Robust Learning from Multiple Information Sources

by

Tianpei Xie

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical and Computing Engineering)
in the University of Michigan
2017

Doctoral Committee:

Professor Alfred O. Hero, Chair
Professor Laura Balzano
Professor Danai Koutra
Professor Nasser M. Nasrabadi, University of West Virginia

tianpei@umich.edu

Orcid iD: 0000-0002-8437-6069

2017

Dedication

To my loving parents, Zhongwei Xie and Xiaoming Zhang. For your patience and love bring me strength and happiness.

Acknowledgments

This thesis marked a milestone for my life-long study. Since entering the master program of Electrical Engineering in the University of Michigan, Ann Arbor (Umich), I have spent more than 6 years in Ann Arbor, with 2 years as a master graduate student and nearly 5 year as a Phd graduate student. This is a unique experience for my life. Surrounded by friendly colleagues, professors and staffs in Umich, I feel myself a member of family in this place. This feelings should not fade in future. Honestly to say, I love this place more than anywhere else. I love its quiescence with bird chirping in the wood. I love its politeness and patience with people slowing walking along the street. I love its openness with people of different races, nationalities work so closely as they were life-long friends. I feel that I owe a debt of gratitude to my friends, my advisors and this university.

Many people have made this thesis possible. I wish to thank the department of Electrical and Computing Engineering in University of Michigan, and the U.S. Army Research Laboratory (ARL) for funding my Phd. I also want to express my sincere gratitude to my committee member: Professor Alfred O Hero, my advisor in Umich; Professor Nasser M. Nasrabadi, my mentor in ARL; Professor Laura Balzano and Professor Danai Koutra in Umich. Without your suggestions and collaborations, I could not have finished this work. Some of chapters are based on discussions with my colleagues in Hero's group. Among them, we wish to thank Sijia Liu, Joel LeBlanc, Brendan Oselio, Yaya Zhai, Kristjan Greenewald, Pin-Yu Chen, Kevin Moon, Yu-Hui Chen, Hamed Firouzi, Zhaoshi Meng, Mark Hsiao, Greg Newstadt, Kevin Xu, Kumar Sricharan, Hye Won Chung and Dennis Wei. Thank you for all your help during my study here. I also feel obliged to mention my roommates Shengtao Wang, Hao Sun and Jiahua Gu. We had enjoyed great times together during these six years. I wish you all lead a brilliant life in future.

I owe my greatest appreciation to my supervisor, Professor Alfred O. Hero, who is a brilliant scholar with deep knowledge in every field that I have ever encountered. Talking with him brings me broad insight in the field. Through his words, my mind traverses through different domains and many related concepts and applications merge together to form a new intuition. I also want to express my gratitude for his great patience to every student. I am a slow thinker. With that, my study cannot continue without Prof. Hero's encouragement and patience. His skill in communications also makes a great example for me. From him, I learned 1) listen more, judge less; 2) write a paper following a single thread of logic; 3) min-max your contribution; 4) go for what you understand, not for what all other ones do. I still remember all the efforts we have made together to write a clear and concise paper. All of these are rare treasure which I will bring along my life a researcher and a developer. Finally, to all colleagues, staffs and professors in Umich, I owe you my deepest gratitude and will end by saying: Go Blue !

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ADMM Alternating direction methods of multipliers

BP-kNN Bipartite k-Nearest-Neighbor

CCP convex-concave procedure

COM-MED Consensus-based Multi-view Maximum Entropy Discrimination

DC difference of convex

DiLat-GGM Decayed-influence Latent variable Gaussian Graphical Model

EM Expectation Maximization

GEM Geometric Entropy Minimization

GGM Gaussian graphical model

GPLVM Gaussian process latent variable model

GSP Graph Signal Processing

GFT Graph Fourier Transforms

GEM-MED Geometric-Entropy-Minimization Maximum-Entropy-Discrimination

KL-divergence Kullback-Leibler divergence

LV-GGM Latent variable Gaussian graphical model

MED Maximum Entropy Discrimination

MM majorization minimization

PSGD projected stochastic gradient descent

ROD Robust-Outlier-Detection

ABSTRACT

In the big data era, the ability to handle high-volume, high-velocity and high-variety information assets has become a basic requirement for data analysts. Traditional learning models, which focus on medium size, single source data, often fail to achieve reliable performance if data come from multiple heterogeneous sources (views). As a result, *robust multi-view data processing* methods that are insensitive to corruptions and anomalies in the data set are needed.

This thesis develops robust learning methods for three problems that arise from real-world applications: robust training on a noisy training set, multi-view learning in the presence of between-view inconsistency and network topology inference using partially observed data. The central theme behind all these methods is the use of *information-theoretic measures*, including entropies and information divergences, as parsimonious representations of uncertainties in the data, as robust optimization surrogates that allows for efficient learning, and as flexible and reliable discrepancy measures for data fusion.

More specifically, the thesis makes the following contributions:

1. We propose a maximum entropy-based discriminative learning model that incorporates the minimal entropy (ME) set anomaly detection technique. The resulting probabilistic model can perform both nonparametric classification and anomaly detection simultaneously. An efficient algorithm is then introduced to estimate the posterior distribution of the model parameters while selecting anomalies in the training data.

2. We consider a multi-view classification problem on a statistical manifold where class labels are provided by probabilistic density functions (p.d.f.) and may not be con-

sistent among different views due to the existence of noise corruption. A stochastic consensus-based multi-view learning model is proposed to fuse predictive information for multiple views together. By exploring the non-Euclidean structure of the statistical manifold, a joint consensus view is constructed that is robust to single-view noise corruption and between-view inconsistency.

3. We present a method for estimating the parameters (partial correlations) of a Gaussian graphical model that learns a sparse sub-network topology from partially observed relational data. This model is applicable to the situation where the partial correlations between pairs of variables on a measured sub-network (internal data) are to be estimated when only summary information about the partial correlations between variables outside of the sub-network (external data) are available. The proposed model is able to incorporate the dependence structure between latent variables from external sources and perform latent feature selection efficiently. From a multi-view learning perspective, it can be seen as a two-view learning system given *asymmetric* information flow from both the *internal view* and the *external view*.

CHAPTER 1

Introduction

In the past decade, the emerging field of data science has attracted significant attention from researchers and developers in the field of statistics, machine learning, information theory, data management and communications. Data in these fields is growing at an unprecedented rate in terms of *volume*, *velocity* and *variety*. Volume means the size of the data set. Velocity corresponds to the speed of data communication and processing. Variety indicates the range of data types and sources.. These three aspects are dimensions that determine the applicability of data processing techniques to increasingly demanding applications. As a result, *Big Data* [1] has gained great popularity and become one of the current and future research frontiers [McAfee et al., 2012, Mayer-Schönberger and Cukier, 2013, Chen and Zhang, 2014]. In this thesis, we primarily focus on the *variety* aspect of the Big data analysis under the condition that the data set may contain corrupted or irregular samples, known as *anomalies*. In particular, we exploit several reliable models and efficient implementations to deal with large-scale data from multiple possibly unreliable sources.

