldsymbol{P}}^*_{t,t-1};$$

and $\tilde{\boldsymbol{C}} = [\boldsymbol{C}, \boldsymbol{u}]$,

$$\tilde{\Upsilon} = \begin{pmatrix} \sum_{t=1}^{\tau}\boldsymbol{\Phi}^*_{[t,t]} & 0 \\ 0 & 0 \end{pmatrix}.$$

Using ELBO$_{JJ}$ in Section III.C.2

It can be shown that (see Appendix V.E.3), given attribute sequence $\boldsymbol{y}$ and BDS $\boldsymbol{\theta} = \{\boldsymbol{S}^{-1}, \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{Q}^{-1}, \boldsymbol{C}, \boldsymbol{u}\}$, $\hat{\mathscr{L}}(\boldsymbol{\theta}, q^*)$ of (III.49) is of the form (using the same notations as in Section III.C.2)

$$\begin{aligned}\hat{\mathscr{L}}_{JJ}(\boldsymbol{\theta}, q^*) = -\frac{1}{2}\Big\{&(\tau-1)\ln|\boldsymbol{Q}| + \mathrm{tr}\Big[\boldsymbol{S}^{-1}(\hat{\boldsymbol{P}}^*_{1,1} - \boldsymbol{\mu}\boldsymbol{m}^{*\intercal}_{[1]} - \boldsymbol{m}^*_{[1]}\boldsymbol{\mu}^\intercal + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal)\Big] \\ &+ \ln|\boldsymbol{S}| + \mathrm{tr}\Big[\boldsymbol{Q}^{-1}(\boldsymbol{\varphi} - \boldsymbol{\Psi}\boldsymbol{A}^\intercal - \boldsymbol{A}\boldsymbol{\Psi}^\intercal + \boldsymbol{A}\boldsymbol{\phi}\boldsymbol{A}^\intercal)\Big] \\ &+ \sum_{t=1}^{\tau_i}\mathrm{tr}\Big[\tilde{\boldsymbol{R}}_t^{-1}\Big(\tilde{\boldsymbol{C}}\hat{\tilde{\boldsymbol{P}}}^*_{t,t}\tilde{\boldsymbol{C}}^\intercal - 2\boldsymbol{\Gamma}_t\tilde{\boldsymbol{C}}^\intercal\Big)\Big]\Big\} + \text{const},\end{aligned} \tag{V.26}$$

which has derivatives

$$\frac{\partial}{\partial \boldsymbol{S}^{-1}}\hat{\mathscr{L}}_{JJ}(\boldsymbol{\theta}, q^*) = \frac{1}{2}\Big(\boldsymbol{S} + \boldsymbol{\mu}\boldsymbol{m}^{*\intercal}_{[1]} + \boldsymbol{m}^*_{[1]}\boldsymbol{\mu}^\intercal - \hat{\boldsymbol{P}}^*_{1,1} - \boldsymbol{\mu}\boldsymbol{\mu}^\intercal\Big), \tag{V.27}$$

$$\frac{\partial}{\partial \boldsymbol{\mu}}\hat{\mathscr{L}}_{JJ}(\boldsymbol{\theta}, q^*) = \boldsymbol{S}^{-1}(\boldsymbol{m}^*_{[1]} - \boldsymbol{\mu}), \tag{V.28}$$

$$\frac{\partial}{\partial \boldsymbol{A}}\hat{\mathscr{L}}_{JJ}(\boldsymbol{\theta}, q^*) = \boldsymbol{Q}^{-1}(\boldsymbol{\Psi} - \boldsymbol{A}\boldsymbol{\phi}), \tag{V.29}$$

$$\frac{\partial}{\partial \boldsymbol{Q}^{-1}}\hat{\mathscr{L}}_{JJ}(\boldsymbol{\theta}, q^*) = \frac{1}{2}\Big[\boldsymbol{\Psi}\boldsymbol{A}^\intercal + \boldsymbol{A}\boldsymbol{\Psi}^\intercal - \boldsymbol{A}\boldsymbol{\phi}\boldsymbol{A}^\intercal - \boldsymbol{\varphi} + (\tau-1)\boldsymbol{Q}\Big], \tag{V.30}$$

$$\frac{\partial}{\partial \tilde{\boldsymbol{C}}}\hat{\mathscr{L}}_{JJ}(\boldsymbol{\theta}, q^*) = \sum_{t=1}^{\tau_i}\tilde{\boldsymbol{R}}_t^{-1}(\boldsymbol{\Gamma}_t - \tilde{\boldsymbol{C}}\hat{\tilde{\boldsymbol{P}}}^*_{t,t}), \tag{V.31}$$

where $\tilde{\boldsymbol{C}} = [\boldsymbol{C}, \boldsymbol{u}]$,

$$\boldsymbol{\varphi} = \sum_{t=2}^{\tau} \hat{\boldsymbol{P}}^*_{t,t},\ \boldsymbol{\phi} = \sum_{t=2}^{\tau} \hat{\boldsymbol{P}}^*_{t-1,t-1},\ \boldsymbol{\Psi} = \sum_{t=2}^{\tau} \hat{\boldsymbol{P}}^*_{t,t-1},$$

and

$$\hat{\hat{\boldsymbol{P}}}^*_{t,t} = \begin{bmatrix} \hat{\boldsymbol{P}}^*_{t,t} & \boldsymbol{m}^*_{[t]} \\ \boldsymbol{m}^{*\,\intercal}_{[t]} & 1 \end{bmatrix},\ \boldsymbol{\Gamma}_t = \boldsymbol{\rho}_t \tilde{\boldsymbol{m}}^{*\,\intercal}_{[t]},\ \ \boldsymbol{\rho}_t = \frac{1}{4}\Big[\frac{2y_{1,t}-1}{\lambda(\xi^*_{1,t})}, \cdots, \frac{2y_{D,t}-1}{\lambda(\xi^*_{D,t})}\Big]^\intercal.$$

The VLADAD is then computed by 1) concatenating (V.21)-(V.25) (for ELBO 1), or (V.27)-(V.31) (for ELBO 2), and 2) aggregating over all attribute sequences extracted from a query video sequence. To improve discrimination, we apply a power-normalization and then $L2$-normalize the VLADAD feature vector, as suggested in [71].

V.C Probabilistic Kernels for Attribute Sequences

Many practical tasks of pattern analysis require a proper relationship characterization between the examples of interest. This is typically implemented with a kernel function that quantifies the similarity of two examples [145]. For sequential data of variable length, where direct comparison is difficult, a common practice is to devise kernel functions via generative models that can explain the data. In this light, we design a p-kernel [57] to encode similarity of two binary sequences via BDS. Note that, unlike the state of the arts [97, 99], we propose a provable positive-definite kernel that can be computed via an efficient *explicit* closed-form feature mapping to the reproducing kernel Hilbert space (RKHS).

Let $p(\theta)$ be a prior distribution for the model parameter θ. The p-kernel

$k(\boldsymbol{y}_1, \boldsymbol{y}_2)$ on the example space $\mathcal{Y}$ is defined via the probabilistic model $p(\boldsymbol{y}|\theta)$ as

$$k(\boldsymbol{y}_1, \boldsymbol{y}_2) = \int_\theta p(\boldsymbol{y}_1|\theta) p(\boldsymbol{y}_2|\theta) p(\theta) d\theta. \tag{V.32}$$

Intuitively, (V.32) evalutes the similarity of two examples by computing their correlation of likelihoods at multiple "probing" models, subject to the model prior distribution. To leverage information from a training set $\mathcal{T}_{\boldsymbol{y}} = \{\boldsymbol{y}^{(i)}\}_{i=1}^N$ for the optimal coverage of $\mathcal{Y}$, an empirical p-kernel is defined as

$$k(\boldsymbol{y}_1, \boldsymbol{y}_2; \mathcal{T}_{\boldsymbol{y}}) = \int_\theta p(\boldsymbol{y}_1|\theta) p(\boldsymbol{y}_2|\theta) p(\theta|\mathcal{T}_{\boldsymbol{y}}) d\theta \approx \frac{1}{N} \sum_i p(\boldsymbol{y}_1|\theta_i) p(\boldsymbol{y}_2|\theta_i), \tag{V.33}$$

where $p(\theta|\mathcal{T}_{\boldsymbol{y}})$ is approximated by $p(\theta|\mathcal{T}_{\boldsymbol{y}}) \approx \frac{1}{N} \sum_i \delta(\theta - \theta_i)$ with $\theta_i = \arg\max_\theta p(\boldsymbol{y}^{(i)}; \theta)$ as the surrogate model of example $\boldsymbol{y}^{(i)}$. To further facilitate the use of classifiers operating in the real-valued vector space with the kernel of (V.33), an *explicit* feature mapping can be constructed as

$$\mathcal{F} : \mathcal{Y} \to \mathbb{R}^N : \boldsymbol{y} \mapsto \frac{1}{\tau} \big[\ln p(\boldsymbol{y}|\theta_1), \cdots, \ln p(\boldsymbol{y}|\theta_N) \big]^\intercal, \tag{V.34}$$

where τ is the length of sequence $\boldsymbol{y}$. Using the explicit mapping of (V.34), a binary sequence is converted to a point in the N-dimensional real vector space with regular dot product as the kernel, where discriminative methods can be readily implemented for classification.

V.D Acknowledgement

The text of Chapter V is, in part, based on the material as it appears in the following publications: The bag-of-model encoding scheme with 0th

and 1st order statistics were originally proposed in W.-X. LI and N. Vasconcelos, "Complex Activity Recognition via Attribute Dynamics," to appear at *International Journal of Computer Vision* (IJCV). The probabilistic kernels for BDS was originally proposed in W.-X. LI, Y. Li and N. Vasconcelos, "Efficient Variational Inference, Learning and Probabilistic Kernels for Binary Dynamic Systems," under review at *Neural Information Processing Systems* (NIPS), 2016. The dissertation author was a primary researcher and an author of the cited material.

V.E Appendix

V.E.1 Convergence of Bag-of-Models Clustering

The bag-of-models clustering procedure of Algorithm 6 is a general framework for clustering examples in a Riemannian manifold $\mathcal{M}$ of statistical models. The goal is to find a preset number of models $\{M_j\}_{j=1}^{K} \subset \mathcal{M}$ in the manifold that best explain a corpora $\mathcal{D} = \{\boldsymbol{z}_i\}_{i=1}^{N}$ ($\boldsymbol{z}_i \in \mathcal{Z}, \forall i$). It is assumed that all models M are parametrized by a set of parameters $\boldsymbol{\theta}$ and have smooth likelihood functions (derivatives of all orders exist and are bounded), and that Algorithm 6 satisfies the following conditions.

Condition 1: the operation $f_{\mathcal{M}}$ of (V.1) consists of estimating the parameters $\boldsymbol{\theta}$ of $\mathcal{M}$ by the *maximum likelihood estimation* (MLE) principle.

Condition 2: the Riemannian metric of the manifold $\mathcal{M}$ defined by the Fisher information $\mathcal{I}_{\boldsymbol{\theta}_z}$ [64, 4] is used as the dissimilarity measure of (V.2). More precisely, the metric of $\mathcal{M}$ in the neighbrhood of model M_z is

$$d_{\mathcal{M}}(M^*,\, M_z) = ||\boldsymbol{\theta}^* - \boldsymbol{\theta}_z||^2_{\mathcal{I}_{\boldsymbol{\theta}_z}}, \tag{V.35}$$

where $||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||^2_{\mathcal{I}} = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\intercal \mathcal{I}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)$, and the Fisher information $\mathcal{I}_{\boldsymbol{\theta}_z}$ is defined as [5]

$$\mathcal{I}_{\boldsymbol{\theta}_z} = -\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x};\boldsymbol{\theta}_z)}\left[\nabla^2_{\boldsymbol{\theta}} \ln p(\boldsymbol{x};\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_z}\right]. \tag{V.36}$$

Given the similarity between Algorithm 6 and k-means, the convergence of the former can be studied with the techniques commonly used to show that the latter converges. This requires the definition of a suitable objective function to quantify the quality of the fit of the set $\{M_i\}_{j=1}^K$ to the corpora $\mathcal{D}$. We rely on the objective

$$\zeta(\{M_i\}_{j=1}^K, \{S_j\}_{j=1}^K) = \sum_j \sum_{\boldsymbol{z}\in S_j} \ln p_{M_j}(\boldsymbol{z}), \tag{V.37}$$

where $p_M(\cdot)$ is the likelihood function of model M, and S_j a subset of $\mathcal{D}$, containing all examples assigned to j-th model. Note that this implies that $\forall i \neq j, S_i \bigcap S_j = \varnothing$ and $\bigcup_j S_j = \mathcal{D}$. From the assumption of smooth models M (*i.e.*, $\forall \boldsymbol{z} \in \mathcal{Z}, M \in \mathcal{M}$, $p_M(\boldsymbol{z}) < \infty$) and the fact that there is only a finite set of assignments $\{S_j\}_{j=1}^K$, the objective function of (V.37) is upper bounded. Since the refinement step of Algorithm 6 updates the models so that

$$M_j^{(t+1)} = f_{\mathcal{M}}(S_j^{(t+1)}) = \underset{M\in\mathcal{M}}{\arg\max} \sum_{\boldsymbol{z}\in S_j^{(t+1)}} \ln p_M(\boldsymbol{z}),$$

the objective either increases or remains constant after each refinement step. It remains to prove that the same holds for each assignment step. If that is the case, Algorithm 6 produces a monotonically increasing and upper-bounded sequence of objective function values. By the monotone convergence theorem, this implies that algorithm converges in a finite number of steps. Note that, as in k-means, there is no guarantee on convergence to the global optimum.

It thus remains to prove that the objective of (V.37) increases with each assignment step. The Riemannian structure of the manifold $\mathcal{M}$, makes this proof more technical than the corresponding one for k-means. In what follows, we provide a sketch of the proof. Let M^* be the model (of parameters $\boldsymbol{\theta}^*$) to which example $\boldsymbol{z}$ is assigned by the assignment step of Algorithm 6, *i.e.*,

$$M^* = \underset{M \in \{M_j^{(t)}\}_{j=1}^K}{\arg\min} \; d_{\mathcal{M}}(M_{\boldsymbol{z}}, M) \tag{V.38}$$

and M° (of parameter $\boldsymbol{\theta}^\circ$) the equivalent model of the previous iteration. It follows from Condition 2 that

$$\begin{aligned} d_{\mathcal{M}}(M^*,\, M_{\boldsymbol{z}}) &= ||\boldsymbol{\theta}^* - \boldsymbol{\theta}_{\boldsymbol{z}}||^2_{\mathcal{I}_{\boldsymbol{\theta}_{\boldsymbol{z}}}} \\ &\leqslant d_{\mathcal{M}}(M^\circ,\, M_{\boldsymbol{z}}) = ||\boldsymbol{\theta}^\circ - \boldsymbol{\theta}_{\boldsymbol{z}}||^2_{\mathcal{I}_{\boldsymbol{\theta}_{\boldsymbol{z}}}}. \end{aligned} \tag{V.39}$$

Note that, $M_{\boldsymbol{z}}$ is the model $p(\boldsymbol{z};\boldsymbol{\theta}_{\boldsymbol{z}})$ onto which $\boldsymbol{z}$ is mapped by (V.1). From Condition 1, $\boldsymbol{\theta}_{\boldsymbol{z}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{z};\boldsymbol{\theta})$ and, using a Taylor series expansion,

$$\ln p(\boldsymbol{z};\boldsymbol{\theta}) \approx \ln p(\boldsymbol{z};\boldsymbol{\theta}_{\boldsymbol{z}}) + \langle \nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{z};\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{z}}}, \boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{z}} \rangle + \frac{1}{2} ||\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{z}}||^2_{H_{\boldsymbol{\theta}_{\boldsymbol{z}}}} \tag{V.40}$$

$$= \ln p(\boldsymbol{z};\boldsymbol{\theta}_{\boldsymbol{z}}) + \frac{1}{2} ||\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{z}}||^2_{H_{\boldsymbol{\theta}_{\boldsymbol{z}}}}, \tag{V.41}$$

where $H_{\boldsymbol{\theta}_{\boldsymbol{z}}} = \nabla^2_{\boldsymbol{\theta}} \ln p(\boldsymbol{z};\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{z}}}$ is the Hessian of $\ln p(\boldsymbol{z};\boldsymbol{\theta})$ at $\boldsymbol{\theta}_{\boldsymbol{z}}$. Since $p(\boldsymbol{z};\boldsymbol{\theta}_{\boldsymbol{z}})$ is the model obtained from a single example $\boldsymbol{z}$, it is a heavily peaky distribution centered at $\boldsymbol{z}$. Hence, the expectation of (V.36) can be approximated by

$$\mathcal{I}_{\boldsymbol{\theta}_{\boldsymbol{z}}} \approx -H_{\boldsymbol{\theta}_{\boldsymbol{z}}}. \tag{V.42}$$

Combining (V.39), (V.41), and (V.42) then results in

$$\begin{aligned}\ln p(\boldsymbol{z};\boldsymbol{\theta}^*) &\approx \ln p(\boldsymbol{z};\boldsymbol{\theta}_z) + \frac{1}{2}||\boldsymbol{\theta}^* - \boldsymbol{\theta}_z||^2_{H_{\boldsymbol{\theta}_z}} \\ &\approx \ln p(\boldsymbol{z};\boldsymbol{\theta}_z) - \frac{1}{2}||\boldsymbol{\theta}^* - \boldsymbol{\theta}_z||^2_{\mathcal{I}_{\boldsymbol{\theta}_z}} \\ &\geqslant \ln p(\boldsymbol{z};\boldsymbol{\theta}_z) - \frac{1}{2}||\boldsymbol{\theta}^\circ - \boldsymbol{\theta}_z||^2_{\mathcal{I}_{\boldsymbol{\theta}_z}} \approx \ln p(\boldsymbol{z};\boldsymbol{\theta}^\circ).\end{aligned}$$

It follows that the objective of (V.37) increases after each assignment step. This is intuitive in the sense that, the closer a model M is to an example's representative model, the better M can explain that example.

V.E.2 The Fisher Vector for BDS Using ELBO$_{SJ}$

In this section, we present the derivation of the Fisher vector for BDS using the tightest variational lower bound $\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta},q^*)$ of (V.26). This consists of computing partial derivatives of $\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta},q^*)$ w.r.t. each of the BDS parameters $\boldsymbol{\theta} = \{\boldsymbol{S}^{-1},\boldsymbol{\mu},\boldsymbol{A},\boldsymbol{Q}^{-1},\boldsymbol{C},\boldsymbol{u}\}$.

