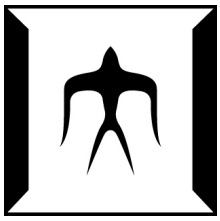
論文 / 著書情報
Article / Book Information

題目(和文)	変化検知のための統計的機械学習
Title(English)	Statistical Machine Learning Approaches to Change Detection
著者(和文)	柳松
Author(English)	Song Liu
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第9555号, 授与年月日:2014年3月26日, 学位の種別:課程博士, 審査員:杉山　将,秋山　泰,篠田　浩一,村田　剛志,藤井　敦
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第9555号, Conferred date:2014/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Statistical Machine Learning Approaches to Change Detection

Song Liu

December 2013



Department of Computer Science
Graduate School of Information Science and Engineering
Tokyo Institute of Technology

Thesis Committee:
Masashi Sugiyama, Chair
Yutaka Akiyama
Koichi Shinoda
Tsuyoshi Murata
Atsushi Fujii

Submitted in partial fulfillment of
the requirements for the degree of

Doctor of Engineering

Keywords: Change Detection, Machine Learning, Density Ratio Estimation

To my parents and grandmother.

Abstract

The development of modern technologies has offered us easier ways of accessing and modifying digital data. The analysis on such changing data challenges our traditional view on machine learning, since the pattern observed today may be altered tomorrow. In contrast, the dynamic view of machine learning allows us to incorporate changing patterns in traditional learning tasks.

In this thesis, we focus on one of the dynamic learning tasks: unsupervised change detection, and propose two novel approaches in distributional and structural change detection respectively. Guided by *Vapnik's Principle* (Vapnik, 1998), both algorithms avoid comparing two separately learned patterns by learning the change directly.

Our first contribution is on non-parametric distributional change detection. We propose a novel statistical change-point detection algorithm based on non-parametric divergence estimation between time-series samples from two retrospective segments. Our method uses the *relative Pearson divergence* as a divergence measure, and it is accurately estimated by a method of density ratio estimation. Through experiments of human-activity sensing, speech and Twitter messages, we demonstrate the usefulness of the proposed method.

Our second contribution is on structural change detection between two Markov networks. We propose a new method of detecting changes by estimating the ratio between Markov network models, instead of learning two Markov networks separately. This density ratio formulation naturally allows us to introduce sparsity in the structural change, which highly contributes to enhancing the interpretability. Furthermore, the computation of the normalization term, which is a critical bottleneck, can be efficiently sample-approximated. Finally, we give the dual objective of the proposed method, which reduces the computational cost on large Markov

networks. The usefulness of the proposed method is examined via toy and real experiments.

As it is shown in this thesis, the unsupervised change detection can successfully capture the changes of patterns in many real-world applications. We believe it will be a very promising field of machine learning in the coming years.

Acknowledgments

First, I would like to express my sincere appreciation to Professor Sugiyama. Without his patient supervision and insightful advice, none of my academic works may be developed and published. I am also very grateful to Professor Akiyama, Professor Shinoda, Professor Fujii, Professor Murata for reviewing and evaluating my thesis.

Second, I would like to thank my funding organizations: JSPS Global COE CompView program, JST PRESTOR program, and JSPS fellowship. Without their generous financial help, I would not be able to support my own living in Japan.

Finally, I am deeply indebted to my family and friends. Without their understanding and support, I can never motivate myself for pursuing a Ph.D. This thesis is dedicated to my mother, father and especially my grandmother, who has passed away during my study in Japan.

I owe many thanks to my (ex-)labmates: Dr. Makoto Yamada, Dr. Hirotaka Hachiya, Dr. Ning Xie, Dr. Gang Niu, Tingting Zhao, Marthinus Christoffel du Plessis, Akihiro Yamashita, from whom I have learnt so much about doing scientific research.

Thanks to my friends, I really lived a happy and enjoyable life in Japan.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis is dedicated to learning the *changes* hidden behind data with statistical machine learning approaches. In this chapter, we explore the fundamental ideas of statistical machine learning and locate our research in this field. The structure of this thesis is illustrated at the end of this section.

1.1 Learning from Data

Data are empirical observations generated by human interactions with the environment. *Learning from data* is a process of extracting *patterns* from such observations (Shawe-Taylor and Cristianini, 2004; Bishop, 2006). It has played an important role in developing human knowledge. One of the great works by Johannes Kepler, unravelled the laws govern planetary bodies moving in the solar system based on Tycho Brahe's data, was an example that how *human learning* revolutionized our understanding of the universe.

With the latest technology, the environment accessible by human has been greatly expanded, and data is now produced and updated at a tremendous speed (Jacobs, 2009). Clearly learning at such a scale is beyond human capability. In this thesis, we concentrate on another type of learning, in which the learning body is switched from human to computers. We refer to such learning process as *machine learning*.

We summarize the procedure of learning in Figure 1.1.

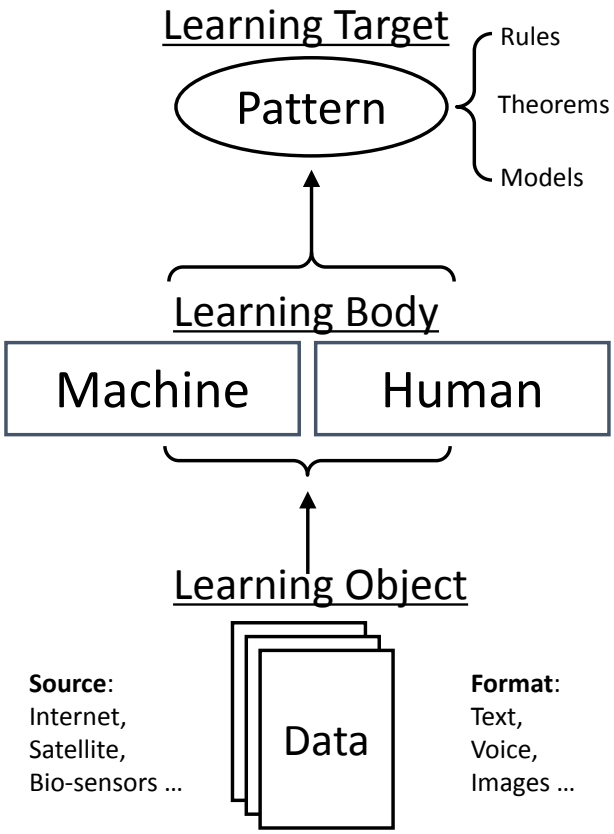


Figure 1.1: The diagram of learning procedure.

1.1.1 Machine Learning from Big Data

Data, as the main object of learning, keeps growing in its scale. For example, one hour of video is uploaded to YouTube every second[1], 1 trillion webpages are indexed by Google[2] and 50 billion photos are handled by Facebook[3] from its users. Without interpretation, such an amount of information is too vast for human to comprehend.

In fact, data is redundant. Imagine a data feed received from an accelerometer

[1] http://www.youtube.com/t/press_statistics.

[2] http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html.

[3] https://www.facebook.com/note.php?note_id=409881258919

sensor of a marathon runner, with sampling rate 30 Hz. We may obtain 1800 data frames per minute. However, it is clear that the data at the current second is not completely irrelevant with the data at the previous second. The pattern of the marathon runner can be expressed through a sequence of transitions where the running behaviour has shifted. Thus, the entire data stream can be compressed into an initial state followed by a series of transition states.

Patterns are regularities that characterise data in compressed rules, and help us understand the mechanism that generates data, make decisions or even predict future (Shawe-Taylor and Cristianini, 2004; Murphy, 2012). However, compressed rules do not necessarily mean "simple".

Most of the human knowledge of physics can be written down explicitly as equations, but it is not true for learned patterns. Consider the task of identifying genes from fragments of human DNA. Though biological laws give us clues, there is no way to find an exact procedure to accomplish such a task. However, given annotated samples from data, it is possible to adapt such information in samples and apply it to new queries. Though this is hard for a human expert, computers are capable of handling complex solutions and making inference.

Though we are entering the era where "data floods", with the help of advanced computing power, learning complex patterns from big data now become possible.

1.1.2 Learning under Uncertainty

There are possibilities that *uncertainties* are recorded within the data. Generally, there are three types of uncertainties concerned during the learning process:

Noise: The term "big data" only refers to the *quantity* of data, rather than its *quality*. On Internet, the owners of information tend publish it as soon as possible so the preciseness of the data is usually compromised.

Outlier (or Missing Data) Due to the failure of instruments, corrupted records or missing out values may frequently occur in the data.

Model Unpredictability In some cases, the data generating source contains certain level of uncertainty. Consider a movie rating dataset collected from amateur

viewers who may rate the same movie with different scores on different days.

With uncertainty lies within, data cannot be described by *exact* patterns. Therefore, we may look for a *good* pattern that fits the data with high probability. If such a model is used for prediction, there is a good chance that such prediction is fulfilled.

Statistical methods help us build such models where the uncertainty is treated as *probability*. Comparing to deterministic models, statistical models are more compact, since the deterministic models require more rules to describe exceptional cases while the statistical models just assign them with lower probability using one single model.

If statistical methods are employed, we refer to such machine learning processes as *Statistical Machine Learning*.

1.1.3 Learning with Changing Data

The modern age has not only produced a large amount of data, but also offers easy ways of accessing and updating data. In fact, *change* is one of the most important properties of big data. For example, satellite images taken on the same spot may be different due to lighting or clouding conditions; the trend of Twitter topics may rapidly shift after breaking news is reported; popular queries sent to search engines may change on an everyday basis. In fact, twenty percent of Google queries everyday have never been seen before [4].

However, the classic machine learning methods are designed for *static* data which remain the same during the learning period. Unfortunately, the patterns obtained today may be invalid tomorrow.

Moreover, it may be more interesting to know the *change of patterns* than the pattern itself in many occasions. As change is the nature of big data, analysing the static pattern only gives the impression what data *used to be*.

After all,

"*Nothing endures but change.*"

-Heraclitus (c. 535 - c. 475 BCE)

[4] `http://certifiedknowledge.org/blog/.....`
`/are-search-queries-becoming-even-more-unique-statistics-from-google.`

1.2 Two Views of Machine Learning

As it was introduced in Section 1.1.3, the changing data has brought analysis tasks huge challenge, while the traditional machine learning only focuses on extracting static rules. There is a need of shifting the idea of machine learning into a more adaptive paradigm so that the change of patterns can be naturally incorporated in the learning framework.

In this section, we discuss machine learning methodologies from two different perspectives.

1.2.1 Static Machine Learning

Traditional machine learning algorithms take a set of instances as their input and discover patterns lies among data (Bishop, 2006). Instances usually consist of *features*, which characterize different aspects of the data. For example, an automated diagnosis system may take each patient as an instance, and their medical examination records as features. As side information, the learning target, *supervision* can also be provided as one of the features. For example, the aforementioned system may use previous medical decisions as supervision to diagnose new patients.

However, in some occasions, such target information may not be available from the data, which makes the task ill-defined. We now introduce two tasks in machine leaning with respect to the availability of their target information.

Supervised Learning Supervised learning is usually divided into two steps: *training* and *testing*. Let the $\{(\boldsymbol{x}, y)\}$ be the set of training instances of our learning algorithm. $y \in \mathbb{R}$ is the learning target and $\boldsymbol{x} \in \mathbb{R}^d$ are predictive features. a learning algorithm is looking for a relationship $f : \mathbb{R}^d \mapsto \mathbb{R}$, so that for each testing instance $\boldsymbol{x}'$, we may assign a prediction $\widehat{y}'$.

From the probabilistic perspective, we assume each instance $(\boldsymbol{x}, y)$ is drawn from a joint probability distribution $p(\boldsymbol{x}, y)$, thus supervised learning is to build a model of *conditional probability density* $p(y|\boldsymbol{x})$ (Hastie et al., 2001).

In this thesis, we only consider continuous instance values. However, patterns can also be learned when the data is discrete. If y have discrete values, for example, $y \in \{-1, 1\}$, such a learning task is referred to *classification*, otherwise, we

call it *regression*.

Unsupervised Learning Unsupervised learning considers a scenario without a specific learning target. The task is simply discovering "interesting" patterns from data. From the view of probability, we may formulate such a task as estimating the density function $p(\boldsymbol{x})$ from training samples $\{\boldsymbol{x}\}$ so that the estimated density $\hat{p}(\boldsymbol{x})$ also holds on testing samples $\{\boldsymbol{x}'\}$ (Murphy, 2012). In fact, once the generating density is found, we may recover any other statistical patterns from it.

From the above discussion, it can be seen that the supervised learning is easier than its unsupervised counterpart. Supervised learning is to obtain an estimate $\hat{p}(y|\boldsymbol{x})$ of the true conditional density $p(y|\boldsymbol{x})$. Consider an unnormalized conditional density model $g(y|\boldsymbol{x};\boldsymbol{\theta})$, we have

$$p(y|\boldsymbol{x}) \approx \hat{p}(y|\boldsymbol{x};\boldsymbol{\theta}) = \frac{g(y|\boldsymbol{x};\boldsymbol{\theta})}{\int_{y\in\mathbb{R}} g(y|\boldsymbol{x};\boldsymbol{\theta})\ \mathrm{d}y},$$

For unsupervised learning, we model $p(\boldsymbol{x})$ using $g(\boldsymbol{x};\boldsymbol{\theta})$, thus

$$p(\boldsymbol{x}) \approx \hat{p}(\boldsymbol{x};\boldsymbol{\theta}) = \frac{g(\boldsymbol{x};\boldsymbol{\theta})}{\int_{\boldsymbol{x}\in\mathbb{R}^d} g(\boldsymbol{x};\ \boldsymbol{\theta})\ \mathrm{d}\boldsymbol{x}}.$$

Clearly normalizing an unsupervised model is more difficult than the supervised model since the integral is in d-dimensional space $\mathbb{R}^d$ rather than one-dimensional space $\mathbb{R}$. The difficulty of normalizing unsupervised model will re-emerge in Chapter 3.

1.2.2 Dynamic Machine Learning and Change Detection

In this section, we introduce a *dynamic* machine learning paradigm, where the change of data is taken into account. Comparing to the static paradigm, such dynamic approaches offer better performance on "shifting data".

Dynamic learning approaches are extensions of traditional approaches. Given the availability of the target feature, it can also be divided as supervised and unsupervised learning.

Supervised Learning (or Transfer Learning) Similarly to traditional supervised learning, the learning algorithm is given a set of training and testing instances each. However, we assume that samples $\{(\boldsymbol{x}, y)\}$ and $\{(\boldsymbol{x}', y')\}$ are drawn from two *different* distributions $p(\boldsymbol{x}, y)$ and $p'(\boldsymbol{x}, y)$ respectively. The target is to learn a model of $p(y|\boldsymbol{x})$ from $\{(\boldsymbol{x}, y)\}$ and have a good accuracy in predicting y' given $\boldsymbol{x}'$. It is a hard problem to learn if two joint distributions $p(\boldsymbol{x}, y)$ and $p'(\boldsymbol{x}, y)$ are completely different, so we assume that only the marginal or conditional distributions are different. If $p(y)$ and $p'(y)$ are different, we consider such problem as learning under *class-prior change* (Saerens et al., 2002; Du Plessis and Sugiyama, 2014); if $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ are different, it is called learning under *covariate shift* (Bickel et al., 2009; Sugiyama et al., 2007). In a generalized setting, the $\{(\boldsymbol{x}, y)\}$ and $\{(\boldsymbol{x}', y')\}$ are not necessarily training and testing datasets, but can be two different learning tasks. If $p(y|\boldsymbol{x})$ and $p'(y|\boldsymbol{x})$ are not exactly the same, but closely related, it is regarded as *multi-task learning*, or *inductive transfer learning* (Raina et al., 2006).

Unsupervised Learning (or Change Detection) The traditional unsupervised learning is to learn interesting patterns from data. However, the unsupervised learning under changing data is to find "interesting changes", or detecting *changes between patterns*. Such a learning task usually involves two sets of data $\{\boldsymbol{x}^P\}$ and $\{\boldsymbol{x}^Q\}$, drawn from $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ respectively. In order to find changes, we have to introduce the measure of changes. In statistics, *divergence* is defined as distance between two distributions (Amari and Nagaoka, 2000; Eguchi and Copas, 2006; Wang et al., 2009). For example, an f-divergence (Ali and Silvey, 1966) is defined as

$$\mathrm{D}[p\|q] = \int q(\boldsymbol{x}) f(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})})\ \mathrm{d}\boldsymbol{x},$$

where f is a convex function and $f(1) = 0$. Other distances are also used as measure of changes, for example, L^2-distance

$$L^2(p, q) = \int \left(p(\boldsymbol{x}) - q(\boldsymbol{x})\right)^2\ \mathrm{d}\boldsymbol{x}.$$

In the above cases, $p(\boldsymbol{x})/q(\boldsymbol{x})$ (or $p(\boldsymbol{x}) - q(\boldsymbol{x})$) describes the changes on the data and we would hope that our model can capture such information. A naive way

is to model $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ separately as $p(\boldsymbol{x};\boldsymbol{\theta})$ and $q(\boldsymbol{x};\boldsymbol{\theta})$ then plug estimated densities into the ratio or difference. However, another way is to model the *change* directly, i.e.

$$\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} := r(\boldsymbol{x};\boldsymbol{\theta}), \quad \text{or} \quad p(\boldsymbol{x}) - q(\boldsymbol{x}) := w(\boldsymbol{x};\boldsymbol{\theta}).$$

In this thesis, we discuss the advantages of the latter approach. An intuitive proof of such criteria can be given by *Vapnik's Principle* (Vapnik, 1998):

> *"When solving a problem of interest, one should not solve a more general problem as an intermediate step."*
>
> -Vladimir Vapnik

However, it should be noticed that we do not claim the direct modelling approach is the *universally* good approach for all applications. In fact, modelling $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ separately may provide useful information on the generating source itself, which is highly beneficial in tasks requires a *description* of the data (e.g. outlier detection).

We summarize two views of machine learning tasks in Table 1.1.

	Static Machine Learning		Dynamic Machine Learning	
	Supervised	**Unsupervised**	**Supervised**	**Unsupervised**
Goal	Predicting future	Learning patterns	Predicting with shifting data	Learning changes of patterns
Statistical Obj.	$p(y\|\boldsymbol{x})$	$p(\boldsymbol{x})$	$p'(y\|\boldsymbol{x})$ when $p(\boldsymbol{x}, y) \neq p'(\boldsymbol{x}, y)$	$p(\boldsymbol{x})/q(\boldsymbol{x})$ or $p(\boldsymbol{x}) - q(\boldsymbol{x})$ when $p(\boldsymbol{x}) \neq q(\boldsymbol{x})$
Learning Tasks	Classification, regression	Clustering, dimensionality reduction	Learning under covariate shift, class-prior change, inductive transfer	Distributional change detection, structural change detection
Examples	Face recognition, stock prediction	News categorization, data visualization	Robotic motion planning in a changing environment	Satellite image change detection, heartbeat anomaly detection

Table 1.1: The comparison of two views of machine learning.

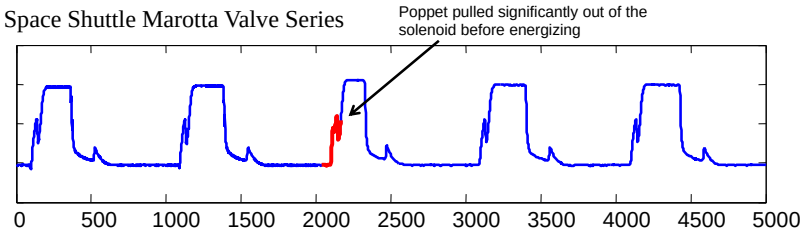


Figure 1.2: The example of space shuttle valve dataset. The red segment is an annotated anomaly (Keogh et al., 2005).

1.2.3 Change Detection Problems

Depending on the motivations of change detection, we divide change detection problems into two categories: distributional change detection and structural change detection.

Distributional Change Detection Given two sets of data, the target of distributional change detection is to determine whether the data generating probability has altered. Always given such two sets of data at any time point, the *change-point detection* problem is to assign a score that indicates how likely a change has happened at the current point.

For example, the detection of anomalies from space shuttle valve (Figure 1.2) raises an alarm when the engine malfunctioning is detected (Keogh et al., 2005).

Structural Change Detection In practice, the input data is usually high-dimensional, with multiple input variables. Variables may be correlated with each other. For example, fMRI data may record correlated activities at different brain regions. Given two sets of multivariate data, it may be interesting to understand what has been changed in such correlations.

The target of structural change detection is to interpret the change of *interactions* among variables from two sets of data.

For example, in controlled experiments of gene profiling, the change of interactions between genes may reveal crucial information that how system responds to external stimuli (Zhang and Wang, 2010).

1.3 Contribution of This Thesis

We contribute two works to change detection problems: Robust nonparametric distributional change detection and interpretable structural change detection.

1.3.1 Two Issues of Change Detection

In order to detect changes, most of the existing methods assume the probability model that generates data is from a given family of distributions, however, in some occasions, the detection of change needs to be made under a highly noisy environment. The generating probability is not from any of the known parametric distribution families. In comparison, the non-parametric detection methods make no assumption on the underlying probability, and benefit from the flexibility of capturing various kinds of changes. Unfortunately, such non-parametric methods usually require large number of samples to achieve an accurate estimation, especially under high-dimensional settings.

In some applications, we are not only interested in detecting changes, but also understanding changes. Though non-parametric methods are flexible, they are less informative in showing how the generating probability has shifted from one to another. Structural change detection is to explore the changes among the correlations between random variables. Existing methods assume Gaussianity on the generating probability. However, such model might be too restrictive in practice. Moreover, to obtain the estimation of model parameters, the existing approaches adopt learning procedures twice on two sets of data separately. In fact, given the change is our only interest, such method is redundant and less accurate.

1.3.2 Robust Nonparametric Distributional Change Detection

In the first part of this thesis, we tackle the issue of improving the robustness of change detector, under non-parametric settings. The resulting algorithm utilizes the latest advances in density ratio estimation and improves the performance of change-point detection via the technique of "relative density ratio estimation" (Yamada et al., 2013).

We present a novel statistical change-point detection algorithm based on non-

parametric divergence estimation between time-series samples from two retrospective segments. Our method uses the *relative Pearson divergence* as a divergence measure, and it is accurately and efficiently estimated by a method of direct density-ratio estimation. Through experiments on artificial and real-world datasets including human-activity sensing, speech, and Twitter messages, we demonstrate the usefulness of the proposed method.

1.3.3 Interpretable Structural Change Detection

In the second part of this thesis, we consider the problem of structural change detection, and present a direct learning method that solves the problem in one shot.

We propose a new method for detecting changes in Markov network structure between two sets of samples. Instead of naively fitting two Markov network models separately to the two data sets and figuring out their difference, we *directly* learn the network structure change by estimating the ratio of Markov network models. This density-ratio formulation naturally allows us to introduce sparsity in the network structure change, which highly contributes to enhancing interpretability. Furthermore, computation of the normalization term, which is a critical computational bottleneck of the naive approach, can be remarkably mitigated. We also give the dual formulation of the optimization problem, which further reduces the computation cost for large-scale Markov networks. Through experiments, we demonstrate the usefulness of our method.

1.4 Organization of This Thesis

This dissertation consists of 4 chapters (see Figure 1.3). In this section we describe the organization of our thesis.

In Chapter 2, we introduce our first contribution on change-point detection. Section 2.1 gives a brief introduction of change-point detection problems and state-of-the-art methods. Section 2.2 formulates the problem of retrospective change-point detection and Section 2.3 reviews the latest development of density ratio estimation methods, which are main building blocks of the proposed

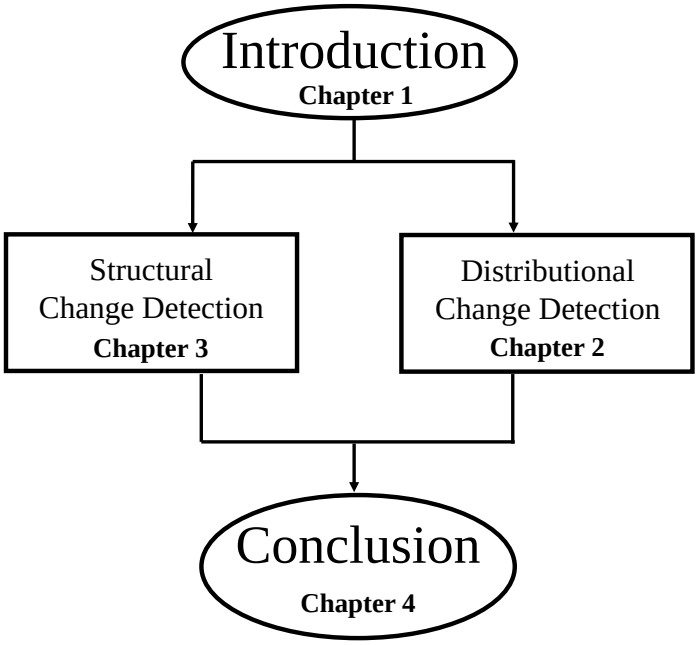


Figure 1.3: Organization of this thesis.

methods. Section 2.4 shows the experiments on both toy and real-world datasets, and the superiority of the proposed method over the other existing methods. We conclude this work in Section 2.5.

Chapter 3 covers our work on structural change detection. Section 3.1 shows the concept of structural change learning and state-of-the-art methods. In Section 3.2 we formulate *pairwise Markov network* and introduce related works. In Section 3.3, we propose the direct structural change learning method, using density ratio estimation. We demonstrate the experiment results comparing with existing methods at Section 3.4 and 3.5 using both toy and real-world datasets, and conclude this work in Section 3.7.

Finally, we summaries this thesis and show possible future works in Chapter 4.