As large-scale data acquisition becomes common and diversified, data quality can become degraded, especially when data collection is conducted without sampling design, when devices are unreliable, or when users are sloppy data collectors. On the other hand, as the size of datasets increases, the existing off-the-shelf models and algorithms in machine learning may not function efficiently and robustly [Szalay and Gray, 2006, Lynch, 2008]. Consequently, robust large-scale data processing from multiple sources has become an important field, a field called *robust multi-view learning*. Here the term *robustness* means that the learning algorithm should be insensitive to noise corruption, sensor failures, or anomalies in the data set. Our main contributions in this thesis to robust multi-view learning are described in the next subsection.

[1] http://blogs.gartner.com/it-glossary/big-data/

1.1 Thesis Outline and Contributions

The thesis addresses the following topics in robust multi-view learning:

1. Robust training on noisy training data sets. In Chapter 3, a robust maximum entropy discrimination method, referred as GEM-MED, is proposed that minimizes the generalization error of the classifier with respect to a *nominal data distribution*[2]. The proposed method exploits the versatility of the kernel method in combination with the power of *minimal-entropy-sets*, which allows one to perform anomaly detection in high dimensions. Instead of focusing on robustifying classification loss functions, GEM-MED combines anomaly detection and classification explicitly as joint constraints in the maximum entropy discrimination framework. This allows GEM-MED to suppress the training outliers more effectively.

2. Multi-view learning on a statistical manifold of probability distributions in the presence of between-view inconsistency. In Chapter 4, we consider a multi-view classification problem where the labels in each view come in the form of probability distributions or histograms, which encodes label uncertainties. Different from the conventional feature fusion and decision fusion approaches, an alternative *model fusion* approach, called COM-MED, is presented that learns a *consensus view* to fuse predictive information from different views. Using information-theoretic divergences as a *stochastic consensus measure*, COM-MED takes into account the intrinsic non-Euclidean geometry of the statistical manifold. Our proposed method is insensitive to both noise corruption in single views and between-view inconsistency.

3. Sub-network topology inference from partially observed relational data. In Chapter 5, we introduce a method to infer the topology of a sub-network, given a partially accessible dataset. We assume that the set of measurements are taken at nodes of a graph whose edges specify pairwise node dependencies. The joint distribution of the measurements is assumed to be Gaussian distributed with a sparse inverse covariance matrix whose zero entries are specified by the topology of the graph. In the sub-net topology inference problem one only directly measures a subset of nodes while only noisy information on the inverse covariance matrix of the remaining nodes is available. The objective is to estimate the (non-marginal) sub-graph associated with the set of directly measured nodes. We propose a solution to this problem that generalizes the existing Latent variable Gaussian graphical model (LV-GGM), which explicitly

[2] A set of data is nominal if it contains no anomalies.

takes into account the local effect of the latent variables. The proposed *Decayed-influence Latent variable Gaussian Graphical Model (DiLat-GGM)* is well-suited for applications such as competitive pricing models where two companies operate in a market where each can only directly measure the behaviors of their own customers.

1.2 Robust Multi-view Learning

Multi-view learning is concerned with the problem of information fusion and learning from multiple feature domains, or *views*. Canonical Correlation analysis (CCA) [Hotelling, 1936, Hardoon et al., 2004] and co-training [Blum and Mitchell, 1998] are two representative algorithms in multi-view learning. Canonical correlation analysis finds maximal-correlated linear representations from two views by learning two individual subspaces jointly. The co-training method seeks to learn multiple classifiers by minimizing their mutual disagreement on a common target. Both of these algorithms function by combining information from multiple feature sets while minimizing the information discrepancy between different views. Following the perspective of co-training, the co-EM was proposed in [Nigam and Ghani, 2000] to handle latent variables in statistical models. In semi-supervised learning, the framework of co-regularization was proposed in [Farquhar et al., 2005, Sindhwani et al., 2005, Sindhwani and Rosenberg, 2008, Sun and Jin, 2011] as generalizations of the co-training algorithm. This framework is referred as *semi-supervised learning with disagreement* in [Zhou and Li, 2010]. Similarly, in [Ganchev et al., 2008], information divergence, such as the *Bhattacharya distance measure* [Bhattachayya, 1943], was proposed as a surrogate disagreement measure for different classifiers. See Figure 1.1 for a comparison of learning procedures of CCA (the left column), co-training (the middle column) and our proposed consensus-based learning framework (the right column, see Chapter 4).

One of the critical drawbacks of these multi-view learning algorithms is their sensitivity to noisy measurements and anomalies in the data set [Schölkopf et al., 1999, Breunig et al., 2000, Zhao and Saligrama, 2009]. To achieve robustness, the co-training algorithm can be extended to incorporate the uncertainties in data or labels [Ganchev et al., 2008, Sun and Jin, 2011]. In [Muslea et al., 2002], a subsampling and active learning strategy is introduced to reduce the influence of corrupted samples. The Bayesian co-training proposed in [Yu et al., 2011] reformulates standard co-training under a Bayesian learning framework using a Gaussian process prior [Rasmussen and Williams, 2006] to provide a confidence level for each decision. In spite of these advances, a unified principle underling the robust design of multi-view learning algorithm is still needed.