Derivative w.r.t. S^{-1}

We have

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{S}^{-1}}\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta},q^*) &= \frac{\partial}{\partial \boldsymbol{S}^{-1}}\,\frac{1}{2}\left\{\ln\left|\boldsymbol{S}^{-1}\right| - \operatorname{tr}\left[\left(\hat{\boldsymbol{P}}^*_{1,1} - 2\boldsymbol{m}^*_{[1]}\boldsymbol{\mu}^\intercal + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal\right)\boldsymbol{S}^{-1}\right]\right\} \\ &= \frac{1}{2}\left(\boldsymbol{S} + 2\boldsymbol{\mu}\boldsymbol{m}^{*T}_{[1]} - \hat{\boldsymbol{P}}^*_{1,1} - \boldsymbol{\mu}\boldsymbol{\mu}^\intercal\right), \end{aligned} \tag{V.43}$$

where $\hat{\boldsymbol{P}}^*_{r,s} = \boldsymbol{\Phi}^*_{[r,s]} + \boldsymbol{m}^*_{[r]}\boldsymbol{m}^{*\intercal}_{[s]}$. Note that, $\boldsymbol{S}^{-1} \in \mathcal{S}^L_{++}$, thus the derivative of (V.43) needs to be projected into the space of symmetric matrices $\mathcal{S}^L$. Since an orthonormal basis of $\mathcal{S}^L$ is $\{\frac{1}{2}(E_{i,j} + E_{j,i}), 1 \leqslant i \leqslant j \leqslant L\}$, where $E_{i,j} \in \mathbb{R}^{L\times L}$ with the

(i, j)-element equal to one and all the rest elements being zero, it can be shown that after the projection, (V.43) becomes

$$\frac{\partial}{\partial \boldsymbol{S}^{-1}} \hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q^*) = \frac{1}{2}\Big(\boldsymbol{S} + \boldsymbol{\mu} \boldsymbol{m}_{[1]}^{*T} + \boldsymbol{m}_{[1]}^{*} \boldsymbol{\mu}^{\intercal} - \hat{\boldsymbol{P}}_{1,1}^{*} - \boldsymbol{\mu}\boldsymbol{\mu}^{\intercal}\Big). \tag{V.44}$$

Derivative w.r.t. μ

We have

$$\frac{\partial}{\partial \boldsymbol{\mu}} \hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q^*) = \frac{\partial}{\partial \boldsymbol{\mu}}\Big[\boldsymbol{\mu}^{\intercal} \boldsymbol{S}^{-1} \boldsymbol{m}_{[1]}^{*} - \frac{1}{2} \boldsymbol{\mu}^{\intercal} \boldsymbol{S}^{-1} \boldsymbol{\mu}\Big] = \boldsymbol{S}^{-1}(\boldsymbol{m}_{[1]}^{*} - \boldsymbol{\mu}). \tag{V.45}$$

Derivative w.r.t. A

We have

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{A}} \hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta}, q^*) &= \frac{\partial}{\partial \boldsymbol{A}}\Bigg[\sum_{t=1}^{\tau-1} \mathrm{tr}\Bigg(\hat{\boldsymbol{P}}_{t,t+1}^{*} \boldsymbol{Q}^{-1} \boldsymbol{A} - \frac{1}{2} \hat{\boldsymbol{P}}_{t,t}^{*} \boldsymbol{A}^{\intercal} \boldsymbol{Q}^{-1} \boldsymbol{A}\Bigg)\Bigg] \\
&= \frac{\partial}{\partial \boldsymbol{A}}\Bigg[\mathrm{tr}\Bigg(\boldsymbol{\Psi}^{\intercal} \boldsymbol{Q}^{-1} \boldsymbol{A} - \frac{1}{2} \boldsymbol{\phi} \boldsymbol{A}^{\intercal} \boldsymbol{Q}^{-1} \boldsymbol{A}\Bigg)\Bigg] \\
&= (\boldsymbol{\Psi}^{\intercal} \boldsymbol{Q}^{-1})^{\intercal} - \frac{1}{2}\Big[\boldsymbol{Q}^{-T} \boldsymbol{A} \boldsymbol{\phi}^{\intercal} + \boldsymbol{Q}^{-1} \boldsymbol{A} \boldsymbol{\phi}\Big] \\
&= \boldsymbol{Q}^{-1}(\boldsymbol{\Psi} - \boldsymbol{A}\boldsymbol{\phi}),
\end{aligned} \tag{V.46}$$

where

$$\boldsymbol{\phi} = \sum_{t=2}^{\tau} \hat{\boldsymbol{P}}_{t-1,t-1}^{*}, \ \boldsymbol{\Psi} = \sum_{t=2}^{\tau} \hat{\boldsymbol{P}}_{t,t-1}^{*}.$$

Derivative w.r.t. Q^{-1}

We have

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{Q}^{-1}}\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta},q^*) &= \frac{\partial}{\partial \boldsymbol{Q}^{-1}}\Bigg[\sum_{t=1}^{\tau-1}\mathrm{tr}\Big(\boldsymbol{A}\hat{\boldsymbol{P}}^*_{t,t+1}\boldsymbol{Q}^{-1} - \frac{1}{2}\boldsymbol{A}\hat{\boldsymbol{P}}^*_{t,t}\boldsymbol{A}^\intercal\boldsymbol{Q}^{-1} \\ &\qquad - \frac{1}{2}\hat{\boldsymbol{P}}^*_{t+1,t+1}\boldsymbol{Q}^{-1}\Big) + (\frac{\tau-1}{2})\ln\Big|\boldsymbol{Q}^{-1}\Big|\Bigg] \\ &= \frac{\partial}{\partial \boldsymbol{Q}^{-1}}\Big[\mathrm{tr}\Big(\boldsymbol{A}\boldsymbol{\Psi}^\intercal\boldsymbol{Q}^{-1} - \frac{1}{2}\boldsymbol{A}\boldsymbol{\phi}\boldsymbol{A}^\intercal\boldsymbol{Q}^{-1} - \frac{1}{2}\boldsymbol{\varphi}\boldsymbol{Q}^{-1}\Big) \\ &\qquad + (\frac{\tau-1}{2})\ln\Big|\boldsymbol{Q}^{-1}\Big|\Big] \\ &= \boldsymbol{\Psi}\boldsymbol{A}^\intercal + \frac{1}{2}\Big[(\tau-1)\boldsymbol{Q} - \boldsymbol{A}\boldsymbol{\phi}\boldsymbol{A}^\intercal - \boldsymbol{\varphi}\Big],\end{aligned} \tag{V.47}$$

where

$$\boldsymbol{\varphi} = \sum_{t=2}^{\tau}\hat{\boldsymbol{P}}^*_{t,t}. \tag{V.48}$$

Again, since $\boldsymbol{Q}^{-1} \in \mathcal{S}_{++}$, the partial derivative of (V.47) is projected into $\mathcal{S}$, giving

$$\frac{\partial}{\partial \boldsymbol{Q}^{-1}}\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta},q^*) = \frac{1}{2}\Big[\boldsymbol{\Psi}\boldsymbol{A}^\intercal + \boldsymbol{A}\boldsymbol{\Psi}^\intercal - \boldsymbol{A}\boldsymbol{\phi}\boldsymbol{A}^\intercal - \boldsymbol{\varphi} + (\tau-1)\boldsymbol{Q}\Big]. \tag{V.49}$$

Derivative w.r.t. $\tilde{C}$

We have

$$\begin{aligned}\frac{\partial}{\partial \tilde{\boldsymbol{C}}}\hat{\mathscr{L}}_{SJ}(\boldsymbol{\theta},q^*) &= \frac{\partial}{\partial \tilde{\boldsymbol{C}}}\Bigg\{\sum_{k,t}\Big[y_{kt}\ln\sigma(\tilde{\boldsymbol{C}}_{k,:}\tilde{\boldsymbol{m}}^*_{[t]}) + (1-y_{kt})\ln\sigma(-\tilde{\boldsymbol{C}}_{k,:}\tilde{\boldsymbol{m}}^*_{[t]})\Big] - \frac{1}{8}\mathrm{tr}\big(\tilde{\boldsymbol{C}}\tilde{\Upsilon}\tilde{\boldsymbol{C}}^\intercal\big)\Bigg\} \\ &= -\frac{1}{4}\Bigg\{\tilde{\boldsymbol{C}}\tilde{\Upsilon} + \sum_{t=1}^{\tau}\begin{bmatrix}\sigma(\tilde{\boldsymbol{C}}_{1,:}\tilde{\boldsymbol{m}}^*_{[t]}) - y_{1t} \\ \vdots \\ \sigma(\tilde{\boldsymbol{C}}_{D,:}\tilde{\boldsymbol{m}}^*_{[t]}) - y_{Dt}\end{bmatrix}\tilde{\boldsymbol{m}}^{*\,\intercal}_{[t]}\Bigg\},\end{aligned} \tag{V.50}$$

where

$$\tilde{\Upsilon} = \begin{pmatrix} \sum_{t=1}^{\tau} \mathbf{\Phi}^*_{[t,t]} & 0 \\ 0 & 0 \end{pmatrix}.$$

V.E.3 The Fisher Vector for BDS Using ELBO$_{JJ}$

In this section, we present the derivation of the Fisher vector for BDS using the tightest ELBO $\tilde{\mathscr{L}}_{JJ}(q^*,\boldsymbol{\xi}^*;\boldsymbol{\theta})$ of (III.49). This consists of computing partial derivatives of $\tilde{\mathscr{L}}_{JJ}(q^*,\boldsymbol{\xi}^*;\boldsymbol{\theta})$ w.r.t. each of the BDS parameters $\boldsymbol{\theta} = \{\boldsymbol{S}^{-1}, \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{Q}^{-1}, \boldsymbol{C}, \boldsymbol{u}\}$.

The derivations of (V.27)-(V.30) are the same as in Section V.E.2. Here we derive the result of (V.31). The first-order derivative of (V.26) w.r.t. $\tilde{\boldsymbol{C}}$ is

$$\frac{\partial}{\partial \tilde{\boldsymbol{C}}}\Big\{ -\frac{1}{2}\sum_{t=1}^{\tau_i} \mathrm{tr}\Big[\tilde{\boldsymbol{R}}_t^{-1}(\tilde{\boldsymbol{C}}\hat{\tilde{\boldsymbol{P}}}^*_{t,t}\tilde{\boldsymbol{C}}^\mathsf{T} - 2\boldsymbol{\Gamma}_t\tilde{\boldsymbol{C}}^\mathsf{T})\Big]\Big\} = \sum_{t=1}^{\tau_i} \tilde{\boldsymbol{R}}_t^{-1}(\boldsymbol{\Gamma}_t - \tilde{\boldsymbol{C}}\hat{\tilde{\boldsymbol{P}}}^*_{t,t}), \tag{V.51}$$

where

$$\hat{\tilde{\boldsymbol{P}}}^*_{t,t} = \begin{bmatrix} \hat{\boldsymbol{P}}^*_{t,t} & \boldsymbol{m}^*_{[t]} \\ \boldsymbol{m}^{*\,\mathsf{T}}_{[t]} & 1 \end{bmatrix}, \;\; \boldsymbol{\Gamma}_t = \boldsymbol{\rho}_t \tilde{\boldsymbol{m}}^{*\,\mathsf{T}}_{[t]}, \;\; \boldsymbol{\rho}_t = \frac{1}{4}\Big[\frac{2y_{1,t}-1}{\lambda(\xi^*_{1,t})}, \cdots, \frac{2y_{D,t}-1}{\lambda(\xi^*_{D,t})}\Big]^\mathsf{T}.$$

Chapter VI

Application: Complex Human Activity Recognition

VI.A Introduction

Understanding human behavior is an important goal for computer vision [3]. While early solutions mostly addressed the recognition of simple movements in controlled environments [17, 14, 142, 51], recent interest has been in more challenging and realistic tasks [93, 130, 111, 85]. In the literature, these tasks are commonly referred to as "action" or "activity" recognition. In this work, we adopt the term "action" to denote behavior at the lowest level of the semantic hierarchy, *e.g.*, "run," "jump," or "kick a ball.' The term "activity" is reserved for behavior of higher level semantics, which can usually be described as a sequence of actions. For example, the Olympic activity "clean and jerk" involves the actions of "grasping a barbell," "raising weights over the athlete's head," and "dropping the bar.' Activities can also be performed by multiple subjects (*i.e.*, be "collective"), or composed of "events" rather than actions (*e.g.*, "wedding ceremony" composed of events such as "walking the bride," "exchange of vows," "opening dance," *etc.*).

Several of the prior works in action and activity recognition have proposed variants of the *bag of visual words* (BoVW), which represents video as a collection of orderless spatiotemporal features and serves as the low-level foundation for many other action analysis frameworks. This family of representations have been shown to consistently achieve state-of-the-art performance for tasks such as action recognition and retrieval [166, 154, 165, 118, 110, 91].

Nevertheless, the BoVW has at least two important limitations. First, it does not account for the fact that most activities are best abstracted as sequences of actions or events. This is illustrated by the activity "packing a box" of Figure VI.1, which most humans would characterize as a sequence of the actions

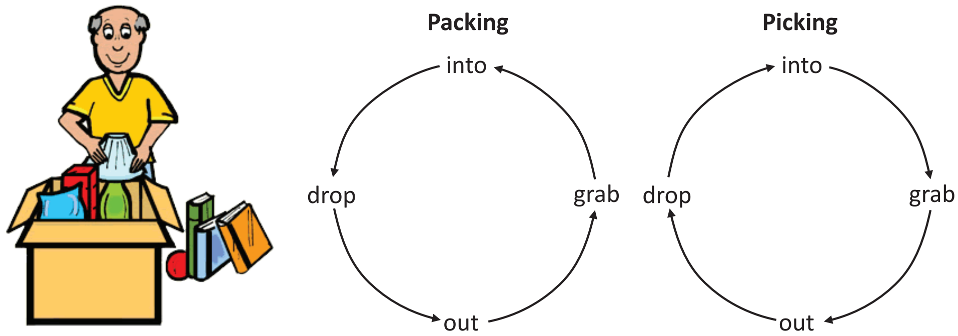


Figure VI.1: The packing example. The actions "move hand into box" (into), "grab object" (grab), "move hand out of box" (out), and "drop object" (drop) are consisting with the activities of "packing a box" and "picking objects from a box.' In the absence of temporal modeling of event semantics, these activities can be quite difficult to distinguish.

"move hand out of box - grab object - move hand into box - drop object." In the absence of an explicit representation of these semantics, it is up to the classifier to learn the importance of concepts such as moving hands, grabbing or dropping objects for the characterization of this activity. While these concepts are not impossible to learn from the evolution of low-level features, this is easier when the classifier is given explicit supervision about the semantics of interest. In result, semantic video modeling has recently began to receive substantial attention. For example, the TRECVID multimedia event *detection* and *recounting* contest [114], one of the major large-scale video analysis research efforts, explicitly states the goal of not only predicting the event category ("detection") of a video sequence, but also identifying its *semantically meaningful and relevant pieces* ("recounting").

Second, the BoVW captures little information about the temporal structure of video. This limits its expressiveness, since a single set of actions (or events) can give rise to multiple activities, depending on the *order* with which the actions are performed. This is again illustrated in Figure VI.1, where the activity of "picking objects from a box" differs from the activity of "packing a box" only in terms of the order of the actions described above, which is now "move hand into box - grab object - move hand out of box - drop object.' Hence, sophisticated modeling of temporal structure can be critical for parsing complex activities. This is beyond the reach of the BoVW.

Recently, there have been various attempts to address the two limitations of the BoVW. On one hand, several authors have proposed richer models of the temporal structure, also known as *dynamics*, of human activity [111, 94, 28, 46]. However, because modeling activity dynamics can be a complex proposition, it is not uncommon for these models to require features specific to certain data sets or activity classes [94, 28], or non-trivial forms of pre-processing, such as tracking [95], per-class manual annotation [46], *etc*. On the other hand, inspired by recent developments in image classification [89, 126], there has been a move towards the representation of action in terms of intermediate-level semantic concepts, such as *attributes* [101, 43]. This introduces a layer of abstraction that improves generalization, enables modeling of contextual relationships [125], and simplifies knowledge transfer across activity classes [101]. However, these models continue to disregard the temporal structure of video.

In this thesis, we exploit the distinct characteristics of complex activities at different temporal granularities, and propose a unified hierarchy for representing these variabilities of human behavior, by combining all these properties via *modeling and encoding the dynamics of human activities in the space of attributes*. The idea is to define each activity as a sequence of *semantic* events, *e.g.*, defining "packing a box" as the *sequence* of the action attributes "remove (hand from box)," "grab (object)," "insert (hand in box)," and "drop (object).' This semantic-level representation is *more robust* to confounding factors, such as diversity of grabbing styles, hand motion speeds, or camera motion, than dynamic representations based on low-level features. It is also *more discriminant* than semantic representations that ignore dynamics, *i.e.*, that simply record the occurrence (or frequency) of the action attributes "remove," "grab," "insert," and "drop.' We already saw that, in the absence of information about the *sequence* in which these attributes occur, the "packing a box" activity cannot be distinguished from the "picking from a box" activity.

To implement this idea, we present novel solutions to the two major technical challenges of using attribute dynamics for activity recognition. The first is the modeling of attribute dynamics itself. As usual in semantics-based recognition [101], video is represented in a semantic feature space, where each feature encodes the probability of occurrence of an action attribute at each time step. We introduce a generative model, the *binary dynamic system* (BDS), to learn *both* the distribution and dynamics of different activities in this space. The BDS is a non-linear dynamic system that combines binary observations with a hidden Gauss-Markov state process. It can be interpreted as either 1) a generalization of *binary principal component analysis* (binary PCA) [139], which accounts for data dynamics; or 2) an extension of the classical *linear dynamic system* (LDS) to a binary observation space.

The second is to account for non-stationary video dynamics. For this, we embed the BDS in the BoVW representation, modeling video sequences as orderless combinations of *short-term video segments of characteristic semantic dynamics*. More precisely, videos are modeled as sequences of short-term segments sampled from a family of BDSs. This representation, the *bag of words for attribute dynamics* (BoWAD), is applicable to more complex activities, *e.g.*, "moving objects across two boxes" which combines the event sequences of "picking objects from a box" and "packing a box," with potentially other events (*e.g.*, "inspecting object") in between. The BoWAD is shown to cope with the semantic noise, content irregularities, and intra-class variation that prevail in video of complex high-level events. These are further complemented by the discriminating feature representation for activity classification, denoted *vector of locally aggregated descriptors for attribute dynamics* (VLADAD), inspired by the recent success of Fisher vectors in image classification [119, 84, 30, 147], which is based on the aggregation of the derivatives of a variational lower-bound of the log-likelihood over attribute sequences.