Chapter 2

Distributional Change Detection

In this section, we propose a novel non-parametric method for distributional change detection tasks. We apply the latest advances in density ratio estimation, to obtain a flexible and robust algorithm. The usefulness of the proposed method is demonstrated via abundant experiments.

This section is organized as follows. Section 2.1 and 2.2 give a brief introduction of background and problem formulation. Section 2.3 describes the proposed methods together with the review of the existing method. Finally, the results of experiments are reported in Section 2.4 and we conclude this research in Section 2.5.

2.1 Introduction

Detecting abrupt changes in time-series data, called *change-point detection*, has attracted researchers in the statistics and data mining communities for decades (Basseville and Nikiforov, 1993; Gustafsson, 2000; Brodsky and Darkhovsky, 1993). Depending on the delay of detection, change-point detection methods can be classified into two categories: *Real-time detection* (Adams and MacKay, 2007; Garnett et al., 2009; Paquet, 2007) and *retrospective detection* (Basseville and Nikiforov, 1993; Takeuchi and Yamanishi, 2006; Moskvina and Zhigljavsky, 2003a).

Real-time change-point detection targets applications that require immediate responses such as robot control. On the other hand, although retrospective

change-point detection requires longer reaction periods, it tends to give more robust and accurate detection. Retrospective change-point detection accommodates various applications that allow certain delays, for example, climate change detection (Reeves et al., 2007), genetic time-series analysis (Wang et al., 2011), signal segmentation (Basseville and Nikiforov, 1993), and intrusion detection in computer networks (Yamanishi et al., 2000). In this chapter, we focus on the retrospective change-point detection scenario and propose a novel non-parametric method.

Having been studied for decades, some pioneer works demonstrated good change-point detection performance by comparing the probability distributions of time-series samples over past and present intervals (Basseville and Nikiforov, 1993). As both the intervals move forward, a typical strategy is to issue an alarm for a change point when the two distributions are becoming significantly different. Various change-point detection methods follow this strategy, for example, the *cumulative sum* (Basseville and Nikiforov, 1993), the *generalized likelihood-ratio method* (Gustafsson, 1996), and the *change finder* (Takeuchi and Yamanishi, 2006). Such a strategy has also been employed in novelty detection (Guralnik and Srivastava, 1999) and outlier detection (Hido et al., 2011).

Another group of methods that have attracted high popularity in recent years is the *subspace* methods (Moskvina and Zhigljavsky, 2003a,b; Ide and Tsuda, 2007; Kawahara et al., 2007). By using a pre-designed time-series model, a subspace is discovered by principal component analysis from trajectories in past and present intervals, and their dissimilarity is measured by the distance between the subspaces. One of the major approaches is called *subspace identification*, which compares the subspaces spanned by the columns of an *extended observability matrix* generated by a state-space model with system noise (Kawahara et al., 2007). Recent efforts along this line of research have led to a computationally efficient algorithm based on *Krylov subspace learning* (Ide and Tsuda, 2007) and a successful application of detecting climate change in south Kenya (Itoh and Kurths, 2010).

The methods explained above rely on pre-designed parametric models, such as underlying probability distributions (Basseville and Nikiforov, 1993; Gustafsson, 1996), auto-regressive models (Takeuchi and Yamanishi, 2006), and state-space

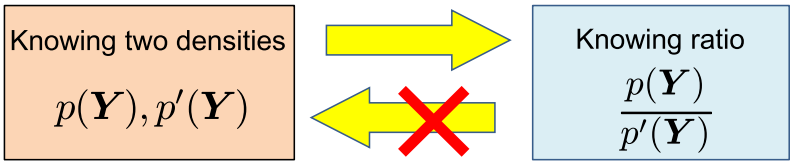


Figure 2.1: Rationale of direct density-ratio estimation.

models (Moskvina and Zhigljavsky, 2003a,b; Ide and Tsuda, 2007; Kawahara et al., 2007), for tracking specific statistics such as the mean, the variance, and the spectrum. As alternatives, non-parametric methods such as *kernel density estimation* (Csörgö and Horváth, 1988; Brodsky and Darkhovsky, 1993) are designed with no particular parametric assumption. However, they tend to be less accurate in high-dimensional problems because of the so-called *curse of dimensionality* (Bellman, 1961; Vapnik, 1998).

To overcome this difficulty, a new strategy was introduced recently, which estimates the *ratio* of probability densities directly without going through density estimation (Sugiyama et al., 2012a). The rationale of this density-ratio estimation idea is that knowing the two densities implies knowing the density ratio, but not vice versa; knowing the ratio does not necessarily imply knowing the two densities because such decomposition is not unique (Figure 2.1). Thus, direct density-ratio estimation is substantially easier than density estimation (Sugiyama et al., 2012a). Following this idea, methods of direct density-ratio estimation have been developed (Sugiyama et al., 2012b), e.g., *kernel mean matching* (Gretton et al., 2009), the *logistic-regression method* (Bickel et al., 2007), and the *Kullback-Leibler importance estimation procedure* (KLIEP) (KLI). In the context of change-point detection, KLIEP was reported to outperform other approaches (Kawahara and Sugiyama, 2012) such as the *one-class support vector machine* (Schölkopf et al., 2001; Desobry et al., 2005) and *singular-spectrum analysis* (Moskvina and Zhigljavsky, 2003b). Thus, change-point detection based on direct density-ratio estimation is promising.

The goal of this paper is to further advance this line of research. More specifically, our contributions in this chapter are two folds. The first contribution is to apply a recently-proposed density-ratio estimation method called the *unconstrained least-squares importance fitting* (uLSIF) (Kanamori et al., 2009) to change-point

detection. The basic idea of uLSIF is to directly learn the density-ratio function in the least-squares fitting framework. Notable advantages of uLSIF are that its solution can be computed analytically (Kanamori et al., 2009), it achieves the optimal non-parametric convergence rate (Kanamori et al., 2012b), it has the optimal numerical stability (Kanamori et al., 2013), and it has higher robustness than KLIEP (Sugiyama et al., 2012b). Through experiments on a range of datasets, we demonstrate the superior detection accuracy of the uLSIF-based change-point detection method.

The second contribution of this paper is to further improve the uLSIF-based change-point detection method by employing a state-of-the-art extension of uLSIF called *relative uLSIF* (RuLSIF) (Yamada et al., 2013). A potential weakness of the density-ratio based approach is that density ratios can be unbounded (i.e., they can be infinity) if the denominator density is not well-defined. The basic idea of RuLSIF is to consider *relative density ratios*, which are smoother and always bounded from above. Theoretically, it was proved that RuLSIF possesses a superior non-parametric convergence property than plain uLSIF (Yamada et al., 2013), implying that RuLSIF gives an even better estimate from a small number of samples. We experimentally demonstrate that our RuLSIF-based change-point detection method compares favorably with other approaches.

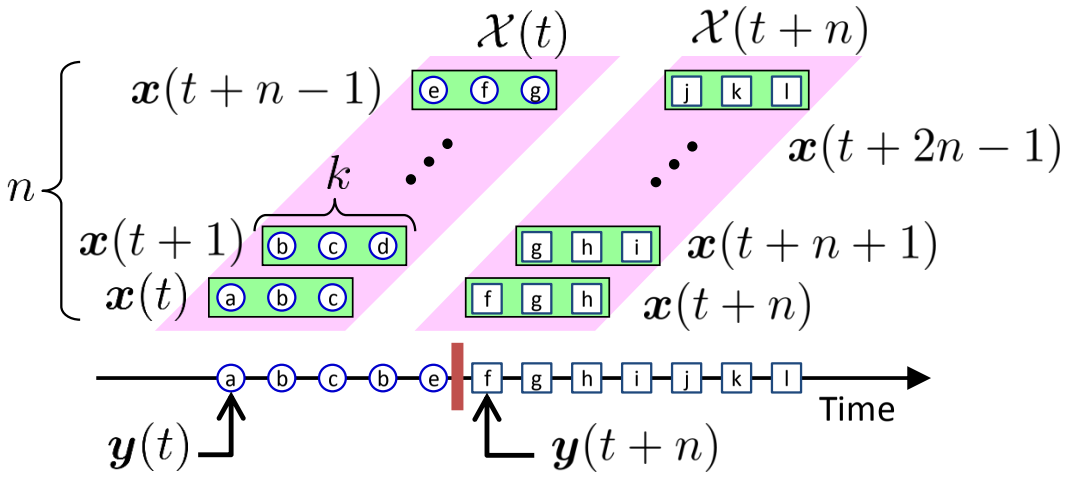
The rest of this paper is structured as follows: In Section 2.2, we formulate our change-point detection problem. In Section 2.3, we describe our proposed change-point detection algorithms based on uLSIF and RuLSIF, together with the review of the KLIEP-based method. In Section 2.4, we report experimental results on various artificial and real-world datasets including human-activity sensing, speech, and Twitter messages from February 2010 to October 2010. Finally, in Section 2.5, conclusions together with future perspectives are stated.

2.2 Problem Formulation

In this section, we formulate our change-point detection problem.

Let $\boldsymbol{y}(t) \in \mathbb{R}^d$ be a d-dimensional time-series sample at time t. Let

$$\boldsymbol{x}(t) := [\boldsymbol{y}(t)^\top, \boldsymbol{y}(t+1)^\top, \ldots, \boldsymbol{y}(t+k-1)^\top]^\top \in \mathbb{R}^{dk}$$



$\boldsymbol{y}(t)$: d-dimensional time-series sample at time t

$\boldsymbol{x}(t)$: subsequence of k time-series samples at time t

$\mathcal{X}(t)$: set of n retrospective subsequences:
$\mathcal{X}(t) := \{\boldsymbol{x}(t), \boldsymbol{x}(t+1), \ldots, \boldsymbol{x}(t+n-1)\}$

Figure 2.2: An illustrative example of notations on one-dimensional time-series data.

be a "subsequence"[1] of time series at time t with length k, where $^\top$ represents the transpose. Following the previous work (Kawahara and Sugiyama, 2012), we treat the subsequence $\boldsymbol{x}(t)$ as a sample, instead of a single d-dimensional time-series sample $\boldsymbol{y}(t)$, by which time-dependent information can be incorporated naturally (see Figure 2.2). Let $\mathcal{X}(t)$ be a set of n retrospective subsequence samples starting at time t:

$$\mathcal{X}(t) := \{\boldsymbol{x}(t), \boldsymbol{x}(t+1), \ldots, \boldsymbol{x}(t+n-1)\}.$$

Note that $[\boldsymbol{x}(t), \boldsymbol{x}(t+1), \ldots, \boldsymbol{x}(t+n-1)] \in \mathbb{R}^{dk \times n}$ forms a *Hankel matrix* and plays a key role in change-point detection based on subspace learning (Moskvina and Zhigljavsky, 2003a; Kawahara et al., 2007).

For change-point detection, let us consider two consecutive segments $\mathcal{X}(t)$ and $\mathcal{X}(t+n)$. Our strategy is to compute a certain dissimilarity measure between $\mathcal{X}(t)$ and $\mathcal{X}(t+n)$, and use it as the plausibility of change points. More specifically, the higher the dissimilarity measure is, the more likely the point is a change point[2].

Now the problems that need to be addressed are what kind of dissimilarity measure we should use and how we estimate it from data. We will discuss these issues in the next section.

2.3 Change-Point Detection via Density-Ratio Estimation

In this section, we first define our dissimilarity measure, and then show methods for estimating the dissimilarity measure.

[1] In fact, only in the case of one-dimensional time-series, $\boldsymbol{x}(t)$ is a subsequence. For higher-dimensional time-series, $\boldsymbol{x}(t)$ concatenates the subsequences of all dimensions into a one-dimensional vector.

[2] Another possible formulation is to compare distributions of samples in $\mathcal{X}(t)$ and $\mathcal{X}(t+n)$ in the framework of *hypothesis testing* (Henkel, 1976). Although this gives a useful threshold to determine whether a point is a change point, computing the p-value is often time consuming, particularly in a non-parametric setup (Efron and Tibshirani, 1993). For this reason, we do not take the hypothesis testing approach in this chapter, although it is methodologically straightforward to extend the proposed approach to the hypothesis testing framework.

2.3.1 Divergence-Based Dissimilarity Measure and Density-Ratio Estimation

In this chapter, we use a dissimilarity measure of the following form:

$$D(P_t\|P_{t+n}) + D(P_{t+n}\|P_t), \tag{2.1}$$

where P_t and P_{t+n} are probability distributions of samples in $\mathcal{X}(t)$ and $\mathcal{X}(t+n)$, respectively. $D(P\|Q)$ denotes the *f-divergence* (Ali and Silvey, 1966; Csiszár, 1967):

$$D(P\|Q) := \int q(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}, \tag{2.2}$$

where f is a convex function such that $f(1) = 0$, and $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ are probability density functions of P and Q, respectively. We assume that $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ are strictly positive. Since the f-divergence is asymmetric (i.e., $D(P\|Q) \neq D(Q\|P)$), we symmetrize it in our dissimilarity measure (2.1) for all divergence-based methods[3].

The f-divergence includes various popular divergences such as the *Kullback-Leibler (KL) divergence* by $f(t) = t\log t$ (Kullback and Leibler, 1951) and the *Pearson (PE) divergence* by $f(t) = \frac{1}{2}(t-1)^2$ (Pearson, 1900):

$$\mathrm{KL}(P\|Q) := \int p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}, \tag{2.3}$$

$$\mathrm{PE}(P\|Q) := \frac{1}{2}\int q(\boldsymbol{x}) \left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} - 1\right)^2 \mathrm{d}\boldsymbol{x}. \tag{2.4}$$

Since the probability densities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ are unknown in practice, we cannot directly compute the f-divergence (and thus the dissimilarity measure). A naive way to cope with this problem is to perform density estimation and plug the estimated densities $\widehat{p}(\boldsymbol{x})$ and $\widehat{q}(\boldsymbol{x})$ in the definition of the f-divergence. However, density estimation is known to be a hard problem (Vapnik, 1998), and thus such a plug-in approach is not reliable in practice.

[3] In the previous work (Kawahara and Sugiyama, 2012), the asymmetric dissimilarity measure $D(P_t||P_{t+n})$ was used. As we numerically illustrate in Section 2.4, the use of the symmetrized divergence contributes highly to improving the performance. For this reason, we decided to use the symmetrized dissimilarity measure (2.1).

Recently, a novel method of divergence approximation based on *direct density-ratio estimation* was explored (Sugiyama et al., 2008; Nguyen et al., 2010; Kanamori et al., 2009). The basic idea of direct density-ratio estimation is to learn the density-ratio function $\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$ without going through separate density estimation of $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$. An intuitive rationale of direct density-ratio estimation is that knowing the two densities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ means knowing their ratio, but not vice versa; knowing the ratio $\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$ does not necessarily mean knowing the two densities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ because such decomposition is not unique (see Figure 2.1). This implies that estimating the density ratio is substantially easier than estimating the densities, and thus directly estimating the density ratio would be more promising[4] (Sugiyama et al., 2012a).

In the rest of this section, we review three methods of directly estimating the density ratio $\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$ from samples $\{\boldsymbol{x}_i^P\}_{i=1}^n$ and $\{\boldsymbol{x}_j^Q\}_{j=1}^n$ drawn from $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$: The *KL importance estimation procedure* (KLIEP) (Sugiyama et al., 2008) in Section 2.3.2, *unconstrained least-squares importance fitting* (uLSIF) (Kanamori et al., 2009) in Section 2.3.3, and *relative uLSIF (RuLSIF)* (Yamada et al., 2013) in Section 2.3.4.

2.3.2 KLIEP

KLIEP (Sugiyama et al., 2008) is a direct density-ratio estimation algorithm that is suitable for estimating the KL divergence.

Density-Ratio Model

Let us model the density ratio $\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$ by the following kernel model:

$$r(\boldsymbol{x};\boldsymbol{\theta}) := \sum_{\ell=1}^{n} \theta_\ell K(\boldsymbol{x}, \boldsymbol{x}_\ell), \tag{2.5}$$

[4] Vladimir Vapnik advocated in his seminal book (Vapnik, 1998) that one should avoid solving a more difficult problem as an intermediate step. The *support vector machine* (Cortes and Vapnik, 1995) is a representative example that demonstrates the usefulness of this principle: It avoids solving a more general problem of estimating data generating probability distributions, and only learns a decision boundary that is sufficient for pattern recognition. The idea of direct density-ratio estimation also follows Vapnik's principle.

where $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_n)^\top$ are parameters to be learned from data samples, and $K(\boldsymbol{x}, \boldsymbol{x}')$ is a kernel basis function. In practice, we use the Gaussian kernel:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\right),$$

where σ (> 0) is the kernel width. In all our experiments, the kernel width σ is determined based on cross-validation.

Learning Algorithm

The parameters $\boldsymbol{\theta}$ in the model $r(\boldsymbol{x}; \boldsymbol{\theta})$ are determined so that the KL divergence from $p(\boldsymbol{x})$ to $r(\boldsymbol{x}; \boldsymbol{\theta})q(\boldsymbol{x})$ is minimized:

$$\begin{aligned}\mathrm{KL} &= \int p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})r(\boldsymbol{x}; \boldsymbol{\theta})}\right) \mathrm{d}\boldsymbol{x} \\ &= \int p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} - \int p(\boldsymbol{x}) \log\left(r(\boldsymbol{x}; \boldsymbol{\theta})\right) \mathrm{d}\boldsymbol{x}\end{aligned}$$

After ignoring the first term which is irrelevant to $r(\boldsymbol{x}; \boldsymbol{\theta})$ and approximating the second term with the empirical estimates, the KLIEP optimization problem is given as follows:

$$\begin{aligned}&\max_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{\ell=1}^{n} \theta_\ell K(\boldsymbol{x}_i^P, \boldsymbol{x}_\ell)\right), \\ &\text{s.t. } \frac{1}{n}\sum_{j=1}^{n}\sum_{\ell=1}^{n} \theta_\ell K(\boldsymbol{x}_j^Q, \boldsymbol{x}_\ell) = 1 \text{ and } \theta_1, \ldots, \theta_n \geq 0.\end{aligned}$$

The equality constraint is for the normalization purpose because $r(\boldsymbol{x}; \boldsymbol{\theta})q(\boldsymbol{x})$ should be a probability density function. The inequality constraint comes from the non-negativity of the density-ratio function. Since this is a convex optimization problem, the unique global optimal solution $\widehat{\boldsymbol{\theta}}$ can be simply obtained, for example, by a gradient-projection iteration. Finally, a density-ratio estimator is given as

$$\widehat{r}(\boldsymbol{x}) = \sum_{\ell=1}^{n} \widehat{\theta}_\ell K(\boldsymbol{x}, \boldsymbol{x}_\ell).$$

KLIEP was shown to achieve the optimal non-parametric convergence rate (Sugiyama et al., 2008; Nguyen et al., 2010).

Change-Point Detection by KLIEP

Given a density-ratio estimator $\widehat{r}(\boldsymbol{x})$, an approximator of the KL divergence is given as

$$\widehat{\mathrm{KL}} := \frac{1}{n}\sum_{i=1}^{n} \log \widehat{r}(\boldsymbol{x}_i^P).$$

In the previous work (Kawahara and Sugiyama, 2012), this KLIEP-based KL-divergence estimator was applied to change-point detection and demonstrated to be promising in experiments.

2.3.3 uLSIF

Recently, another direct density-ratio estimator called uLSIF was proposed (Kanamori et al., 2009, 2012b), which is suitable for estimating the PE divergence.

Learning Algorithm

In uLSIF, the same density-ratio model as KLIEP is used (see Section 2.3.2). However, its training criterion is different; the density-ratio model is fitted to the true density-ratio under the squared loss. More specifically, the parameter $\boldsymbol{\theta}$ in the model $r(\boldsymbol{x};\boldsymbol{\theta})$ is determined so that the following squared loss $J(\boldsymbol{x})$ is minimized:

$$\begin{aligned} J(\boldsymbol{x}) &= \frac{1}{2}\int \left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} - r(\boldsymbol{x};\boldsymbol{\theta})\right)^2 q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\ &= \frac{1}{2}\int \left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)^2 q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int p(\boldsymbol{x}) r(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x} + \frac{1}{2}\int r(\boldsymbol{x};\boldsymbol{\theta})^2 q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \end{aligned}$$

Since the first term is a constant, we focus on the last two terms. By substituting $r(\boldsymbol{x};\boldsymbol{\theta})$ with our model stated in (2.5) and approximating the integrals by the empirical averages, the uLSIF optimization problem is given as follows:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^n}\left[\frac{1}{2}\boldsymbol{\theta}^\top \widehat{\boldsymbol{H}}\boldsymbol{\theta} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\theta} + \frac{\lambda}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right], \tag{2.6}$$

where the penalty term $\frac{\lambda}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta}$ is included for a regularization purpose. $\lambda\ (\geq 0)$ denotes the regularization parameter, which is chosen by cross-validation (Sugiyama et al., 2008). $\widehat{\boldsymbol{H}}$ is the $n \times n$ matrix with the (ℓ, ℓ')-th element given by

$$\widehat{H}_{\ell,\ell'} := \frac{1}{n}\sum_{j=1}^{n} K(\boldsymbol{x}_j^Q, \boldsymbol{x}_\ell) K(\boldsymbol{x}_j^Q, \boldsymbol{x}_{\ell'}). \tag{2.7}$$

$\widehat{\boldsymbol{h}}$ is the n-dimensional vector with the ℓ-th element given by

$$\widehat{h}_\ell := \frac{1}{n}\sum_{i=1}^{n} K(\boldsymbol{x}_i^P, \boldsymbol{x}_\ell).$$

It is easy to confirm that the solution $\widehat{\boldsymbol{\theta}}$ of (2.6) can be analytically obtained as

$$\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_n)^{-1}\widehat{\boldsymbol{h}}, \tag{2.8}$$

where $\boldsymbol{I}_n$ denotes the n-dimensional identity matrix. Finally, a density-ratio estimator is given as

$$\widehat{r}(\boldsymbol{x}) = \sum_{\ell=1}^{n} \widehat{\theta}_\ell K(\boldsymbol{x}, \boldsymbol{x}_\ell).$$

Change-Point Detection by uLSIF

Given a density-ratio estimator $\widehat{r}(\boldsymbol{x})$, an approximator of the PE divergence can be constructed as

$$\widehat{\mathrm{PE}} := -\frac{1}{2n}\sum_{j=1}^{n} \widehat{r}(\boldsymbol{x}_j^Q)^2 + \frac{1}{n}\sum_{i=1}^{n} \widehat{r}(\boldsymbol{x}_i^P) - \frac{1}{2}.$$

This approximator is derived from the following expression of the PE divergence (Sugiyama et al., 2010, 2011b):

$$\mathrm{PE}(P\|Q) = -\frac{1}{2}\int \left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)^2 q(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int \left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \frac{1}{2}. \tag{2.9}$$

The first two terms of (2.9) are actually the negative uLSIF optimization objective without regularization. This expression can also be obtained based on the fact

that the f-divergence $D(P\|Q)$ is lower-bounded via the *Legendre-Fenchel convex duality* (Rockafellar, 1970) as follows (Keziou, 2003; Nguyen et al., 2007):

$$D(P||Q) = \sup_h \left(\int p(\boldsymbol{x}) h(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - \int q(\boldsymbol{x}) f^*(h(\boldsymbol{x})) \, \mathrm{d}\boldsymbol{x} \right), \tag{2.10}$$

where f^* is the convex conjugate of convex function f defined at (2.2). The PE divergence corresponds to $f(t) = \frac{1}{2}(t-1)^2$, for which convex conjugate is given by $f^*(t^*) = \frac{(t^*)^2}{2} + t^*$. For $f(t) = \frac{1}{2}(t-1)^2$, the supremum can be achieved when $\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} = h(\boldsymbol{x}) + 1$. Substituting $h(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} - 1$ into (2.10), we can obtain (2.9).

uLSIF has some notable advantages: Its solution can be computed analytically (Kanamori et al., 2009) and it possesses the optimal non-parametric convergence rate (Kanamori et al., 2012b). Moreover, it has the optimal numerical stability (Kanamori et al., 2013), and it is more robust than KLIEP (Sugiyama et al., 2012b). In Section 2.4, we will experimentally demonstrate that uLSIF-based change-point detection compares favorably with the KLIEP-based method.

2.3.4 RuLSIF

Depending on the condition of the denominator density $q(\boldsymbol{x})$, the density-ratio value $\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$ can be unbounded (i.e., they can be infinity). This is actually problematic because the non-parametric convergence rate of uLSIF is governed by the "sup"-norm of the true density-ratio function: $\max_{\boldsymbol{x}} \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$. To overcome this problem, *relative density-ratio estimation* was introduced (Yamada et al., 2013).

Relative PE Divergence

Let us consider the α-*relative PE-divergence* for $0 \leq \alpha < 1$:

$$\begin{aligned} \mathrm{PE}_\alpha(P\|Q) &:= \mathrm{PE}(P\|\alpha P + (1-\alpha)Q) \\ &= \int q_\alpha(\boldsymbol{x}) \left(\frac{p(\boldsymbol{x})}{q_\alpha(\boldsymbol{x})} - 1 \right)^2 \mathrm{d}\boldsymbol{x}, \end{aligned}$$

where $q_\alpha(\boldsymbol{x}) = \alpha p(\boldsymbol{x}) + (1-\alpha) q(\boldsymbol{x})$ is the α-*mixture density*. We refer to

$$g_\alpha(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{\alpha p(\boldsymbol{x}) + (1-\alpha) q(\boldsymbol{x})}$$

as the α-*relative density-ratio*. The α-relative density-ratio is reduced to the plain density-ratio if $\alpha = 0$, and it tends to be "smoother" as α gets larger. Indeed, one can confirm that the α-relative density-ratio is bounded above by $1/\alpha$ for $\alpha > 0$, even when the plain density-ratio $\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$ is unbounded. This was proved to contribute to improving the estimation accuracy (Yamada et al., 2013).