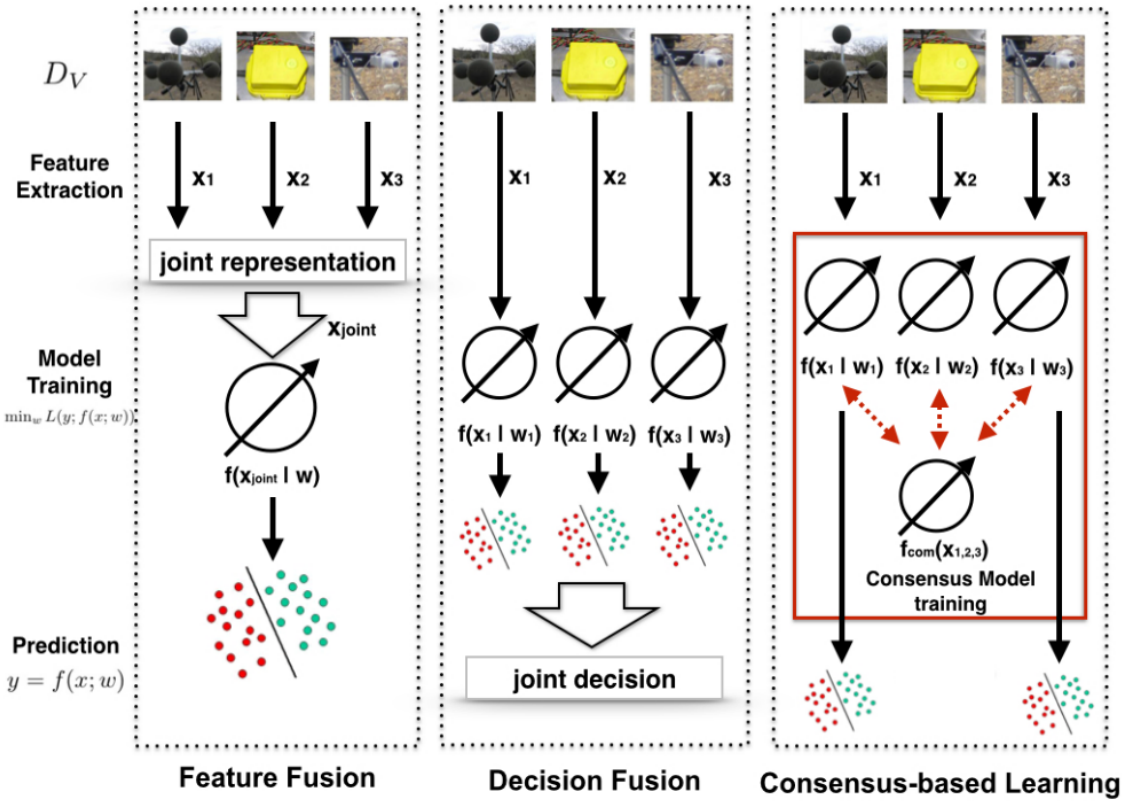


Figure 1.1: A classification of multi-view learning methods according to the information fusion strategy. At the top of each column are three sensors (acoustic, seismic and optical) that provide different views of a common scene. The left column corresponds to the feature fusion or early fusion approach, where the fusion stage takes place *before* the learning stage. The middle column corresponds to the decision fusion or late fusion approach, where the final decisions take place *after* each individual learner has made its own decision. The right column corresponds to the proposed consensus-based method in Chapter 4. Note that the proposed method iteratively retrains each individual learner based on their mutual disagreement.

On the other hand, the field of robust learning provides methods that systematically address the stability and robustness of the learning algorithm, e.g. [Kearns, 1998, Bousquet and Elisseeff, 2002, Song et al., 2002, Bartlett and Mendelson, 2003, Xu et al., 2006, Tyler, 2008, Wang et al., 2008, Masnadi-Shirazi and Vasconcelos, 2009, Yang et al., 2010]. Among these works, the *entropy-based learning* methods have recently drawn significant attention [Eguchi and Kato, 2010, Basseville, 2013]. In [Grünwald and Dawid, 2004, Cover and Thomas, 2012], it is shown that the distribution that maximizes the entropy over a family of distributions also minimizes the worst-case expected log-loss. Other researchers [Grünwald and Dawid, 2004, Nock and Nielsen, 2009] have shown that the Bregman divergence [Bregman, 1967] can be used as a discrepancy function to reduce regret in robust

decision-making. In this thesis, we focus on the entropy-based robust multi-view learning framework in which entropy and information divergences are used to define the robust surrogate function and the *nominal region* [3] (in Chapter 3) and the multi-view discrepancy (in Chapter 4). Other related works are summarized in Chapter 2.

1.3 Entropy-based Learning and Anomaly Detection

As basic measures of uncertainty and information in physics and information theory [Cover and Thomas, 2012], entropies and information divergences are known to be invariant under data transformations such as transition, rotation and geometric distortions [Skilling and Bryan, 1984, Maes et al., 1997, Swaminathan et al., 2006]. Such invariances makes them natural candidates for robust learning. In this section, we discuss both parametric inference via entropy maximization, and nonparametric anomaly detection via entropy estimation. The former treats the information divergences as surrogate loss functions for learning problems, as in [Jaakkola et al., 1999, Nock and Nielsen, 2009, Basseville, 2013] and the latter defines the nominal region based on the concept of *minimal entropy (ME) set* [Hero, 2006, Sricharan and Hero, 2011]. We will discuss the maximum entropy learning models in detail in Chapter 2.

1.3.1 Parametric Inference via Maximum Entropy and Statistical Manifold

The maximum entropy principle states that the best representative of a class of distributions that describe the current state of observation given prior data is the one with largest entropy [Cover and Thomas, 2012]. As a generative learning framework, this principle embodies the Bayesian integration of prior information with observed constraints. Since first introduced by E.T. Jaynes in [Jaynes, 1957a,b], maximum entropy learning models have become popular in natural language processing [Berger et al., 1996, Manning et al., 1999, Ratnaparkhi et al., 1996, Charniak, 2000, Malouf, 2002, Jurafsky and Martin, 2014], object recognition [Jeon and Manmatha, 2004, Lazebnik et al., 2005], image restoration [Minerbo, 1979, Burch et al., 1983, Skilling and Bryan, 1984, Gull and Skilling, 1984] and structured learning in computer vision [Nowozin et al., 2011]. These models are well-studied in branches of machine learning such as probabilistic graphical models [Lafferty et al., 2001, Wainwright et al., 2008], neural networks [Ackley et al., 1985, Hinton and Salakhutdinov, 2006], boosting [Murata et al., 2004, Rätsch et al., 2007, Schapire and Freund, 2012],

[3] The nominal region is the set of all possible regular data in the data set.

nonparametric Bayesian learning [Jaakkola et al., 1999, Zhu et al., 2014], multi-task and multi-view learning [Jebara, 2011], anomaly detection [Jaakkola et al., 1999, Xie et al., 2017] and model selection [Hastie et al., 2009].

In [Jaakkola et al., 1999], T. Jaakkola proposed a *discriminative* learning framework based on the maximum entropy principle, namely *Maximum Entropy Discrimination (MED)*, which allows for training of both parameters and the structure of the joint probability model. Relying on the choice of discriminative functions and margin prior, Maximum Entropy Discrimination (MED) incorporates large-margin classification into the Bayesian learning framework and it subsumes the support vector machine (SVM). MED can also be used to handle the *parametric anomaly detection problem* when the nominal region is defined by the level sets of the underlying parametric data distribution. Furthermore, due to its flexible formulation, MED can be extended to nonparametric Bayesian inference [Zhu et al., 2011, Chatzis, 2013, Zhu et al., 2014], which robustly captures local nonlinearity of complex data. In [Jebara, 2011], the multi-task MED was proposed to combine multiple datasets in learning. MED can also be used as a parametric anomaly detection method, which is introduced in Chapter 3. MED serves as a prototype in our development in Chapter 3 and Chapter 4. In Chapter 2, we give a more detailed derivation of the MED problem.