VI.B Related Work

Many approaches to action recognition have been proposed in the last decades [3, 162]. Early methods aimed to detect a small number of short-term atomic movements in distractor-free environments. These methods relied extensively on operations such as tracking [113, 19, 105], or filtering [17, 121, 174, 29], that do not generalize well to more complex environments.

Over the last decade, there has been an increased focus on effective and scalable automatic analysis of video involving complicated motion, distractor-ridden scenes, complex backgrounds, unconstrained camera motion, *etc*. Various representations have been proposed to address these challenges, including BoVW [142, 92], spatio-temporal pyramid matching [93, 90], decomposable segments [111, 47], trajectories [103, 74, 164, 165], attributes [101], fusion with depth-maps [176], holistic volume encoding [51, 130, 144], neural networks [73, 148, 109, 167], and so forth. In this context, the BoVW and its variants have consistently achieved state-of-the-art performance for tasks like action recognition and retrieval, specially when combined with informative descriptors

[92, 166, 83, 165] and advanced encoding schemes [93, 154, 117, 143]. In fact, even sophisticated deep learning models, which capture hierarchical structure and have obliterated the performance of the state of the art in areas such as image and speech analysis [36, 132, 153], have failed to match the most recent BoVW schemes based on hand-crafted features [117, 118, 110, 91], in the context of action recognition from video [148, 109, 167][1].

The main justification for the robustness of the BoVW, *i.e.*, that it reduces video to an orderless collection of spatiotemporal descriptors, also limits the applicability of this representation to fine-grained activity discrimination, where it is important to account for precise temporal structure. A number of approaches have been proposed to characterize this structure. One possibility is to represent activities in terms of limb or torso motion, spatiotemporal shape models, or motion templates [51, 63]. Since they require detection, segmentation, tracking, or 3D structure recovery of body parts, these representations can be fragile.

A more robust alternative is to model the temporal structure of the BoVW. This can be achieved with generalizations of popular still image recognition methods. For example, Laptev *et al.* extend pyramid matching to video, using a 3D binning scheme that roughly characterizes the spatio-temporal structure of video [93]. Niebles *et al.* employ a latent support vector machine (SVM) that augments the BoVW with temporal context, which they show to be critical for understanding realistic motion [111]. These approaches have relatively coarse modeling of dynamics. More elaborate models are usually based on generative representations. For example, Laxton *et al.* model a combination of object

[1] There is an ongoing debate on how deep architectures can capture long-term low-level motion information. While early models failed to achieve competitive performance [73, 81], recent works [148, 109, 167] show promising results, albeit still inferior to those of the best hand-crafted features [117, 118, 110, 91]. It is worth noting that this issue is orthogonal to the contributions of this work, since the proposed method is built on a space of attribute responses which could be computed with a convolutional neural network (CNN).

contexts and motion sequences with a dynamic Bayesian network [94], while Gaidon *et al.* reduce each activity to three atomic actions and model their temporal distributions [46]. These methods rely on activity-class specific features and require detailed manual supervision. Alternatively, several researchers have proposed to model BoVW dynamics with LDSs. For example, Kellokumpu *et al.* combine dynamic textures [39] and local binary patterns [82], Li *et al.* perform a discriminant canonical correlation analysis on the space of activity dynamics [95], and [28] map frame-wise motion histograms to a reproducing kernel Hilbert space, where they learn a kernel dynamic system (KDS).

Due to their success in areas like handwriting [53] and speech recognition [52], *recurrent neural networks* (RNN) have recently started to receive substantial attention for action recognition. In this context, they are usually learned from features extracted with a low-level visual representation (BoVW, CNN, *etc*). For example, Baccouche *et al.* use an RNN to learn temporal dynamics of either hand-drafted [6], or CNN [7] features. More recently, Donahue *et al.* combine a CNN and the *long short-term memory* (LSTM) model of [60] to optimize both the low-level visual activation and dynamic components of an action recognition system [38]. Alternatively, Ng *et al.* study temporal aggregation strategies for video classification by either pooling over time or using LSTMs over frame-wise CNN activations [109]. So far, RNN-based methods for action recognition have failed to outperform even approaches without temporal order modeling, *e.g.*, the convolutional pooling of [109] or the two stream method of [148]. A major obstacle to these approaches is temporal scalability. Since the temporal depth of a RNN is linear in the number of input frames, most methods operate on a small number of video frames, *e.g.*, 9 frames in [6], a few seconds in [7], 16 and 30 frames for [38] and [109], respectively. This limits discrimination for complex, longer-term, activities. Finally, current RNNs model the entire content of a video sequence. This is problematic when the video contains sub-regions that do not depict the specific activity of interest, a common occurrence for open-source videos of complex activities.

Recent research in image recognition has shown that various limitations of the BoVW are overcome by representations of higher semantic level [126]. The features that underly these representations are confidence scores for the appearance of pre-defined visual concepts in images. These can be object attributes [89], object classes [124, 122, 70], contextual classes [125], or generic visual concepts [123]. Lately, semantic attributes have been used for action recognition [101, 72], demonstrating the benefits of mid-level semantic representations for the analysis of complex human activities. However, all these representations ignore the temporal structure of video, representing actions as orderless feature collections and reducing an *entire* video sequence to an attribute vector. For this reason, we denote them *holistic attribute* representations.

The evolution of semantic concepts has not been thoroughly exploited as a clue for activity understanding, although there have been a few efforts in this direction since our early work of [97]. For example, hidden Markov models (HMM) have been employed to capture the temporal structure of the projection of a video sequence into a space of clusters of visual features [155] or a space of supervised attribute detectors [151]. [11] have instead proposed to represent complex activities by the spectrum (or some other harmonic signature) of a model of attribute dynamics derived from the control literature. Finally, [152] extract discriminative segments from the video and characterize them by temporal transitions of attribute scores.

Figure VI.2: Evolution of activity in the attribute space. Left: key frames of activities "hurdle race" (top) and "long jump" (bottom); Right: attribute transition probabilities of the two activities ("hurdle race" / "long jump") for attributes "run," "jump," and "land.'

VI.C Activity Representation via Attribute Dynamics

In this section, we discuss the representation of activities with attribute dynamics.

VI.C.1 Action Attributes

Attribute representations are members of the class of semantic representations [123, 101] for image and video. These are representations defined on feature spaces with explicit semantics, *i.e.*, where features are visual concepts, scene classes, *etc*. Images or video are mapped into these spaces by classifiers trained to detect the semantics of interest. For attribute representations, these are binary detectors of video attributes $\{c_k\}_{k=1}^{D}$ that map a video $\boldsymbol{x} \in \mathcal{X}$ into a binary vector

$$\boldsymbol{y} = [y_1, \cdots, y_D]^\intercal \in \{0,1\}^D, \tag{VI.1}$$

indicating the presence/absence of each attribute in $\boldsymbol{x}$. Classifier output y_k is a Bernoulli random variable, whose probability parameter $\pi_k(\boldsymbol{x})$ is a confidence score for the presence of attribute c_k in $\boldsymbol{x}$. This is usually an estimate of the *posterior probability* of attribute c given video $\boldsymbol{x}$, *i.e.*, $\pi_c(\boldsymbol{x}) = p(c|\boldsymbol{x})$. The *semantic space* $\mathcal{S}$ is the space of such scores, defined by

$$\boldsymbol{\pi} : \mathcal{X} \to \mathcal{S} = [0,1]^K,\ \boldsymbol{\pi}(\boldsymbol{x}) = (\pi_1(\boldsymbol{x}), \cdots, \pi_K(\boldsymbol{x}))^\intercal. \tag{VI.2}$$

The benefits of attribute representations for recognition, namely a higher level of abstraction (which enables better generalization than appearance-based representations), robustness to classification errors, and ability to account for contextual relationships between concepts, have been previously documented in [89, 125, 115, 101, 72].

VI.C.2 Temporal Structure in Attribute Space

Since existing attribute representations do not account for temporal structure, they have limited applicability to video analysis. Temporal structure cannot be captured by representations that are either holistic, such as (VI.2), or reduce video to an orderless collection of instantaneous descriptors, such as histograms. We propose to overcome this problem by introducing models of the *dynamics, i.e.*, temporal evolution, of video attributes. This relies on the mapping of each video into a sequence of semantic vectors

$$\mathbf{\Pi} = \{\boldsymbol{\pi}_t(\boldsymbol{x})\} \subset \mathcal{S}, \tag{VI.3}$$

where $\pi_{tk}(\boldsymbol{x})$ is the confidence score for presence, in $\boldsymbol{x}$, of attribute k at time t. These scores are obtained by application of attribute detectors to a sliding video window. Fig. VI.2 motivates the modeling of attribute dynamics, by depicting two activity categories ("long jump" and "hurdle race") that instantiate the same attributes with roughly equal probabilities, but span two very different trajectories in $\mathcal{S}$. While hurdle racing involves a rhythmic transition between short patterns of racing, jumping, and landing, a long jump starts with a longer running sequence, followed by a single jump, and ends with a landing.

It is important to distinguish short- and long-term dynamics. The charac-

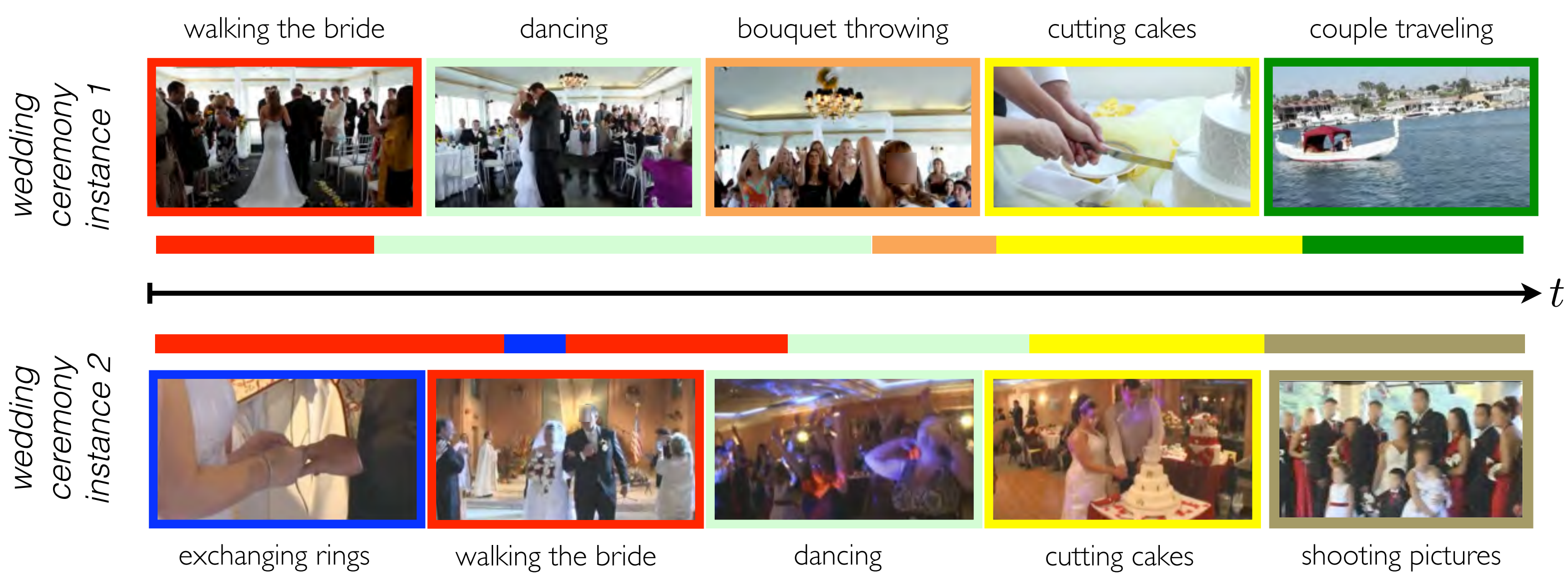


Figure VI.3: Composition of complex video events. Video sequences of complex activities, such as "wedding ceremony," are composed by several actions, *e.g.*, "walking the bride," or "cutting cake"). These actions and/or the corresponding durations (indicated by color boxes/bars in the figure) can differ significantly across sequences and are not always informative (*e.g.*, "couple traveling") of the activity class.

terization of short-term dynamics can substantially enhance the expressiveness of a video model. For example, decomposing the activity “long-jump” into the short term events “run-run,” “run-jump” and “jump-land,” is sufficient to discriminate it from the activity “triple-jump,” which is composed of short-term events “run-jump,” “jump-jump” and “jump-land.′ The presence (or absence) of the “jump-jump” segment is the essential difference between the two activities, which are otherwise very similar. In this work, we capture these short-term dynamics with a dynamic Bayesian network, the *binary dynamic system* (BDS), which extends classical linear dynamical systems [131] to semantic observations.

Long-term temporal structure, on the other hand, can be less predictable, since attributes of complex activities are highly non-stationary. There are at least three major sources of non-stationarity. First, complex activities are frequently composed of atomic actions with different dynamics. For example, the “wedding ceremony” sequences of Fig. VI.3 are composed of several events (*e.g.*, “dancing,” “cutting the cake,” or “bouquet throwing”). Since the dynamics of these events can be quite distinct, it is very challenging to capture the long-term dynamics of the activity with a single model. Second, and more importantly, the training data available is usually too sparse to cover the intra-class variations of high-level activities. For example, while some wedding videos involve scenes of an honeymoon trip, most do not. In this case, attempting to model long-range dynamics is prone to overfitting. Finally, the most discriminant video segments for event recognition are frequently embedded in video that is only marginally informative of the activity class. For example, the discriminant (for weddings) “bouquet toss” sequence can be surrounded by “dancing” sequences (which appear equally in wedding and birthday videos). The ability to identify these discriminant segments, while ignoring the surrounding “action noise” (non-

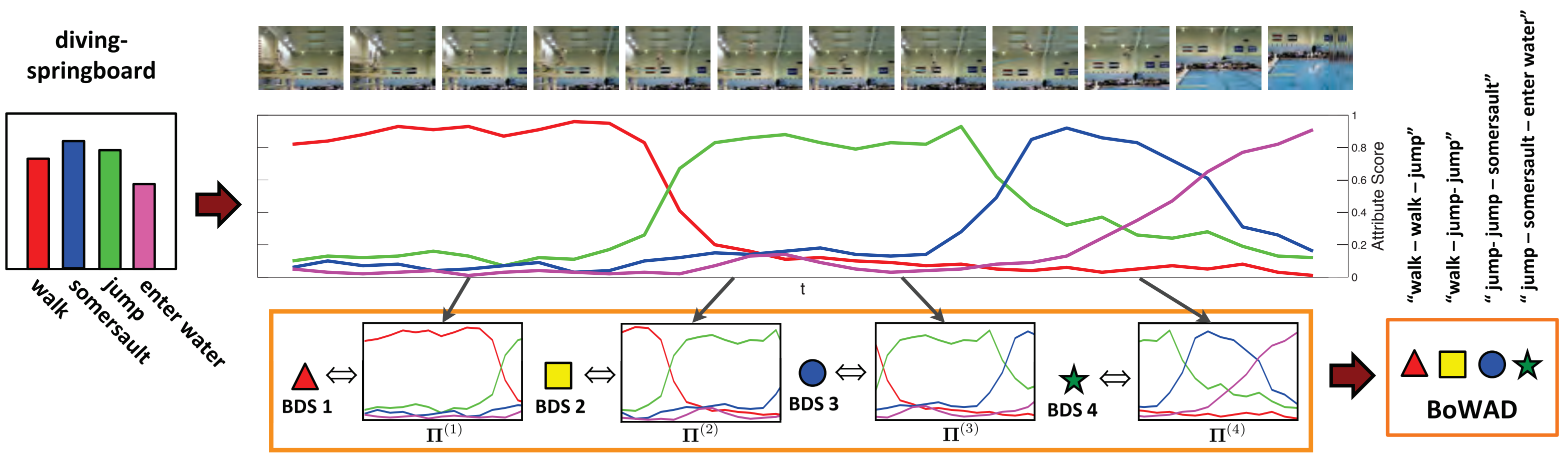


Figure VI.4: Illustration for bag of words for attribute dynamics. BoWAD representation of a video of the activity "diving-springboard" is exemplified. (Top) video sequence. (Middle) The classic (holistic) representation of the video on a space of four attributes (represented by four colors) is shown in the left. The proposed representation of the video as a trajectory in the attribute space (four colored functions) is shown at the center. The trajectory is split into overlapping sort-term segments. (Bottom) each segment is assigned to the BDS, in a previously learned dictionary, that best explains it. Dictionary BDS's, denoted WADs, are models of short-term behavior, such as "walk-walk-jump," "walk-jump-jump," "jump-jump-somersault" and "jump-somersault-enter water.' The activity is represented by a BoWAD, which is a histogram of assignments of segments to WADs.

informative segments) are critical for robust event recognition.

These observations suggest that the modeling of dynamics involves a trade-off between gains in discrimination *v.s.* potential for overfitting. Modeling short-term dynamics increases discrimination with small overfitting potential. However, the latter increases with the temporal support of the video sequences. In result, there is an optimal support, beyond which the benefits of dynamic models start to vanish. This suggests the combination of dynamic models, such as the BDS, for short-term dynamics and representations that may be less discriminant but more robust, such as the BoVW, for long-term dynamics. To accomplish this goal, we propose to encode activity sequences with a BoVW representation that uses the BDS as descriptor of short-term attribute dynamics.