As explained in Section 2.3.1, we use symmetrized divergence

$$\mathrm{PE}_\alpha(P\|Q) + \mathrm{PE}_\alpha(Q\|P)$$

as a change-point score, where each term is estimated separately.

Learning Algorithm

For approximating the α-relative density ratio $g_\alpha(\boldsymbol{x})$, we still use the same kernel model $r(\boldsymbol{x};\boldsymbol{\theta})$ given by (2.5). In the same way as the uLSIF method, the parameter $\boldsymbol{\theta}$ is learned by minimizing the squared loss between true and estimated relative ratios:

$$\begin{aligned} J(\boldsymbol{x}) &= \frac{1}{2}\int q_\alpha(\boldsymbol{x})\Big(g_\alpha(\boldsymbol{x}) - r(\boldsymbol{x};\boldsymbol{\theta})\Big)^2 \,\mathrm{d}\boldsymbol{x} \\ &= \frac{1}{2}\int q_\alpha(\boldsymbol{x}) g_\alpha^2(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} - \int q_\alpha(\boldsymbol{x}) g_\alpha(\boldsymbol{x}) r(\boldsymbol{x};\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{x} \\ &\quad + \frac{\alpha}{2}\int p(\boldsymbol{x}) r(\boldsymbol{x};\boldsymbol{\theta})^2 \,\mathrm{d}\boldsymbol{x} + \frac{1-\alpha}{2}\int q(\boldsymbol{x}) r(\boldsymbol{x};\boldsymbol{\theta})^2 \,\mathrm{d}\boldsymbol{x}. \end{aligned}$$

Again, by ignoring the constant and approximating the expectations by sample averages, the α-relative density-ratio can be learned in the same way as the plain density-ratio. Indeed, the optimization problem of a relative variant of uLSIF, called RuLSIF, is given as the same form as uLSIF; the only difference is the definition of the matrix $\widehat{\boldsymbol{H}}$:

$$\widehat{H}_{\ell,\ell'} := \frac{\alpha}{n}\sum_{i=1}^{n} K(\boldsymbol{x}_i^P, \boldsymbol{x}_\ell) K(\boldsymbol{x}_i^P, \boldsymbol{x}_{\ell'}) + \frac{(1-\alpha)}{n}\sum_{j=1}^{n} K(\boldsymbol{x}_j^Q, \boldsymbol{x}_\ell) K(\boldsymbol{x}_j^Q, \boldsymbol{x}_{\ell'}).$$

Thus, the advantages of uLSIF regarding the analytic solution, numerical stability, and robustness are still maintained in RuLSIF. Furthermore, RuLSIF possesses an even better non-parametric convergence property than uLSIF (Yamada et al., 2013).

Change-Point Detection by RuLSIF

By using an estimator $\widehat{r}(\boldsymbol{x})$ of the α-relative density-ratio, the α-relative PE divergence can be approximated as

$$\widehat{\mathrm{PE}}_\alpha := -\frac{\alpha}{2n}\sum_{i=1}^{n}\widehat{r}(\boldsymbol{x}_i^P)^2 - \frac{1-\alpha}{2n}\sum_{j=1}^{n}\widehat{r}(\boldsymbol{x}_j^Q)^2 + \frac{1}{n}\sum_{i=1}^{n}\widehat{r}(\boldsymbol{x}_i^P) - \frac{1}{2}.$$

In Section 2.4, we will experimentally demonstrate that the RuLSIF-based change-point detection performs even better than the plain uLSIF-based method.

2.4 Experiments

In this section, we experimentally investigate the performance of the proposed and existing change-point detection methods on artificial and real-world datasets including human-activity sensing, speech, and Twitter messages. The MATLAB implementation of the proposed method is available at

`http://sugiyama-www.cs.titech.ac.jp/~song/change_detection/`.

For all experiments, we fix the parameters at $n = 50$ and $k = 10$. α in the RuLSIF-based method is fixed to 0.1. Sensitivity to different parameter choices and more issues regarding algorithm-specific parameter tuning will be discussed below.

2.4.1 Artificial Datasets

As mentioned in Section 2.3.1, we use the symmetrized divergence for change-point detection. We first illustrate how symmetrization of the PE divergence affects the change-point detection performance.

The top graph in Figure 2.3 shows an artificial time-series signal that consists of three segments with equal length of 200. The samples are drawn from $\mathcal{N}(0, 2^2)$, $\mathcal{N}(0, 1^2)$, and $\mathcal{N}(0, 2^2)$, respectively, where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. Thus, the variances change at time 200 and 400. In this experiment, we consider three types of divergence measures:

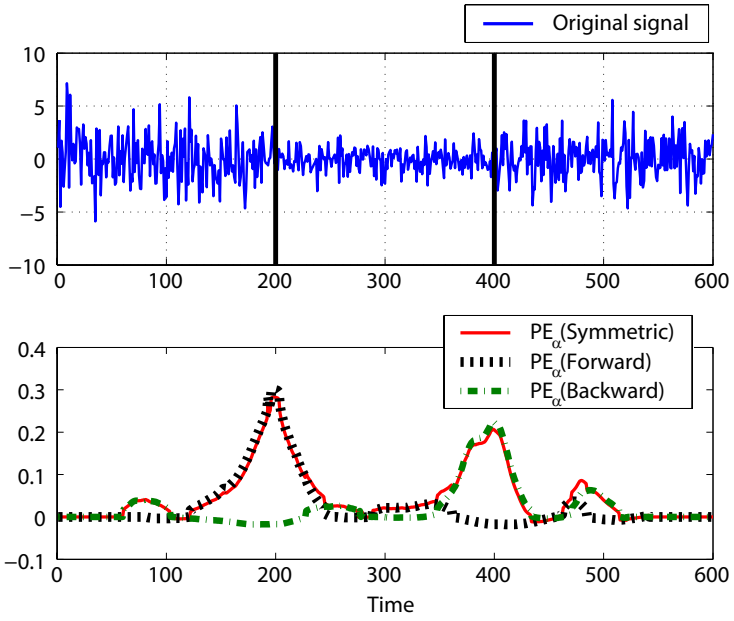


Figure 2.3: (Top) The original signal (blue) is segmented into 3 sections with equal length. Samples are drawn from the normal distributions $\mathcal{N}(0, 2^2)$, $\mathcal{N}(0, 1^2)$, and $\mathcal{N}(0, 2^2)$, respectively. (Bottom) Symmetric (red) and asymmetric (black and green) PE_α divergences.

- $\mathrm{PE}_\alpha(\mathrm{Symmetric}) : \mathrm{PE}_\alpha(P_t||P_{t+n}) + \mathrm{PE}_\alpha(P_{t+n}||P_t)$,
- $\mathrm{PE}_\alpha(\mathrm{Forward}) : \mathrm{PE}_\alpha(P_t||P_{t+n})$,
- $\mathrm{PE}_\alpha(\mathrm{Backward}) : \mathrm{PE}_\alpha(P_{t+n}||P_t)$.

Three divergences are compared in the bottom graph of Figure 2.3.

As we can see from the graphs, $\mathrm{PE}_\alpha(\mathrm{Forward})$ detects the first change point successfully, but not the second one. On the other hand, $\mathrm{PE}_\alpha(\mathrm{Backward})$ behaves oppositely. This implies that combining forward and backward divergences can improve the overall change-point detection performance. For this reason, we only use $\mathrm{PE}_\alpha(\mathrm{Symmetric})$ as the change-point score of the proposed method from here on.

Next, we illustrate the behavior of our proposed RuLSIF-based method, and then compare its performance with the uLSIF-based and KLIEP-based methods.

In our implementation, two sets of candidate parameters,

- $\sigma = 0.6d_{\mathrm{med}}$, $0.8d_{\mathrm{med}}$, d_{med}, $1.2d_{\mathrm{med}}$, and $1.4d_{\mathrm{med}}$,
- $\lambda = 10^{-3}$, 10^{-2}, 10^{-1}, 10^{0}, and 10^{1},

are provided to the cross-validation procedure, where d_{med} denotes the median distance between samples. The best combination of these parameters is chosen by grid search via cross-validation. We use 5-fold cross-validation for all experiments.

We use the following 4 artificial time-series datasets that contain manually inserted change-points:

- **Dataset 1 (Jumping mean):** The following 1-dimensional auto-regressive model borrowed from Takeuchi and Yamanishi (2006) is used to generate 5000 samples (i.e., $t = 1, \ldots, 5000$):

  $$y(t) = 0.6y(t-1) - 0.5y(t-2) + \epsilon_t,$$

  where ϵ_t is a Gaussian noise with mean μ and standard deviation 1.5. The initial values are set as $y(1) = y(2) = 0$. A change point is inserted at every 100 time steps by setting the noise mean μ at time t as

  $$\mu_N = \begin{cases} 0 & N = 1, \\ \mu_{N-1} + \frac{N}{16} & N = 2, \ldots, 49, \end{cases}$$

  where N is a natural number such that $100(N-1) + 1 \leq t \leq 100N$.

- **Dataset 2 (Scaling variance):** The same auto-regressive model as Dataset 1 is used, but a change point is inserted at every 100 time steps by setting the noise standard deviation σ at time t as

  $$\sigma = \begin{cases} 1 & N = 1, 3, \ldots, 49, \\ \ln(e + \frac{N}{4}) & N = 2, 4, \ldots, 48. \end{cases}$$

- **Dataset 3 (Switching covariance):** 2-dimensional samples of size 5000 are drawn from the origin-centered normal distribution, and a change point is

inserted at every 100 time steps by setting the covariance matrix $\boldsymbol{\Sigma}$ at time t as

$$\boldsymbol{\Sigma} = \begin{cases} \begin{pmatrix} 1 & -\frac{4}{5} - \frac{N-2}{500} \\ -\frac{4}{5} - \frac{N-2}{500} & 1 \end{pmatrix} & N = 1, 3, \ldots, 49, \\ \begin{pmatrix} 1 & \frac{4}{5} + \frac{N-2}{500} \\ \frac{4}{5} + \frac{N-2}{500} & 1 \end{pmatrix} & N = 2, 4, \ldots, 48. \end{cases}$$

- **Dataset 4 (Changing frequency):** 1-dimensional samples of size 5000 are generated as

$$y(t) = \sin(\omega x) + \epsilon_t,$$

where ϵ_t is a origin-centered Gaussian noise with standard deviation 0.8. A change point is inserted at every 100 points by changing the frequency ω at time t as

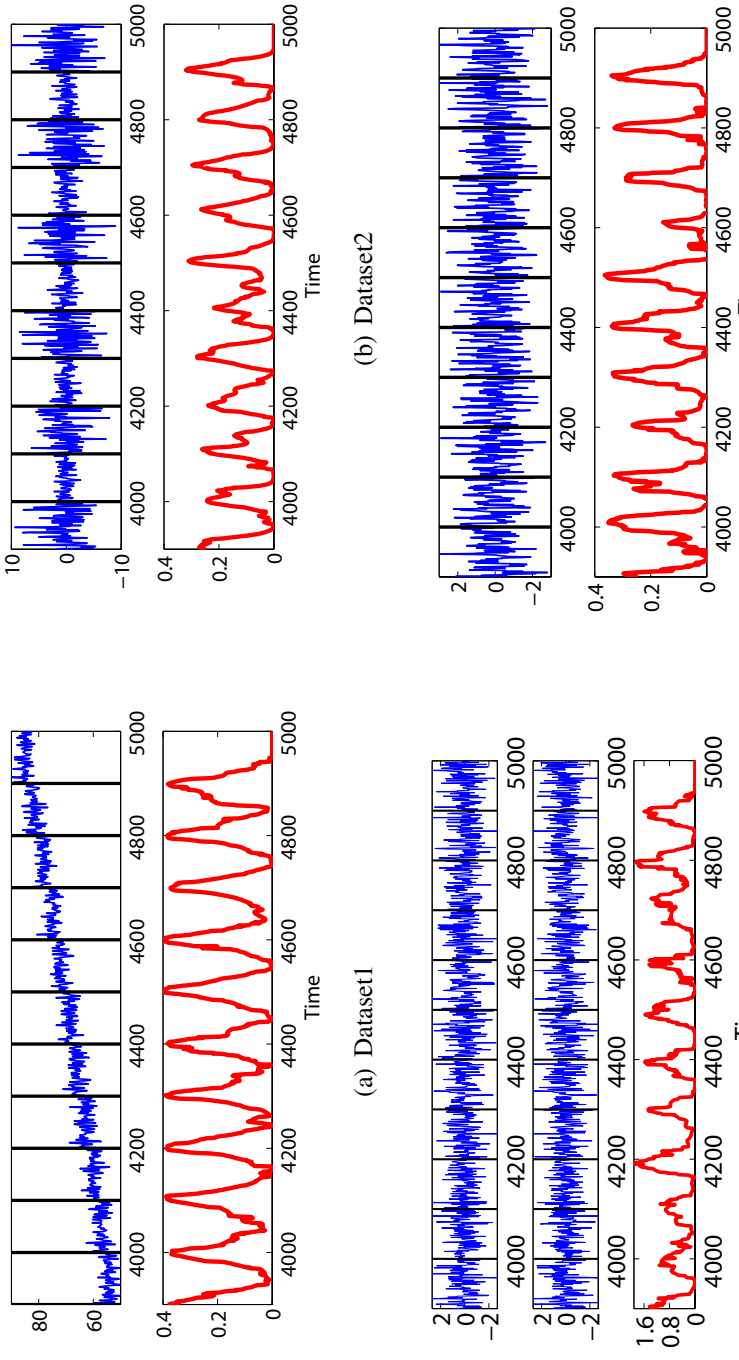
$$\omega_N = \begin{cases} 1 & N = 1, \\ \omega_{N-1} \ln(e + \frac{N}{2}) & N = 2, \ldots, 49. \end{cases}$$

Note that, to explore the ability of detecting change points with different significance, we purposely made latter change-points more significant than earlier ones in the above datasets.

Figure 2.4 shows examples of these datasets for the last 10 change points and corresponding change-point score obtained by the proposed RuLSIF-based method. Although the last 10 change points are the most significant, we can see from the graphs that, for Dataset 3 and Dataset 4, these change points can be even hardly identified by human. Nevertheless, the change-point score obtained by the proposed RuLSIF-based method increases rapidly after changes occur.

Next, we compare the performance of RuLSIF-based, uLSIF-based, and KLIEP-based methods in terms of the *receiver operating characteristic (ROC) curves* and the area under the ROC curve (AUC) values. We define the *true positive rate* and *false positive rate* in the following way (Kawahara and Sugiyama, 2012):

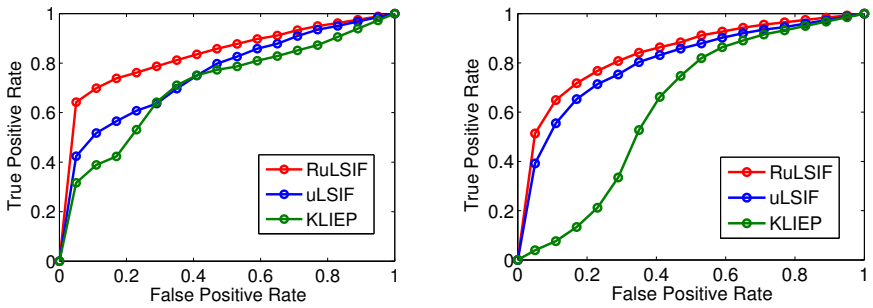
- True positive rate (TPR): $n_{\mathrm{cr}}/n_{\mathrm{cp}}$,
- False positive rate (FPR): $(n_{\mathrm{al}} - n_{\mathrm{cr}})/n_{\mathrm{al}}$,



(a) Dataset1 (b) Dataset2

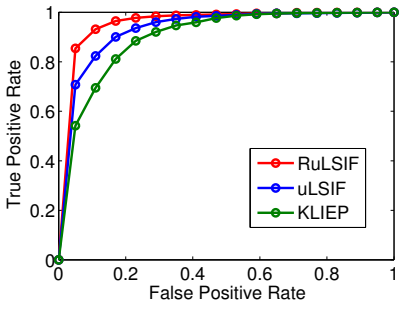
(c) Dataset3 (d) Dataset4

Figure 2.4: Illustrative time-series samples (upper) and the change-point score obtained by the RuLSIF-based method (lower). The true change-points are marked by black vertical lines in the upper graphs.

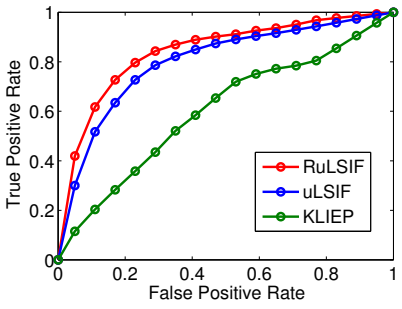


(a) Dataset1

(b) Dataset2



(c) Dataset3



(d) Dataset4

Figure 2.5: Average ROC curves of RuLSIF-based, uLSIF-based, and KLIEP-based methods.

Table 2.1: The AUC values of RuLSIF-based, uLSIF-based, and KLIEP-based methods. The best and comparable methods by the t-test with significance level 5% are described in boldface.

	RuLSIF	uLSIF	KLIEP
Dataset 1	**.848(.023)**	.763(.023)	.713(.036)
Dataset 2	**.846(.031)**	.806(.035)	.623(.040)
Dataset 3	**.972(.012)**	.943(.015)	.904(.017)
Dataset 4	**.844(.031)**	.801(.024)	.602(.036)

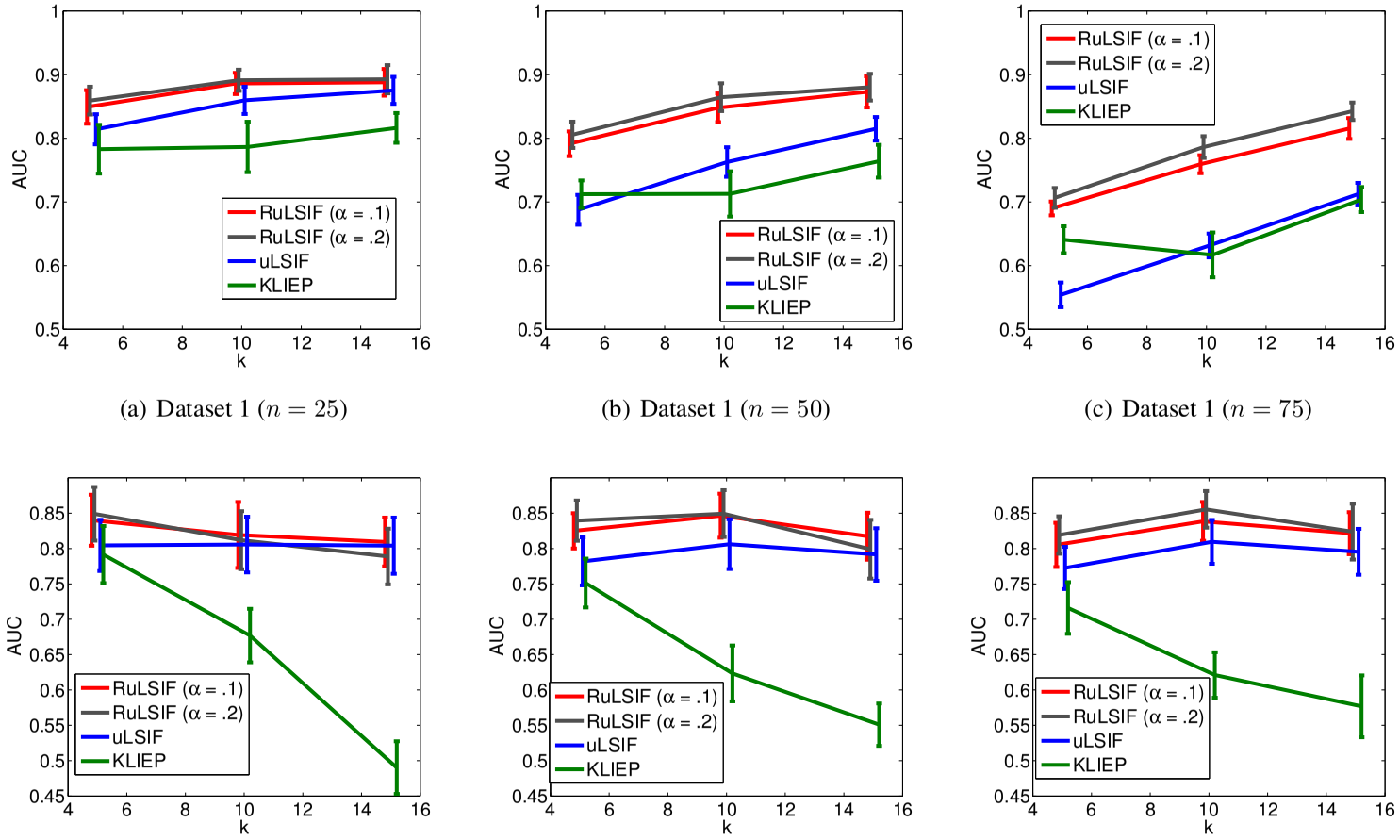
where n_{cr} denotes the number of times change points are correctly detected, n_{cp} denotes the number of all change points, and n_{al} is the number of all detection alarms.

Following the strategy of the previous researches (Desobry et al., 2005; Harchaoui et al., 2009), peaks of a change-point score are regarded as detection alarms. More specifically, a detection alarm at step t is regarded as correct if there exists a true alarm at step t^* such that $t \in [t^* - 10, t^* + 10]$. To avoid duplication, we remove the kth alarm at step t_k if $t_k - t_{k-1} < 20$.

We set up a threshold η for filtering out all alarms whose change-point scores are lower than or equal to η. Initially, we set η to be equal to the score of the highest peak. Then, by lowering η gradually, both TPR and FPR become non-decreasing. For each η, we plot TPR and FPR on the graph, and thus a monotone curve can be drawn.

Figure 2.5 illustrates ROC curves averaged over 50 runs with different random seeds for each dataset. Table 2.1 describes the mean and standard deviation of the AUC values over 50 runs. The best and comparable methods by the t-test with significance level 5% are described in boldface. The experimental results show that the uLSIF-based method tends to outperform the KLIEP-based method, and the RuLSIF-based method even performs better than the uLSIF-based method.

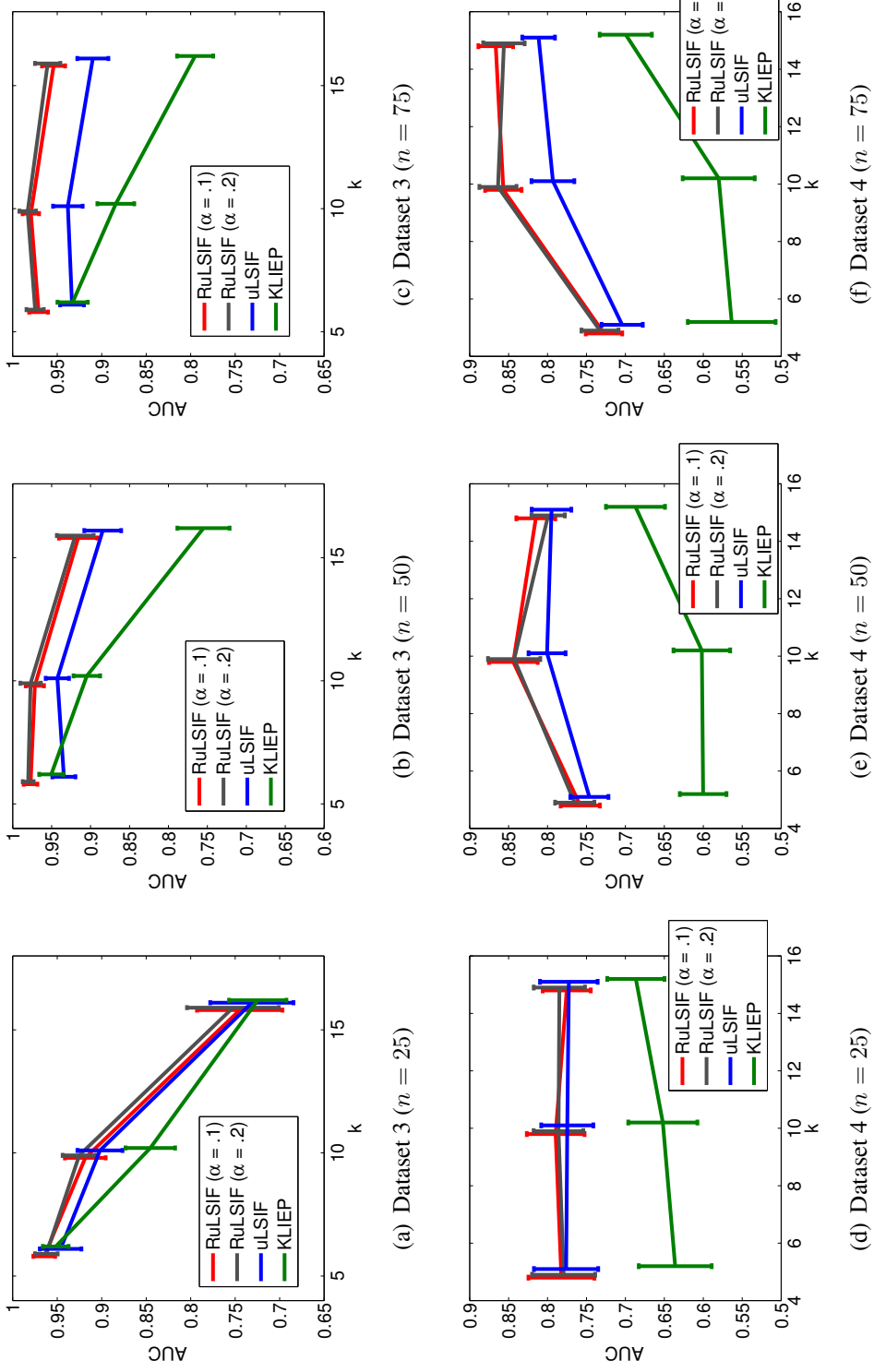
Finally, we investigate the sensitivity of the performance on different choices of n and k in terms of AUC values. In Figure 2.6, 2.7 , the AUC values of RuLSIF ($\alpha = 0.1$ and 0.2), uLSIF (which corresponds to RuLSIF with $\alpha = 0$), and KLIEP were plotted for $k = 5$, 10, and 15 under a specific choice of n in each graph. We generate such graphs for all 4 datasets with $n = 25$, 50, and 75. The result shows that the proposed method consistently performs better than the other methods, and the order of the methods according to the performance is kept unchanged over various choices of n and k. Moreover, the RuLSIF methods with $\alpha = 0.1$ and 0.2 perform rather similarly. For this reason, we keep using the medium parameter values among the candidates in the following experiments: $n = 50$, $k = 10$, and $\alpha = 0.1$.