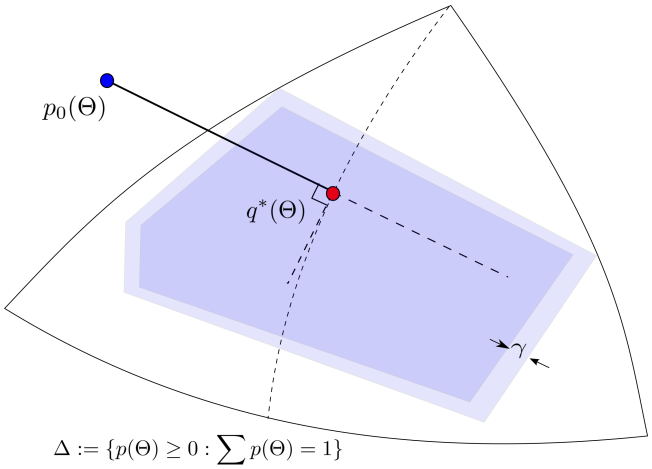


Figure 1.2: Maximum entropy learning relies on an information projection of the prior distribution $p_0(\Theta)$ onto the feasible region (shaded region). The margin variable γ allows for adjustment of the feasible region. Note that the projection $q^*(\Theta)$ is unique due to the Pythagorean property of Bregman divergences [Amari and Nagaoka, 2007]. The information divergence also induces an non-Euclidean structure of the feasible region, which forms a sub-manifold of the set of all probability distributions.

It is worth mentioning that in information geometry [Amari and Nagaoka, 2007], maximum entropy learning can be interpreted as information projection [4] over a feasible region

[4] In [Amari and Nagaoka, 2007], it is called the *e-projection*.

defined by a set of linear constraints, as shown in Figure 1.2. This justifies that maximum entropy learning will yield a unique efficient solution that lies in the exponential family [Kupperman, 1958, Wainwright et al., 2008]. Furthermore, the divergence function $\mathbb{D}(\cdot \parallel \cdot)$ also induces a non-Euclidean geometry on the feasible region, which forms a *finite-dimensional statistical sub-manifold* [5] [Amari and Nagaoka, 2007]. These geometric properties help to build intuition about the minimum entropy discrimination approaches.

1.3.2 Nonparametric Entropy Estimation and Anomaly Detection

Anomaly detection [Chandola et al., 2009] is another important application that addresses identification of anomalies in corrupted data. Here information-theoretical measures such as entropy and information divergences can also be used to evaluate the anomalies in a data set [Lee and Xiang, 2001, Noble and Cook, 2003, Chandola et al., 2009]. The main advantage of entropy-based anomaly detection methods is that they do not make any assumptions about the underlying statistical distribution for the data except that the nominal data are i.i.d. Furthermore, entropy and information divergences can be estimated efficiently using nonparametric approaches, e.g., [Beirlant et al., 1997, Hero and Michel, 1999, Hero et al., 2002, Sricharan et al., 2012, Sricharan and Hero, 2012, Moon and Hero, 2014a]. Hero [Hero, 2006] proposed the *Geometric Entropy Minimization (GEM)* approach for non-parametric anomaly detection based on the the concept of a *minimal-entropy set*. In [Sricharan and Hero, 2011], the computational complexity of GEM is improved using the *bipartite k-Nearest-Neigbor-Graph (BP-kNNG)*. These will be used to construct our joint anomaly detection and classification method (GEM-MED) for learning to classify in the presence of possible sensor failiures.

Note that entropy-based anomaly detection and robust supervised learning adopt two different perspectives in handling anomalous data. The former evaluates the underlying distribution of covariate data and the latter investigates the relationship between the covariate and the response. In Chapter 3, a new model that combines both of these two perspectives is proposed. We will demonstrate better performance than the state-of-the-art algorithm in robust supervised learning.

[5] A *finite-dimensional statistical sub-manifold* is the space of probability distributions that are parameterized by some smooth, continuously-varying parameter Θ.

1.4 Multi-view Interpretation of Graph Signal Processing

In many applications, data is provided as records in a relational database that are sampled in an irregular manner. For instance, in a social network such as Facebook, each record consists of the account information for each user and his/her friendship connections. Facebook users annotate their profiles in different ways and with different levels of care, leading to noisy relational data. This dataset can represented by a graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ with the node set $\mathcal{V}$ being the set of personal records for each individual and the edge set $\mathcal{E}$ being the set of all relationships in the dataset. Similarly, in a sensor network, each node represents the measurements taken from one sensor and the link describes the conditional dependency relationship between two sensor variables given data from all other sensors. See Figure 1.3.

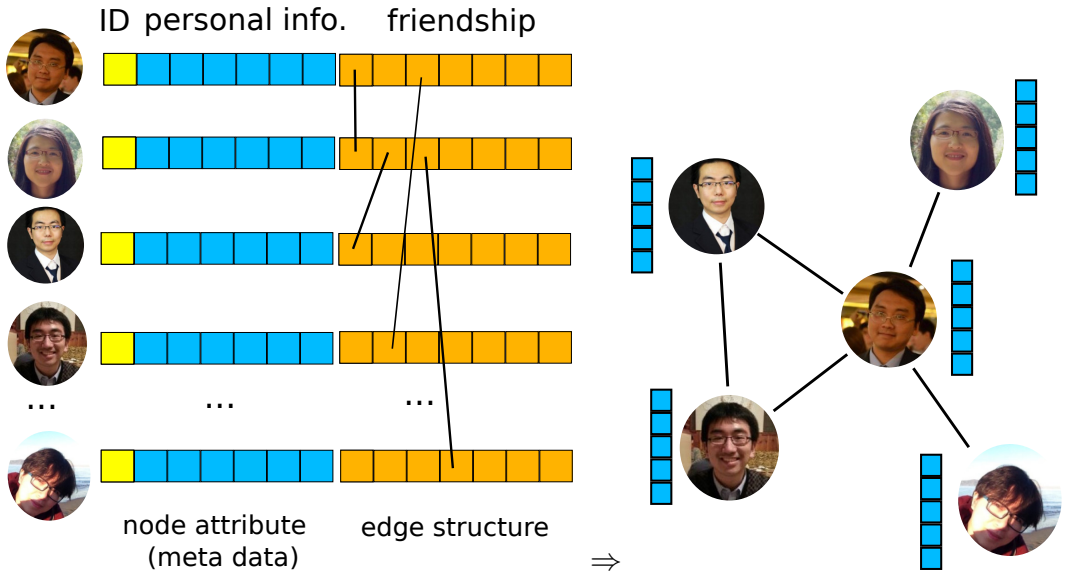


Figure 1.3: The facebook social media can be described as a network with node (personal information) and link (the friendship connection).

A graph provides a generic *two-view data representation*: the node view (vertex domain) represents the information content of the node attributes and the link view (edge domain) represents the structure of connectivity between different nodes. In such terms, *learning on graphs* or *Graph Signal Processing (GSP)* [Gori et al., 2005, Ando and Zhang, 2007, Shuman et al., 2013, Sandryhaila and Moura, 2013, 2014a,b, Zhang et al., 2015] can be seen as a generalization of multi-view learning when the samples interact through their connections in a graph. Figure 1.4 illustrates differences between multiview learning and learning on graphs in terms of the graph topology. Note that the probabilistic graph signal processing models involve both centralized model in Figure 1.4 (left) or a decentralized model in Figure 1.4 (right). The topology of network depends on the data of interest. In

this section, we provide an overview of the graph signal processing.