The proposed video representation is illustrated in Fig. VI.4. A video $\boldsymbol{x}$ is split into segments $\{\boldsymbol{s}^{(i)}\}_{i=1}^{N}$ of τ_i frames (possibly overlapping in time)[2]. The attribute mapping of (VI.3) is then applied to each segment, producing an attribute sequence $\boldsymbol{\Pi}^{(i)} = \{\boldsymbol{\pi}_t\}_{t=t_i}^{t_i+\tau_i-1}$, where t_i is the starting time of the i-th segment. $\boldsymbol{x}$ is finally represented by the *bag of attribute sequences* (BoAS) $\{\boldsymbol{\Pi}^{(i)}\}$ shown in the orange box. This generalizes the BoVW image representation. A dictionary of representative BDSs, denoted *words for attributes dynamics* (WAD), is learned by clustering a collection of BoAS from a set of training attribute sequences. The WAD dictionary is then used to encode the attribute sequences extracted from $\boldsymbol{x}$ as a feature vector for final video classification. This is implemented by either 1) the histogram of WAD counts, denoted a *bag of words for attribute dynamics* (BoWAD), or 2) a descriptor of the first order statistics of attribute sequences after clustering with a WAD mixture, denoted the *vector of local aggregated descriptors*

[2]The optimization of the lengths $\{\tau_i\}$ of the video segments $\{\boldsymbol{s}^{(i)}\}$ is left for further research. In this work, we simply considered segments of equal length $\{\tau_i\} = \tau, \forall i$, chosen from a finite set of segment lengths τ, selected so as to achieved good empirical performance on the datasets considered. The specific values of τ used are discussed in the experimental section.

for attribute dynamics (VLADAD).

VI.D Models of Attribute Dynamics

In this section, we address the modeling of the dynamics of attribute sequences. We start by considering binary attributes and then generalize the discussion to account for confidence scores.

VI.D.1 Soft Binary PCA

By mapping each video into a sequence of vectors $\{\boldsymbol{\pi}_t\}$ of attribute probabilities, the semantic representation of (VI.3) is much richer than a sequence of binary attribute vectors $\boldsymbol{y}_t$. This, however, prevents the direct application of binary PCA. A solution is nevertheless possible if, instead of the conventional ML criterion, we resort to the maximization of the *expected* log-likelihood of the binary observations $\boldsymbol{y}_t$. This equates parameter learning to the optimization problem

$$\begin{aligned}\boldsymbol{\theta}^* &= \arg\max_{\boldsymbol{\theta}} \; \langle \ln \mathcal{L}(\boldsymbol{\theta}) \rangle_{p(\boldsymbol{y};\boldsymbol{\pi})} && \text{(VI.4)}\\ &= \arg\max_{\boldsymbol{\theta}} \; \langle \ln p(y;\boldsymbol{\theta}) \rangle_{p(\boldsymbol{y};\boldsymbol{\pi})}. && \text{(VI.5)}\end{aligned}$$

Since $\langle \boldsymbol{y}_t \rangle_{p(\boldsymbol{y};\boldsymbol{\pi})} = \boldsymbol{\pi}_t$, it follows from (IV.2) that

$$\langle \mathcal{L} \rangle_{p(\boldsymbol{y};\boldsymbol{\pi})} = \sum_{k,t} \Big[\pi_{kt} \ln \sigma(\boldsymbol{\Theta}_{kt}) + (1-\pi_{kt}) \ln \sigma(-\boldsymbol{\Theta}_{kt}) \Big], \qquad \text{(VI.6)}$$

and (VI.5) can be solved with the binary PCA algorithm.

It should be noted that this solution is identical to the ML estimate of binary PCA in the case of infinite data since, by the law of large numbers,

$$\frac{1}{N}\sum_{i=1}^{N}\ln p(y^{(i)};\boldsymbol{\theta}) \xrightarrow[N\to\infty]{} \langle \ln p(y;\boldsymbol{\theta})\rangle_{p(\boldsymbol{y};\boldsymbol{\pi})},$$

where $\{y^{(i)}\}_{i=1}^{N}$ are N *independent and identically distributed* (*i.i.d.*) examples from from $p(\boldsymbol{y};\boldsymbol{\pi})$. The solution of (VI.5) also minimizes the KL divergence between $p(\boldsymbol{y};\boldsymbol{\pi})$ and the model $p(\boldsymbol{y};\boldsymbol{\theta})$, since

$$\mathrm{KL}(p(\boldsymbol{y};\boldsymbol{\pi})||p(\boldsymbol{y};\boldsymbol{\theta})) = \langle \ln p(y;\boldsymbol{\pi})\rangle_{p(\boldsymbol{y};\boldsymbol{\pi})} - \langle \ln p(y;\boldsymbol{\theta})\rangle_{p(\boldsymbol{y};\boldsymbol{\pi})} \geqslant 0, \tag{VI.7}$$

and the first term is independent of $\boldsymbol{\theta}$.

VI.D.2 Variational Inference for Expected Log-likelihood

The variational setting for learning BDS parameters is slightly different from the standard variational setting because, in (VI.5), the goal is to maximize the expected log-likelihood with regards to a reference distribution $\tilde{p}(y) = p(\boldsymbol{y};\boldsymbol{\pi})$, *i.e.*

$$\langle \ln \mathcal{L}(\theta, y)\rangle_{\tilde{p}(y)} = \langle \ln p(y;\theta)\rangle_{\tilde{p}(y)}. \tag{VI.8}$$

In this case

$$\begin{aligned}\langle \ln \mathcal{L}(\theta, y)\rangle_{\tilde{p}(y)} &= \mathscr{L}(\theta, q) + \langle \mathrm{KL}(q(x)||p(x|y;\theta))\rangle_{\tilde{p}(y)} \\ &\geqslant \mathscr{L}(\theta, q)\end{aligned} \tag{VI.9}$$

with lower bound

$$\mathscr{L}(\theta,q) = \langle \mathscr{L}(\theta,y,q) \rangle_{\tilde{p}(y)} \tag{VI.10}$$

$$= \int_x q(x) \langle \ln p(y,x;\theta) \rangle_{\tilde{p}(y)} dx + \mathrm{H}_q(X), \tag{VI.11}$$

where $\mathrm{H}_q(X) = -\int_x q(x) \ln q(x) dx$ is the entropy of X under distribution $q(x)$. This bound is tightest at

$$q^*(x) = \arg\max_{q \in \mathcal{D}_q} \mathscr{L}(\theta,q) \tag{VI.12}$$

$$= \arg\min_{q \in \mathcal{D}_q} \langle \mathrm{KL}(q(x)||p(x|y;\theta)) \rangle_{\tilde{p}(y)}. \tag{VI.13}$$

Note that, by Jensen's inequality,

$$\mathscr{L}(\theta,q^*) = \max_{q \in \mathcal{D}_q} \langle \mathscr{L}(\theta,y,q) \rangle_{\tilde{p}(y)} \tag{VI.14}$$

$$\leqslant \left\langle \max_{q \in \mathcal{D}_q} \mathscr{L}(\theta,y,q) \right\rangle_{\tilde{p}(y)} \tag{VI.15}$$

$$= \left\langle \mathscr{L}(\theta,y,q_y^*) \right\rangle_{\tilde{p}(y)}. \tag{VI.16}$$

Hence, the tightest bound of the expected log-likelihood lower bounds the average tightest log-likelihood bounds across observation sequences. Intuitively, (VI.14) lower bounds the log-likelihood over all samples from $\tilde{p}(y)$ that share the same hidden variable, distributed according to $q^*(x)$. On the other hand, (VI.16) uses the distribution $q_y^*(\boldsymbol{x})$ that best explains each sample y.

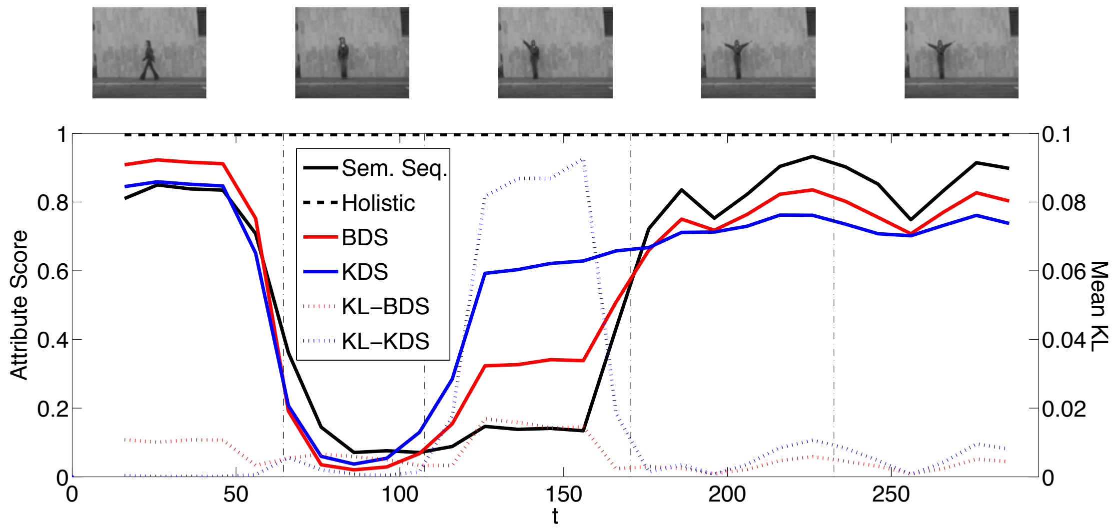


Figure VI.5: Synthetic sequences of binary dynamic systems. Top: key frames from activity sequence class "walk- pjump-wave1-wave2-wave2" in Syn-4/5/6. Bottom: score of "two-arms-motion" attribute. True scores in black, and scores sampled from BDS (red) and KDS (blue). Also shown is the KL-divergence between sampled and true scores, for both models.

VI.E Experiments: Event Recognition

In this section, we discuss experiments designed to evaluate the performance of the proposed BDS, BoWAD, and VLADAD. Three benchmarks from various perspectives are adopted to assess the behavior of these approaches: the *Weizmann Complex Activity* is a synthetic benchmark with comprehensive simulated challenges; *Olympic Sports* contains weakly cropped and aligned complex sport sequences; and *Multimedia Event Detection* features high level events with instances from open-source repositories.

VI.E.1 Attribute Classifiers

The VLADAD can be computed for any implementation of attribute classifiers. Since the goal was not attribute detection *per se*, we used two popular methods to produce attribute sequences. The first attribute classifier extracted

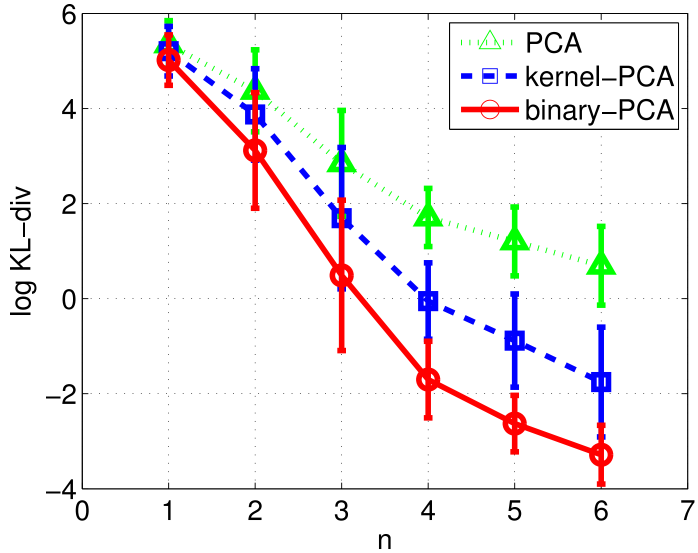


Figure VI.6: Fitting of attribute data. Log KL-divergence between original and reconstructed attribute scores, *v.s.* number of PCA components n, on Syn-4/5/6 for PCA, KPCA, and binary PCA are shown.

space-time interest points (STIP) of [92] and computed at each interest point a descriptor combining a histogram of oriented gradients (HoG) and a histogram of optical flow (HoF). The second classifier was based on the improved trajectory feature (ITF) of [165], using a descriptor composed of HoG, HoF, frame-wise trajectory (FWT), and motion boundary histogram (MBH), which has been shown to achieve state-of-the-art performance in action recognition even superior than features by deep learning [81, 118, 148, 167]. All features were extracted with the binary or source code provided by its authors [3].

In all experiments, attribute detection was based on the BoVW. For each descriptor, a codebook of size V was learned by k-means, over the entire training set, and used to quantize features. Different ITF descriptors were processed separately and merged by averaging kernel matrices during prediction. The

[3] Binary for STIP available at `http://www.di.ens.fr/~laptev/download`; source code for ITF available at `http://lear.inrialpes.fr/~wang/download`.

attribute annotations of [101] were used for Weizmann and Olympic Sports and those of [10] for MED. Appendix VI.I.2 provides details on attribute definitions and annotations. On Weizmann, attribute detectors were implemented with a linear SVM, using LIBSVM [26] with probability outputs. However, we found this to have scalability problems for the larger Olympics and MED datasets. On these datasets attribute classifiers were logistic regressors, implemented with LIBLINEAR [42]. To maximize attribute detection accuracy, while retaining the efficiency of linear classification, we used an additive kernel mapping of the histogram intersection kernel (HIK), as suggested in [160]. The attribute trajectory $\{\boldsymbol{\pi}_t\}$ of a video sequence was computed with a sliding window, where attribute detectors predicted attribute scores at each window anchoring position. An holistic attribute vector, encoding the presence of attributes in the entire video sequence, was also constructed by max-pooling $\{\boldsymbol{\pi}_t\}$ over time.

VI.E.2 Weizmann Complex Activity

The first set of experiments aimed to systematically compare the ability of different models to capture the dynamics of attribute sequences. A non-trivial difficulty of such a study is the need for datasets with classes that 1) differ only in terms of attributes dynamics, and 2) enable a quantification of these differences. It is critical that such datasets do not include discriminant information beyond attribute dynamics, such as discriminant scene backgrounds, objects, or scene durations. Unfortunately, these conditions are not met by existing action datasets. For example, the “making a sandwich” activity of the MED dataset is the only one to include the “sandwich” object. This enables the use of object recognition as a proxy for action recognition, an alternative that would not be viable if the dataset also contained an “eating a sandwich” activity. To avoid these problems, we assembled a synthetic dataset of complex sequences, which were synthesized from the atomic actions of the popular Weizmann dataset [51].

Weizmann contains 10 *atomic* action classes (*e.g.*, skipping, walking) performed by 9 people and was annotated with 30 low-level attributes (*e.g.*, "one-arm-motion") by [101]. Attribute sequences were computed over 30-frame sliding video windows with 10-frame stride. STIP features were used with a 1000-word vocabulary for low-level descriptor quantization. The availability of attribute ground truth for all atomic actions enables learning of clean attribute models. Hence, performance variations can be attributed to the quality of the attribute-based inference of the different approaches.

Three subsets of synthetic sequences were created by concatenating Weizmann actions (see Appendix VI.I.1 for some examples). These subsets vary in the variability and complexity of temporal structure of their video sequences. They target the study of different hypotheses regarding the role of dynamics in action recognition. The first, denoted "Syn-4/5/6" evaluates the ability of different models to capture dynamics of varying complexity, when all video segments are informative of the action class, *i.e.*, when the dynamics have no noise. The remaining two evaluate robustness to "noisy dynamics.' "Syn20 $\times$ 1" consists of actions of homogeneous dynamics, which are buried in additional video segments of dynamics uncharacteristic of the action class. "Syn10 $\times$ 2" consists of discontinuous actions of homogenous dynamics, which are interleaved with segments of "noisy dynamics.'

Complex Dynamics

In the first subset, "Syn-4/5/6", a sequence of *degree n* ($n = 4$,5,6) is composed of n atomic actions, performed by the same person. The row of images at the top of Figure VI.5 presents keyframes of an activity sequence of degree 5, composed by the atomic actions "walk," "pjump," "wave1," "wave2," and "wave2.' The black curve (labeled "Sem. Seq") in the plot at the bottom of the figure shows the score of the "two-arms-motion" attribute over time. 40 activity categories were defined per degree n (total of 120 activity categories), and the dataset was assembled per category, containing one activity sequence per person (9 people, 1080 sequences in total). Overall, the activity sequences differ in the number, category, and temporal order of atomic actions.

We started by comparing the binary PCA that underlies the BDS to the PCA and KPCA decompositions of the LDS and KDS. In all cases, a set of attribute score vectors $\{\boldsymbol{\pi}_t\}$ was projected into the low-dimensional PCA subspace, the reconstructed score vectors $\{\hat{\boldsymbol{\pi}}_t\}$ were computed and the KL divergence between $B(\boldsymbol{y}, \boldsymbol{\pi}_t)$ and $B(\boldsymbol{y}, \hat{\boldsymbol{\pi}}_t)$ was measured. The logit kernel $K(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) = \sigma^{-1}(\boldsymbol{\pi}_1)^\intercal \sigma^{-1}(\boldsymbol{\pi}_2)$, where $\sigma^{-1}(\cdot)$ is the element-wise logit function, was used for KPCA. Fig. VI.6 shows the average log-KL divergence, over the entire dataset, as a function of the number of PCA components used in the reconstruction. Binary PCA outperformed both PCA and KPCA. The improvements over KPCA are particularly interesting, since the latter uses the logistic transformation that distinguishes binary PCA from PCA. This is explained by the Euclidean similarity measure that underlies the assumption of Gaussian noise in KPCA, as discussed in Section IV.A.2.

To gain some more insight on the different models, a KDS and a BDS were learned from the 30 dimensional attribute score vectors of the activity sequence in Figure VI.5. A new set of attribute score vectors were then sampled from each model. The evolution of the scores sampled for the "two-arms-motion" attribute are shown in the figure (in red/blue for BDS/KDS). Note how the scores sampled

Table VI.1: Accuracy on Syn-4/5/6.

method		accuracy
BoVW [93]	(x1y1t1)	57.8%
	(x1y1t3)	78.8%
	(x1y1t6)	92.5%
holistic attribute		72.6%
DTM [13]		84.6%
ToT [168]		88.2%
KDS [28]		90.2%
BDS		94.8%

from the BDS approximate the original attribute scores better than those sampled from the KDS. This was quantified by computing the KL-divergences between the original attribute scores and those sampled from the two models, which are also shown in the figure.

We next evaluated the benefits of different representations of dynamics for activity recognition. Recognition rates were obtained with a 9-fold leave-one-out-cross-validation (LOOCV), where, per trial, the activities of one subject were used as test set and those of the remaining 8 as training set. We compared the performance of classifiers based on the KDS and BDS to those of a BoVW classifier with temporal pyramid (TP) matching [93], a holistic attribute classifier that ignores attribute dynamics, the dynamic topic model (DTM) [13] and the topic over time (ToT) model [168] from the text literature. For the latter, topics were equated to the activity attributes and learned with supervision (using the SVMs for attribute detection). Unsupervised versions of the topic models had worse performance and are omitted. Classification was performed with Bayes' rule for topic models, and a nearest-neighbor classifier for the remaining methods. BDS distances were measured with (V.8), while for the KDS we adopted the logit kernel. The dimension of the BDS state space was 5. The $\mathcal{X}^2$ distance was used for all BoVW and holistic attribute classifiers. In an attempt to match the pooling mechanism of temporal pyramid matching to the structure of the synthetic Weizmann sequences, we considered a variant with 6 temporal bins. This is denoted BoVW-x1y1t6.