(a) Dataset 1 ($n = 25$) (b) Dataset 1 ($n = 50$) (c) Dataset 1 ($n = 75$)

(d) Dataset 2 ($n = 25$) (e) Dataset 2 ($n = 50$) (f) Dataset 2 ($n = 75$)

Figure 2.6: AUC plots for $n = 25, 50, 75$ and $k = 5, 10, 15$. The horizontal axes denote k, while the vertical axes denote AUC values (on Dataset 1 and 2).



(a) Dataset 3 ($n = 25$) (b) Dataset 3 ($n = 50$) (c) Dataset 3 ($n = 75$)

(d) Dataset 4 ($n = 25$) (e) Dataset 4 ($n = 50$) (f) Dataset 4 ($n = 75$)

Figure 2.7: AUC plots for $n = 25, 50, 75$ and $k = 5, 10, 15$. The horizontal axes denote k, while the vertical axes denote AUC values (on Dataset 3 and 4).

2.4.2 Real-World Datasets

Next, we evaluate the performance of the density-ratio estimation based methods and other existing change-point detection methods using two real-world datasets: Human-activity sensing and speech.

We include the following methods in our comparison.

- **Singular spectrum transformation (SST) (Moskvina and Zhigljavsky, 2003a; Ide and Tsuda, 2007; Itoh and Kurths, 2010):** Change-point scores are evaluated on two consecutive trajectory matrices using the distance-based singular spectrum analysis. This corresponds to a state-space model with no system noise. For this method, we use the first 4 eigenvectors to compare the difference between two subspaces, which was confirmed to be reasonable choice in our preliminary experiments.

- **Subspace identification (SI) (Kawahara et al., 2007):** SI identifies a subspace in which time-series data is constrained, and evaluates the distance of target sequences from the subspace. The subspace spanned by the columns of an observability matrix is used for estimating the distance from the subspace spanned by subsequences of time-series data. For this method, we use the top 4 significant singular values according to our preliminary experiment results.

- **Auto regressive (AR) (Takeuchi and Yamanishi, 2006):** AR first fits an AR model to time-series data, and then auxiliary time-series is generated from the AR model. With an extra AR model-fitting, the change-point score is given by the log-likelihood. The order of the AR model is chosen by Schwarz's Bayesian information criterion (Schwarz, 1978).

- **One-class support vector machine (OSVM) (Desobry et al., 2005):** Change-point scores are calculated by OSVM using two sets of descriptors of signals. The kernel width σ is set to the median value of the distances between samples, which is a popular heuristic in kernel methods (Schölkopf and Smola, 2002). Another parameter ν is set to 0.2, which indicates the proportion of outliers.

First, we use a human activity dataset. This is a subset of the *Human Activity Sensing Consortium (HASC) challenge 2011*[5], which provides human activity information collected by portable three-axis accelerometers. The task of change-point detection is to segment the time-series data according to the 6 behaviors: "*stay*", "*walk*", "*jog*", "*skip*", "*stair up*", and "*stair down*". The starting time of each behavior is arbitrarily decided by each user. Because the orientation of accelerometers is not necessarily fixed, we take the ℓ_2-norm of the 3-dimensional (i.e., x-, y-, and z-axes) data.

In Figure 2.8(a), examples of original time-series, true change points, and change-point scores obtained by the RuLSIF-based method are plotted. This shows that the change-point score clearly captures trends of changing behaviors, except the changes around time 1200 and 1500. However, because these changes are difficult to be recognized even by human, we do not regard them as critical flaws. Figure 2.8(b) illustrates ROC curves averaged over 10 datasets, and Figure 2.8(c) describes AUC values for each of the 10 datasets. The experimental results show that the proposed RuLSIF-based method tends to perform better than other methods.

In Figure 2.9 and 2.10, we pick up two fractions of the original datasets to demonstrate the performance between methods on illustrative results. As we can see from both graphs, the RuLSIF-based score is very stable and gives clear indication of all change-points, while in comparison, the SST-based score gives a few false alarms due to outliers and misses the first change-points on Figure 2.9(c) and the KLIEP-based score misses the second change-point and shows a very fluctuated result on Figure 2.10(c).

Next, we use the *IPSJ SIG-SLP Corpora and Environments for Noisy Speech Recognition* (CENSREC) dataset provided by National Institute of Informatics (NII)[6], which records human voice in a noisy environment. The task is to extract speech sections from recorded signals. This dataset offers several voice recordings with different background noises (e.g., noise of highway and restaurant). Segmentation of the beginning and ending of human voice is manually annotated. Note that we only use the annotations as the ground truth for the final performance

[5] `http://hasc.jp/hc2011/`

[6] `http://research.nii.ac.jp/src/eng/list/index.html`

evaluation, not for change-point detection (i.e., this experiment is still completely unsupervised).

Figure 2.11(a) illustrates an example of the original signals, true change-points, and change-point scores obtained by the proposed RuLSIF-based method. This shows that the proposed method still gives clear indications for speech segments. Figure 2.11(b) and Figure 2.11(c) show average ROC curves over 10 datasets and AUC values for each of the 10 datasets. The results show that the proposed method significantly outperforms other methods.

2.4.3 Twitter Dataset

Finally, we apply the proposed change-point detection method to the *CMU Twitter dataset*[7], which is an archive of Twitter messages collected from February 2010 to October 2010 via the Twitter application programming interface.

Here we track the degree of popularity of a given topic by monitoring the frequency of selected keywords. More specifically, we focus on events related to "*Deepwater Horizon oil spill in the Gulf of Mexico*" which occurred on April 20, 2010[8], and was widely broadcast among the Twitter community. We use the frequencies of 10 keywords: "*gulf*", "*spill*", "*bp*", "*oil*", "*hayward*", "*mexico*", "*coast*", "*transocean*", "*halliburton*", and "*obama*" (see Figure 2.12(a)). We perform change-point detection directly on the 10-dimensional data, with the hope that we can capture correlation changes between multiple keywords, in addition to changes in the frequency of each keyword.

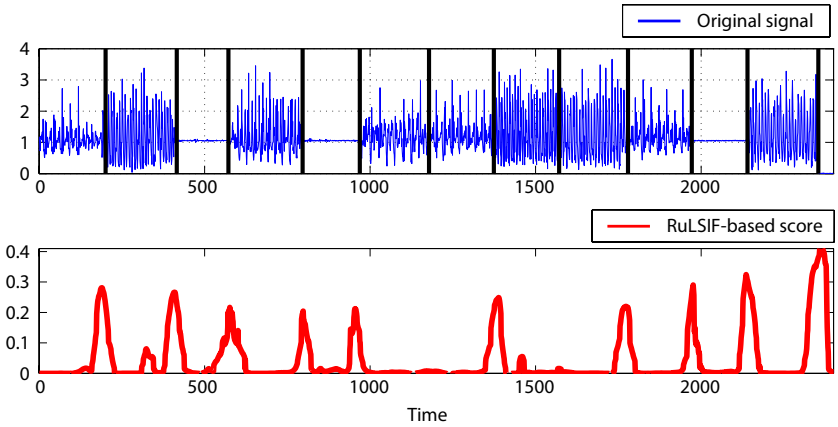
For quantitative evaluation, we referred to the Wikipedia entry "Timeline of the Deepwater Horizon oil spill"[9] as a real-world event source. The change-point score obtained by the proposed RuLSIF-based method is plotted in Figure 2.12(b), where four occurrences of important real-world events show the development of this news story.

As we can see from Figure 2.12(b), the change-point score increases immediately after the initial explosion of the deepwater horizon oil platform and soon

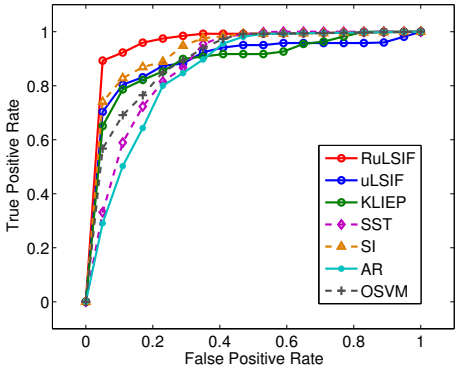
[7] http://www.ark.cs.cmu.edu/tweets/

[8] http://en.wikipedia.org/wiki/Deepwater_Horizon_oil_spill

[9] http://en.wikipedia.org/wiki/Timeline_of_the_Deepwater_Horizon_oil_spill



(a) One of the original signals and change-point scores obtained by the RuLSIF-based method

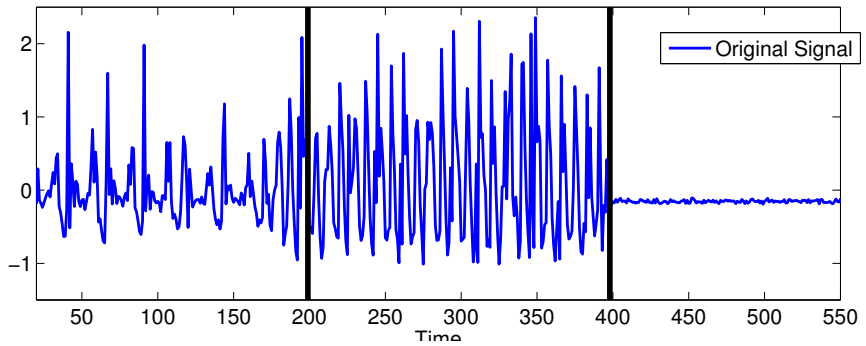


(b) Average ROC curves

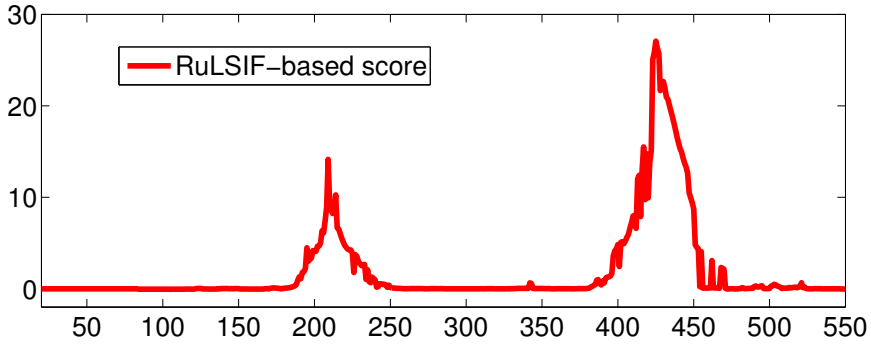
ID	RuLSIF	uLSIF	KLIEP	AR	SI	SST	OSVM
1001	.974	.853	.838	.899	.958	.903	.900
1002	.996	.963	.909	.872	.969	.880	.905
1003	.989	.854	.929	.869	.895	.851	.937
1004	.996	.868	.890	.881	.941	.886	.891
1005	.938	.952	.972	.849	.972	.915	.943
1006	.933	.918	.889	.778	.890	.925	.842
1007	.972	.857	.834	.850	.941	.817	.891
1008	.995	.922	.930	.892	.981	.860	.907
1009	.987	.880	.907	.833	.979	.842	.951
1010	.991	.952	.889	.821	.915	.867	.903
Ave.	**.977**	.902	.900	.854	.944	.875	.907
Std.	**.024**	.044	.042	.037	.034	.034	.032

(c) AUC values. The best and comparable methods by the t-test with significance level 5% are described in boldface.

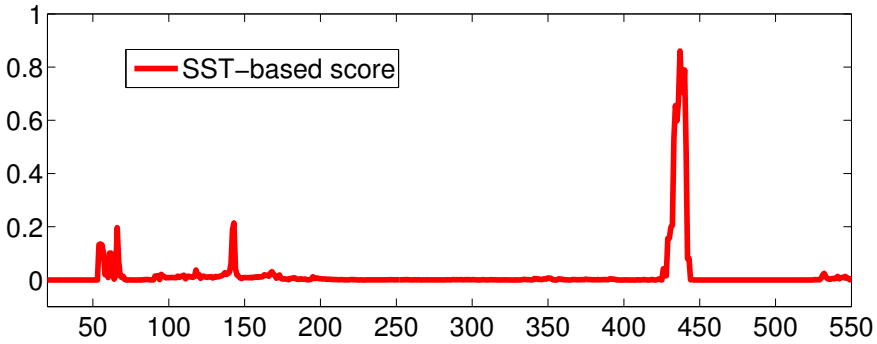
Figure 2.8: Results on HASC human-activity dataset.



(a) a fraction of the original signal

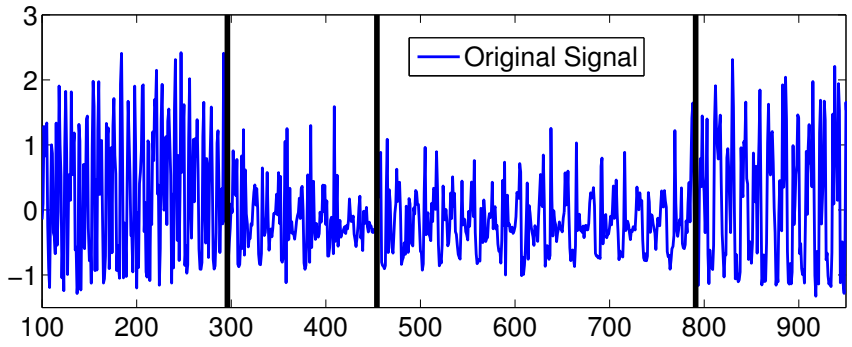


(b) RuLSIF-based score

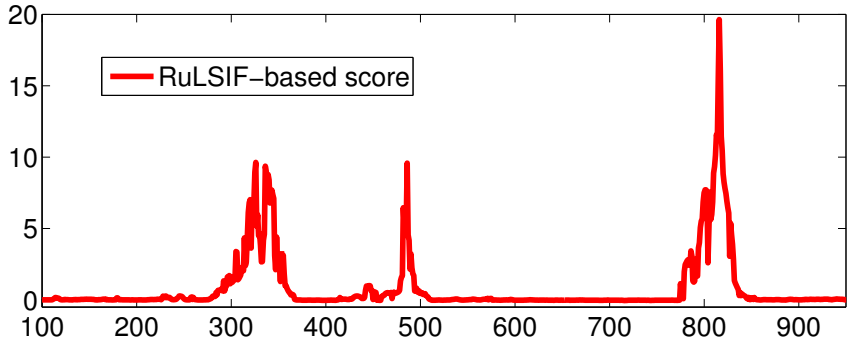


(c) SST-based score

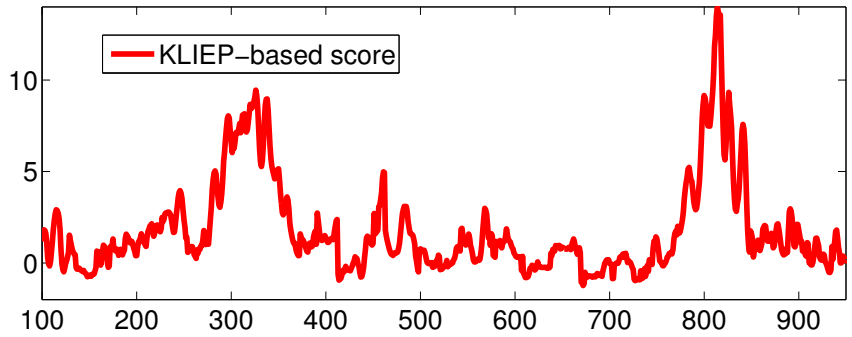
Figure 2.9: Results on HASC human-activity dataset.



(a) a fraction of the original signal

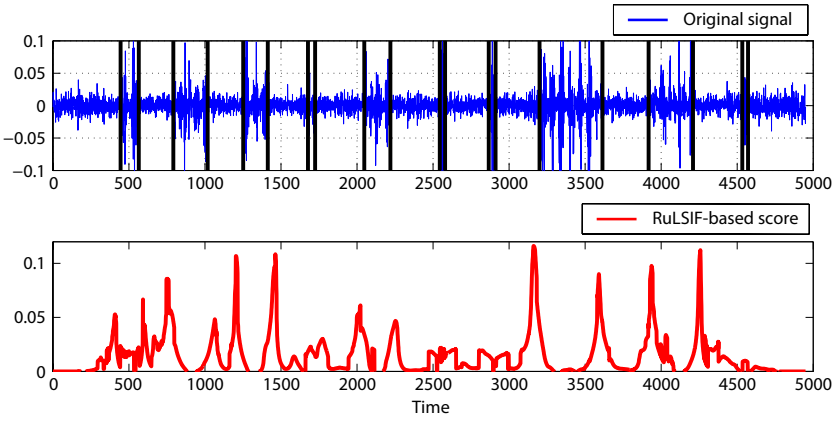


(b) RuLSIF-based score

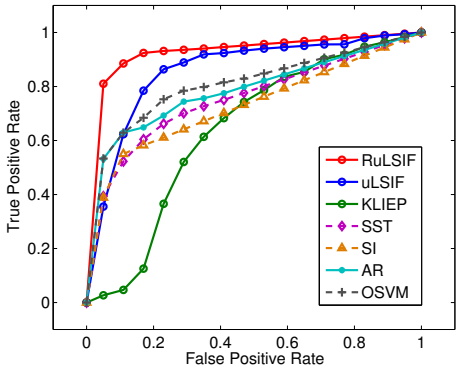


(c) KLIEP-based score

Figure 2.10: Results on HASC human-activity dataset.



(a) One of the original signals and change-point scores obtained by the RuLSIF-based method

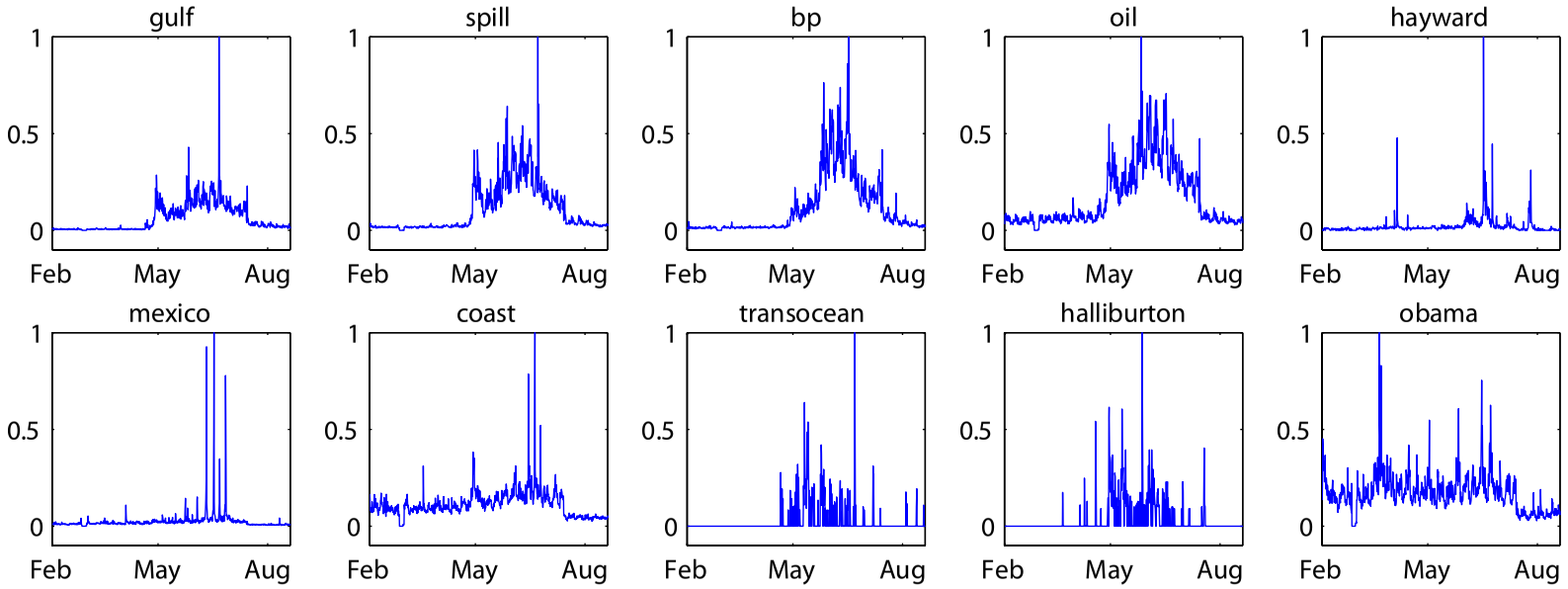


(b) Average ROC curves

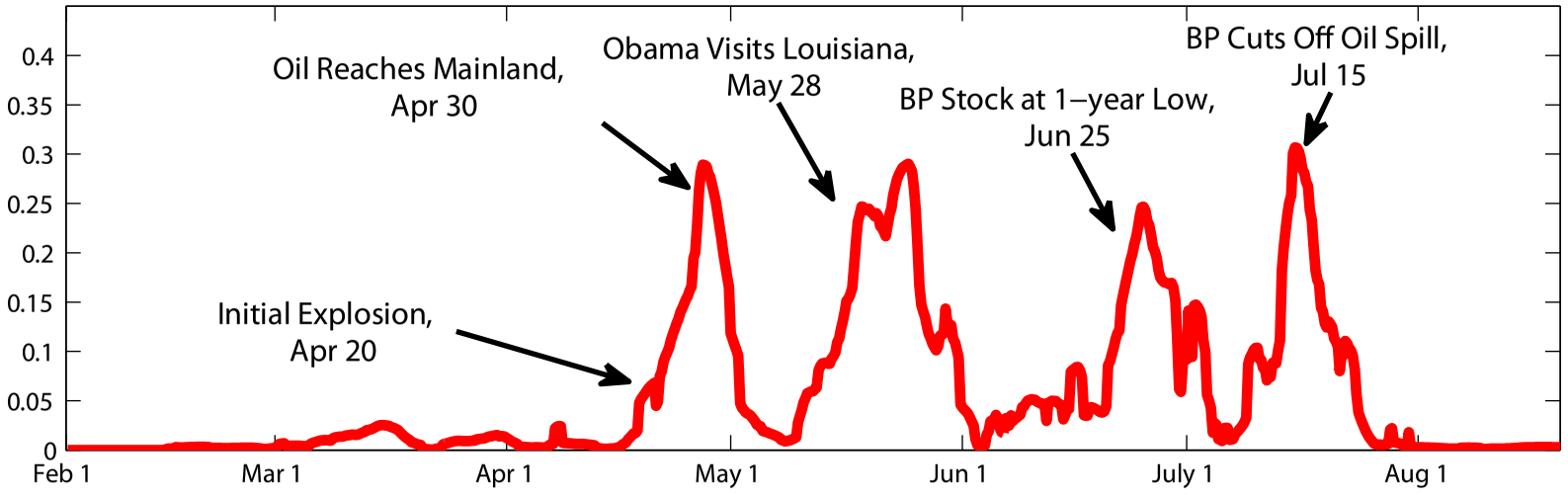
ID	RuLSIF	uLSIF	KLIEP	AR	SI	SST	OSVM
01	1.00	.902	.650	.860	.690	.806	.800
02	.911	.845	.712	.733	.800	.745	.725
03	.963	.931	.708	.910	.899	.807	.932
04	.903	.813	.587	.816	.735	.685	.751
05	.927	.907	.565	.831	.823	.809	.840
06	.857	.913	.676	.868	.740	.736	.838
07	.987	.797	.657	.807	.759	.797	.829
08	.962	.757	.581	.629	.704	.682	.800
09	.924	.913	.693	.738	.744	.781	.790
10	.966	.856	.554	.796	.725	.790	.850
Ave.	**.940**	.863	.638	.798	.762	.764	.815
Std.	**.044**	.059	.061	.081	.063	.049	.057

(c) AUC values. The best and comparable methods by the t-test with significance level 5% are described in boldface.

Figure 2.11: Results on CENSREC speech dataset.