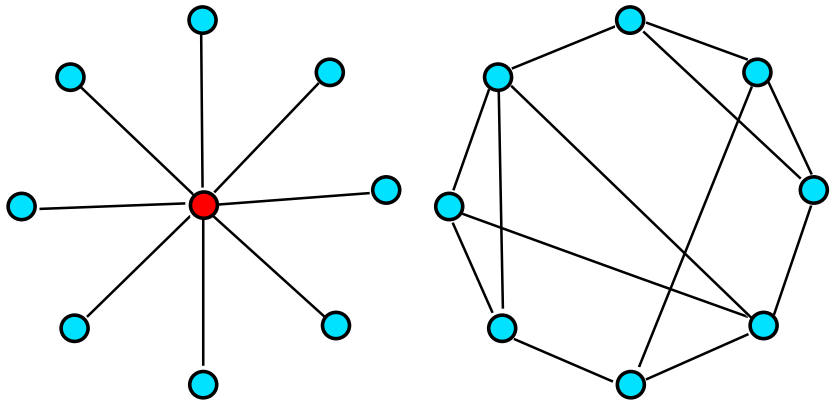


Figure 1.4: The graph structure for multi-view learning (left) and the graph signal processing (right). Note that for multi-view learning, all nodes (views) are connected to the central node (consensus view), while for the graph signal processing, the structure could be more general. Specifically, it can be a centralized network (left) or a decentralized network (right).

1.4.1 Graph Signal, Graph Laplacian and Graph Fourier Transform

The primary subject of interest in GSP is the *graph signal*. Defined as a real-valued function over the vertex domain of a given graph, a graph signal summarizes the relational information of data in vertex domain. The GSP is a field that is concerned with the approximation, representation and transformation of graph signals, especially when the data are corrupted by noises [Shuman et al., 2013, Sandryhaila and Moura, 2013, 2014a,b]. To analyze graph signal, the *graph Laplacian matrix* is introduced. The graph Laplacian matrix plays an important role in spectral graph theory [Chung, 1997, Agaskar and Lu, 2013]. It also proves useful in machine learning, such as spectral clustering [Ng et al., 2002, Von Luxburg, 2007], manifold learning [Belkin and Niyogi, 2003, Coifman and Lafon, 2006] and manifold regularization [Belkin et al., 2006, Belkin and Niyogi, 2008]. From spectral graph theory, both eigenvectors and eigenvalues of the graph Laplacian have interpretations: the eigenvalues are associated with the connectivity, invariance and various geometrical properties regarding the network topology. The innovation of graph signal processing was to identify the eigenvectors of the graph Laplacian as an orthonormal basis of the function space of the graph signals. The role of eigenvectors resembles the role of Fourier basis in digital signal processing (DSP). This interpretation led to the definition of Graph Fourier Transforms (GFT) as the fundamental building block in spectral analysis of graph signal. Similar to Discrete Fourier transform, the GFT defines an orthogonal transformation over the space of graph signals via the eigenspace of the underling graph Laplacian matrix. With the GFT,

it is thus natural to extend the traditional signal processing techniques into the graph domain, creating the field of graph filter design [Chen et al., 2015, Wang et al., 2015], graph signal interpolation [Zhu and Rabbat, 2012, Narang et al., 2013b, Anis et al., 2015, Wang et al., 2015, Zhang et al., 2015, Tsitsvero et al., 2016] and sampling theorems for graph signal [Agaskar and Lu, 2013, Narang et al., 2013b, Anis et al., 2014, Chen et al., 2015, Wang et al., 2015, Tsitsvero et al., 2016]. For instance, in [Narang et al., 2013a, Anis et al., 2014, Chen et al., 2015], a set of sampling theorems for graph signal were developed to reconstruct band-limited graph signals based on the graph spectral analysis [Chung, 1997] and harmonic analysis [Kim et al., 2016]. In [Narang et al., 2013b], Narang et al. proposed localized graph filtering based methods for interpolating signals defined on arbitrary graphs.

All of these methods assume that a graph signal is *smooth* on given graph, i.e., that the GFT of the graph signal concentrates on the low frequencies. Under a graph signal smoothness assumption, an edge between two nodes indicates the presence of correlation between the signals at these nodes. The dichotomy between the nodes that generate signals and the edges that correlate the signals reflects synergy between node and link view: samples collected in a small neighborhood in edge domain tend to embody similar information content in vertex domain. However, such assumption may not hold in practice, especially when some of data are missing, the remaining data may not be smooth over the sub-network. In Chapter 5, we seeks a *relaxation* of the existing *universal smoothness* condition that is imposed upon *every* vertex and its neighbors. The proposed model is generative and it allows data in a subset of the network to be *non-smooth* with respect to the underlying topology.

1.4.2 Statistical Graph Signal Processing

Conventional GSP only cares about the deterministic graph signals [Shuman et al., 2013, Sandryhaila and Moura, 2013, 2014a,b]. One of the major drawbacks of these methods is that they lack of ability to represent the uncertainty inherited in the observations, making them sensitive to the noises and anomalies in the data set. This motivates the introduction of *statistical graph signal processing*, which incorporates the GSP into probabilistic graphical models [Koller and Friedman, 2009], inverse covariance estimation [Friedman et al., 2008, Rothman et al., 2008, Yuan, 2010, Wiesel et al., 2010, Chen et al., 2011, Hsieh et al., 2011, Wiesel and Hero, 2012, Danaher et al., 2014] and Bayesian inference. In [Marques et al., 2016], the author extended the classical definition of stationary random process to random graph signals. They also proposed a number of nonparametric methods to estimate the power spectral density of random graph signal. [Mei and Moura, 2016] pre-

sented an efficient algorithm to estimate a directed weighted graph that captures the causal spatial-temporal relationship among multiple time series. In [Xu and Hero, 2014], Xu et al introduced a state-space model for dynamic network that extended the well-known stochastic blockmodel for static networks to the dynamic setting. An extended Kalman filter based model was proposed that achieved a near-optimal performance in terms of estimation accuracy.

Probabilistic graphical models [Lauritzen, 1996, Wainwright et al., 2008, Koller and Friedman, 2009] provides a systematic framework in representation, inference and learning of high-dimensional data. It have deep connections with information theory, convex analysis as well as graph theory. In Chapter **2**, we review several connections between graphical models, information geometry and maximum entropy learning. This serves as a preliminary for Chapter **5** in which we will discuss the applications of graphical models in robust learning and multi-view learning. In particular, we consider the situation where the graph signal is generated by a Gaussian graphical model (GGM). That is, the joint distribution of graph signal is Gaussian distributed that factorizes according to the underlying network. To infer the underling network topology, the GGM provides a convenient tool that associates the model selection problem with a inverse covariance estimation problem [Lauritzen, 1996, Rue and Held, 2005, Banerjee et al., 2008, Friedman et al., 2008, Rothman et al., 2008, Wainwright et al., 2008, Yuan, 2010, Chen et al., 2011, Pavez and Ortega, 2016]. The latter is convex and is much easier to solve.