The accuracy of all classifiers is reported in Table VI.1. BDS achieved the best performance, followed by BoVW-x1y1t6, KDS, the dynamic topic models, and BoVW-x1y1t1 and holistic attribute. Note the large difference between the holistic attribute and the best dynamic model ($\approx 22\%$). This shows that while attributes are important (14.8% improvement over BoVW without temporal pooling), they are not the whole story. Problems involving *fine-grained* activity classification, *i.e.*, discrimination between activities composed of similar actions executed in different sequence, requires modeling of attribute dynamics. This is reflected by both the improvement of BoVW with $x1y1t3$ and $x1y1t6$ temporal pyramids over naive BoVW, and that of models of attribute dynamics over the holistic attribute vector. Among the dynamic models, the BDS outperformed the KDS, the topic models DTM and ToT, and BoVW with pyramids $x1y1t3/t6$. It is also worth noting the sensitivity of pyramid matching to the number of temporal bins, with performance varying between 57.8% (x1y1t1) and 92.5% (x1y1t6).

Noisy dynamics

The remaining two datasets evaluated the robustness of the different methods to noise, poor segmentation, and alignment. The second dataset, "Syn20$\times$1" was composed of activity classes of large variability. Each activity was defined as a sequence of 20 *consecutive* atomic actions. This sequence was inserted at a *random* temporal location of a larger sequence of 40 atomic actions. The remaining 20 actions in the larger sequence were randomly selected from Weizmann. The third

Table VI.2: Accuracy on Syn20×1 and Syn10×2.

method		Syn20×1	Syn10×2
BoVW [93]	(x1y1t1)	23.3%	28.9%
	(x1y1t3)	36.7%	31.1%
	(x1y1t6)	55.6%	24.4%
holistic attribute		17.8%	16.7%
DTM [13]		49.3%	46.5%
ToT [168]		57.2%	55.9%
KDS [28]		61.6%	63.1%
BDS		64.4%	65.6%
BoWAD	(BMC)	100%	100%
	(MDS-*k*M)	100%	98.9%
VLADAD	(BMC)	100%	100%
	(MDS-*k*M)	100%	100%

dataset, "Syn 10×2," tested the detection of *discontinuous* activities. Each activity was defined by two subsequences, each with 10 consecutive atomic actions. The two subsequences were randomly inserted at non-overlapping locations of the larger (40 atomic actions) sequence. For both sets, 20 activities were synthesized for each of 9 subjects, producing 180 sequences per set.

In addition to the classifiers of Table VI.1, both the BoWAD and VLADAD were evaluated on these datasets. For both, short-term attribute sequences consisted of attribute vectors from 12 consecutive windows. The dimension of the BDS state space was again 5. WAD dictionaries were learned with both BMC and the MDS-*k*M algorithm of [128]. One-versus-all SVMs were used for BoVW and BoWAD classification, using a χ^2 kernel. VLADAD was implemented with a linear kernel, KDS and BDS used the kernel $K(\mathbf{\Omega}_a, \mathbf{\Omega}_b) = \exp(-\frac{1}{\gamma} d^2(\mathbf{\Omega}_a, \mathbf{\Omega}_b))$ where d is the distance used in Syn-4/5/6. These kernels achieved the best performance for each of the methods in our preliminary experiments.

Table VI.3: Mean average precisions on Olympic Sports.

method		w/o LA fusion		w/ LA fusion	
		STIP	ITF	STIP	ITF
BoVW	(x1y1t1)	59.0%	83.7%	-	-
[93]	(x1y1t3)	53.2%	81.6%	-	-
DMS [111]		62.5%	-	-	-
holistic attribute		62.6%	82.1%	64.2%	84.9%
VD-HMM [155]		66.8%	-	-	-
HMM-FV [151]		65.3%	84.7%	66.4%	86.7%
CTR [11]		64.9%	85.5%	67.1%	87.3%
BDS		67.8%	86.1%	68.7%	88.6%
BoWAD	(BMC)	73.5%	90.3%	74.9%	91.2%
	(MDS-*k*M)	71.2%	88.2%	72.6%	89.8%
VLADAD	(BMC)	**76.9%**	**91.7%**	**77.2%**	**93.1%**
	(MDS-*k*M)	71.7%	90.6%	73.4%	91.4%

Table VI.2 summarizes the performance of the different methods. Both BoVW and the holistic attribute vector performed poorly. Note, in particular, how BoVW-x1y1t6 now underperformed the two other implementations of temporal pyramid matching. This highlights the difficulty of designing universal pooling schemes, that can withstand significant intra class variability. This problem also affected the dynamics models, which performed substantially worse than in Table VI.1. While the BDS significantly outperformed the other methods, its performance was still lackluster. This is explained by the underlying assumption of a single dynamic process, a severe mismatch on Syn20$\times$1 and Syn10$\times$3, where the activities of interest are 1) not temporally aligned and 2) immersed in irrelevant video content. It is thus not surprising that the BoWAD and VLADAD achieved substantially better performance on these datasets, reaching perfect classification. With respect to BoWAD clustering, both strategies achieved excellent results, with BMC performing slightly better than MDS-*k*M. Overall,

Table VI.4: Performance on Olympic Sports.

method	mAP
[156]	82.9%
[69]	83.2%
[68]	85.3%
[98]	84.5%
[165]	91.1%
[78]	74.6%
[110]	92.3%
[91]	92.9%
VLADAD	**93.1%**

these results demonstrate the robustness of the proposed BoWAD and VLADAD representations to intra-class variation and noise.

VI.E.3 Olympic Sports

The second set of experiments was performed on Olympic Sports [111]. This contains YouTube videos of 16 sport activities, with a total of 783 sequences. Some activities are sequences of atomic actions, whose temporal structure is critical for discrimination from other classes (*e.g.*, "clean and jerk" *v.s.* "snatch," and "long-jump" *v.s.* "triple-jump"). Since the attribute labels of [101] are only available for whole sequences, the attribute classifiers are much noisier than in the previous experiment, degrading the quality of attribute models. We followed the train-test split proposed by [111] and used per-category average precision (AP) and mean AP (mAP) to measure recognition performance. In all cases, low-level feature quantization was based on 4000-word codebooks, learned with k-means. Attribute sequences were computed with a 30-frame sliding window, implemented with a stride of 4 frames.

The proposed approaches were compared to BoVW-TP, the decomposable

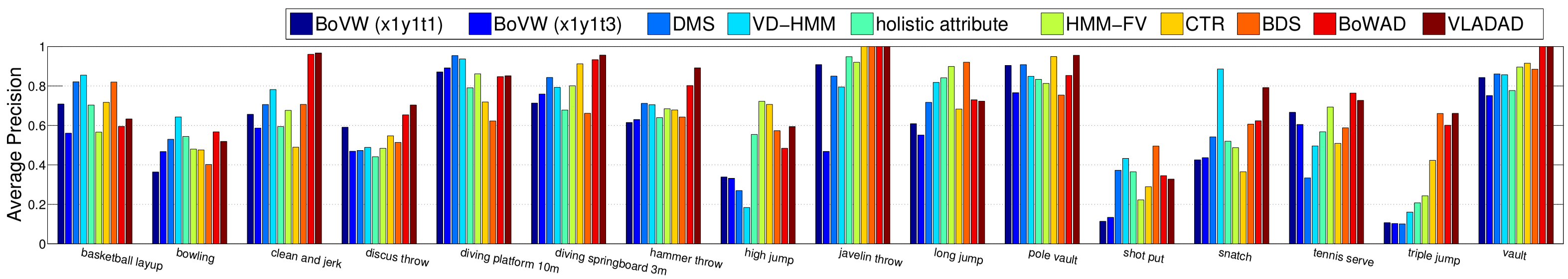


Figure VI.7: Average precisions on Olympic Sports with STIP as the low-level feature.

motion segments model (DMS) of [111], the hidden Markov model with latent states of variable duration (VD-HMM) [155], the holistic attribute, and two recent approaches that also model attribute dynamics: the HMM fisher vector (HMM-FV) of [151] and the combined temporal representation (CTR) of [11]. Classification was performed with SVMs using a χ^2 or Jensen-Shannon kernel for histogram-based methods (BoVW, holistic attribute, BoWAD); SVMs using a radial basis function (RBF) kernel $K_\alpha(i,j) = \exp(-\frac{1}{\alpha}d^2(i,j))$ for HMM-FV and CTR; a nearest neighbor classifier or SVM using the RBF kernel for BDS; and a linear SVM for VLADAD. For each method, the best classifier parameters were chosen by 4-fold cross-validation on the training set. The number of PCA components L of the BDS was selected from $\{2,4,6,8\}$, and the length τ of the attribute sequences of BoWAD and VLADAD from $\{4,6,8,10,12,16\}$ by cross-validation on the training set.

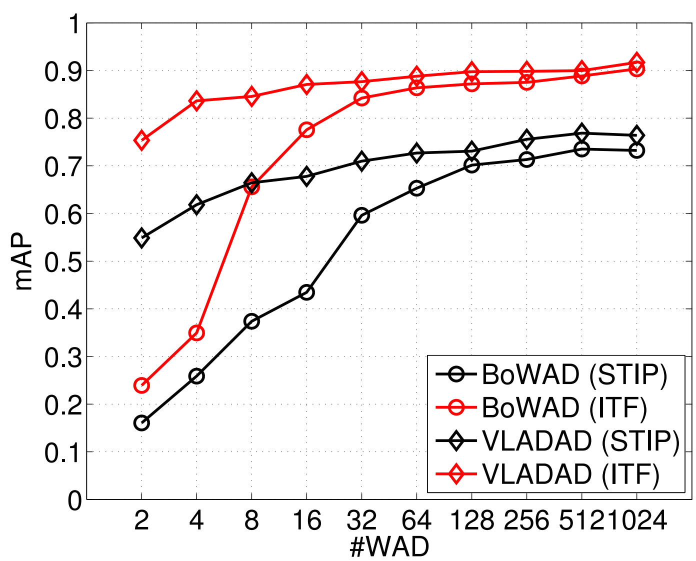


Figure VI.8: Mean average precision *v.s.* size of WAD dictionary on Olympic Sports.

The performance of the different approaches is summarized in Table VI.3[4].

[4]Note that the version of Olympic Sports used in [111] is different from that released publicly.

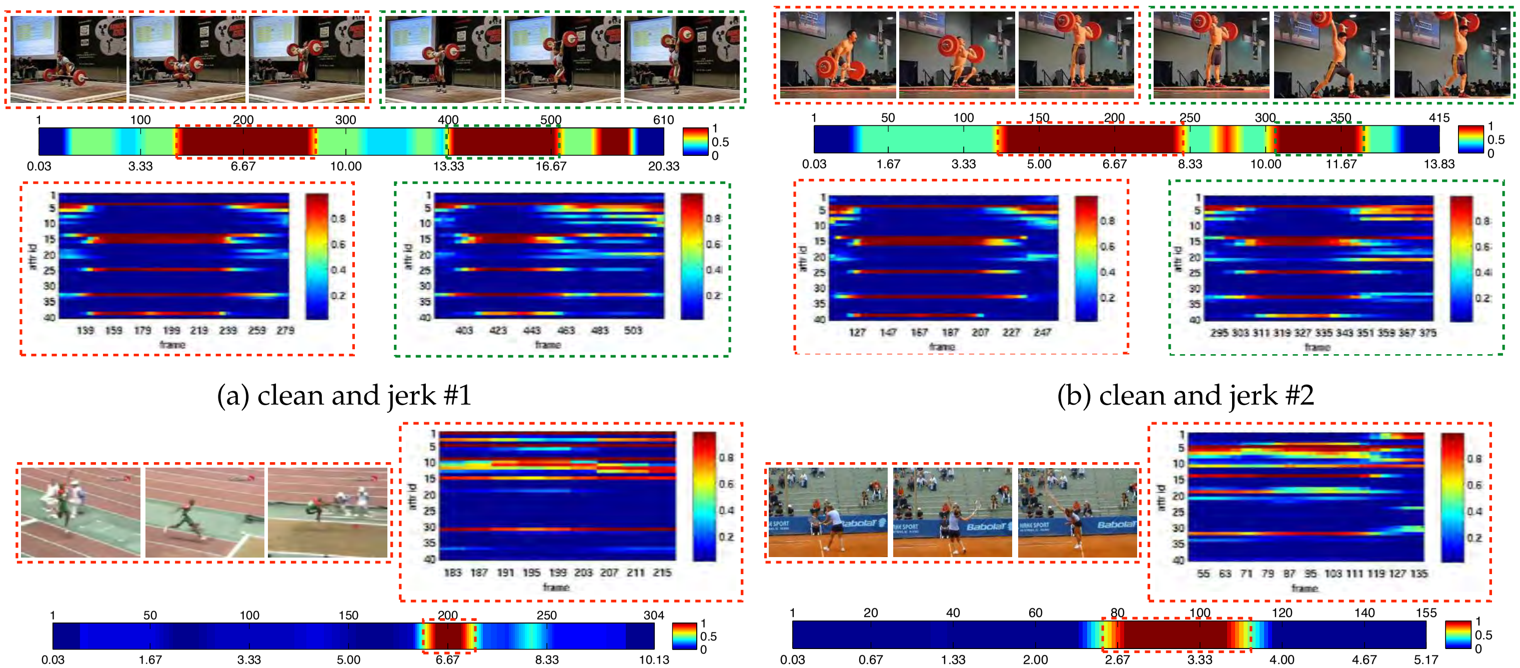
Several conclusions can be drawn. First, all models benefit strongly from the ITF features. The increased performance of BDS, BoWAD, and VLADAD with these features suggests that a more discriminant set of low-level features, and thus cleaner attributes, can significantly simplify the problem of modeling of attribute dynamics.

Second, the BDS again outperforms all other models. The gains are larger over methods that do not account for dynamics (*e.g.*, the holistic attribute vector) but substantial even over the alternative models of attribute dynamics, such as HMM-FV or CTR. This is likely due to the richer characterization of the hidden state space by the BDS and its modeling of low-dimensional attribute subspaces. An interesting observation is that BoVW-x1y1t3 underperforms the vanilla BoVW significantly, reflecting the fact that its rigid temporal cells with fixed temporal anchor points 1) are coarse for capturing finer structure within each cell, and 2) cannot adapt to intra-class variation. This vulnerability of BoVW with augmented "rigidity" to over-fitting is also confirmed by other works in literature [90].

Third, the BDS gains are smaller than in Weizmann. This is due, in part, to the increased difficulty of modeling dynamics because annotations are noisy and, in part, to the nature of the dataset. While Weizmann requires fine-grained temporal discrimination for most classes, this is not the case in Olympic. For example, the holistic attribute vector suffices to discriminate classes that are very distinctive, *e.g.*, that have *unique* motion. An example is "diving platform 10m," which can be singled out by its distinctive patterns of fast downward motion. This is visible in the per-category average-precision plot of Fig. VI.7, where the holistic attribute vector performs very well for this class. On the other hand, finer

---

DMS performance on the latter was reported in [155].



(a) clean and jerk #1

(b) clean and jerk #2

(c) triple jump

(d) tennis serve

Figure VI.9: Recounting by BoWAD on Olympic sports. The normalized score, for activity recognition, of each video segment is shown as a bar (time in seconds displayed at the bottom, frame id at the top). As shown in the color key, red corresponds to a score of 1 (most relevant), blue to a score of 0 (less relevant). The dashed lines identify the most significant events. Associated key-frames are shown at the top, corresponding attribute sequences at the bottom. Same setting applies to all recounting illustrations. Best viewed in color.

grained temporal analysis is required to distinguish between similar classes, *e.g.*, "long-jump" *v.s.*"triple jump," or "clean and jerk" *v.s.*"snatch.' Fig. VI.7 clearly shows that these classes 1) pose a greater challenge to previous methods, and 2) lead to the largest gains by the BDS, BoWAD, and VLADAD.

Fourth, while the BDS performs quite well for classes with reasonably well segmented and aligned sequences (*e.g.*, "long jump"), the assumption of a single dynamic process again limits its performance for categories with larger variability (*e.g.*, "snatch," "clean and jerk," "tennis serve," *etc*). Both BoWAD and VLADAD perform better in this case, improving BDS performance by 4% to 9% overall. Fig. VI.7 shows that this improvement is particularly significant for categories, such as "clean and jerk" and "tennis serve," whose discriminant events are scattered throughout the video sequence.

Fifth, regarding encoding schemes there is now a clear gap between BoWAD (73.5% for STIP, 90.3% for ITF) and VLADAD (76.9% for STIP, 91.7% for ITF). This confirms many previous observations for the effectiveness of Fisher scores in image and video classification. Fig. VI.8 shows that the VLADAD gains hold across a substantial range of WAD codebook size. Note that a 16-word VLADAD codebook already has mAP (around 87%) superior to most methods in Table VI.4. Similarly, we observed a consistent advantage of BMC over MDS-*k*M clustering, with differences of 1% to 5% in mAP (see Table VI.3).

Sixth, Fig. VI.7 shows that even methods with low overall performance, *e.g.*, the holistic attribute vector, can have good performance for some classes. This suggests that there is some complementarity in the different video representations, and it may be beneficial to combine them [150, 154, 175]. We have investigated this by combining representations based on low-level features, holistic attributes, and dynamic modeling, using the late fusion scheme of [154], which uses the geometric mean of scores of different classifiers as the final prediction score. The combination of multiple representations, denoted "LA fusion" in Table VI.3), does not change the conclusions above. Again, the BDS outperforms previous models of temporal structure, based on either low-level motion (DMS and VD-HMM) or attributes (HMM-FV and CTR), the BoWAD outperforms the BDS, and the VLADAD has top performance.

Seventh, all methods benefit from late fusion. This confirms that some discriminant information might be discarded by attribute modeling (gains by inclusion of low-level features) and holistic modeling can sometimes be useful. However, the effect is small, with a gain less than 1% for most the best performing methods.

Finally, Table VI.4 compares the VLADAD-BMC with ITF features and late fusion to previous approaches in the literature. The proposed representation achieves state-of-the-art performance on this dataset, surpassing the previous best results by [165, 110, 91]. Note that all these three benchmarks are based on ITF encoded with Fisher vector, which is a stronger baseline than ours (ITF with vanilla BoVW). This enhancement could be incorporated into our attribute detectors, potentially leading to even better performance.