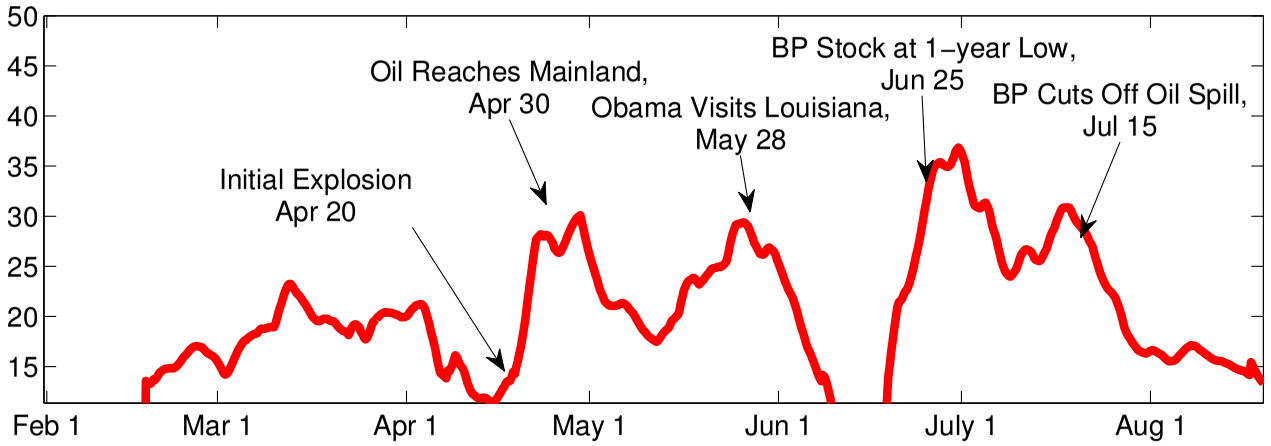


(a) Normalized frequencies of 10 keywords

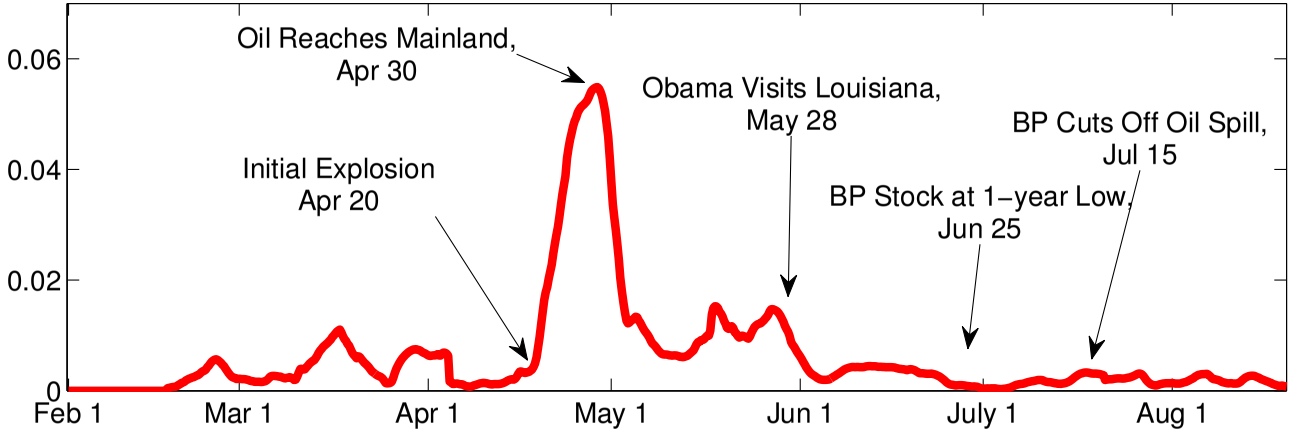


(b) Change-point score obtained by the RuLSIF-based method and exemplary real-world events

Figure 2.12: Results on Twitter dataset.



(a) Change-point score obtained by the KLIEP-based method and exemplary real-world events



(b) Change-point score obtained by the SST-based method and exemplary real-world events

Figure 2.13: Results on Twitter dataset.

reaches the first peak when oil was found on the sea shore of Louisiana on April 30. Shortly after BP announced its preliminary estimation on the amount of leaking oil, the change-point score rises quickly again and reaches its second peak at the end of May, at which time President Obama visited Louisiana to assure local residents of the federal government's support. On June 25, the BP stock was at its one year's lowest price, while the change-point score spikes at the third time. Finally, BP cut off the spill on July 15, as the score reaches its last peak.

As comparison, we also illustrate the change-point score generated by KLIEP and SST based method with the same real-world news labelled aside, on Figure 2.13(a) and 2.13(b). In general, the KLIEP generated score roughly shows the similar trend with the RuLSIF generated score, while the overall pattern looks more fluctuated. In contrast, the SST generated score, only spikes right after the initial explosion, and remains insensitive after that.

2.5 Conclusion

In this chapter, we first formulated the problem of retrospective change-point detection as the problem of comparing two probability distributions over two consecutive time segments. We then provided a comprehensive review of state-of-the-art density-ratio and divergence estimation methods, which are key building blocks of our change-point detection methods. Our contributions in this chapter were to extend the existing KLIEP-based change-point detection method (Kawahara and Sugiyama, 2012), and to propose to use uLSIF as a building block. uLSIF has various theoretical and practical advantages, for example, the uLSIF solution can be computed analytically, it possesses the optimal non-parametric convergence rate, it has the optimal numerical stability, and it has higher robustness than KLIEP. We further proposed to use RuLSIF, a novel divergence estimation paradigm emerged in the machine learning community recently. RuLSIF inherits good properties of uLSIF, and moreover it possesses an even better non-parametric convergence property. Through extensive experiments on artificial datasets and real-world datasets including human-activity sensing, speech, and Twitter messages, we demonstrated that the proposed RuLSIF-based change-point detection method is promising.

Though we estimated a density ratio between two consecutive segments, some earlier researches (Basseville and Nikiforov, 1993; Gustafsson, 1996, 2000) introduced a hyper-parameter that controls the size of a margin between two segments. In our preliminary experiments, however, we did not observe significant improvement by changing the margin. For this reason, we decided to use a straightforward model that two segments have no margin in between.

Chapter 3

Structural Change Detection

In this chapter, we investigate the problem of structural change detection, and propose a *direct* approach to learn changes between Markov Networks.

We first formulate the problem as estimating the ratio between two Markov Networks which encodes the changes of conditional independence. Our contributions are:

- the number of parameters to be estimated is halved comparing to other state-of-the-art methods;
- the normalization term can be sample-approximated for non-Gaussian distributions;
- the dual objective is derived, so problems on large Markov Networks can be solved efficiently.

This chapter will be organized as follows. First, we introduce the background and problem formulation in Section 3.1 and 3.2. The proposed method is illustrated in Section 3.3. After demonstrating the results of toy and real-world experiments in Section 3.4 and Section 3.5, we conclude our work in Section 3.7.

3.1 Introduction

Changes in interactions between random variables are interesting in many real-world phenomena. For example, genes may interact with each other in different ways when external stimuli change, co-occurrence between words may appear/disappear when the domains of text corpora shift, and correlation among pixels may change when a surveillance camera captures anomalous activities. Discovering such changes in interactions is a task of great interest in machine learning and data mining, because it provides useful insights into underlying mechanisms in many real-world applications.

In this chapter, we consider the problem of detecting changes in conditional independence among random variables between two sets of data. Such conditional independence structure can be expressed via an undirected graphical model called a *Markov network* (MN) (Bishop, 2006; Wainwright and Jordan, 2008; Koller and Friedman, 2009), where nodes and edges represent variables and their conditional dependencies, respectively. As a simple and widely applicable case, the pairwise MN model has been thoroughly studied recently (Ravikumar et al., 2010; Lee et al., 2007). Following this line, we also focus on the pairwise MN model as a representative example.

A naive approach to change detection in MNs is the two-step procedure of first estimating two MNs separately from two sets of data by *maximum likelihood estimation* (MLE), and then comparing the structure of the learned MNs. However, MLE is often computationally intractable due to the normalization factor included in the density model. Therefore, Gaussianity is often assumed in practice for computing the normalization factor analytically (Hastie et al., 2001), though this Gaussian assumption is highly restrictive in practice. We may utilize *importance sampling* (Robert and Casella, 2005) to numerically compute the normalization factor, but an inappropriate choice of the instrumental distribution may lead to an estimate with high variance (Wasserman, 2010); for more discussions on sampling techniques, see Gelman (1995) and Hinton (2002). Hyvärinen (2005) and Gutmann and Hyvärinen (2012) have explored an alternative approach to avoid computing the normalization factor which are not based on MLE.

However, the two-step procedure has the conceptual weakness that structure change is not directly learned. This indirect nature causes a crucial problem:

Suppose that we want to learn a sparse structure change. For learning sparse changes, we may utilize ℓ_1-regularized MLE (Banerjee et al., 2008; Friedman et al., 2008; Lee et al., 2007), which produces sparse MNs and thus the change between MNs also becomes sparse. However, this approach does not work if each MN is dense but only change is sparse.

To mitigate this indirect nature, the *fused-lasso* (Tibshirani et al., 2005) is useful, where two MNs are simultaneously learned with a sparsity-inducing penalty on the *difference* between two MN parameters (Zhang and Wang, 2010; Danaher et al., 2013). Although this fused-lasso approach allows us to learn sparse structure change naturally, the restrictive Gaussian assumption is still necessary to obtain the solution in a computationally tractable way.

The *nonparanormal* assumption (Liu et al., 2009, 2012) is a useful generalization of the Gaussian assumption. A nonparanormal distribution is a *semiparametric Gaussian copula* where each Gaussian variable is transformed by a monotone non-linear function. Nonparanormal distributions are much more flexible than Gaussian distributions thanks to the feature-wise non-linear transformation, while the normalization factors can still be computed analytically. Thus, the fused-lasso method combined with nonparanormal models would be one of the state-of-the-art approaches to change detection in MNs. However, the fused-lasso method is still based on separate modeling of two MNs, and its computation for more general non-Gaussian distributions is challenging.

In this chapter, we propose a more direct approach to structural change learning in MNs based on *density ratio estimation* (DRE) (Sugiyama et al., 2012a). Our method does not separately model two MNs, but directly models the *change* in two MNs. This idea follows Vapnik's principle (Vapnik, 1998):

> If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem.

This principle was used in the development of *support vector machines* (SVMs): rather than modeling two classes of samples, SVM directly learns a decision

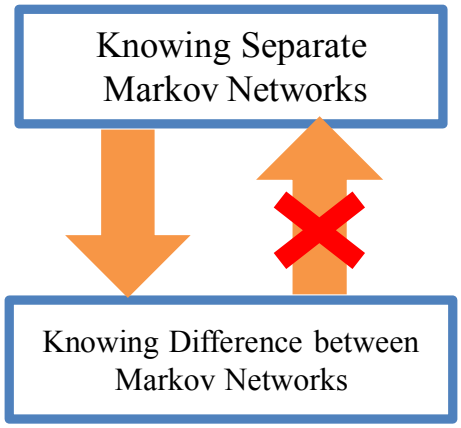


Figure 3.1: The rationale of direct structural change learning: finding the difference between two MNs is a more specific task than finding the entire structures of those two networks, and hence should be possible to learn with less data.

boundary that is sufficient for performing pattern recognition. In the current context, estimating two MNs is more general than detecting changes in MNs (Figure 3.1). By directly detecting changes in MNs, we can also halve the number of parameters, from two MNs to one MN-difference.

Another important advantage of our DRE-based method is that the normalization factor can be approximated efficiently, because the normalization term in a density ratio function takes the form of the expectation over a data distribution and thus it can be simply approximated by the sample average without additional sampling. Through experiments on gene expression and Twitter data analysis, we demonstrate the usefulness of our proposed approach.

The remainder of this paper is structured as follows. In Section 3.2, we formulate the problem of detecting structural changes and review currently available approaches. We then propose our DRE-based structural change detection method in Section 3.3. Results of illustrative and real-world experiments are reported in Section 3.4 and Section 3.5, respectively. Finally, we conclude our work and show the future direction in Section 3.7.

3.2 Problem Formulation and Related Methods

In this section, we formulate the problem of change detection in Markov network structure and review existing approaches.

3.2.1 Problem Formulation

Consider two sets of independent samples drawn separately from two probability distributions P and Q on $\mathbb{R}^d$:

$$\{\boldsymbol{x}_i^P\}_{i=1}^{n_P} \overset{\text{i.i.d.}}{\sim} P \text{ and } \{\boldsymbol{x}_i^Q\}_{i=1}^{n_Q} \overset{\text{i.i.d.}}{\sim} Q.$$

We assume that P and Q belong to the family of *Markov networks* (MNs) consisting of univariate and bivariate factors[1], i.e., their respective probability densities p and q are expressed as

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_{u,v=1,u\geq v}^{d} \boldsymbol{\theta}_{u,v}^\top \boldsymbol{f}(x^{(u)}, x^{(v)})\right), \tag{3.1}$$

where $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})^\top$ is the d-dimensional random variable, $\top$ denotes the transpose, $\boldsymbol{\theta}_{u,v}$ is the parameter vector for the elements $x^{(u)}$ and $x^{(v)}$, and

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_{1,1}^\top, \ldots, \boldsymbol{\theta}_{d,1}^\top, \boldsymbol{\theta}_{2,2}^\top, \ldots, \boldsymbol{\theta}_{d,2}^\top, \ldots, \boldsymbol{\theta}_{d,d}^\top)^\top$$

is the entire parameter vector. $\boldsymbol{f}(x^{(u)}, x^{(v)})$ is a bivariate vector-valued basis function. $Z(\boldsymbol{\theta})$ is the normalization factor defined as

$$Z(\boldsymbol{\theta}) = \int \exp\left(\sum_{u,v=1,u\geq v}^{d} \boldsymbol{\theta}_{u,v}^\top \boldsymbol{f}(x^{(u)}, x^{(v)})\right) \mathrm{d}\boldsymbol{x}.$$

$q(\boldsymbol{x};\boldsymbol{\theta})$ is defined in the same way.

Given two densities which can be parameterized using $p(\boldsymbol{x};\boldsymbol{\theta}^P)$ and $q(\boldsymbol{x};\boldsymbol{\theta}^Q)$, our goal is to discover *the changes in parameters* from P to Q, i.e., $\boldsymbol{\theta}^P - \boldsymbol{\theta}^Q$.

[1] Note that the proposed algorithm itself can be applied to *any* MNs containing more than two elements in each factor.

3.2.2 Sparse Maximum Likelihood Estimation and Graphical Lasso

Maximum likelihood estimation (MLE) with group ℓ_1-regularization has been widely used for estimating the sparse structure of MNs (Schmidt and Murphy, 2010; Ravikumar et al., 2010; Lee et al., 2007):

$$\max_{\boldsymbol{\theta}} \left[\frac{1}{n_P} \sum_{i=1}^{n_P} \log p(\boldsymbol{x}_i^P; \boldsymbol{\theta}) - \lambda \sum_{u,v=1, u \geq v}^{d} \|\boldsymbol{\theta}_{u,v}\| \right], \tag{3.2}$$

where $\|\cdot\|$ denotes the ℓ_2-norm. As λ increases, $\|\boldsymbol{\theta}_{u,v}\|$ may drop to 0. Thus, this method favors an MN that encodes more conditional independencies among variables.

Computation of the normalization term $Z(\boldsymbol{\theta})$ in Eq.(3.1) is often computationally intractable when the dimensionality of $\boldsymbol{x}$ is high. To avoid this computational problem, the Gaussian assumption is often imposed (Friedman et al., 2008; Meinshausen and Bühlmann, 2006). More specifically, the following zero-mean Gaussian model is used:

$$p(\boldsymbol{x}; \boldsymbol{\Theta}) = \frac{\det(\boldsymbol{\Theta})^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2} \boldsymbol{x}^\top \boldsymbol{\Theta} \boldsymbol{x} \right),$$

where $\boldsymbol{\Theta}$ is the inverse covariance matrix (a.k.a. the precision matrix) and $\det(\cdot)$ denotes the determinant. Then $\boldsymbol{\Theta}$ is learned as

$$\max_{\boldsymbol{\Theta}} \left[\log\det(\boldsymbol{\Theta}) - \mathrm{tr}(\boldsymbol{\Theta} \boldsymbol{S}^P) - \lambda \|\boldsymbol{\Theta}\|_1 \right],$$

where $\boldsymbol{S}^P$ is the sample covariance matrix of $\{\boldsymbol{x}_i^P\}_{i=1}^n$. $\|\boldsymbol{\Theta}\|_1$ is the ℓ_1-norm of $\boldsymbol{\Theta}$, i.e., the absolute sum of all elements. This formulation has been studied intensively in Banerjee et al. (2008), and a computationally efficient algorithm called the *graphical lasso* (Glasso) has been proposed (Friedman et al., 2008).

Sparse changes in conditional independence structure between P and Q can be detected by comparing two MNs estimated separately using sparse MLE. However, this approach implicitly assumes that two MNs are sparse, which is not necessarily true even if the change is sparse.

3.2.3 Fused-Lasso (Flasso) Method

To more naturally handle sparse changes in conditional independence structure between P and Q, a method based on *fused-lasso* (Tibshirani et al., 2005) has been developed (Zhang and Wang, 2010). This method directly sparsifies the *difference* between parameters.

The original method conducts *feature-wise neighborhood regression* (Meinshausen and Bühlmann, 2006) jointly for P and Q, which can be conceptually understood as maximizing the local conditional Gaussian likelihood jointly on each feature (Ravikumar et al., 2010). A slightly more general form of the learning criterion may be summarized as

$$\max_{\boldsymbol{\theta}_s^P, \boldsymbol{\theta}_s^Q} \left[\ell_s^P(\boldsymbol{\theta}_s^P) + \ell_s^Q(\boldsymbol{\theta}_s^Q) - \lambda_1(\|\boldsymbol{\theta}_s^P\|_1 + \|\boldsymbol{\theta}_s^Q\|_1) - \lambda_2\|\boldsymbol{\theta}_s^P - \boldsymbol{\theta}_s^Q\|_1 \right],$$

where $\ell_s^P(\boldsymbol{\theta})$ is the log conditional likelihood for the s-th element $x^{(s)} \in \mathbb{R}$ given the rest $\boldsymbol{x}^{(-s)} \in \mathbb{R}^{d-1}$:

$$\ell_s^P(\boldsymbol{\theta}) = \frac{1}{n_P} \sum_{i=1}^{n_P} \log p(x_i^{(s)P} | \boldsymbol{x}_i^{(-s)P}; \boldsymbol{\theta}).$$

$\ell_s^Q(\boldsymbol{\theta})$ is defined in the same way as $\ell_s^P(\boldsymbol{\theta})$.

Since the Flasso-based method directly sparsifies the change in MN structure, it can work well even when each MN is not sparse. However, using other models than Gaussian is difficult because of the normalization issue described in Section 3.2.2.

3.2.4 Nonparanormal Extensions

In the above methods, Gaussianity is required in practice to compute the normalization factor efficiently, which is a highly restrictive assumption. To overcome this restriction, it has become popular to perform structure learning under the *nonparanormal* settings (Liu et al., 2009, 2012), where the Gaussian distribution is replaced by a *semi-parametric Gaussian copula*.

A random vector $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})^\top$ is said to follow a *nonparanormal* distribution, if there exists a set of monotone and differentiable functions, $\{h_i(x)\}_{i=1}^d$, such that $\boldsymbol{h}(\boldsymbol{x}) = (h_1(x^{(1)}), \ldots, h_d(x^{(d)}))^\top$ follows the Gaussian distribution.

Nonparanormal distributions are much more flexible than Gaussian distributions thanks to the non-linear transformation $\{h_i(x)\}_{i=1}^d$, while the normalization factors can still be computed in an analytical way.

However, the nonparanormal transformation is restricted to be element-wise, which is still restrictive to express complex distributions.

3.2.5 Maximum Likelihood Estimation for Non-Gaussian Models by Importance-Sampling

A numerical way to obtain the MLE solution under general non-Gaussian distributions is *importance sampling*.

Suppose that we try to maximize the log-likelihood[2]:

$$\begin{aligned}\ell_{\mathrm{MLE}}(\boldsymbol{\theta}) &= \frac{1}{n_P}\sum_{i=1}^{n_P}\log p(\boldsymbol{x}_i^P;\boldsymbol{\theta}) \\ &= \frac{1}{n_P}\sum_{i=1}^{n_P}\sum_{u\geq v}\boldsymbol{\theta}_{u,v}^\top \boldsymbol{f}(x_i^{(u)P}, x_i^{(v)P}) - \log\int \exp\left(\sum_{u\geq v}\boldsymbol{\theta}_{u,v}^\top \boldsymbol{f}(x^{(u)}, x^{(v)})\right)\,\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{3.3}$$

The key idea of importance sampling is to compute the integral by the expectation over an easy-to-sample *instrumental density* $p'(\boldsymbol{x})$ (e.g., Gaussian) weighted according to the *importance* $1/p'(\boldsymbol{x})$. More specifically, using i.i.d. samples $\{\boldsymbol{x}'_i\}_{i=1}^{n'} \overset{\text{i.i.d.}}{\sim} p'(\boldsymbol{x})$, the last term of Eq.(3.3) can be approximately computed as follows:

$$\begin{aligned}\log\int \exp\left(\sum_{u\geq v}\boldsymbol{\theta}_{u,v}^\top \boldsymbol{f}(x^{(u)}, x^{(v)})\right)\,\mathrm{d}\boldsymbol{x} &= \log\int p'(\boldsymbol{x})\frac{\exp\left(\sum_{u\geq v}\boldsymbol{\theta}_{u,v}^\top \boldsymbol{f}(x^{(u)}, x^{(v)})\right)}{p'(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} \\ &\approx \log\frac{1}{n'}\sum_{i=1}^{n'}\frac{\exp\left(\sum_{u\geq v}\boldsymbol{\theta}_{u,v}^\top \boldsymbol{f}(x_i'^{(u)}, x_i'^{(v)})\right)}{p'(\boldsymbol{x}'_i)}.\end{aligned}$$

We refer to this implementation of Glasso as IS-Glasso below.

However, importance sampling tends to produce an estimate with large variance if the instrumental distribution is not carefully chosen. Although it is often

[2]From here on, we simplify $\sum_{u,v=1,u\geq v}^d$ as $\sum_{u\geq v}$.

suggested to use a density whose shape is similar to the function to be integrated but with thicker tails as p', it is not straightforward in practice to decide which p' to choose, especially when the dimensionality of $\boldsymbol{x}$ is high (Wasserman, 2010).

We can also consider an importance-sampling version of the Flasso method (which we refer to as IS-Flasso)[3]

$$\max_{\boldsymbol{\theta}^P,\boldsymbol{\theta}^Q}\left[\ell^P_{\mathrm{MLE}}(\boldsymbol{\theta}^P)+\ell^Q_{\mathrm{MLE}}(\boldsymbol{\theta}^Q)-\lambda_1(\|\boldsymbol{\theta}^P\|^2+\|\boldsymbol{\theta}^Q\|^2)-\lambda_2\sum_{u\geq v}\|\boldsymbol{\theta}^P_{u,v}-\boldsymbol{\theta}^Q_{u,v}\|\right],$$

where both $\ell^P_{\mathrm{MLE}}(\boldsymbol{\theta}^P)$ and $\ell^Q_{\mathrm{MLE}}(\boldsymbol{\theta}^Q)$ are approximated by importance sampling for non-Gaussian distributions. However, in the same way as IS-Glasso, the choice of instrumental distributions is not straightforward.

3.3 Direct Learning of Structural Changes via Density Ratio Estimation

The Flasso method can more naturally handle sparse changes in MNs than separate sparse MLE. However, the Flasso method is still based on separate modeling of two MNs, and its computation for general high-dimensional non-Gaussian distributions is challenging. In this section, we propose to directly learn structural changes based on *density ratio estimation* (Sugiyama et al., 2012a). Our approach does not involve separate modeling of each MN and allows us to approximate the normalization term efficiently for *any* distributions.

3.3.1 Density Ratio Formulation for Structural Change Detection

Our key idea is to consider the ratio of p and q:

$$\frac{p(\boldsymbol{x};\boldsymbol{\theta}^P)}{q(\boldsymbol{x};\boldsymbol{\theta}^Q)}\propto\exp\left(\sum_{u\geq v}(\boldsymbol{\theta}^P_{u,v}-\boldsymbol{\theta}^Q_{u,v})^\top\boldsymbol{f}(x^{(u)},x^{(v)})\right).$$

[3]For implementation simplicity, we maximize the joint likelihood of p and q, instead of its feature-wise conditional likelihood. We also switch the first penalty term from ℓ_1 to ℓ_2.

Here $\boldsymbol{\theta}^P_{u,v} - \boldsymbol{\theta}^Q_{u,v}$ encodes the difference between P and Q for factor $\boldsymbol{f}(x^{(u)}, x^{(v)})$, i.e., $\boldsymbol{\theta}^P_{u,v} - \boldsymbol{\theta}^Q_{u,v}$ is zero if there is no change in the factor $\boldsymbol{f}(x^{(u)}, x^{(v)})$.