1.4.3 Graph Topology Inference

Most tasks in GSP require a full knowledge of the network topology. However, in many applications such as recommendation systems [Aggarwal et al., 1999] and artificial intelligence [Ferber, 1999], sensor networks [Hall and Llinas, 1997] and market prediction [Choi et al., 2010a], a complete network topology may not be available. For these applications, the main task is to infer the network topology given measurements on vertices.

Learning graph topology given data requires additional assumption on the data. In GSP [Hammond et al., 2011, Zhu and Rabbat, 2012, Narang et al., 2013a, Sandryhaila and Moura, 2013, 2014b, Shuman et al., 2013], smoothness conditions are necessary in order to make sure that the data contain sufficient graph information to perform reliable network inference. Under various smoothness assumptions, several algorithms are proposed to solve the network inference problem. For instance, Done et al. [Dong et al., 2016] propose to learning Laplacian matrix by solving a regression problem with graph regularization [Belkin et al., 2006]. Similarly, Liu et al. [Chepuri et al., 2016] learns the graph topology

by solving a convex relaxation of edge selection problem in the context of signal recovery. The regularization and penalties used in these approaches amount to imposing different degrees of smoothness on the solution. Graph Laplacians can also be learned implicitly in the context of multiple kernel learning [Argyriou et al., 2005, Shivaswamy and Jebara, 2010], where additional feature transformation are used to learn a convex combination of graph Laplacians.

For probabilistic GSP [Zhang et al., 2015], learning of graph topology is closely associated with graphical model selection [Lauritzen, 1996, Koller and Friedman, 2009], based on the assumption that the graphical model factorizes according to the underlying network $\mathcal{G}$. For instance, Ravikumar et al [Ravikumar et al., 2010] propose a high-dimensional Ising model selection method based on ℓ_1-regularized logistic regression. Anandkumar et al. [Anandkumar et al., 2011] introduce a threshold-based algorithm for structure learning of high-dimensional Ising and Gaussian models based on condition mutual information. For Gaussian graphical models (GGM), the *sparse inverse covariance (precision) estimation* has attracted a lot of attention in the field of statistics and machine learning [Lauritzen, 1996, Rue and Held, 2005, d'Aspremont et al., 2008, Banerjee et al., 2008, Friedman et al., 2008, Rothman et al., 2008, Wainwright et al., 2008, Yuan, 2010, Chen et al., 2011, Pavez and Ortega, 2016]. Finding the sparse precision matrix from sample covariance involves solving a ℓ_1-regularized Log-Determinant (LogDet) problem [Wang et al., 2010], which can be achieved in polynomial time via interior point methods [Boyd and Vandenberghe, 2004], or by fast coordinate descent [Banerjee et al., 2008, d'Aspremont et al., 2008, Friedman et al., 2008, Mazumder and Hastie, 2012]. For instance, the graphical Lasso [Friedman et al., 2008] is among the most popular algorithm and it is often solved using descent methods such as Newton's algorithm, as in the QUIC algorithm of [Hsieh et al., 2013, 2014], or coordinate descent, as in the ℓ_0 approach of [Marjanovic and Hero, 2015]. In [Pavez and Ortega, 2016], a generalized Laplacian matrix is learned based on a modified dual graphical Lasso [Mazumder and Hastie, 2012] which can be used in spectrum analysis of graph signal. In [Ravikumar et al., 2008], it is shown that, under some incoherence conditions, the support of estimated precision matrix recovers the edge set of the underlying network with high probability.

If the latent variables are present, the marginal precision matrix is no longer sparse due to the marginalization effect. Chandrasekaran et al. [Chandrasekaran et al., 2011, 2012] introduced the latent variable Gaussian graphical model (LV-GGM) which effectively represent the marginal precision matrix using a sparse plus low-rank structure. The LV-GGM is a convex problem and can be solved via interior point methods. Fast implementations include the LogdetPPA in [Wang et al., 2010], the ADMM in [Ma et al., 2013] or AltGD in

[Xu et al., 2017]. The sign consistency and rank consistency for LV-GGM are also proved in [Chandrasekaran et al., 2012], under some conditions addressing the identifiability issues. The identifiability of LV-GGM implies that the latent variables have global influence regardless its position on the network. Additional properties of the LV-GGM were established in [Meng et al., 2014]. In particular, they obtained Frobenius norm error bounds for estimating the precision matrix of an LV-GGM under weaker conditions than [Chandrasekaran et al., 2011, 2012]. A more flexible assumption is based on a decayed-influence latent variable model, which associates the strength of latent effect with a distance measure between the corresponding latent vertices and observed vertices. In Chapter **5**, we proposed the DiLat-GGM, which accommodates noisy side information about the unobserved latent variables.

1.5 List of Publications

1. Xie, Tianpei, Nasser M. Nasrabadi, and Alfred O. Hero. "Learning to Classify With Possible Sensor Failures." IEEE Transactions on Signal Processing 65, no. 4 (2017): 836-849. [Xie et al., 2017]

2. Xie, Tianpei, Nasser M. Nasrabadi, and Alfred O. Hero. "Semi-supervised multisensor classification via consensus-based Multi-View Maximum Entropy Discrimination." In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 1936-1940. IEEE, 2015. [Xie et al., 2015]

3. Xie, Tianpei, Nasser M. Nasrabadi, and Alfred O. Hero. "Learning to classify with possible sensor failures." In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 2395-2399. IEEE, 2014. [Xie et al., 2014]

4. Xie, Tianpei, Nasser M. Nasrabadi, and Alfred O. Hero. "Multi-view learning on statistical manifold via stochastic consensus constraints." in preparation.

5. Xie, Tianpei, Sijia Liu, and Alfred O. Hero. "Collaborative network topology learning from partially observed relational data." in preparation.

CHAPTER 2

Background: Information Theory, Graphical Models and Optimization in Robust Learning

2.1 Introduction

This chapter provides background material to facilitate the understanding of the rest of the thesis. The purpose is to discuss several important concepts and methods that have influence to our work but are not explained in detail in the following chapters.

The *central* theme behind all the methods developed in the thesis is the use of *information-theoretic measures*, including entropies and various divergence measures, as parsimonious representations of uncertainties in the data, as robust optimization surrogates that allows for efficient learning, and as flexible and reliable discrepancy measures in data fusion. Since its introduction in 1948 [Shannon, 1948], information theory has played an important role in the field of digital communications [Shannon, 1948, Gallager, 1968, Shannon, 2001], physics [Jaynes, 1957a,b] and statistics [Kullback, 1997, Akaike, 1998, Cover and Thomas, 2012]. In theoretical machine learning, information theory has been widely used as measure of capacity [MacKay, 2003] and sample complexity [Devroye et al., 2013, Mohri et al., 2012]. Combinations of non-parametric estimation theory and information theory also provides efficient and robust estimators for Bayes error [Hero and Michel, 1999, Hero et al., 2001, 2002, Sricharan et al., 2010, Sricharan and Hero, 2012, Sricharan et al., 2012, Moon and Hero, 2014a,b].