VI.E.4 TRECVID-MED11

The third set of experiments used the 2011 TRECVID multimedia event detection (MED11) open source data-set [114]. This is one of the most challenging datasets for activity or event recognition due to 1) the vaguely defined high-level event categories (*e.g.*, "birthday party"); 2) the large intra-class variation in terms of event composition (*e.g.*, temporal duration, organization), stage setting, illumination, cutting, resolution, *etc*; 3) large negative samples, and so forth.

Table VI.5: Mean average precisions (in percentage) on MED11.

method		DEVT				DEVO			
		w/o LA fusion		w/ LA fusion		w/o LA fusion		w/ LA fusion	
		STIP	ITF	STIP	ITF	STIP	ITF	STIP	ITF
random guess		0.98				0.37			
BoVW	(x1y1t1)	15.70	32.68	-	-	8.31	18.53	-	-
[93]	(x1y1t3)	15.50	31.86	-	-	9.66	18.92	-	-
DMS [111]		5.72	-	-	-	2.52	-	-	-
holistic attribute		10.62	25.03	16.31	33.42	4.93	12.45	8.93	19.67
VD-HMM [155]		11.25	-	-	-	4.77	-	-	-
HMM-FV [151]		8.15	21.82	16.50	33.77	4.49	11.64	9.52	20.08
CTR [11]		9.46	22.42	17.14	33.61	4.62	11.08	9.61	19.72
BDS		6.75	16.72	16.33	33.49	3.67	9.21	9.16	19.21
BoWAD	(BMC)	13.38	26.20	18.05	35.02	7.49	14.36	10.25	20.91
	(MDS-*k*M)	12.70	25.08	17.37	34.11	6.92	13.68	9.94	20.30
VLADAD	(BMC)	14.19	27.04	18.56	**35.40**	7.91	15.61	10.92	**21.84**
	(MDS-*k*M)	13.41	26.16	17.93	34.62	7.33	14.84	10.15	20.89

We followed the protocol suggested by the TRECVID evaluation guidelines for performance evaluation. Specifically, the event collection (EC) set was used for training. EC contains 2,392 training samples of 15 high-level events (see Table VI.6 for the full list), with 100-200 positive examples per event. Two evaluation sets, DEV-T and DEV-O, were used for testing. DEV-T has 10,723 samples (370 hours of video in total), approximately 1% of which is from events 1-5 and the remaining 99% are negative samples; while DEV-O has 32,061 samples (1200 hours in total), with around 0.5% from events 6-15 and 99.5% negative samples.

Attribute classification was based on 103 attributes defined by [10]. 8,000-word codebooks were learned with k-means for low-level feature quantization. Attribute scores were computed with a 180-frame sliding window and a 30-frame stride. All classifier settings were as in Section VI.E.3, with the exception of the length τ of attribute sequences for BoWAD and VLADAD, which was selected from $\{5, 10, 15, 20\}$, corresponding to roughly 5, 10, 15 and 20 seconds. To account

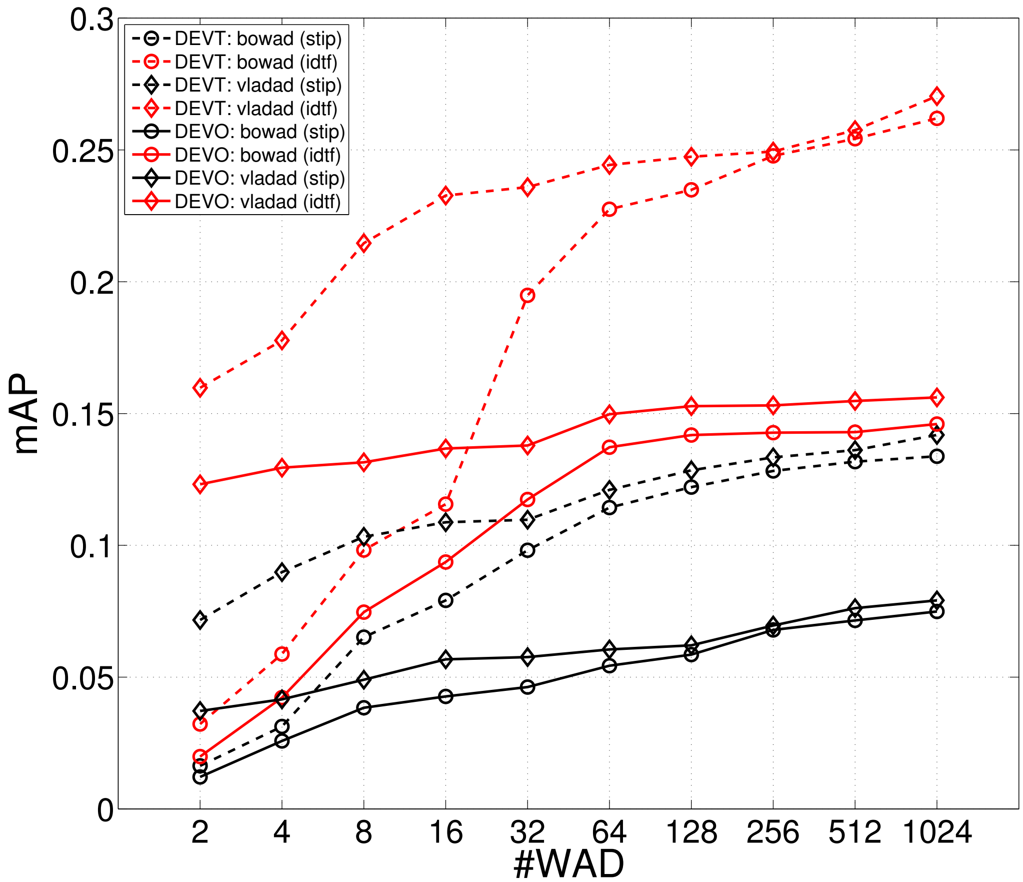


Figure VI.10: Mean average precision *v.s.* size of WAD dictionary on MED11.

for the variability of instances from the same event, both the BoWAD histograms and VLADAD were computed with different τ and concatenated into the feature used for event prediction.

Table VI.5 summarizes the event detection performance of the different methods. Most of these results are in line with those of the previous section. For example, the VLADAD again outperformed the BoWAD, especially for small codebook sizes. This is shown in greater detail in Fig. VI.10. Similarly, clustering with the BMC again outperformed MDS-kM. Finally, Fig. VI.11 shows the APs of VLADAD for different attribute sequence lengths τ. Not surprisingly, different

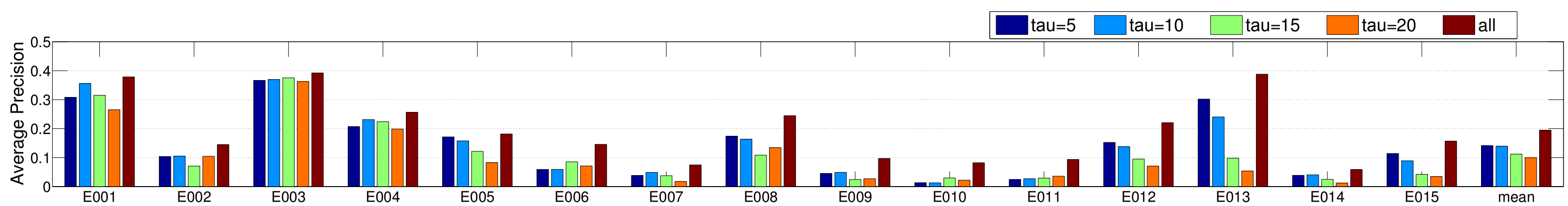


Figure VI.11: Average precision of VLADAD on MED11. ITF is used.

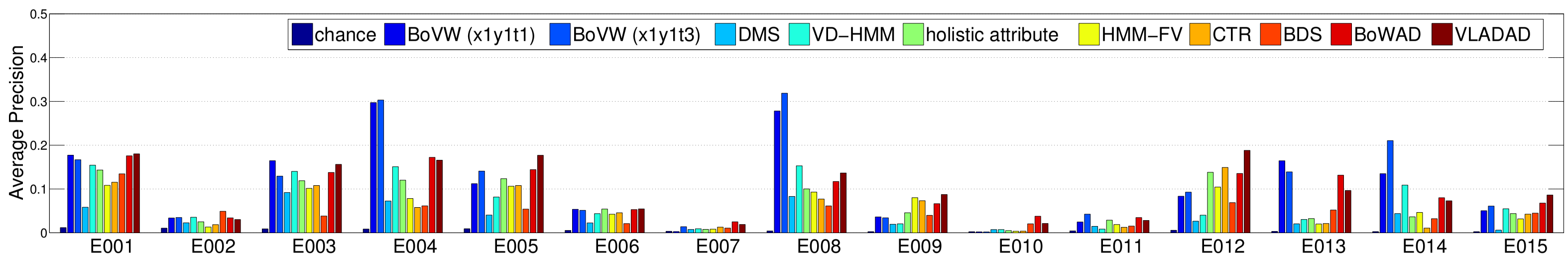


Figure VI.12: Comparison of average precisions on MED11. STIP is used.

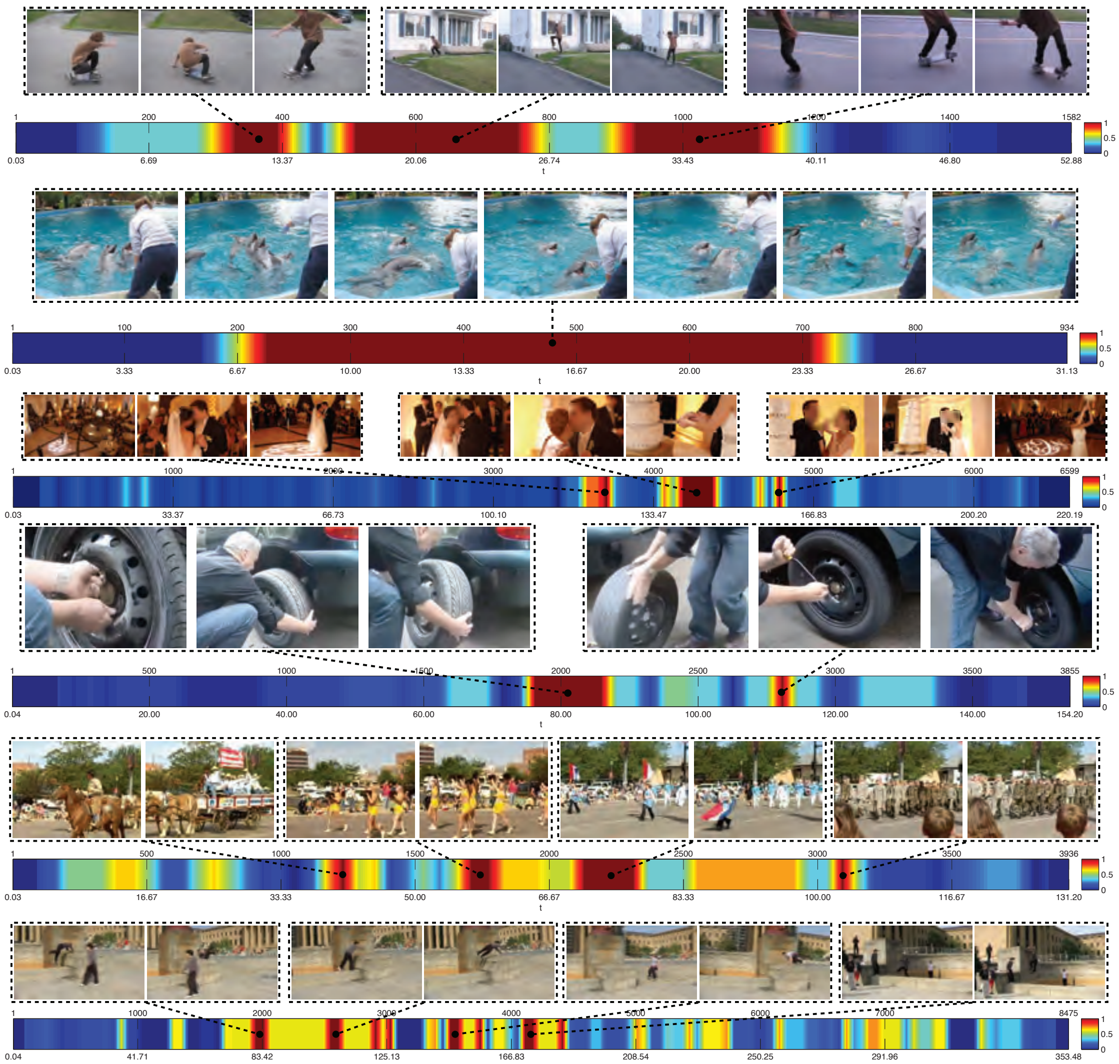


Figure VI.13: Recounting by BoWAD on MED11. Sequences of "attempt a board trick," "feed an animal," "wedding ceremony," "change a vehicle tyre," "parade," and "parkour" (top to bottom) are shown. Snapshots from the most significant clips of each sequence are also shown.

Table VI.6: Event list for MED11.

ID	Event Name	ID	Event Name
E001	attempt a board trick	E009	get a vehicle unstuck
E002	feed an animal	E010	groom an animal
E003	land a fish	E011	make a sandwich
E004	wedding ceremony	E012	parade
E005	work on a wood project	E013	parkour
E006	birthday party	E014	repair an appliance
E007	change a vehicle tyre	E015	work on a sewing project
E008	flash mob gathering		

lengths performed best for different events. For example, while in “parkour” (E013) the discriminant motion of “rush-jump-climb-land” takes about 5 seconds, in “land a fish” the distinctive motion of “pull-throw-catch” lasts between 5 and 20 seconds. Combing different attribute sequence lengths achieved the best performance for all event classes.

However, there were also some significant differences. First, the previously proposed models of temporal structure, either for low-level features (DMS and VD-HMM) or attributes (BDS, HMM-FV, CTR), performed worse or, at most, on par with the holistic attribute vector. This can be justified by the complexity and variability of the MED events. The BDS was particularly affected by this problem, performing $1\% - 5\%$ worse than the other models of attribute dynamics. Together with Section VI.E.3, these results confirm that, while the BDS is a better model of dynamics for segmented and aligned video, it has difficulties for video containing multiple dynamic processes. The fact that the BoWAD and VLADAD outperform both the holistic attribute vector, and the previous models of low-level (DMS, VD-HMM) and attribute (HMM-FV, CTR) dynamics shows that they effectively address this problem.

Second, and more surprising, attribute-based models underperformed the

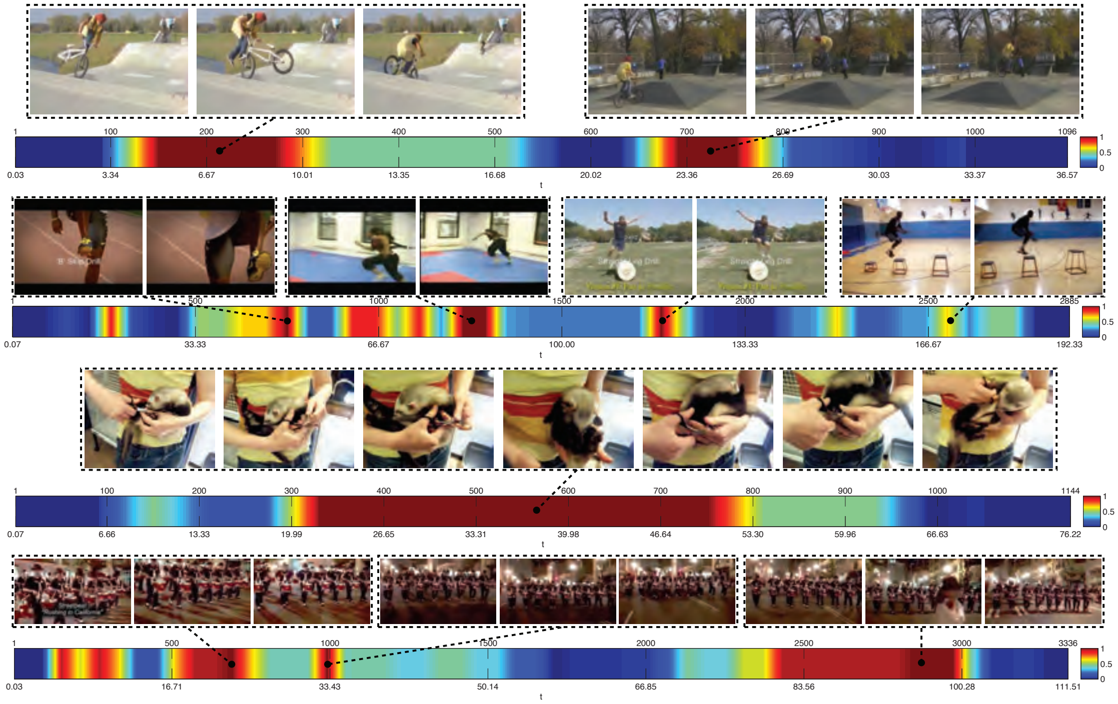


Figure VI.14: False positives of recounting on MED11. Examples come from top 0.1% detections for "attempt a board trick," "parkour," "groom an animal," and "parade" (top to bottom).

BoVW. This could be due to 1) noisy attribute classification, or 2) limited attribute vocabulary. Since, as shown in Fig. VI.12, attribute-based approaches handled some events better than the BoVW we believe that the latter is the main problem. In any case, since this shows that attribute representations capture information complementary to that of the BoW, the fusion of attribute models and the BoVW should lead to the best performance. Table VI.5 shows that this is indeed the case, as all attribute representations improve on the BoVW when combined with it by late fusion. In fact, when fused with BoVW and holistic attribute, the VLADAD achieves 21.84% mAP on MED11 DEV-O. In comparison to other benchmarks, this is substantially higher than the 15.69% of [158], 16.02% of [87], 15.35% of [56], and comparable to 22.13% (best results for a single low-level feature) by [173].