Once we consider the ratio of p and q, we actually do not have to estimate $\boldsymbol{\theta}^P_{u,v}$ and $\boldsymbol{\theta}^Q_{u,v}$; instead estimating their difference $\boldsymbol{\theta}_{u,v} = \boldsymbol{\theta}^P_{u,v} - \boldsymbol{\theta}^Q_{u,v}$ is sufficient for change detection:

$$r(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{N(\boldsymbol{\theta})} \exp\left(\sum_{u \geq v} \boldsymbol{\theta}^\top_{u,v} \boldsymbol{f}(x^{(u)}, x^{(v)})\right), \tag{3.4}$$

where

$$N(\boldsymbol{\theta}) = \int q(\boldsymbol{x}) \exp\left(\sum_{u \geq v} \boldsymbol{\theta}^\top_{u,v} \boldsymbol{f}(x^{(u)}, x^{(v)})\right) \mathrm{d}\boldsymbol{x}.$$

The normalization term $N(\boldsymbol{\theta})$ guarantees[4]

$$\int q(\boldsymbol{x}) r(\boldsymbol{x}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} = 1.$$

Thus, in this density ratio formulation, we are no longer modeling p and q separately, but we model the change from p to q *directly*. This direct nature would be

[4] If the model $q(\boldsymbol{x}; \boldsymbol{\theta}^Q)$ is correctly specified, i.e., there exists $\boldsymbol{\theta}^{Q^*}$ such that $q(\boldsymbol{x}; \boldsymbol{\theta}^{Q^*}) = q(\boldsymbol{x})$, then $N(\boldsymbol{\theta})$ can be interpreted as importance sampling of $Z(\boldsymbol{\theta}^P)$ via instrumental distribution $q(\boldsymbol{x})$. Indeed, since

$$Z(\boldsymbol{\theta}^P) = \int q(\boldsymbol{x}) \frac{\exp\left(\sum_{u \geq v} {\boldsymbol{\theta}^P_{u,v}}^\top \boldsymbol{f}(x^{(u)}, x^{(v)})\right)}{q(\boldsymbol{x}; \boldsymbol{\theta}^{Q^*})} \mathrm{d}\boldsymbol{x},$$

where $q(\boldsymbol{x}; \boldsymbol{\theta}^{Q^*}) = q(\boldsymbol{x})$, we have

$$N(\boldsymbol{\theta}^P - \boldsymbol{\theta}^{Q^*}) = \frac{Z(\boldsymbol{\theta}^P)}{Z(\boldsymbol{\theta}^{Q^*})} = \int q(\boldsymbol{x}) \exp\left(\sum_{u \geq v} (\boldsymbol{\theta}^P_{u,v} - {\boldsymbol{\theta}^Q_{u,v}}^*)^\top \boldsymbol{f}(x^{(u)}, x^{(v)})\right) \mathrm{d}\boldsymbol{x}.$$

This is exactly the normalization term $N(\boldsymbol{\theta})$ of the ratio $p(\boldsymbol{x}; \boldsymbol{\theta}^P)/q(\boldsymbol{x}; \boldsymbol{\theta}^{Q^*})$. However, we note that the density ratio estimation method we use in this chapter is consistent to the optimal solution in the model even without the correct model assumption (Kanamori et al., 2010). An alternative normalization term,

$$N'(\boldsymbol{\theta}, \boldsymbol{\theta}^Q) = \int q(\boldsymbol{x}; \boldsymbol{\theta}^Q) r(\boldsymbol{x}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{x},$$

may also be considered, as in the case of MLE. However, this alternative form requires an extra parameter $\boldsymbol{\theta}^Q$ which is not our main interest.

more suitable for change detection purposes according to Vapnik's principle that encourages avoidance of solving more general problems as an intermediate step (Vap). This direct formulation also allows us to halve the number of parameters from both $\boldsymbol{\theta}^P$ and $\boldsymbol{\theta}^Q$ to only $\boldsymbol{\theta}$.

Furthermore, the normalization factor $N(\boldsymbol{\theta})$ in the density ratio formulation can be easily approximated by the sample average over $\{\boldsymbol{x}_i^Q\}_{i=1}^{n_Q} \overset{\text{i.i.d.}}{\sim} q(\boldsymbol{x})$, because $N(\boldsymbol{\theta})$ is the expectation over $q(\boldsymbol{x})$:

$$N(\boldsymbol{\theta}) \approx \frac{1}{n_Q} \sum_{i=1}^{n_Q} \exp\left(\sum_{u \geq v} \boldsymbol{\theta}_{u,v}^\top \boldsymbol{f}(x_i^{(u)Q}, x_i^{(v)Q}) \right).$$

3.3.2 Direct Density-Ratio Estimation

Density ratio estimation has been recently introduced to the machine learning community and is proven to be useful in a wide range of applications (Sugiyama et al., 2012a). Here, we concentrate on the density ratio estimator called the *Kullback-Leibler importance estimation procedure* (KLIEP) for log-linear models (Sugiyama et al., 2008; Tsuboi et al., 2009).

For a density ratio model $r(\boldsymbol{x}; \boldsymbol{\theta})$, the KLIEP method minimizes the Kullback-Leibler divergence from $p(\boldsymbol{x})$ to $\widehat{p}(\boldsymbol{x}) = q(\boldsymbol{x}) r(\boldsymbol{x}; \boldsymbol{\theta})$:

$$\begin{aligned} \mathrm{KL}[p \| \widehat{p}] &= \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x}) r(\boldsymbol{x}; \boldsymbol{\theta})} \mathrm{d}\boldsymbol{x} \\ &= \text{Const.} - \int p(\boldsymbol{x}) \log r(\boldsymbol{x}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{3.5}$$

Note that our density-ratio model (3.4) automatically satisfies the non-negativity and normalization constraints:

$$r(\boldsymbol{x}; \boldsymbol{\theta}) \geq 0 \text{ and } \int q(\boldsymbol{x}) r(\boldsymbol{x}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} = 1.$$

In practice, we maximize the empirical approximation of the second term in

Eq.(3.5):

$$\begin{aligned}\ell_{\mathrm{KLIEP}}(\boldsymbol{\theta}) &= \frac{1}{n_P}\sum_{i=1}^{n_P}\log r(\boldsymbol{x}_i^P;\boldsymbol{\theta}) \\ &= \frac{1}{n_P}\sum_{i=1}^{n_P}\sum_{u\geq v}\boldsymbol{\theta}_{u,v}^\top \boldsymbol{f}(x_i^{(u)P}, x_i^{(v)P}) \\ &\quad - \log\left(\frac{1}{n_Q}\sum_{i=1}^{n_Q}\exp\left(\sum_{u\geq v}\boldsymbol{\theta}_{u,v}^\top \boldsymbol{f}(x_i^{(u)Q}, x_i^{(v)Q})\right)\right).\end{aligned}$$

Because $\ell_{\mathrm{KLIEP}}(\boldsymbol{\theta})$ is concave with respect to $\boldsymbol{\theta}$, its global maximizer can be numerically found by standard optimization techniques such as gradient ascent or quasi-Newton methods. The gradient of ℓ_{KLIEP} with respect to $\boldsymbol{\theta}_{u,v}$ is given by

$$\begin{aligned}\nabla_{\boldsymbol{\theta}_{u,v}}\ell_{\mathrm{KLIEP}}(\boldsymbol{\theta}) =& \frac{1}{n_P}\sum_{i=1}^{n_P}\boldsymbol{f}(\boldsymbol{x}_i^{(u)P}, \boldsymbol{x}_i^{(v)P}) \\ &- \frac{\frac{1}{n_Q}\sum_{i=1}^{n_Q}\exp\left(\sum_{u'\geq v'}\boldsymbol{\theta}_{u',v'}^\top \boldsymbol{f}(x_i^{(u')Q}, x_i^{(v')Q})\right)\boldsymbol{f}(x_i^{(u)Q}, x_i^{(v)Q})}{\frac{1}{n_Q}\sum_{j=1}^{n_Q}\exp\left(\sum_{u''\geq v''}\boldsymbol{\theta}_{u'',v''}^\top \boldsymbol{f}(x_j^{(u'')Q}, x_j^{(v'')Q})\right)},\end{aligned}$$

which can be computed in a straightforward manner for *any* feature vector $\boldsymbol{f}(x^{(u)}, x^{(v)})$.

3.3.3 Sparsity-Inducing Norm

To find a sparse change between P and Q, we propose to regularize the KLIEP solution with a sparsity-inducing norm $\sum_{u\geq v}\|\boldsymbol{\theta}_{u,v}\|$. Note that the MLE approach sparsifies both $\boldsymbol{\theta}^P$ and $\boldsymbol{\theta}^Q$ so that the difference $\boldsymbol{\theta}^P - \boldsymbol{\theta}^Q$ is also sparsified, while we directly sparsify the difference $\boldsymbol{\theta}^P - \boldsymbol{\theta}^Q$; thus our method can still work well even if $\boldsymbol{\theta}^P$ and $\boldsymbol{\theta}^Q$ are dense.

In practice, we may use the following *elastic-net* penalty (Zou and Hastie, 2005) to better control overfitting to noisy data:

$$\max_{\boldsymbol{\theta}}\left[\ell_{\mathrm{KLIEP}}(\boldsymbol{\theta}) - \lambda_1\|\boldsymbol{\theta}\|^2 - \lambda_2\sum_{u\geq v}\|\boldsymbol{\theta}_{u,v}\|\right], \tag{3.6}$$

where $\|\boldsymbol{\theta}\|^2$ penalizes the magnitude of the entire parameter vector.

3.3.4 Dual Formulation for High-Dimensional Data

The solution of the optimization problem (3.6) can be easily obtained by standard sparse optimization methods. However, in the case where the input dimensionality d is high (which is often the case in our setup), the dimensionality of parameter vector $\boldsymbol{\theta}$ is large, and thus obtaining the solution can be computationally expensive. Here, we derive a dual optimization problem (Boyd and Vandenberghe, 2004), which can be solved more efficiently for high-dimensional $\boldsymbol{\theta}$ (Figure 3.2).

As detailed in Appendix, the dual optimization problem is given as

$$\begin{aligned}&\min_{\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_{n_Q})^\top} \sum_{i=1}^{n_Q} \alpha_i \log \alpha_i + \frac{1}{\lambda_1} \sum_{u \geq v} \max(0, \|\boldsymbol{\xi}_{u,v}\| - \lambda_2)^2 \\ &\text{subject to } \alpha_1, \ldots, \alpha_{n_Q} \geq 0 \text{ and } \sum_{i=1}^{n_Q} \alpha_i = 1,\end{aligned} \tag{3.7}$$

where

$$\begin{aligned}\boldsymbol{\xi}_{u,v} &= \boldsymbol{g}_{u,v} - \boldsymbol{H}_{u,v}\boldsymbol{\alpha}, \\ \boldsymbol{H}_{u,v} &= [\boldsymbol{f}(x_1^{(u)Q}, x_1^{(v)Q}), \ldots, \boldsymbol{f}(x_{n_Q}^{(u)Q}, x_{n_Q}^{(v)Q})], \\ \boldsymbol{g}_{u,v} &= \frac{1}{n_P} \sum_{i=1}^{n_P} \boldsymbol{f}(x_i^{(u)P}, x_i^{(v)P}).\end{aligned}$$

The primal solution can be obtained from the dual solution as

$$\boldsymbol{\theta}_{u,v} = \begin{cases} \dfrac{1}{\lambda_1}\left(1 - \dfrac{\lambda_2}{\|\boldsymbol{\xi}_{u,v}\|}\right)\boldsymbol{\xi}_{u,v} & \text{if } \|\boldsymbol{\xi}_{u,v}\| > \lambda_2, \\ \mathbf{0} & \text{if } \|\boldsymbol{\xi}_{u,v}\| \leq \lambda_2. \end{cases} \tag{3.8}$$

Note that the dimensionality of the dual variable $\boldsymbol{\alpha}$ is equal to n_Q, while that of $\boldsymbol{\theta}$ is quadratic with respect to the input dimensionality d, because we are considering pairwise factors. Thus, if d is not small and n_Q is not very large (which is often the case in our experiments shown later), solving the dual optimization problem would be computationally more efficient. Furthermore, the dual objective (and its gradient) can be computed efficiently in parallel for each (u, v), which is a useful property when handling large-scale MNs. Note that the dual objective is differentiable everywhere, while the primal objective is not.

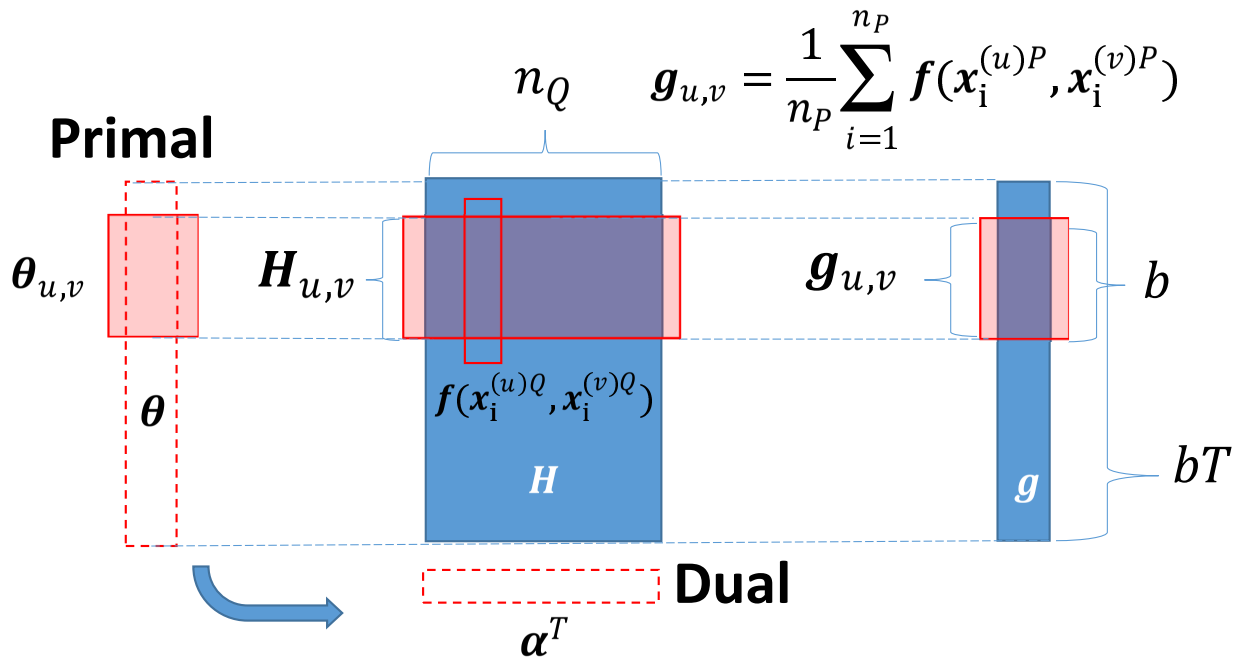


Figure 3.2: Schematics of primal and dual optimization. b denotes the number of basis functions and T denotes the number of factors. Because we are considering pairwise factors, $T = \mathcal{O}(d^2)$ for input dimensionality d.

3.4 Numerical Experiments

In this section, we compare the performance of the proposed KLIEP-based method, the Flasso method, and the Glasso method for Gaussian models, nonparanormal models, and non-Gaussian models. Results are reported on datasets with three different underlying distributions: multivariate Gaussian, nonparanormal, and non-Gaussian "diamond" distributions. We also investigate the computation time of the primal and dual formulations as a function of the input dimensionality. The MATLAB implementation of the primal and dual methods are available at

`http://sugiyama-www.cs.titech.ac.jp/~song/SCD.html`.

3.4.1 Gaussian Distribution

First, we investigate the performance of each method under Gaussianity.

Consider a 40-node sparse Gaussian MN, where its graphical structure is characterized by precision matrix $\mathbf{\Theta}^P$ with diagonal elements equal to 2. The off-diagonal elements are randomly chosen[5] and set to 0.2, so that the overall sparsity of $\mathbf{\Theta}^P$ is 25%. We then introduce changes by randomly picking 15 edges and reducing the corresponding elements in the precision matrix by 0.1. The resulting precision matrices $\mathbf{\Theta}^P$ and $\mathbf{\Theta}^Q$ are used for drawing samples as

$$\{\boldsymbol{x}_i^P\}_{i=1}^{n_P} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, (\mathbf{\Theta}^P)^{-1}) \text{ and } \{\boldsymbol{x}_i^Q\}_{i=1}^{n_Q} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, (\mathbf{\Theta}^Q)^{-1}),$$

where $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ denotes the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. Datasets of size $n = n_P = n_Q = 50, 100$ are tested.

We compare the performance of the KLIEP, Flasso, and Glasso methods. Because all methods use the same Gaussian model, the difference in performance is caused only by the difference in estimation methods. We repeat the experiments 20 times with randomly generated datasets and report the results in Figure 3.3.

The top 6 graphs are examples of regularization paths[6]. The dashed lines represent changed edges in the ground truth, while the solid lines represent unchanged edges. The top row is for $n = 100$ while the middle row is for $n = 50$.

[5] We set $\Theta_{u,v} = \Theta_{v,u}$ for not breaking the symmetry of the precision matrix.

[6] Paths of univariate factors are omitted for clear visibility.

The bottom 3 graphs are the data generating distribution and averaged precision-recall (P-R) curves with standard error over 20 runs. The P-R curves are plotted by varying the group-sparsity control parameter λ_2 with $\lambda_1 = 0$ in KLIEP and Flasso, and by varying the sparsity control parameters as $\lambda = \lambda^P = \lambda^Q$ in Glasso.

In the regularization path plots, solid vertical lines show the regularization parameter values picked based on hold-out data $\{\widetilde{\boldsymbol{x}}_i^P\}_{i=1}^{3000} \overset{\text{i.i.d.}}{\sim} P$ and $\{\widetilde{\boldsymbol{x}}_i^Q\}_{i=1}^{3000} \overset{\text{i.i.d.}}{\sim} Q$ as follows:

- **KLIEP:** The *hold-out log-likelihood* (HOLL) is maximized:

$$\frac{1}{\widetilde{n}_P}\sum_{i=1}^{\widetilde{n}_P}\log\frac{\exp\left(\sum_{u\geq v}\widehat{\boldsymbol{\theta}}_{u,v}^{\top}\boldsymbol{f}(\widetilde{x}_i^{(u)P},\widetilde{x}_i^{(v)P})\right)}{\frac{1}{\widetilde{n}_Q}\sum_{j=1}^{\widetilde{n}_Q}\exp\left(\sum_{u'\geq v'}\widehat{\boldsymbol{\theta}}_{u',v'}^{\top}\boldsymbol{f}(\widetilde{x}_j^{(u')Q},\widetilde{x}_j^{(v')Q})\right)}.$$

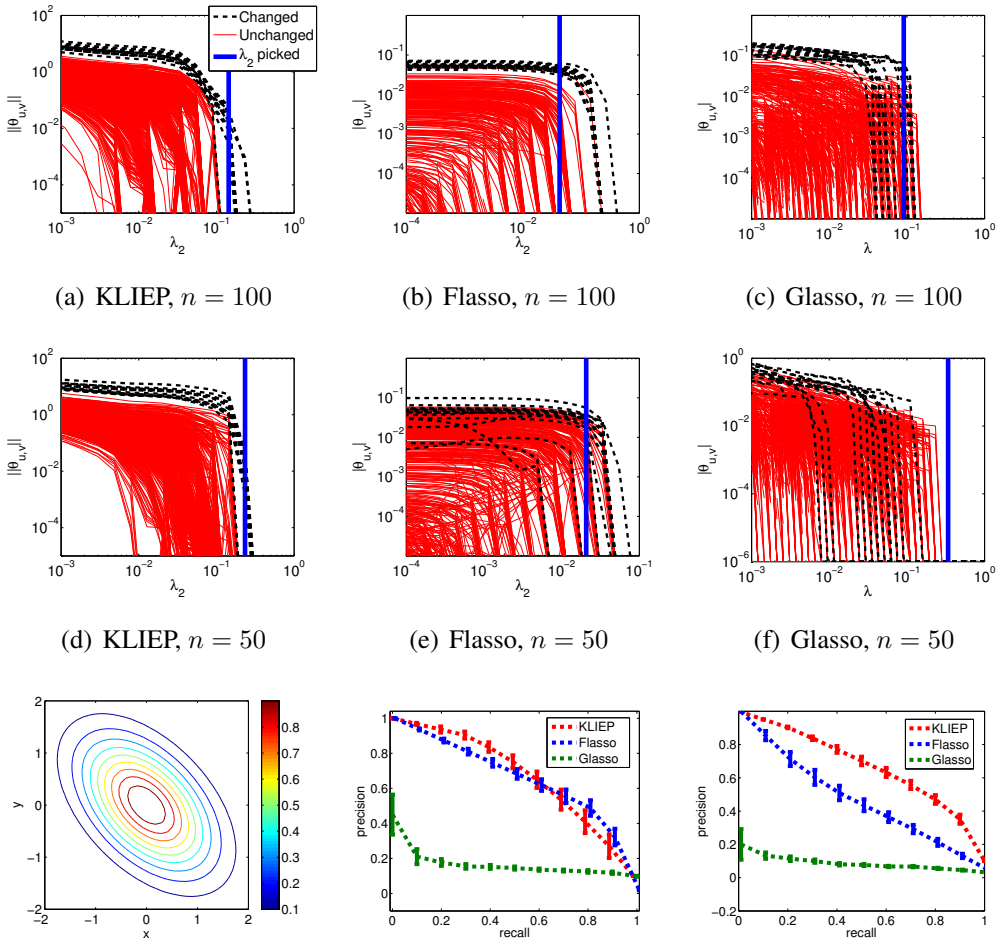
- **Flasso:** The sum of feature-wise conditional HOLLs for $p(x^{(s)}|\boldsymbol{x}^{(-s)};\boldsymbol{\theta}_s)$ and $q(x^{(s)}|\boldsymbol{x}^{(-s)};\boldsymbol{\theta}_s)$ over all nodes is maximized:

$$\frac{1}{\widetilde{n}_P}\sum_{i=1}^{\widetilde{n}_P}\sum_{s=1}^{d}\log p(\widetilde{x}_i^{(s)P}|\widetilde{\boldsymbol{x}}_i^{(-s)P};\widehat{\boldsymbol{\theta}}_s^P)+\frac{1}{\widetilde{n}_Q}\sum_{i=1}^{\widetilde{n}_Q}\sum_{s=1}^{d}\log q(\widetilde{x}_i^{(s)Q}|\widetilde{\boldsymbol{x}}_i^{(-s)Q};\widehat{\boldsymbol{\theta}}_s^Q).$$

- **Glasso:** The sum of HOLLs for $p(\boldsymbol{x};\boldsymbol{\theta})$ and $q(\boldsymbol{x};\boldsymbol{\theta})$ is maximized:

$$\frac{1}{\widetilde{n}_P}\sum_{i=1}^{\widetilde{n}_P}\log p(\widetilde{\boldsymbol{x}}_i^P;\widehat{\boldsymbol{\theta}}^P)+\frac{1}{\widetilde{n}_Q}\sum_{i=1}^{\widetilde{n}_Q}\log q(\widetilde{\boldsymbol{x}}_i^Q;\widehat{\boldsymbol{\theta}}^Q).$$

When $n = 100$, KLIEP and Flasso clearly distinguish changed (dashed lines) and unchanged (solid lines) edges in terms of parameter magnitude. However, when the sample size is halved to $n = 50$, the separation is visually rather unclear in the case of Flasso. In contrast, the paths of changed and unchanged edges are still almost disjoint in the case of KLIEP. The Glasso method performs rather poorly in both cases. A similar tendency can be observed also in the P-R curve plot: When the sample size is $n = 100$, KLIEP and Flasso work equally well, but KLIEP gains its lead when the sample size is reduced to $n = 50$. Glasso does not perform well in both cases.



(a) KLIEP, $n = 100$ (b) Flasso, $n = 100$ (c) Glasso, $n = 100$

(d) KLIEP, $n = 50$ (e) Flasso, $n = 50$ (f) Glasso, $n = 50$

(g) Gaussian distribution (h) P-R curve, $n = 100$ (i) P-R curve, $n = 50$

Figure 3.3: Experimental results on the Gaussian dataset.

3.4.2 Nonparanormal Distribution

We post-process the Gaussian dataset used in Section 3.4.1 to construct nonparanormal samples. More specifically, we apply the power function,

$$h_i^{-1}(x) = \mathrm{sign}(x)|x|^{\frac{1}{2}},$$

to each dimension of $\boldsymbol{x}^P$ and $\boldsymbol{x}^Q$, so that $\boldsymbol{h}(\boldsymbol{x}^P) \sim \mathcal{N}(\mathbf{0}, (\boldsymbol{\Theta}^P)^{-1})$ and $\boldsymbol{h}(\boldsymbol{x}^Q) \sim \mathcal{N}(\mathbf{0}, (\boldsymbol{\Theta}^Q)^{-1})$.

To cope with the non-linearity in the KLIEP method, we use the power nonparanormal basis functions with power $k = 2$, 3, and 4:

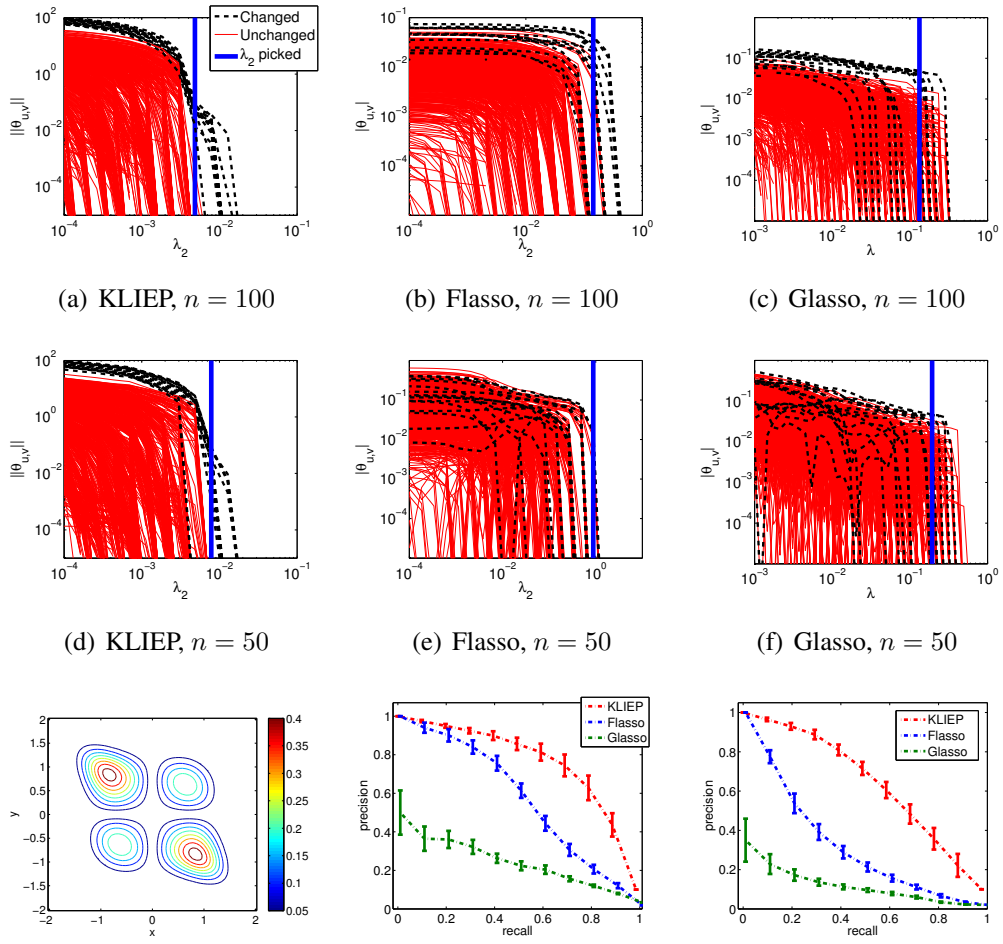
$$\boldsymbol{f}(x_i, x_j) = (\mathrm{sign}(x_i)|x_i|^k, \mathrm{sign}(x_j)|x_j|^k, 1)^\top.$$

Model selection of k is performed together with the regularization parameter by HOLL maximization. For Flasso and Glasso, we apply the nonparanormal transform as described in Liu et al. (2009) before the structural change is learned.