Much of this thesis builds on the concept of the *maximum entropy learning models*. The class of maximum entropy models have deep connections with the *exponential family* of distributions and probabilistic graphical models [Wainwright et al., 2008, Koller and Friedman, 2009]. From the perspective of convex analysis, [Wainwright et al., 2008] show that the maximum entropy estimation problem and the maximal likelihood estimation problem are *conjugate dual* to each other for exponential families. This leads to an alternative interpretation of a given statistical model, which is often useful in formulating alternative

Table 2.1: The various concepts discussed in different chapters of this thesis (•)

Index terms	Chapter 3	Chapter 4	Chapter 5
information theory	•	•	•
KL-divergence	•	•	•
latent variable models	•	•	•
exponential family	•	•	•
minimum discrimination information	•	•	
maximum entropy discrimination	•	•	
regularized Bayesian inference	•	•	
information projection	•	•	
convex duality	•	•	
statistical manifold		•	
posterior regularization		•	
Hellinger distance		•	
Bhattacharyya distance		•	
graphical models			•
matrix Bregman divergence			•

computation and optimization methods. Maximum entropy learning and the exponential family are also studied in the field of information geometry [Amari and Nagaoka, 2007], which investigates the geometric properties of the space of parametric probability distributions.

In Section 2.2, we review a variety of information-theoretic measures and discuss their application to maximum likelihood estimation, Bayesian inference and robust statistics. In Section 2.3, we discuss the formulation a maximum entropy learning method, the method of minimum entropy discrimination, especially MED. MED is associated with graphical models, which are discussed in Section 2.4. We then introduce some concepts in information geometry in Section 2.5, which provides geometric interpretations of maximum entropy models. We also establish the convex duality between maximum likelihood and maximum entropy for exponential families in Section 2.5. In Table 2.1, we list the set of concepts and algorithms discussed in this chapter as well as their occurrence in various chapters of the thesis.

2.2 Information-theoretic Measures

For more details on information theory and associated measured of uncertainty and information the reader may refer to [Cover and Thomas, 2012]. Let x, y be continuous random variables, whose joint distribution is $P(x, y)$ with Lebesgue continuous density $p(x, y)$. Define the marginal probability density over x as $p(x)$. The *Shannon entropy* [Shannon,

1948] over a single variable x is defined as

$$H(x) := H(p) = \mathbb{E}_p[-\log p] := -\int p(x) \log p(x) dx.$$

The Shannon entropy is a measure of uncertainty of random variable [Cover and Thomas, 2012] and $H(x) \geq 0$. The joint Shannon entropy over (x, y) is $H(x, y) = -\int p(x, y) \log p(x, y) dx dy$ and the conditional entropy is $H(y|x) := -\int p(x, y) \log p(y|x) dy dx = \mathbb{E}_{p(x)}[H(p(y|x))]$ where $H(p(y|x)) = \mathbb{E}_{p(y|x)}[-\log p(y|x)] = -\int p(y|x) \log p(y|x) dy$. An important property of Shannon entropy is the chain rule: $H(y, x) = H(x) + H(y|x)$.

The Shannon entropy is a concave functional [Gelfand et al., 2000] over the space of probability density functions $\{p \geq 0 : \int p = 1\}$, and the expectation operator $\mathbb{E}_p[\cdot]$ is a linear functional with respect to p.

One of the most important concepts for our work is the *Kullback-Leibler divergence (KL-divergence)*, also known as the *relative entropy* [Kullback, 1997]. Given two probability densities p and q for random variable x, the KL-divergence from q to p is defined as

$$\mathbb{KL}(p \parallel q) = -\mathbb{E}_p\left[\log\left(\frac{q}{p}\right)\right] = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx. \tag{2.1}$$

KL-divergence is a measure (but not a metric) of the non-symmetric difference between distributions p and q. KL-divergence is non-symmetric, $\mathbb{KL}(p \parallel q) \neq \mathbb{KL}(q \parallel p)$, and $\mathbb{KL}(p \parallel q) \geq 0$ for all (p, q) distributions, where the equality holds if and only if $p = q$. $\mathbb{KL}(p \parallel q)$ is *convex* in the pair (p, q); that is, for any two pairs of distributions (p_1, q_1) and (p_2, q_2),

$$\mathbb{KL}(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2) \leq \lambda \mathbb{KL}(p_1 \parallel q_1) + (1-\lambda)\mathbb{KL}(p_2 \parallel q_2)$$

for any $\lambda \in [0, 1]$. The entropy of a random variable with density p with finite support can be viewed as a special case of the KL divergence between p and a uniform density u over the support set. Specifically, $H(p) = \log|\mathcal{X}| - \mathbb{KL}(p \parallel u)$, where $\mathcal{X}$ is the support of distributions p and u, u is uniform distribution on $\mathcal{X}$, and $|\mathcal{X}|$ is the Lebesgue measure of the support set.

Using KL-divergence, we can reformulate the maximum likelihood estimation as the solution to a geometric projection problem. Consider a set of $i.i.d$ data $\{x_i\}_{i=1}^n$ generated by a parametric distribution $p(x; \theta)$ and let the empirical distribution be $p_n(x) := \sum_{i=1}^n \delta_{x_i}(x)$. The maximum likelihood estimate $\hat{\theta}$ of $\theta \in \Omega$ is the optimal solution of the

following problem

$$\min_{\theta\in\Omega} \quad \mathbb{KL}\left(p_n(x) \parallel p(x;\theta)\right) = -\sum_{i=1}^{n} \log p(x_i;\theta). \tag{2.2}$$

Here the empirical distribution $p_n(x)$ represents the distribution from data and $p(x;\theta)$ represents the distribution from model. Thus maximum likelihood estimation can be seen as minimizing the divergence from the model distribution to the data distribution, where the divergence quantifies the model fitting error.

In the thesis, we consider the KL-divergence $\mathbb{KL}\left(p \parallel q\right)$ from a Bayesian perspective. For a Bayesian statistician, $\mathbb{KL}\left(p \parallel q\right)$ describes the amount of information gain about the random parameter θ if one's belief is revised from prior distribution $p(\theta)$ to the posterior distribution $q(\theta)$ as a result of observing data from $p(x;\theta)$. In particular, the posterior distribution $p(\theta|x_1,\ldots,x_n) = \frac{q(\theta)\prod_{i=1}^{n} p(x_i;\theta)}{p(x_1,\ldots,x_n)}$ from the Bayes' theorem can be obtained alternatively by solving

$$\min_{q(\theta)\in\Delta} \quad \mathbb{KL}\left(q(\theta) \parallel p(\theta)\right) - \sum_{i=1}^{n} \int_{\theta} \log p(x_i;\theta) q(\theta) d\theta, \tag{2.3}$$