VI.F Experiments: Event Recounting

An interesting property of the BoWAD is that it can be easily combined with "recounting" procedures to support semantic video segmentation, summarization, and activity identification. This follows from the fact that the contribution of a particular WAD to the score of an activity classifier can be seen as a measure of the importance of the corresponding pattern of attribute dynamics for the detection of the target activity. We used the recounting procedure of [177], quantifying the significance of a video segment (for event detection) by the weighted sum of the similarities between the corresponding BoWAD histogram bin and those of the SVM support vectors. More specifically, let x be the BoWAD histogram and consider a prediction rule based on an additive kernel, *e.g.*, an SVM with HIK. In this case,

$$h(x) = \sum_i \alpha_i g(x, z^{(i)}) + c, \tag{VI.17}$$

where $z^{(i)}$ is the i-th support vector, α_i the corresponding SVM weight, c a constant, and $g(x, z^{(i)}) = \sum_j g_j(x_j, z^{(i)})$ measures the similarity between z_i and x. The prediction rule then can be rewritten as

$$h(x) = \sum_{j,i} \alpha_i g_j(x_j, z^{(i)}) + c = \sum_j h_j(x_j) + c, \tag{VI.18}$$

where $h_j(x_j) = \sum_i \alpha_i g_j(x_j, z^{(i)})$ is the contribution of histogram bin x_i to the classification score of the BoWAD histogram. Note that, unlike the holistic attributes of [177], for which temporal localization intractable, each video segment is associated with a WAD in the BoWAD, which corresponds to a short-term pattern of activity. This allows the quantification of the contribution of the video segment to event detection by $h_j(x_j)$, where x_j is the bin of the corresponding WAD. This enables a precise characterization of the temporal duration and anchor points of different event evidence.

Four examples are illustrated in Fig. VI.9. In both instances of “clean and jerk,” the BoWAD discovers the two signature motion of “lifting barbell to chest level” and “lifting barbell over head.’ Note the variation in temporal location and duration of these events in the two sequences. On the other hand, the signature events discovered for “triple jump” and “tennis serve,” are “large step forward followed by jump,” and “toss ball into the air followed by hit,” respectively. These results illustrate the robustness of the BoWAD to video uninformative of the target activity, and its ability to zoom in on the discriminant events. This is critical for accurate activity recognition from realistic video.

Another important task in TRECVID is recounting of multimedia events, which we implemented as in Section VI.E.3. Several BoWAD recounting examples are illustrated in Fig. VI.13, again showing that modeling local signature behavior is sufficient for accurate detection of complex activities. Specifically, the BoWAD captures a somersault by a subject riding a skateboard in “attempt a board trick,” the action of throwing food to dolphins in “feeding an animal,” the scattered scenes of “dancing,” “cutting cake,” and “bouquet toss” in “wedding ceremony,” the marching crowd on “parade,” and so on. On the other hand, as shown in Fig. VI.14, recounting results also reveal two major reasons for detection false positives. The first is the existence of visual content (*e.g.*, motion) confusable with that of the target event. The top sequence of Fig. VI.14, a sequence of “attempt a board trick” where a bike rider performs somersaults similar to those executed by skateboard riders in the background, is an example of this problem. Similarly, the second sequence shows a false positive for "parkour," where several athletes perform plyometric activities or other forms of training, which involve running, jumping over obstacles, and climbing. The second reason for false positives is the ambiguity of certain activities, which lead to inconsistent ground-truth on MED11. For example, the third and fourth sequences of Fig. VI.14 are labeled as background events for "groom an animal" and "parade," respectively. However, the recounting results show that both sequences are indeed instances of these events.

VI.G Summary and Discussion

In this work, we have proposed a novel representation for video, based on the modeling of action attribute dynamics. The core of this representation is the binary dynamic system (BDS), a joint model for attribute appearance and dynamics. This model was shown to be effective for video sequences that display a single activity, of homogeneous dynamics. To address the challenges of complex activity recognition, where video sequences can be composed of multiple atomic events or actions, the BDS was embedded in a BoVW-style representation, denoted the BoWAD. This is based on a BDS codebook, representing video as an histogram of assignments to BDSs that characterize temporally localized attribute dynamics. To enhance discrimination, this representation was extended into a Fisher-like encoding that characterizes the first order distribution of local behavior in the BDS manifold. This generalizes the popular VLAD representation and was denoted the VLADAD. Experiments have shown that the BDS, the BoWAD, and the VLADAD have state of the art performance for activity recognition in video whose segments range from precisely segmented and well aligned to unsegmented and scattered within larger video streams. The ability of these representations to capture signature events of different activity classes was demonstrated through various recounting examples.

VI.H Acknowledgements

The text of Chapter VI is, in part, based on the material as it appears in the following publications: W.-X. LI and N. Vasconcelos, “Complex Activity Recognition via Attribute Dynamics,” to appear at *International Journal of Computer Vision* (IJCV), and W.-X. LI, Y. Li and N. Vasconcelos, “Efficient Variational Inference, Learning and Probabilistic Kernels for Binary Dynamic Systems,” under review at *Neural Information Processing Systems* (NIPS), 2016. The dissertation author was a primary researcher and an author of the cited material.

VI.I Appendix

VI.I.1 Weizmann Complex Activity

Synthetic Datasets

The synthetic dataset contains three sets: Syn-4/5/6, Syn20$\times$1 and Syn10$\times$2, which are generated using the 10 atomic actions (per person) from the original Weizmann dataset by [51]. Exemplar activities in Syn-4/5/6, Syn20$\times$1, and Syn10$\times$2 are shown in Table VI.7, Table VI.8, and Table VI.9, respectively. For Syn20$\times$1, and Syn10$\times$2, two of the 9 instances for an activity (each instance is assembled from each of the 9 people’s atomic actions).

Table VI.7: Examples for Syn-4/5/6.

Syn-4	skip-run-walk-wave1
Syn-5	jack-wave1-bend-walk-walk
Syn-6	wave2-run-walk-wave1-jump-wave2

Table VI.8: Examples for Syn20×1.

Ground-truth Activity	wave1-wave1-wave2-walk-walk-wave1-walk-wave2-wave2-walk-jack-skip-wave2-bend-bend-jump-run-skip-jack-wave1
Noisy Instances[1]	side-wave2-walk-skip-run-wave1-bend-bend-walk-walk-wave1-wave1-wave2-walk-walk-wave1-walk-wave2-wave2-walk-jack-skip-wave2-bend-bend-jump-run-skip-jack-wave1-side-bend-side-walk-run-side-walk-jack-bend-walk;
	jump-run-wave1-wave1-wave2-walk-walk-wave1-walk-wave2-wave2-walk-jack-skip-wave2-bend-bend-jump-run-skip-jack-wave1-wave1-walk-side-jump-side-jump-jump-run-jack-side-wave1-run-run-skip-wave1-jack-side-bend;

[1] ground-truth activities are composed of actions in red.

VI.I.2 Attribute Definition

Weizmann Complex Activity

Attribute definitions from [101] on Weizmann complex activity are shown in Table VI.10.

Olympic Sports

Attribute definitions from [101] on Olympic Sports dataset [111] are shown in Table VI.3.

Table VI.9: Examples for Syn10×2

Ground-truth Activity	jack-jump-side-jump-pjump-run-jack-side-bend-wave1; run-side-side-skip-run-jump-walk-jack-run-skip
Noisy Instances[2]	wave2-run-wave1-bend-jump-wave1-skip-side-jack-jump-side-jump-pjump-run-jack-side-bend-wave1-walk-wave2-wave2-wave1-side-pjump-wave2-run-side-side-skip-run-jump-walk-jack-run-skip-jack-pjump-pjump-pjump-pjump;
	jump-jack-jump-side-jump-pjump-run-jack-side-bend-wave1-jump-side-skip-jack-run-side-bend-jump-pjump-side-run-side-side-skip-run-jump-walk-jack-run-skip-side-pjump-wave2-walk-run-pjump-wave2-wave2-walk;

[2] ground-truth activities are composed of actions in red.

VI.I.3 TRECVID MED11

Attribute definitions from [10] on TRECVID MED11 dataset [114] are shown in Table VI.12.

Table VI.10: Attributes for Weizmann Actions

attribute	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
arm-hand-alternate-move-forward	0	0	0	0	1	0	0	1	0	0
arm-hand-hang-down-swing-back-forward	0	0	0	0	0	0	0	1	0	0
arm-hand-swing-move-back-forward-motion	0	0	1	0	1	0	1	1	0	0
arm-intense-motion	0	1	1	0	0	0	0	0	0	0
arm-shape-fold	0	0	1	0	1	0	1	0	1	1
arm-shape-straight	1	1	1	1	0	1	0	1	1	1
arm-side-open-up-down-motion	0	1	0	0	0	0	0	0	0	1
arm-small-swing-motion-left-right-up-down	0	1	0	0	0	0	0	0	1	1
arm-synchronized-arm-motion	0	1	1	0	0	0	1	0	0	0
arm-up-motion-over-shoulder	0	1	1	0	0	0	1	0	1	1
chest-level-arm-motion	0	0	0	0	1	0	0	0	0	0
cyclic-motion	0	1	1	1	1	1	1	1	1	1
huge-wave motion-up-down	0	0	1	0	1	1	1	0	0	0
intense-motion	0	1	1	1	1	1	1	0	0	0
leg-alternate-move-forward	0	0	0	0	1	1	0	1	0	0
leg-feet-small-moving-motion	0	0	0	0	0	0	0	1	0	0
leg-intense-motion	0	1	1	1	1	1	1	0	0	0
leg-motion	0	1	1	1	1	1	1	0	0	0
leg-side-stretch-motion	0	1	0	0	0	1	0	0	0	0
leg-two-leg-synchronized-motion	0	1	1	1	0	0	0	0	0	0
leg-up-forward-motion	0	0	1	0	1	0	1	0	0	0
one-arm-motion	1	0	0	0	0	0	1	0	1	0
small-wave-motion-up-down	0	0	0	0	0	0	0	1	0	0
torso-bend-motion	1	0	0	0	0	0	0	0	0	0
torso-vertical-shape-down-forward-motion	0	0	1	0	1	0	1	0	0	0
torso-vertical-shape-down-motion	0	1	0	1	0	0	0	0	0	0
torso-vertical-shape-up-forward-motion	0	0	1	0	1	0	1	0	0	0
torso-vertical-shape-up-motion	0	1	0	1	0	0	0	0	0	0
translation-motion	0	0	1	0	1	1	1	1	0	0
two-arms-motion	0	1	1	0	1	0	0	1	0	1

Table VI.11: Attributes for Olympic Sports

attribute	basketball-layup	bowl	clean-jerk	discus-throw	diving-platform-10m	diving-spring-3m	hammer-throw	high-jump	javelin-throw	long-jump	pole-vault	shot-put	snatch	tennis-serve	triple-jump	vault
ball	1	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0
bend	0	1	1	0	0	0	0	0	0	0	0	1	1	1	0	0
big-ball	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
big-step	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
crouch	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0
down-motion-in-air	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0
fast-run	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1
indoor	1	1	1	0	1	1	0	0	0	0	0	0	1	0	0	1
jump	1	0	0	0	1	1	0	1	0	1	1	0	0	0	1	1
jump-forward	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
lift-something	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
local-jump-up	1	0	0	0	0	1	0	1	0	0	1	0	0	1	0	0
motion-in-the-air	0	0	0	0	1	1	0	1	0	1	1	0	0	0	1	1
one-arm-open	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
one-arm-swing	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0
one-hand-holding-pole	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
open-arm-lift	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
outdoor	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1
raise-arms	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
run	1	0	0	0	0	0	0	1	1	1	1	0	0	0	1	1
run-in-air	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
slow-run	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
small-ball	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0
small-local-jump	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
somersault-in-air	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1
spring-platform	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1
standup	0	0	1	0	0	0	0	0	0	1	1	0	1	0	1	0
throw-away	0	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0
throw-up	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
track	0	1	0	0	0	0	0	0	0	1	1	0	0	0	1	1
turn-around	0	0	0	1	0	0	1	0	0	0	0	1	0	1	0	0
turn-around-with-two-arms-open	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
two-arms-open	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0
two-arms-swing-overhead	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
two-hand-holding-pole	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
up-down-motion-local	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
up-motion-in-air	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1
water	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
with-pat	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
with-pole	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0

Table VI.12: Attribute List for TRECVID MED11[1]

Animal-approaching	Animal-eating	Blowdrying	Close-door	Flash-photography
Hands-visible	Machine-carving	Machine-drilling	Machine-planing	Machine-sawing
Open-box	Open-door	People-dancing	People-marching	Perosn-rolling
Person-bending	Person-blowing-candles	Person-carving	Person-casting	Person-cheering
Person-clapping	Person-cleaning	Person-climbing	Person-close-trunk	Person-crying
Person-cutting	Person-cutting-cake	Person-cutting-fabric	Person-dancing	Person-dragging
Person-drilling	Person-drinking	Person-eating	Person-erasing	Person-falling
Person-fitting-bolts	Person-flipping	Person-gluing	Person-hammering	Person-hitting
Person-holding-sword	Person-hugging	Person-inserting-key	Person-jacking-car	Person-jumping
Person-kicking	Person-kissing	Person-laughing	Person-lifting	Person-lighting
Person-lighting-candle	Person-marching	Person-measuring	Person-opening-door	Person-open-trunk
Person-packaging	Person-painting	Person-petting	Person-picking	Person-planing
Person-playing-instrument	Person-pointing	Person-polishing	Person-pouring	Person-pullingout-candles
Person-pushing	Person-pushing	Person-reeling	Person-riding	Person-rolling
Person-running	Person-sawing	Person-sewing	Person-shaking-hands	Person-singing
Person-sketching	Person-sliding	Person-spraying	Person-squatting	Person-standing-up
Person-steering	Person-surfing	Person-taking-pictures	Person-throwing	Person-turning-wrench
Person-twist	Person-twisting-wood	Person-using-knife	Person-using-tire-tube	Person-walking
Person-washing	Person-waving	Person-welding	Person-wetting-wood	Person-whistling
Person-wiping	Person-writing	Shake	Spreading-cream	Stir
Taking-pictures	Vehicle-moving	Wheel-rotating		

[1] About 10,000 short-term clips are annotated for attribute training.

Chapter VII

Conclusion

In this thesis, we study the problem of modeling temporal structure of human behavior via dynamics modeling. We propose a temporal structure hierarchy for human behavior representation that accounts for the distinct properties of an activity at different different temporal granularity. While primitive motion residing at the low-level of the hierarchy can be captured by data-driven schemes such as BoVW representation, we propose to model the temporal structure of human behavior at midium level on a robust, stable yet general platform that encodes some semantically meaningful concepts (denoted attributes). This representation platform bridges the gap between low-level visual feature and the high-level logic reasoning, which is also shown to bring in benefits such as better generalization, knowledge transfer, and so forth. While attributes take care of abstracting semantic information from low-level visual signal, the dynamic model focuses on charactering the evolution patterns in this space. To cope with long-term non-stationarity and intra-class variation for complex behavior at the high level, we derive several encoding schemes that capture the statistics of the attribute dynamics in local snippets, instead of precise characterization of the whole sequence, which is prone to over-fitting due to the sparse nature of complex event instantiation.

The proposed framework is implemented via a series of novel models, together with the corresponding technical tools for inference, parameter estimation, similarity measure, statistics encoding, and so on. In particular, a dynamic model is proposed to capture the evolution pattern in sequential binary data, denoted the binary dynamic system (BDS), which is comprised of a binary principal component analysis for modeling appearance and Gauss-Markov process to encode dynamics. A mixture model is further deduced from BDS to capture multiple evolution patterns in a large data corpus. Accurate and efficient approximate inference schemes are developed for the posterior based on the variational methods to handle the intrinsic intractability; and a variational expectation-maximization algorithm is also proposed for parameter estimation. Relying on these tools, measurements that quantify the similarity or dissimilarity of two binary sequences are developed from the perspective of control theory, information geometry, and kernel methods. Encoding schemes for the zeroth and first order statistics of sequential binary data in the model manifold are also proposed, resulting in the bag-of-words for attribute dynamics and vector of locally aggregated descriptor for attribute dynamics.

Empirical study on several challenging tasks of complex human activity analysis justifies the effectiveness of the proposed solution. This has not only produced the state-of-the-art results for event detection, but also recounting results that provides the visual evidence anchored over time in the video for the prediction, which enables tasks like semantic video segmentation, content based video summarization, and so forth.