The experiments are conducted on 20 randomly generated datasets with $n = 50$ and 100, respectively. The regularization paths, data generating distribution, and averaged P-R curves are plotted in Figure 3.4. The results show that Flasso clearly suffers from the performance degradation compared with the Gaussian case, perhaps because the number of samples is too small for the complicated nonparanormal distribution. Due to the two-step estimation scheme, the performance of Glasso is poor. In contrast, KLIEP separates changed and unchanged edges still clearly for both $n = 50$ and $n = 100$. The P-R curves also show the same tendency.

3.4.3 "Diamond" Distribution with No Pearson Correlation

In the experiments in Section 3.4.2, though samples are non-Gaussian, the *Pearson correlation* is not zero. Therefore, methods assuming Gaussianity can still capture some linear correlation between random variables. Here, we consider a more challenging case with a diamond-shaped distribution within the exponential family that has zero Pearson correlation between variables. Thus, the methods assuming Gaussianity cannot extract any information in principle from this dataset.



(a) KLIEP, $n = 100$ (b) Flasso, $n = 100$ (c) Glasso, $n = 100$

(d) KLIEP, $n = 50$ (e) Flasso, $n = 50$ (f) Glasso, $n = 50$

(g) Nonparanormal distribution (h) P-R curve, $n = 100$ (i) P-R curve, $n = 50$

Figure 3.4: Experimental results on the nonparanormal dataset.

The probability density function of the diamond distribution is defined as follows (Figure 3.5(a)):

$$p(\boldsymbol{x}) \propto \exp\left(-\sum_{i=1}^{d} 2x_i^2 - \sum_{(i,j):A_{i,j}\neq 0} 20x_i^2 x_j^2\right), \tag{3.9}$$

where the adjacency matrix $\boldsymbol{A}$ describes the MN structure. Note that this distribution cannot be transformed into a Gaussian distribution by any nonparanormal transformations.

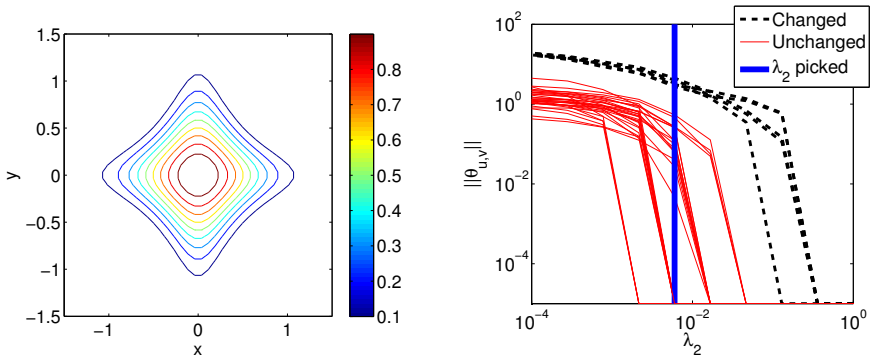
We set $d = 9$ and $n_P = n_Q = 5000$. $\boldsymbol{A}^P$ is randomly generated with 35% sparsity, while $\boldsymbol{A}^Q$ is created by randomly removing edges in $\boldsymbol{A}^P$ so that the sparsity level is dropped to 15%. Samples from the above distribution are drawn by using a *slice sampling* method (Neal, 2003). Since generating samples from high-dimensional distributions is non-trivial and time-consuming, we focus on a relatively low-dimensional case. To avoid sampling error which may mislead the experimental evaluation, we also increase the sample size, so that the erratic points generated by accident will not affect the overall population.

In this experiment, we compare the performance of KLIEP, Flasso, and Glasso with the Gaussian model, the power nonparanormal model, and the polynomial model:

$$\boldsymbol{f}(x_i, x_j) = (x_i^k, x_j^k, x_i x_j^{k-1}, \ldots, x_i^{k-1}x_j, x_i^{k-1}, x_j^{k-1}, \ldots, x_i, x_j, 1)^\top \text{ for } i \neq j.$$

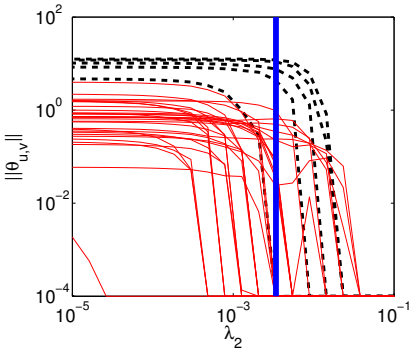
The univariate polynomial transform is defined as $\boldsymbol{f}(x_i, x_i) = \boldsymbol{f}(x_i, 0)$. We test $k = 2, 3, 4$ and choose the best one in terms of HOLL. The Flasso and Glasso methods for the polynomial model are computed by importance sampling, i.e., we use the IS-Flasso and IS-Glasso methods (see Section 3.2.5). Since these methods are computationally very expensive, we only test $k = 4$ which we found to be a reasonable choice. We set the instrumental distribution p' as the standard normal $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and use sample $\{\boldsymbol{x}'_i\}_{i=1}^{70000} \sim p'$ for approximating integrals. p' is purposely chosen so that it has a similar "bell" shape to the target densities but with larger variance on each dimension.

The averaged P-R curves over 20 datasets are shown in Figure 3.5(e). KLIEP with the polynomial model significantly outperforms all the other methods, while the IS-Glasso and especially IS-Flasso give better result than the KLIEP, Flasso,

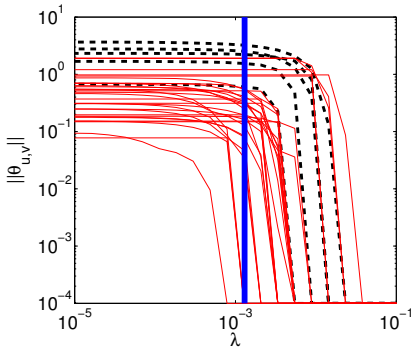


(a) Diamond distribution

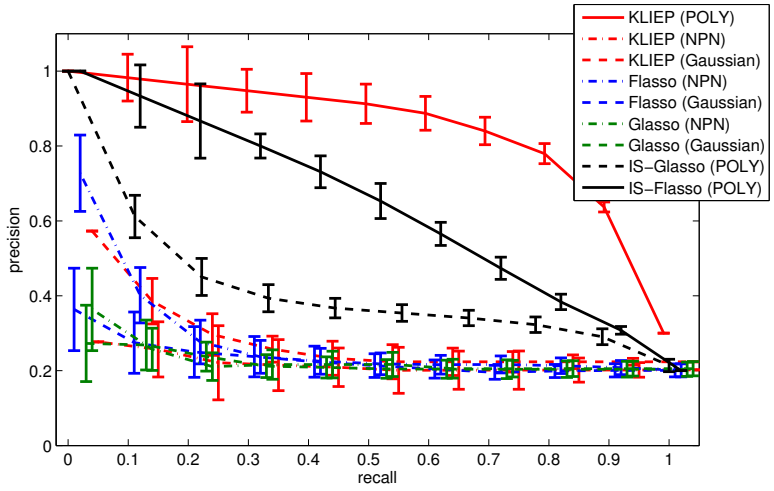
(b) KLIEP



(c) IS-Flasso



(d) IS-Glasso



(e) P-R curve

Figure 3.5: Experimental results on the diamond dataset. "NPN" and "POLY" denote the nonparanormal and polynomial models, respectively. Note that the precision rate of 100% recall for a random guess is approximately 20%.

and Glasso methods with the Gaussian and nonparanormal models. This means that the polynomial basis function is indeed helpful in handling completely non-Gaussian data. However, as discussed in Section 3.2.2, it is difficult to use such a basis function in Glasso and Flasso because of the computational intractability of the normalization term. Although IS-Glasso can approximate integrals, the result shows that such approximation of integrals does not lead to a very good performance. In comparison, the result of the IS-Flasso method is much improved thanks to the coupled sparsity regularization, but it is still not comparable to KLIEP.

The regularization paths of KLIEP with the polynomial model illustrated in Figure 3.5(b) show the usefulness of the proposed method in change detection under non-Gaussianity. We also give regularization paths obtained by the IS-Flasso and IS-Glasso methods on the same dataset in Figures 3.5(c) and 3.5(d), respectively. The graphs show that both methods do not separate changed and unchanged edges well, though the IS-Flasso method works slightly better.

3.4.4 Computation Time: Dual versus Primal Optimization Problems

Finally, we compare the computation time of the proposed KLIEP method when solving the dual optimization problem (3.7) and the primal optimization problem (3.6). Both the optimization problems are solved by using the same convex optimizer *minFunc*[7]. The datasets are generated from two Gaussian distributions constructed in the same way as Section 3.4.1. 150 samples are separately drawn from two distributions with dimension $d = 40, 50, 60, 70, 80$. We then perform change detection by computing the regularization paths using 20 choices of λ_2 ranging from 10^{-4} to 10^{0} and fix $\lambda_1 = 0.1$. The results are plotted in Figure 3.6.

It can be seen from the graph that as the dimensionality increases, the computation time for solving the primal optimization problem is sharply increased, while that for solving the dual optimization problem grows only moderately: when $d = 80$, the computation time for obtaining the primal solution is almost 10 times more than that required for obtaining the dual solution. Thus, the dual formulation

[7] `http://www.di.ens.fr/~mschmidt/Software/minFunc.html`

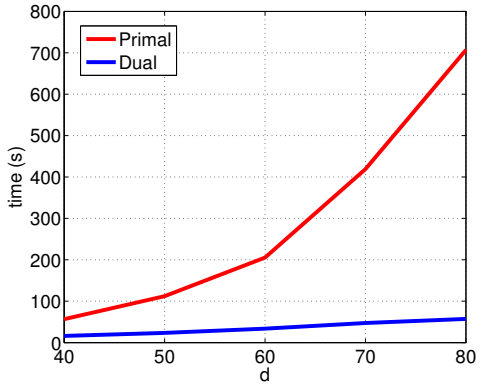


Figure 3.6: Comparison of computation time for solving primal and dual optimization problems.

is computationally much more efficient than the primal formulation.

3.5 Applications

In this section, we report the experimental results on a synthetic gene expression dataset and a Twitter dataset.

3.5.1 Synthetic Gene Expression Dataset

A gene regulatory network encodes interactions between DNA segments. However, the way genes interact may change due to environmental or biological stimuli. In this experiment, we focus on detecting such changes. We use *SynTReN*, which is a generator of gene regulatory networks used for benchmark validation of bioinformatics algorithms (Van den Bulcke et al., 2006).

We first choose a sub-network containing 13 nodes from an existing signaling network in *Saccharomyces cerevisiae* (shown in Figure 3.7(a)). Three types of interactions are modeled: activation (ac), deactivation (re), and dual (du). 50 samples are generated in the first stage, after which we change the types of interactions in 6 edges, and generate 50 samples again. Four types of changes are considered: ac $\rightarrow$ re, re $\rightarrow$ ac, du $\rightarrow$ ac, and du $\rightarrow$ re.

We use KLIEP and IS-Flasso with the polynomial transform function for $k \in \{2, 3, 4\}$. The regularization parameter λ_1 in KLIEP and Flasso is tested with choices $\lambda_1 \in \{0.1, 1, 10\}$. We set the instrumental distribution p' as the standard normal $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and use sample $\{\boldsymbol{x}'_i\}_{i=1}^{70000} \sim p'$ for approximating integrals in IS-Flasso.

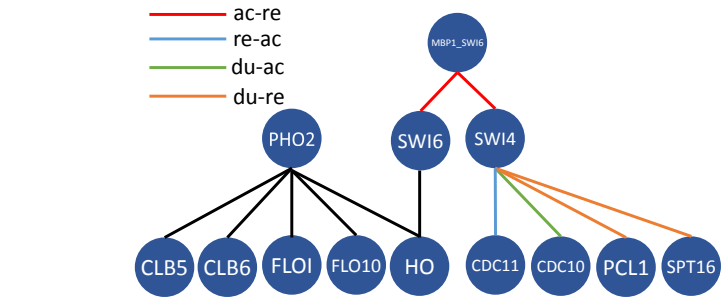
The regularization paths on one example dataset for KLIEP, IS-Flasso, and the plain Flasso with the Gaussian model are plotted in Figures 3.7(b), 3.7(c), and 3.7(d), respectively. Averaged P-R curves over 20 simulation runs are shown in Figure 3.7(e). We can see clearly from the KLIEP regularization paths shown in Figure 3.7(b) that the magnitude of estimated parameters on the changed pairwise interactions is much higher than that of the unchanged edges. IS-Flasso also achieves rather clear separation between changed and unchanged interactions, though there are a few unchanged interactions drop to zero at the final stage. Flasso gives many false alarms by assigning non-zero values to the unchanged edges, even after some changed edges hit zeros.

Reflecting a similar pattern, the P-R curves plotted in Figure 3.7(e) show that the proposed KLIEP method has the best performance among all three methods. We can also see that the IS-Flasso method achieves significant improvement over the plain Flasso method with the Gaussian model. The improvement from Flasso to IS-Flasso shows that the use of the polynomial basis is useful on this dataset, and the improvement from IS-Flasso to KLIEP shows that the direct estimation can further boost the performance.

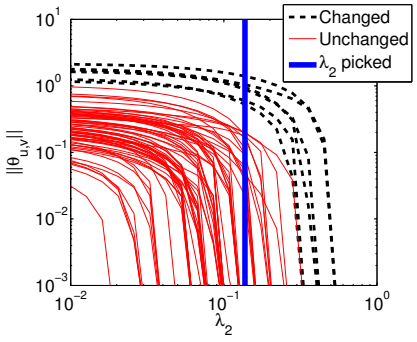
3.5.2 Twitter Story Telling

Finally, we use KLIEP with the polynomial transform function for $k \in \{2, 3, 4\}$ and Flasso as event detectors from Twitter. More specifically, we choose the *Deepwater Horizon oil spill*[8] as the target event, and we hope that our method can recover some story lines from Twitter as the news events develop. Counting the frequencies of 10 keywords (BP, oil, spill, Mexico, gulf, coast, Hayward, Halliburton, Transocean, and Obama), we obtain a dataset by sampling 4 times per day from February 1st, 2010 to October 15th, 2010, resulting in 1061 data

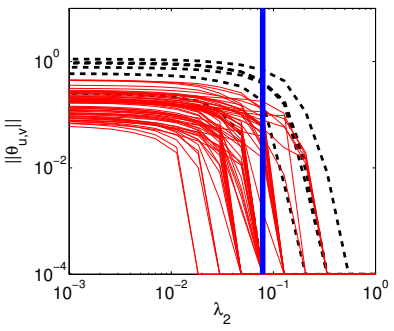
[8] `http://en.wikipedia.org/wiki/Deepwater_Horizon_oil_spill`



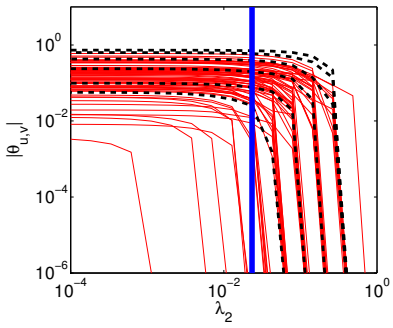
(a) Gene regulatory network



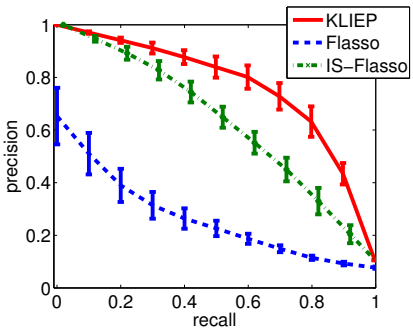
(b) KLIEP



(c) IS-Flasso



(d) Flasso



(e) P-R curve

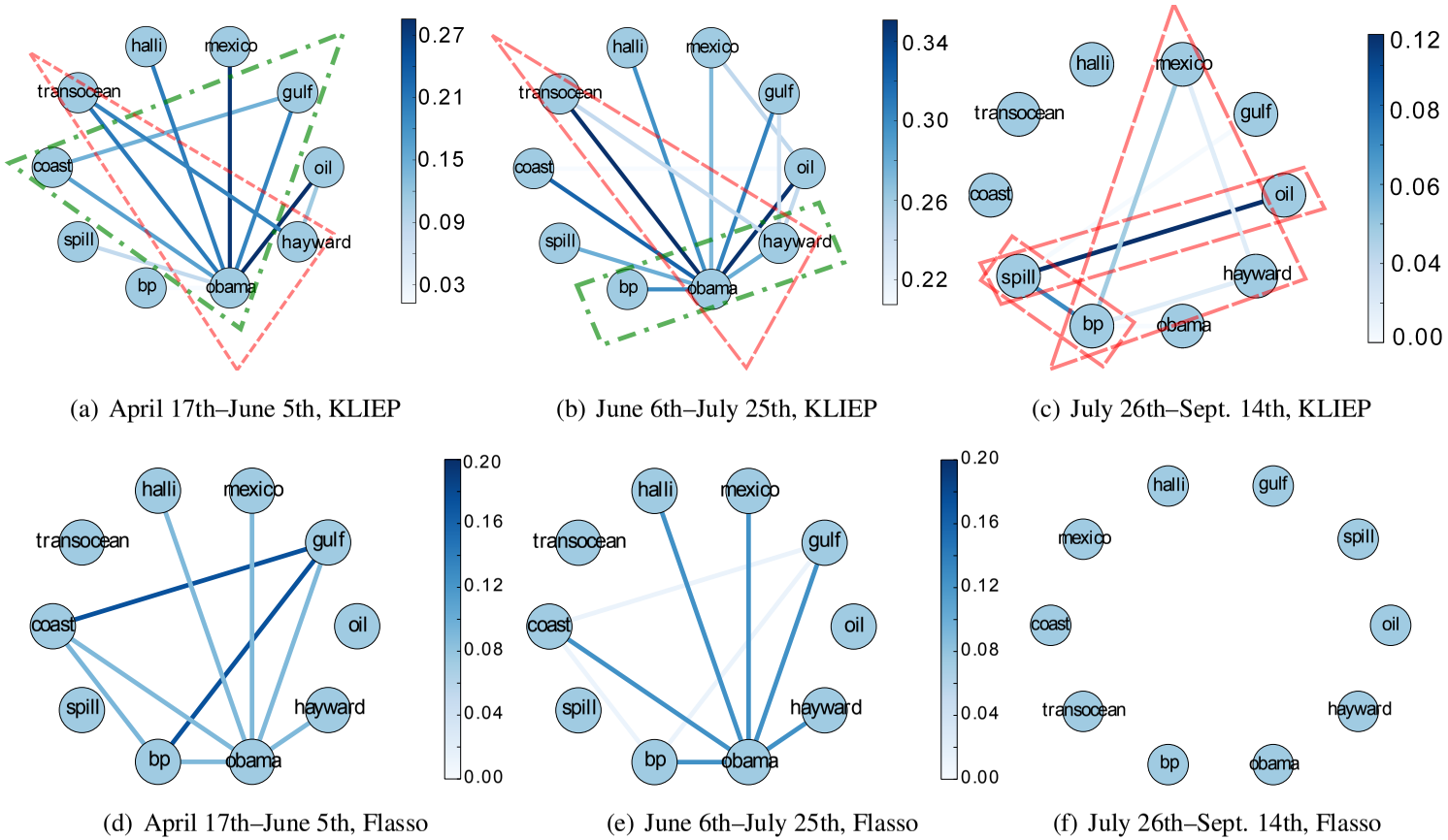
Figure 3.7: Experiments on synthetic gene expression datasets.

samples.

We segment the data into two parts: the first 300 samples collected before the day of oil spill (April 20th, 2010) are regarded as conforming to a 10-dimensional joint distribution Q, while the second set of samples that are in an arbitrary 50-day window after the oil spill accident happened is regarded as following distribution P. Thus, the MN of Q encodes the original conditional independence of frequencies between 10 keywords, while the underlying MN of P has changed since an event occurred. We expect that unveiling changes in MNs between P and Q can recover the drift of popular topic trends on Twitter in terms of the dependency among keywords.

The detected change graphs (i.e., the graphs with only detected changing edges) on 10 keywords are illustrated in Figure 3.8. The edges are selected at a certain value of λ_2 indicated by the maximal *cross-validated log-likelihood* (CVLL). Since the edge set that is picked by CVLL may not be sparse in general, we sparsify the graph based on the permutation test as follows: we randomly shuffle the samples between P and Q and repeatedly run change detection algorithms for 100 times; then we observe detected edges by CVLL. Finally, we select the edges that are detected using the original non-shuffled dataset and remove those that were detected in the shuffled datasets for more than 5 times (i.e., the significance level 5%). For KLIEP, k is also tuned by using CVLL. In Figure 3.8, we plot detected change graphs which are generated using samples of P starting from April 17th, July 6th, and July 26th, respectively.

The initial explosion happened on April 20th, 2010. Both methods discover dependency changes between keywords. Generally speaking, KLIEP captures more conditional independence changes between keywords than the Flasso method, especially when comparing Figure 3.8(c) and Figure 3.8(f). At the first two stages (Figures 3.8(a), 3.8(b), 3.8(d) and 3.8(e)), the keyword "Obama" is very well connected with other keywords in the results given by both methods. Indeed, at the early development of this event, he lies in the center of the news stories, and his media exposure peaks after his visit to the Louisiana coast (May 2nd, May 28nd, and June 5th) and his meeting with BP CEO Tony Hayward on June 16th. Notably, both methods highlight the "gulf-obama-coast" triangle in Figures 3.8(a) and 3.8(d) and the "bp-obama-hayward" chain in Figures 3.8(b)



(a) April 17th–June 5th, KLIEP

(b) June 6th–July 25th, KLIEP

(c) July 26th–Sept. 14th, KLIEP

(d) April 17th–June 5th, Flasso

(e) June 6th–July 25th, Flasso

(f) July 26th–Sept. 14th, Flasso

Figure 3.8: Change graphs captured by the proposed KLIEP method (top) and the Flasso method (bottom). The date range beneath each figure indicates when P was sampled, while Q is fixed to dates from February 1st to April 20th. Notable structures shared by the graph of both methods are surrounded by the dash-dotted lines. Unique structures that only appear in the graph of the proposed KLIEP method are surrounded by the dashed lines.

and 3.8(e).

However, there are some important differences worth mentioning. First, the Flasso method misses the "transocean-hayward-obama" triangle in Figures 3.8(d) and 3.8(e). Transocean is the contracted operator in the Deepwater Horizon platform, where the initial explosion happened. On Figure 3.8(c), the chain "bp-spill-oil" may indicate that the phrase "bp spill" or "oil spill" has been publicly recognized by the Twitter community since then, while the "hayward-bp-mexico" triangle, although relatively weak, may link to the event that Hayward stepped down from the CEO position on July 27th.

It is also noted that Flasso cannot find any changed edges in Figure 3.8(f), perhaps due to the Gaussian restriction.

3.6 Derivation of the Dual Optimization Problem

First, we rewrite the optimization problem (3.6) as

$$\min_{\boldsymbol{\theta},\boldsymbol{w}} \left[\log\left(\sum_{i=1}^{n_Q} \exp\left(w_i\right) \right) - \boldsymbol{\theta}^\top \boldsymbol{g} + \frac{\lambda_1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} + \lambda_2 \sum_{u \geq v} \|\boldsymbol{\theta}_{u,v}\| - C \right] \tag{3.10}$$
$$\text{subject to } \boldsymbol{w} = \boldsymbol{H}^\top \boldsymbol{\theta},$$

where

$$\begin{aligned}
\boldsymbol{w} &= (w_1, \ldots, w_{n_Q})^\top, \\
\boldsymbol{H} &= (\boldsymbol{H}_{1,1}^\top, \ldots, \boldsymbol{H}_{d,1}^\top, \boldsymbol{H}_{2,2}^\top, \ldots, \boldsymbol{H}_{d,2}^\top, \ldots, \boldsymbol{H}_{d,d}^\top)^\top, \\
\boldsymbol{H}_{u,v} &= [\boldsymbol{f}(x_1^{(u)Q}, x_1^{(v)Q}), \ldots, \boldsymbol{f}(x_{n_Q}^{(u)Q}, x_{n_Q}^{(v)Q})], \\
\boldsymbol{g} &= (\boldsymbol{g}_{1,1}^\top, \ldots, \boldsymbol{g}_{d,1}^\top, \boldsymbol{g}_{2,2}^\top, \ldots, \boldsymbol{g}_{d,2}^\top, \ldots, \boldsymbol{g}_{d,d}^\top)^\top, \\
\boldsymbol{g}_{u,v} &= \frac{1}{n_P} \sum_{i=1}^{n_P} \boldsymbol{f}(x_i^{(u)P}, x_i^{(v)P}), \\
C &= \log n_Q.
\end{aligned}$$

With Lagrange multipliers $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n_Q})^\top$, the Lagrangian of (3.10) is given as

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\alpha}) &= \min_{\boldsymbol{w},\boldsymbol{\theta}} \left[\log \sum_{i=1}^{n_Q} \exp(w_i) - \boldsymbol{\theta}^\top \boldsymbol{g} + \frac{\lambda_1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} + \lambda_2 \sum_{u \geq v} \|\boldsymbol{\theta}_{u,v}\| - (\boldsymbol{w} - \boldsymbol{H}^\top \boldsymbol{\theta})^\top \boldsymbol{\alpha} \right] - C \\
&= \min_{\boldsymbol{w}} \left[\log \sum_{i=1}^{n_Q} \exp(w_i) - \boldsymbol{w}^\top \boldsymbol{\alpha} \right] \\
&\quad + \min_{\boldsymbol{\theta}} \left[\boldsymbol{\theta}^\top (\boldsymbol{H}\boldsymbol{\alpha} - \boldsymbol{g}) + \frac{\lambda_1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} + \lambda_2 \sum_{u \geq v} \|\boldsymbol{\theta}_{u,v}\| \right] - C \\
&= \min_{\boldsymbol{w}} \psi_1(\boldsymbol{w}) + \min_{\boldsymbol{\theta}} \psi_2(\boldsymbol{\theta}) - C. \qquad (3.11)
\end{aligned}$$

A few lines of algebra can show that $\psi_1(\boldsymbol{w})$ reaches the minimum $-\sum_{i=1}^{n_Q} \alpha_i \log \alpha_i$ at

$$\alpha_i = \frac{\exp(w_i)}{\sum_{i=1}^{n_Q} \exp(w_i)}, \quad i = 1, \ldots, n_Q.$$

Note that extra constraints are implied from the above equation:

$$\alpha_1, \ldots, \alpha_{n_Q} \geq 0 \text{ and } \sum_{i=1}^{n_Q} \alpha_i = 1.$$

Since $\psi_2(\boldsymbol{\theta})$ is not differentiable at $\boldsymbol{\theta}_{u,v} = \mathbf{0}$, we can only obtain its sub-gradient:

$$\nabla_{\boldsymbol{\theta}_{u,v}} \psi_2(\boldsymbol{\theta}) = -\boldsymbol{\xi}_{u,v} + \lambda_1 \boldsymbol{\theta} + \lambda_2 \nabla_{\boldsymbol{\theta}_{u,v}} \|\boldsymbol{\theta}_{u,v}\|,$$

where

$$\boldsymbol{\xi}_{u,v} = \boldsymbol{g}_{u,v} - \boldsymbol{H}_{u,v}\boldsymbol{\alpha},$$

$$\nabla_{\boldsymbol{\theta}_{u,v}} \|\boldsymbol{\theta}_{u,v}\| = \begin{cases} \dfrac{\boldsymbol{\theta}_{u,v}}{\|\boldsymbol{\theta}_{u,v}\|} & \text{if } \boldsymbol{\theta}_{u,v} \neq \mathbf{0}, \\ \{\boldsymbol{y} \mid \|\boldsymbol{y}\| \leq 1\} & \text{if } \boldsymbol{\theta}_{u,v} = \mathbf{0}. \end{cases}$$

By setting $\nabla_{\boldsymbol{\theta}_t} \psi_2(\boldsymbol{\theta}) = \mathbf{0}$, we can obtain the solution to this minimization problem by Eq.(3.8).

Substituting the solutions of the above two minimization problems with respect to $\boldsymbol{\theta}$ and $\boldsymbol{w}$ into (3.11), we obtain the dual optimization problem (3.7).

3.7 Conclusion

In this chapter, we proposed a *direct* approach to learning sparse changes in MNs by density ratio estimation. Rather than fitting two MNs separately to data and comparing them to detect a change, we estimated the ratio of the probability densities of two MNs where changes can be naturally encoded as sparsity patterns in estimated parameters. This direct modeling allows us to halve the number of parameters and approximate the normalization term in the density ratio model by a sample average without sampling. We also showed that the number of parameters to be optimized can be further reduced with the dual formulation, which is highly useful when the dimensionality is high. Through experiments on artificial and real-world datasets, we demonstrated the usefulness of the proposed method over state-of-the-art methods including nonparanormal-based methods and sampling-based methods.

Chapter 4

Conclusions and Future Works

4.1 Conclusions

This thesis is devoted to statistical machine learning approaches to unsupervised change detection problems. In Chapter 1, We introduced machine learning tasks under the static view and dynamic view respectively. Under the dynamic view, we focus on detecting the changes of patterns from two sets of data. Two tasks were investigated: distributional change detection and structural change detection. For each task, we considered tackling one of the major issues.

In Chapter 2, the distributional change detection was formulated as testing the statistical divergence between two consecutive segments of time-series data. To improve the accuracy of the distributional change detection, we employed the latest advances of density ratio estimation and proposed a flexible and robust algorithm. The proposed algorithm extended the previous effort of non-parametric change-point detection method, and used *unconstrained Least-Square Importance Fitting (uLSIF)* as a building block. Comparing to the previous density ratio estimation method, uLSIF enjoys various theoretical and practical advantages. Furthermore, we proposed to use *Relative Pearson Divergence*, which has been proposed recently, as change-point score. Through experiments on toy and real-world datasets, we demonstrated the proposed method was promising.

In Chapter 3, to solve the problem of structural change detection in pairwise Markov network, we formulated the problem into density ratio estimation, and estimated the density ratio directly using log-linear model. In order to obtain in-

terpretable results, the group sparse regularizer was adopted which helped produce a result with sparsity. We also derived a dual objective function, and experimental results showed that optimizing the dual objective was much faster than the primal objective. Comparing to existing methods, the proposed method adopted a single-shot procedure instead of twice estimation. Moreover, its novel density ratio formulation allowed us to consider a far richer distribution family rather than only Gaussian distributions. Through experiments on both toy and real-world datasets, we demonstrated the usefulness of our methods.

4.2 Discussions

The focus of this thesis was purely on *statistical changes*, i.e. changes between probability density functions. In order to capture changes, we estimated density ratio function using two sets of samples. However, it should be noticed that the definition of *changes* heavily relies on applications where the statistical change detection may or may not be applied.

Particularly, in the distributional change detection, the type of change we may detect depends on the construction of samples. In the estimation algorithm, we do not limit our approach to long-term or short-term changes, however, such time-scope information is already encoded into samples, as part of the problem setting (see Figure 2.2). Recently, Yamanaka et al. (2013) has shown that changes in different time-scopes can be captured by varying problem settings (e.g. n, k).

In practice, looking for a proper choice of such problem setting is crucial, however, due to the research focus, we regarded them as known issues, and focused on investigating the statistical properties of each method.

By quoting the Vapnik's principle, we compared the direct and separated learning methods in this thesis. Experiment results showed that the direct method demonstrated better performance. However, it should not be understood as the direct method is always better in all applications.

The direct learning method does not provide information regarding to the generating source itself, so it may not be used if the model "before" or "after" the change is part of the learning target.

Moreover, since the separated generating probability is not modelled, the do-

main/expert knowledge on the separated patterns may not be used in our approach.

4.3 Future Works

Our research in this thesis has demonstrated that under the dynamic view of machine learning, the unsupervised change detection is a very promising area. However, limited by the scope of this thesis, we did not investigate many other important tasks. In this section, we will illustrate the future works.

4.3.1 Future Works for Distributional Change Detection

First, some enhancements of the proposed method in Chapter 2 need to be developed.

Through the experiment illustrated in Figure 2.6, 2.7 in Section 2.4.1, we can see that the performance of the proposed method is affected by the choice of hyper-parameters n and k. However, discovering optimal values for these parameters remains a challenge, which will be investigated in our future work.

RuLSIF was shown to possess a better convergence property than uLSIF (Yamada et al., 2013) in terms of density ratio estimation. However, how this theoretical advantage in density ratio estimation can be translated into practical performance improvement in change detection is still not clear, beyond the intuition that a better divergence estimator gives a better change score. We will address this issue more formally in the future work.

In addition, it is also interesting to discover the physical meanings of (Relative) Pearson divergence. As the *Relative Entropy*, Kullback-Leibler divergence plays an important role in Information Theory. However, similar interpretation for Pearson divergence is not yet known. Understanding such physical meaning of Pearson divergence would help us find a guideline of choosing appropriate statistical distance in change-point detection.

Second, to improve the performance of the proposed method on more challenging data, we may consider several advanced techniques.

Although the proposed RuLSIF-based change-point detection was shown to work well even for multi-dimensional time-series data, its accuracy may be fur-

ther improved by incorporating *dimensionality reduction*. Recently, several attempts were made to combine dimensionality reduction with direct density-ratio estimation (Sugiyama et al., 2010, 2011b; Yamada and Sugiyama, 2011). Our future work will apply these techniques to change-point detection and evaluate their practical usefulness. A more ambitious plan is to consider multi-modal data where more than one class of time-series are considered. Such high-dimensional mixed-source data appears in many applications, such as text-audio change-detection.

Compared with other approaches, methods based on density ratio estimation tend to be computationally more expensive because of the cross-validation procedure for model selection. However, thanks to the analytic solution, the RuLSIF- and uLSIF-based methods are computationally more efficient than the KLIEP-based method that requires an iterative optimization procedure (see Figure 9 in Kanamori et al. (2009) for the detailed time comparison between uLSIF and KLIEP). Our important future work is to further improve the computational efficiency of the RuLSIF-based method.

In Chapter 2, we focused on computing the change-point score that represents the plausibility of change points. Another possible formulation is hypothesis testing, which provides a useful threshold to determine whether a point is a change point. Methodologically, it is straightforward to extend the proposed method to produce the p-values, following the recent literatures (Sugiyama et al., 2011a; Kanamori et al., 2012a). However, computing the p-value is often time consuming, particularly in a non-parametric setup. Thus, overcoming the computational bottleneck is an important future work for making this approach more practical. Moreover, the model-based method may include certain prior knowledge and offer an appropriate threshold based on learnt models. Combining these two methods may lead to a highly efficient algorithm for determining the threshold.

In this research, our interest was developing a generalized learning method, however, the proposed method also has many potential applications in reality.

Recent reports pointed out that Twitter messages can be indicative of real-world events (Petrović et al., 2010; Sakaki et al., 2010). Following this line, we showed in Section 2.4.3 that our change-detection method can be used as a novel tool for analyzing Twitter messages. An important future challenge along this line includes automatic keyword selection for topics of interests.

4.3.2 Future Works for Structural Change Detection

In Chapter 3, we only considered MNs with pairwise factors. However, such a model may be misspecified when higher order interactions exist. For example, combination with the idea of *hierarchical log-linear models* presented in Schmidt and Murphy (2010) may lead to a promising solution to this problem.

If the purpose is only for detecting changes in parameters, our method can still be directly applied to models that contains higher-order interactions, and adopt group lasso regularizer to induce group sparsity on each factors. However, unlike pairwise models, such group sparsity in higher order log-linear model does not directly correspond to the structural information of changes.

Consider the structure of a single MN, variable set A is conditionally independent with variable set B if and only if $\boldsymbol{\theta}_{\mathrm{C}}$ on variable set C is zero, for all C that C contains at least one variable from A and at least one variable from B (Whittaker, 1990). Generally speaking, a non-zero factor induces a complete connected graphical structural among corresponding random variables, even the factors defined on a subset of random variables with lower orders are zero. This phenomena is demonstrated in Figure 4.1. To solve this problem, *hierarchical* log-linear model (Wasserman, 2010; Schmidt and Murphy, 2010) is introduced:

Theorem 4.1. *A log-linear model*

$$p = \frac{1}{Z} \exp\left(\sum_{C \subseteq X} \phi_C\right),$$

is hierarchical if $\phi_A = 0$ *and* $A \subseteq B$ *implies* $\phi_B = 0$*, where* ϕ_A *and* ϕ_B *are functions defined only on subsets of variable set* X*.*

Such hierarchy enforces that whenever a higher order factor does not equal to $\mathbf{0}$, all factors defined on lower order subsets of variables much be non-zero. The sparsity in such log-linear model directly reflects the network structure. Schmidt and Murphy (2010) showed that we can enforce such hierarchical sparsity pattern by using hierarchical regularization.

A problem caused by introducing higher order factors is the exponentially large number of factors. Via *active set selection*, an effective method can be developed for solving the maximal likelihood estimation under weak optimality (Schmidt and Murphy, 2010).

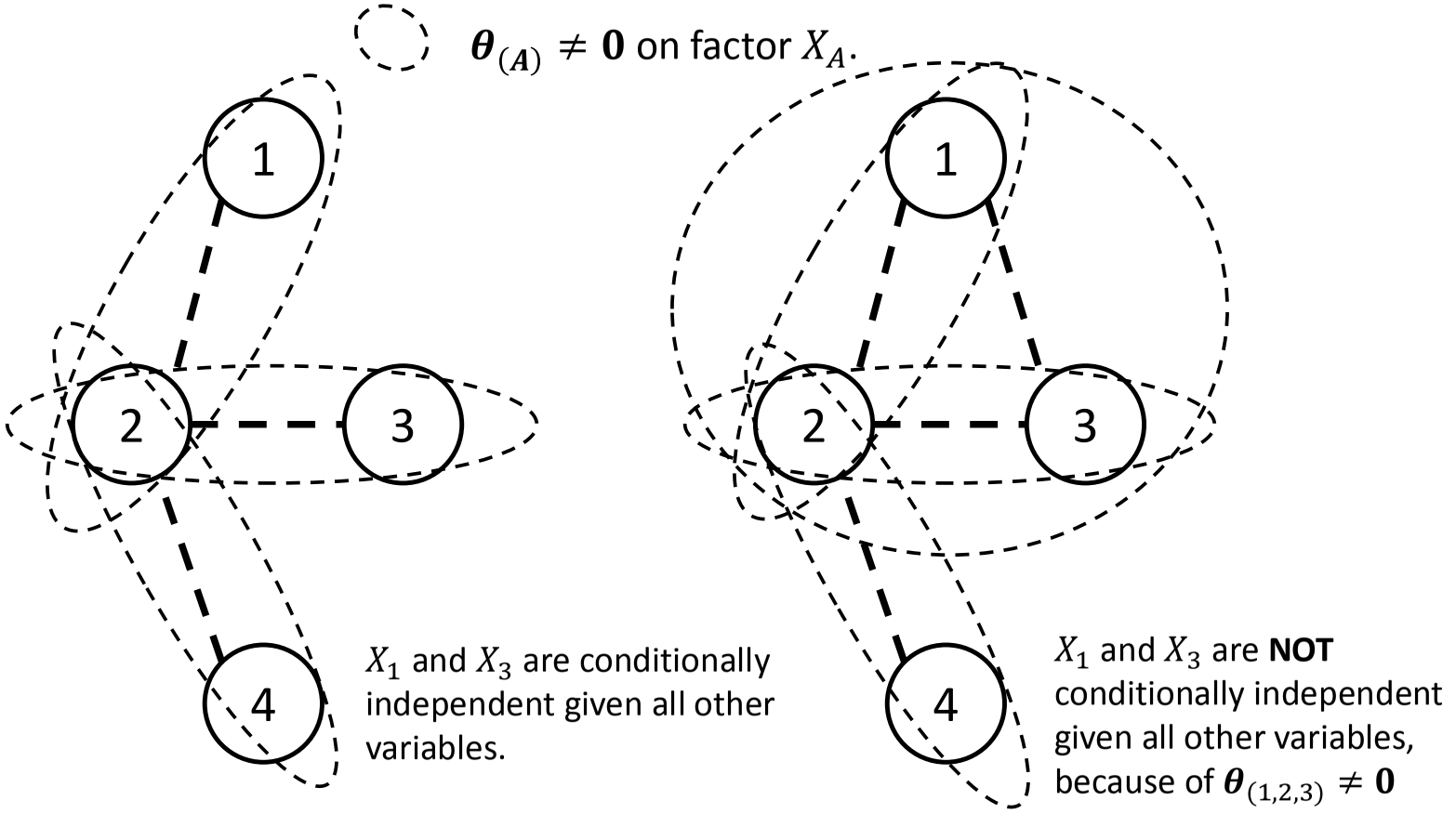


Figure 4.1: An illustrative example shows the sparsity in the lower order log-linear model does not necessarily reflect the structural of Markov network.

Beyond the modelling of higher order interactions, some theoretical issues also need to be analysed. For example, how to theoretically elucidate the advantage of the proposed method, beyond the Vapnik's principle of solving the target problem directly. Such theoretical results may also give insights on how to interpret Vapnik's principle in the change detection context. Moreover, the relation to *score matching* (Hyvärinen, 2005), which avoids computing the normalization term in density estimation, is also an interesting issue to be further investigated.

In the context of change detection, we are mainly interested in the situation where p and q are close to each other (if p and q are completely different, it is straightforward to detect changes). When p and q are similar, density ratio estimation for $p(\boldsymbol{x})/q(\boldsymbol{x})$ or $q(\boldsymbol{x})/p(\boldsymbol{x})$ perform similarly. However, given the asymmetry of density ratios, the solutions for $p(\boldsymbol{x})/q(\boldsymbol{x})$ or $q(\boldsymbol{x})/p(\boldsymbol{x})$ are generally different. The choice of the numerator and denominator in the ratio is left for future investigation.

Bibliography

R. P. Adams and D. J. C. MacKay. Bayesian online changepoint detection. Technical report, arXiv, 2007. arXiv:0710.3742v1 [stat.ML].

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28 (1):131–142, 1966.

S. Amari and H. Nagaoka. *Methods of Information Geometry*. Oxford University Press, Providence, RI, USA, 2000.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, March 2008.

M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, NJ, USA, 1961.

S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning*, pages 81–88, 2007.

S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning under covariate shift. *The Journal of Machine Learning Research*, 10:2137–2155, 2009.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, USA, 2006.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

B. Brodsky and B. Darkhovsky. *Nonparametric Methods in Change-Point Problems*. Kluwer Academic Publishers, Dordrecht, the Netherlands, 1993.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2: 229–318, 1967.

M. Csörgö and L. Horváth. 20 nonparametric methods for changepoint problems. In P. R. Krishnaiah and C. R. Rao, editors, *Handbook of Statistics*, volume 7, pages 403–425. Elsevier, Amsterdam, the Netherlands, 1988.

P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2013.

F. Desobry, M. Davy, and C. Doncarli. An online kernel change detection algorithm. *IEEE Transactions on Signal Processing*, 53(8):2961–2974, 2005.

M. C. Du Plessis and M. Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Networks*, 50(0):110 – 119, 2014.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York, NY, USA, 1993.

S. Eguchi and J. Copas. Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma. *Journal of Multivariate Analysis*, 97(9):2034–2040, 2006.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

R. Garnett, M. A. Osborne, and S. J. Roberts. Sequential Bayesian prediction in the presence of changepoints. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 345–352, 2009.

A. Gelman. Method of moments using Monte Carlo simulation. *Journal of Computational and Graphical Statistics*, 4(1):36–54, 1995.

A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. In J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence, editors, *Dataset Shift in Machine Learning*, chapter 8, pages 131–160. MIT Press, Cambridge, MA, USA, 2009.

V. Guralnik and J. Srivastava. Event detection from time series data. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 33–42, 1999.

F. Gustafsson. The marginalized likelihood ratio test for detecting abrupt changes. *IEEE Transactions on Automatic Control*, 41(1):66–78, 1996.

F. Gustafsson. *Adaptive Filtering and Change Detection*. Wiley, Chichester, UK, 2000.

M. U. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361, 2012.

Z. Harchaoui, F. Bach, and E Moulines. Kernel change-point analysis. In *Advances in Neural Information Processing Systems 21*, pages 609–616, 2009.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, USA, 2001.

R. E. Henkel. *Tests of Significance*. SAGE Publication, Beverly Hills, CA, USA, 1976.

S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori. Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*, 26(2):309–336, 2011.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

T. Ide and K. Tsuda. Change-point detection using Krylov subspace learning. In *Proceedings of the SIAM International Conference on Data Mining*, pages 515–520, 2007.

N. Itoh and J. Kurths. Change-point detection of climate time series by non-parametric method. In *Proceedings of the World Congress on Engineering and Computer Science 2010*, volume 1, 2010.

A. Jacobs. The pathologies of big data. *Communications of ACM*, 52(8):36–44, 2009.

T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

T. Kanamori, T. Suzuki, and M. Sugiyama. Theoretical analysis of density ratio estimation. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E93-A(4):787–798, 2010.

T. Kanamori, T. Suzuki, and M. Sugiyama. f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Transactions on Information Theory*, 58(2):708–720, 2012a.

T. Kanamori, T. Suzuki, and M. Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2012b.

T. Kanamori, T. Suzuki, and M. Sugiyama. Computational complexity of kernel-based density-ratio estimation: A condition number analysis. *Machine Learning*, 90:431–460, 2013.

Y. Kawahara and M. Sugiyama. Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*, 5(2):114–127, 2012.

Y. Kawahara, T. Yairi, and K. Machida. Change-point detection in time-series data based on subspace identification. In *Proceedings of the 7th IEEE International Conference on Data Mining*, pages 559–564, 2007.

E. Keogh, J. Lin, and A. Fu. Hot sax: efficiently finding the most unusual time series subsequence. In *The Fifth IEEE International Conference on Data Mining*, pages 8 pp.–, 2005.

A. Keziou. Dual representation of ϕ-divergences and applications. *Comptes Rendus Mathematique*, 336(10):857–862, 2003.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

S.-I. Lee, V. Ganapathi, and D. Koller. Efficient structure learning of Markov networks using l_1-regularization. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 817–824, Cambridge, MA, 2007. MIT Press.

H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10:2295–2328, 2009.

H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. The nonparanormal skeptic. In *Proceedings of the 29th International Conference on Machine Learning (ICML2012)*, 2012.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

V. Moskvina and A. Zhigljavsky. Application of the singular-spectrum analysis to change-point detection in time series. *Journal of Sequential Analysis*, 2003a. In submission.

V. Moskvina and A. Zhigljavsky. Change-point detection algorithm based on the singular-spectrum analysis. *Communications in Statistics: Simulation and Computation*, 32:319–352, 2003b.

K. P. Murphy. *Machine learning: a probabilistic perspective*. Cambridge, MA, 2012.

R. M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–741, 2003.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Nonparametric estimation of the likelihood ratio and divergence functionals. In *Proceedings of IEEE International Symposium on Information Theory*, pages 2016–2020, Nice, France, 2007.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

U. Paquet. Empirical Bayesian change point detection. *Graphical Models*, 1995: 1–20, 2007.

K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50: 157–175, 1900.

S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to Twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189, 2010.

R. Raina, A. Y. Ng, and D. Koller. Constructing informative priors using transfer learning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 713–720, 2006.

P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38 (3):1287–1319, 2010.

J. Reeves, J. Chen, X. L. Wang, R. Lund, and Q. Lu. A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46(6):900–915, 2007.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, Secaucus, NJ, USA, 2005.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, USA, 1970.

M. Saerens, P. Latinne, and C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1): 21–41, 2002.

T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860, 2010.

M. W. Schmidt and K. P. Murphy. Convex structure learning in log-linear models: Beyond pairwise potentials. *Journal of Machine Learning Research - Proceedings Track*, 9:709–716, 2010.

B. Schölkopf and A. J. Smola. *Learning with Kernels Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004. ISBN 0521813972.

M. Sugiyama, M. Krauledat, and K. R. Müller. Covariate shift adaptation by importance weighted cross validation. *The Journal of Machine Learning Research*, 8:985–1005, 2007.

M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

M. Sugiyama, M. Kawanabe, and P. L. Chui. Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks*, 23(1):44–59, 2010.

M. Sugiyama, T. Suzuki, Y. Itoh, T. Kanamori, and M. Kimura. Least-squares two-sample test. *Neural Networks*, 24(7):735–751, 2011a.

M. Sugiyama, M. Yamada, P. von Bünau, T. Suzuki, T. Kanamori, and M. Kawanabe. Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. *Neural Networks*, 24(2):183–198, 2011b.

M. Sugiyama, T. Suzuki, and T. Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, Cambridge, UK, 2012a.

M. Sugiyama, T. Suzuki, and T. Kanamori. Density ratio matching under the Bregman divergence: A unified framework of density ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64:1009–1044, 2012b.

J. Takeuchi and K. Yamanishi. A unifying framework for detecting outliers and change points from non-stationary time series data. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):482–492, 2006.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Y. Tsuboi, H. Kashima, S. Hido, S. Bickel, and M. Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155, 2009.

T. Van den Bulcke, K. Van Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. De Moor, and K. Marchal. SynTReN: A generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 7(1):43, 2006.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, USA, 1998.

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, 2008.

Q. Wang, S. R. Kulkarni, and S. Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.

Y. Wang, C. Wu, Z. Ji, B. Wang, and Y. Liang. Non-parametric change-point method for differential gene expression detection. *PLoS ONE*, 6(5):e20060, 2011.

L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated, 2010.

J. Whittaker. Graphical models in applied multivariate statistics. 1990.

M. Yamada and M. Sugiyama. Direct density-ratio estimation with dimensionality reduction via hetero-distributional subspace analysis. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 549–554, Aug. 7–11 2011.

M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25(5):1324–1370, 2013.

M. Yamanaka, M. Matsugu, and M. Sugiyama. Salient object detection based on direct density-ratio estimation. *IPSJ Online Transactions*, 6(0):96–103, 2013.

K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 320–324, 2000.

B. Zhang and Y.J. Wang. Learning structural changes of Gaussian graphical models in controlled experiments. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI2010)*, pages 701–708, 2010.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, 2005.