Bibliography

[1] J. Adam. Virtual reality. *IEEE Spectrum*, 30(10):22–29, 1993. 2

[2] B. Afsari, R. Chaudhry, A. Ravichandran, and R. Vidal. Group action induced distances for averaging and clustering linear dynamical systems with applications to the analysis of dynamic scenes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 91

[3] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(16):1–16, 2011. 2, 6, 110, 114

[4] S. Amari and H. Nagaoka. *Methods of Information Geometry*. American Mathematical Society, 2000. 96, 102

[5] S.-i. Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2):251–276, 1998. 103

[6] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Action Classification in Soccer Videos with Long Short-Term Memory Recurrent Neural Networks. *Proceedings of the International Conference on Artificial Neural Networks*, 2010. 116

[7] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential Deep Learning for Human Action Recognition. *2nd International Workshop on Human Behavior Understanding*, 2011. 116

[8] D. Barber and W. Wiegerinck. Tractable variational structures for approximating graphical models. *Advances in Neural Information Processing Systems*, 1999. 30

[9] B. Berelson and G. A. Steiner. *Human behavior: An inventory of scientific findings*. Harcourt, Brace & World, 1964. 5, 7

[10] S. Bhattacharya. Recognition of complex events in open-source web-scale videos: a bottom up approach. *ACM International Conference on Multimedia*, 2013. 130, 144, 156

[11] S. Bhattacharya, M. M. Kalayeh, R. Sukthankar, and M. Shah. Recognition of complex events: Exploiting temporal dynamics between underlying concepts. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 117, 136, 139, 144

[12] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006. 16, 25, 30

[13] D. M. Blei and J. D. Lafferty. Dynamic topic models. *Proceedings of the International Conference on Machine Learning*, 2006. 133, 135

[14] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001. 2, 110

[15] A. F. Bobick and I. C. Society. A State-Based Approach to the Representation and Recognition of Gesture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(12):1325–1337, 1997. 2

[16] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 51, 54, 55

[17] C. Bregler. Learning and recognizing human dynamics in video sequences. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1997. 110, 114

[18] L. Buesing, J. H. Macke, and M. Sahani. Spectral learning of linear dynamics from generalised-linear observations with application to neural population data. *Advances in Neural Information Processing Systems*, 2012. 68

[19] L. Campbell and A. Bobick. Recognition of human body motion using phase space constraints. *Proceedings of the IEEE International Conference on Computer Vision*, 1995. 2, 114

[20] I. Carter. *Human Behavior in the Social Environment: A Social Systems Approach*. Transaction Publishers, 2011. 5, 7

[21] A. B. Chan and N. Vasconcelos. Probabilistic kernels for the classification of auto-regressive visual processes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 92

[22] A. B. Chan and N. Vasconcelos. Classifying video with kernel dynamic textures. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 62

[23] A. B. Chan and N. Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):909–926, 2008. 22, 68, 93

[24] A. B. Chan and N. Vasconcelos. Layered Dynamic Textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1862–1879, 2009. 30, 72, 86

[25] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15, 2009. 2

[26] C. Chang and C. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011. 130

[27] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–659, 2013. 2

[28] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 6, 7, 62, 112, 116, 133, 135

[29] O. Chomat and J. L. Crowley. Probabilistic recognition of activity using local appearance. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1999. 114

[30] R. G. Cinbis, J. Verbeek, and C. Schmid. Image categorization using fisher kernels of non-iid image models. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 97, 114

[31] M. Collins, S. Dasgupta, and R. E. Schapire. A generalization of principal component analysis to the exponential family. *Advances in Neural Information Processing Systems*, 2002. 60

[32] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 3rd edition, 2009. 35

[33] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley, New York, 2 edition, 2006. 25

[34] T. Darrell and A. Pentland. Space-time gestures. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1993. 2

[35] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977. 21, 22, 62, 64

[36] L. Deng and D. Yu. Deep Learning: Methods and Applications. *Foundations and Trends in Signal Processing*, 7(3-4):197–387, 2014. 115

[37] A. Dix, J. Finlay, G. Abowd, and R. Beale. *Human–Computer Interaction*. Prentice Hall, 2003. 2

[38] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 116

[39] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto. Dynamic textures. *International Journal of Computer Vision*, 51(2):91–109, 2003. 19, 62, 116

[40] J. Durbin and S. Koopman. *Time Series Analysis By State Space Methods*. Oxford University Press, 2001. 16

[41] J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2012. 16

[42] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008. 130

[43] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 112

[44] D. A. Forsyth and J. Ponce. *Computer Vision, A Modern Approach*. Prentice Hall, 2003. 2

[45] B. Friedland. *Control System Design: An Introduction to State-Space Methods*. Dover, 2005. 16

[46] A. Gaidon, Z. Harchaoui, and C. Schmid. Actom sequence models for efficient action detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 112, 116

[47] A. Gaidon, Z. Harchaoui, and C. Schmid. Temporal localization of actions with actoms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2782–2795, 2013. 114

[48] A. Gaidon, Z. Harchaoui, and C. Schmid. Activity representation with motion hierarchies. *International Journal of Computer Vision*, 107(3):219–238, 2014. 3

[49] Y. Gao, L. Buesing, K. V. Shenoy, and J. P. Cunningham. High-dimensional neural spike train analysis with generalized count linear dynamical systems. In *Advances in Neural Information Processing Systems*, 2015. 16, 42

[50] Z. Ghahramani and G. E. Hinton. Parameter estimation for linear dynamical systems. Technical Report CRG-TR-96-2, University of Toronto, 1996. 72, 86

[51] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, 2007. 2, 110, 114, 115, 131, 154

[52] A. Graves, A.-r. Mohamed, and G. Hinton. Speech Recognition with Deep Recurrent Neural Networks. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013. 116

[53] A. Graves and J. Schmidhuber. Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. *Advances in Neural Information Processing Systems*, 2009. 116

[54] A. Gunawardana and W. Byrne. Convergence Theorems for Generalized Alternating Minimization Procedures. *Journal of Machine Learning Research*, 6:2049–2073, 2005. 64

[55] B. Haasdonk. Feature space interpretation of svms with indefinite kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):482–492, 2005. 92

[56] H. Hajimirsadeghi, W. Yan, A. Vahdat, and G. Mori. Visual Recognition by Counting Instances: A Multi-Instance Cardinality Potential Kernel. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 150

[57] D. Haussler. Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, University of California, Santa Cruz, 1999. 100

[58] F. C. Heilbron, V. Escorcia, B. Ghanem, J. C. Niebles, and U. Norte. ActivityNet : A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 3

[59] D. Helbing and P. Molnár. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282–4286, 1995. 5

[60] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. 116

[61] P. A. Højen-Sørensen, O. Winther, and L. K. Hansen. Mean-Field Approaches to Independent Component Analysis. *Neural Computation*, 14(4):889–918, 2002. 30

[62] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79(8):2554–2558, 1982. 16

[63] N. Ikizler and D. A. Forsyth. Searching for complex human activities with no visual examples. *International Journal of Computer Vision*, 80(3):337–357, 2008. 115

[64] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. *Advances in Neural Information Processing Systems*, 1999. 95, 102

[65] T. S. Jaakkola. *Tutorial on Variational Approximation Methods*. MIT Press, 2001. 30

[66] T. S. Jaakkola and M. I. Jordan. A variational approach to bayesian logistic regression models and their extensions. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 1996. 41

[67] T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000. 41

[68] A. Jain, A. Gupta, M. Rodriguez, and L. S. Davis. Representing videos using mid-level discriminative patches. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 137

[69] M. Jain, H. Jegou, and P. Bouthemy. Better exploiting motion for better action recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 137

[70] M. Jain, J. C. van Gemert, and C. G. M. Snoek. What do 15,000 Object Categories Tell Us About Classifying and Localizing Actions? *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 117

[71] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1718, 2012. 94, 97, 100

[72] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. *Proceedings of the IEEE International Conference on Computer Vision*, 2013. 6, 117, 120

[73] S. Ji, W. Xu, M. Yang, and K. Yu. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, Jan. 2013. 7, 114, 115

[74] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo. Trajectory-based modeling of human actions with motion reference points. *Proceedings of the European Conference on Computer Vision*, 2012. 114

[75] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, 1973. 2

[76] G. Johansson. Spatio-temporal differentiation and integration in visual motion perception. *Psychological Research*, 38(4):379–393, 1976. 2

[77] G. Johansson, S. S. Bergström, and W. Epstein. *Perceiving events and objects*. L. Erlbaum Associates, 1994. 3

[78] S. Jones and L. Shao. A multigraph representation for improved unsupervised/semi-supervised learning of human actions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 137

[79] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999. 21, 25, 27, 62, 97

[80] H. J. Kappen and F. B. Rodriguez. Efficient learning in Boltzmann machines using linear response theory. *Neural Computation*, 10(5):1137–1156, 1998. 30

[81] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-Scale Video Classification with Convolutional Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2, 3, 7, 115, 129

[82] V. Kellokumpu, G. Zhao, and M. Pietikäinen. Human activity recognition using a dynamic texture based method. *Proceedings of the British Machine Vision Conference*, 2008. 6, 7, 116

[83] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 115

[84] J. Krapac, J. Verbeek, and F. Jurie. Modeling spatial layout with fisher vectors for image categorization. *Proceedings of the IEEE International Conference on Computer Vision*, 2011. 114

[85] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. *Proceedings of the IEEE International Conference on Computer Vision*, 2011. 3, 110

[86] S. Kullback. *Information Theory and Statistics*. Courier Dover Publications, 1997. 17, 18

[87] K. Lai, D. Liu, M. Chen, and S. Chang. Recognizing Complex Events in Videos by Learning Key Static-Dynamic Evidences. *Proceedings of the European Conference on Computer Vision*, 2014. 150

[88] K. Lai, F. Yu, M. Chen, and S. Chang. Video Event Detection by Inferring Temporal Instance Labels. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 5

[89] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 6, 112, 117, 120

[90] Z. Lan, X. Li, and A. G. Hauptmann. Temporal Extension of Scale Pyramid and Spatial Pyramid Matching for Action Recognition, 2014. 5, 114, 140

[91] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj. Beyond Gaussian Pyramid: Multi-skip Feature Stacking for Action Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 110, 115, 137, 143

[92] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005. 3, 7, 114, 115, 129

[93] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 3, 5, 93, 110, 114, 115, 133, 135, 136, 144

[94] B. Laxton, J. Lim, and D. Kriegman. Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 112, 116

[95] B. Li, M. Ayazoglu, T. Mao, O. Camps, and M. Sznaier. Activity recognition using dynamic subspace angles. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 6, 7, 112, 116

[96] W.-X. LI, V. Mahadevan, and N. Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):18–32, 2014. 2

[97] W.-X. LI and N. Vasconcelos. Recognizing activities by attribute dynamics. *Advances in Neural Information Processing Systems*, 2012. 20, 68, 100, 117

[98] W.-X. LI, Q. Yu, A. Divakaran, and N. Vasconcelos. Dynamic pooling for complex event recognition. *Proceedings of the IEEE International Conference on Computer Vision*, 2013. 5, 137

[99] W.-X. LI, Q. Yu, H. Sawhney, and N. Vasconcelos. Recognizing activities via bag of words for attribute dynamics. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 5, 100

[100] Y. Li, W.-X. LI, V. Mahadevan, and N. Vasconcelos. VLAD^3: Encoding Dynamics of Deep Features for Action Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 88

[101] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 6, 16, 112, 113, 114, 117, 119, 120, 130, 131, 137, 155

[102] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press, 2010. 2

[103] P. Matikainen, M. Hebert, and R. Sukthankar. Representing Pairwise Spatial and Temporal relations for action recognition. *Proceedings of the European Conference on Computer Vision*, 2010. 114

[104] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 5

[105] D. J. Moore, I. A. Essa, and M. H. H. III. Exploiting human actions and object context for recognition tasks. *Proceedings of the IEEE International Conference on Computer Vision*, 1999. 114

[106] P. J. Moreno, P. P. Ho, and N. Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. *Advances in Neural Information Processing Systems*, 2004. 92

[107] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. 16

[108] R. M. Neal and G. E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models*, 89:355–368, 1998. 26, 62

[109] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond Short Snippets: Deep Networks for Video Classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 114, 115, 116

[110] B. Ni, P. Moulin, X. Yang, and S. Yan. Motion Part Regularization: Improving Action Recognition via Trajectory Selection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 110, 115, 137, 143

[111] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. *Proceedings of the European Conference on Computer Vision*, 2010. 3, 110, 112, 114, 115, 136, 137, 139, 144, 155

[112] T. D. Nielsen and F. V. Jensen. *Bayesian Networks and Decision Graphs*. Springer-Verlag New York, 2007. 22

[113] S. Niyogi and E. Adelson. Analyzing and recognizing walking figures in xyt. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1994. 114

[114] P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Michel, A. F. Smeaton, W. Kraaij, and G. Quenot. Trecvid 2011 – an overview of the goals, tasks, data, evaluation mechanisms, and metrics. *Proceedings of TRECVID 2011*, 2011. 3, 5, 111, 143, 156

[115] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. Zero-shot learning with semantic output codes. *Advances in Neural Information Processing Systems*, 2009. 6, 120

[116] G. Parisi. *Statistical Field Theory*. Perseus Books, 1998. 30

[117] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice. 2014. 3, 7, 115

[118] X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action Recognition with Stacked Fisher Vectors. *Proceedings of the European Conference on Computer Vision*, 2014. 7, 110, 115, 129

[119] F. Perronnin, J. Sanchez, and T. Mensink. Improving the Fisher Kernel for Large-Scale Image Classification. *Proceedings of the European Conference on Computer Vision*, 2010. 114

[120] C. Peterson and J. R. Anderson. A Mean Field Theory Learning Algorithm for Neural Networks. *Complex Systems*, 1:995–1019, 1987. 30

[121] C. Pinhanez and A. Bobick. Human action detection using pnf propagation of temporal constraints. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1998. 2, 114

[122] A. Quattoni, M. Collins, and T. Darrell. Learning visual representations using images with captions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 117

[123] N. Rasiwasi, P. J. Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. *IEEE Transactions on Multimedia*, 9(5):923–938, 2007. 117, 119

[124] N. Rasiwasia and N. Vasconcelos. Scene classification with low-dimensional semantic spaces and weak supervision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 117

[125] N. Rasiwasia and N. Vasconcelos. Holistic context modeling using semantic co-occurrences. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 6, 112, 117, 120

[126] N. Rasiwasia and N. Vasconcelos. Holistic context models for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):902–917, 2012. 112, 117

[127] A. Ravichandran, R. Chaudhry, and R. Vidal. Categorizing Dynamic Textures using a Bag of Dynamical Systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):342–353, 2012. 88

[128] A. Ravichandran, R. Chaudhry, and R. Vidal. Categorizing dynamic textures using a bag of dynamical systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):342–353, 2012. 91, 135

[129] K. Reisz. *The Technique of Film Editing*. Focal Press, 2010. 9

[130] M. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 3, 110, 114

[131] S. Roweis and Z. Ghahramani. A unifying review of linear gaussian models. *Neural Computation*, 11(2):305–345, 1999. 21, 33, 39, 45, 72, 86, 122

[132] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. 2015. 115

[133] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2003. 16, 29

[134] H. Sagan. *Introduction to the Calculus of Variations*. McGraw-Hill, 1969. 54

[135] R. Salakhutdinov and G. E. Hinton. Deep Boltzmann Machines. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2009. 30

[136] L. K. Saul and M. I. Jordan. Exploiting tractable substructures in intractable networks. *Advances in Neural Information Processing Systems*, 1996. 30

[137] L. K. Saul and M. I. Jordan. Mean Field Theory for Sigmoid Belief Networks. *Journal of Artificial Intelligence Research*, 4:61–76, 1996. 30

[138] L. K. Saul and M. I. Jordan. Attractor dynamics in feedforward neural networks. *Neural Computation*, 12:1313–1335, 2000. 37

[139] A. I. Schein, L. K. Saul, and L. H. Ungar. A generalized linear model for principal component analysis of binary data. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2003. 20, 60, 61, 113

[140] K. Schindler and L. Van Gool. Action snippets: How many frames does human action recognition require? *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 3

[141] B. Schölkopf, A. Smola, and K. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998. 62

[142] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. *Proceedings of the International Conference on Pattern Recognition*, 2004. 2, 3, 5, 110, 114

[143] L. Shao, L. Liu, and M. Yu. Kernelized Multiview Projection for Robust Action Recognition. *International Journal of Computer Vision*, 2015. 115

[144] L. Shao, X. Zhen, D. Tao, and X. Li. Spatio-temporal Laplacian pyramid coding for action recognition. *IEEE Transactions on Cybernetics*, 44(6):817–827, 2014. 114

[145] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004. 100

[146] R. H. Shumway and D. S. Stoffer. An approach to time series smoothing and forecasting using the em algorithm. *Journal of Time Series Analysis*, 3(4):253–264, 1982. 19, 21, 33, 72, 86

[147] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Fisher Networks for Large-Scale Image Classification. *Advances in Neural Information Processing Systems*, 2013. 114

[148] K. Simonyan and A. Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. *Advances in Neural Information Processing Systems*, 2014. 7, 114, 115, 116, 129

[149] B. F. Skinner. *Science And Human Behavior*. Free Press, 1965. 5, 9

[150] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. *ACM International Conference on Multimedia*, 2005. 142

[151] C. Sun and R. Nevatia. Active: Activity concept transitions in video event classification. *Proceedings of the IEEE International Conference on Computer Vision*, 2013. 117, 136, 139, 144

[152] C. Sun and R. Nevatia. Discover: Discovering important segments for classification of video events and recounting. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 117

[153] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 115

[154] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, and H. Sawhney. Evaluation of low-level features and their combinations for complex event detection in open source videos. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 5, 110, 115, 142

[155] K. Tang, F.-F. Li, and D. Koller. Learning latent temporal structure for complex event detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 3, 117, 136, 139, 140, 144

[156] S. Todorovic. Human activities as stochastic kronecker graphs. *Proceedings of the European Conference on Computer Vision*, 2012. 137

[157] A. Tözeren. *Human Body Dynamics: Classical Mechanics and Human Movement*. Springer, 2000. 6

[158] A. Vahdat, K. Cannons, G. Mori, S. Oh, and I. Kim. Compositional Models for Video Event Detection: A Multiple Kernel Learning Latent Variable Approach. *Proceedings of the IEEE International Conference on Computer Vision*, 2013. 150

[159] N. Vasconcelos, P. Ho, and P. Moreno. The kullback-leibler kernel as a framework for discriminant and localized representations for visual recognition. *Proceedings of the European Conference on Computer Vision*, 2004. 92

[160] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):480–492, 2012. 130

[161] S. V. N. Vishwanathan, A. J. Smola, and R. Vidal. Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. *International Journal of Computer Vision*, 73(1):95–119, 2006. 91

[162] M. Vrigkas, C. Nikou, and I. Kakadiaris. A Review of Human Activity Recognition Methods. *Frontiers in Robotics and AI*, 2:1–28, 2015. 114

[163] M. J. Wainwright and M. I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2007. 30

[164] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013. 114

[165] H. Wang and C. Schmid. Action recognition with improved trajectories. *Proceedings of the IEEE International Conference on Computer Vision*, 2013. 3, 7, 110, 114, 115, 129, 137, 143

[166] H. Wang, M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. *Proceedings of the British Machine Vision Conference*, 2009. 5, 93, 110, 115

[167] L. Wang, Y. Qiao, and X. Tang. Action Recognition With Trajectory-Pooled Deep-Convolutional Descriptors. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 114, 115, 129

[168] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2006. 133, 135

[169] M. West, P. J. Harrison, and H. S. Migon. Dynamic generalized linear models and bayesian forecasting. *Journal of the American Statistical Association*, 80(389):73–83, 1985. 16

[170] J. Winn and C. M. Bishop. Variational message passing. *Journal of Machine Learning Research*, 6:661–694, 2005. 30

[171] E. Xing, M. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2003. 30

[172] D. Xu and S.-f. Chang. Video event recognition using kernel methods with multilevel temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1985–97, 2008. 2

[173] Z. Xu, I. Tsang, Y. Yang, Z. Ma, and A. Hauptmann. Event Detection using Multi-Level Relevance Labels and Multiple Features. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 150

[174] Y. Yacoob and M. J. Black. Parameterized modeling and recognition of activities. *Proceedings of the IEEE International Conference on Computer Vision*, 1998. 114

[175] G. Ye, D. Liu, I.-h. Jhuo, and S.-f. Chang. Robust late fusion with rank minimization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2012. 142

[176] M. Yu, L. Liu, and L. Shao. Structure-Preserving Binary Representations for RGB-D Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015. 114

[177] Q. Yu, J. Liu, H. Cheng, A. Divakaran, and H. Sawhney. Multimedia event recounting with concept based representation. *ACM International Conference on Multimedia*, 2012. 151

[178] J. Zhang. The Mean Field Theory in EM Procedures for Markov Random Fields. *IEEE Transactions on Signal Processing*, 40(10):2750–2583, 1992. 30