where $\Delta := \{q(\theta) > 0, \theta \in \Omega : \int q(\theta)d\theta = 1\}$. Note that compared to (2.2), the unknown variable in (2.3) is on the first argument. The problem (2.3) is referred as the *variational formulation* of Bayes' theorem in [Zhu et al., 2014]. In [Ganchev et al., 2010], the scheme of *posterior regularization* is proposed to incorporate additional prior information in the semi-supervised learning process. The KL-divergence is used as a regularizer, which separates out the model complexity and the complexity of structural constraints. In [Zhu et al., 2014], this framework is generalized to Bayesian inference with the non-parametric Bayesian priors [Müller and Quintana, 2004]. They proposed the *regularized Bayesian inference* of θ as follows

$$\begin{aligned} \min_{q(\boldsymbol{\theta}),q(\mu)\in\Delta} \quad & \mathbb{KL}\left(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta})\right) - \sum_{i=1}^{n} \int_{\boldsymbol{\theta}} \log p(x_i;\boldsymbol{\theta}) q(\boldsymbol{\theta}) d\boldsymbol{\theta} + \mathbb{KL}\left(q(\mu) \parallel p(\mu)\right), \\ \text{s.t.} \quad & \mathbb{E}_q\left[\phi(\boldsymbol{x},\boldsymbol{\theta}) - \mu\right] \leq 0 \end{aligned} \tag{2.4}$$

where $\boldsymbol{x} := [x_i]_{i=1}^{n}$ and $\phi : \mathcal{X} \times \Omega \to \mathbb{R}$ is a feature mapping. Note that the parameter $\boldsymbol{\theta} := \boldsymbol{\theta}(x)$ is infinite-dimensional and depends on the data. The variational formulation of Bayesian inference is the basis for our development of robust and multi-view learning in Chapter 3 and Chapter 4. Note that since $\mathbb{KL}\left(p \parallel q\right)$ is a convex function in (p,q), the problems (2.3) and (2.4) are both convex and have an unique global optima.

Besides the KL-divergence, several other divergence measures will appear in this thesis, including the Hellinger distance [Hellinger, 1909], the Bhattacharyya distance [Bhattacharyya, 1946] and the f-divergence [Csisz et al., 1963, Ali and Silvey, 1966] as their generalization. These measures are popular in robust statistics [Beran, 1977, Lindsay, 1994, Cutler and Cordero-Brana, 1996], physics [Braunstein and Caves, 1994], signal processing [Beigi, 2011, Nielsen and Boltz, 2011] and data mining [Cieslak et al., 2012]. In the presence of noisy mixture components, these measures are more robust as compared to KL-divergence. The *Hellinger distance* between two continuous distributions $p(x)$ and $q(x)$ is defined as

$$\begin{aligned} H(p,q) &:= \sqrt{\frac{1}{2}\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx} \\ &= \sqrt{1 - \int \sqrt{p(x)q(x)}dx}, \end{aligned}$$

where $\int \sqrt{p(x)q(x)}dx$ is referred to the *Bhattacharyya coefficient*. $H(p,q) \in [0,1]$ and it provides lower and upper bounds for the *total variation distance* between two distributions; $H^2(p,q) \leq \|p-q\|_1 \leq \sqrt{2}H(p,q)$. With the Bhattacharyya coefficient, the *Bhattacharyya distance* is defined as

$$B(p,q) := -\log\left(\int \sqrt{p(x)q(x)}dx\right).$$

The Bhattacharyya distance $B(p,q) \in [0,\infty]$ and it measures the amount of overlap between p and q. In Chapter 4, we use a variational formulation of the Bhattacharyya distance

$$B(p,q) = \min_{r\in\Delta} \quad \mathbb{KL}\left(r(x) \,\|\, p(x)\right) + \mathbb{KL}\left(r(x) \,\|\, q(x)\right), \tag{2.5}$$

where p, q are predictive distributions over x in two different *views*.

A generalization of the KL-divergence and the Hellinger distance is the *f-divergence*, which is defined as

$$\mathbb{D}^f\left(p \,\|\, q\right) := \int f\left(\frac{p(x)}{q(x)}\right) p(x)dx.$$

where $f : \mathbb{R} \to \mathbb{R}$ is convex function and $f(1) = 0$. The KL-divergence corresponds to the case where $f(x) = x\log x$ and the square of Hellinger distance corresponds to the case when $f(x) = (\sqrt{x}-1)^2$. Note that for all pairs (p,q), the f-divergence $\mathbb{D}^f\left(p \,\|\, q\right)$ is *non-negative* and *convex*.

The variational formulations in (2.3), (2.4) and (2.5) can be seen as learning maximum entropy models. We next discuss maximum entropy learning in Section 2.3.

2.3 Maximum Entropy Discrimination

As discussed in Section 1.3, maximum entropy learning has many applications. In this section, we discuss its natural extension, called the *Principle* of *Minimum Discrimination Information (MDI)*. When dealing with continuous random variable with non-uniform prior distribution, a maximum entropy model follows the MDI, which states that given new data, the information gain of a new distribution q from the original distribution p should be as small as possible; that is, $\mathbb{KL}\left(q \,\|\, p\right)$ is minimized. Specifically, in MDI one solves the convex optimization problem

$$\begin{aligned} \min_{q \geq 0} \quad & \mathbb{KL}\left(q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta})\right) \\ \text{s.t.} \quad & \mathbb{E}_q\left[\eta_j(\boldsymbol{\theta})\right] := \int \eta_j(\boldsymbol{\theta}) q(\boldsymbol{\theta}) d\boldsymbol{\theta} = T_j(\boldsymbol{x}), \quad j = 1, \ldots, s. \\ & \int q(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1, \end{aligned} \tag{2.6}$$

where $\eta_j : \boldsymbol{\theta} \mapsto \eta_j(\boldsymbol{\theta}) \in \mathbb{R}$ corresponds to a mapping of parameters $\boldsymbol{\theta}$ and $T_j(\boldsymbol{x})$ is a constant that depends only on data.

The Lagrangian functional associated with (2.6) is

$$\mathcal{L}(q, \boldsymbol{\lambda}) = \int q \log \frac{q}{p} + \lambda_0 \int q + \sum_{j=1}^{s} \lambda_j \int q\, \eta_j. \tag{2.7}$$

Since (2.6) is a convex optimization, calculus of variations over $q(\boldsymbol{\theta})$ asserts that the solution to (2.6) must satisfy

$$\frac{\partial \mathcal{L}}{\partial q(\boldsymbol{\theta})} = \log q(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}) + 1 + \lambda_0 + \sum_{j=1}^{s} \lambda_j \eta_j(\boldsymbol{\theta}) = 0.$$

The stationary point condition yields the global optimal solution

$$q^*(\boldsymbol{\theta}) = p(\boldsymbol{\theta}) \exp\left(- \sum_{j=1}^{s} \lambda_j \eta_j(\boldsymbol{\theta}) - \lambda_0 + 1 \right) \tag{2.8}$$

where $\lambda_0, \ldots, \lambda_s$ are chosen so that the equality constraints are satisfied. This implies

that $h(\boldsymbol{x}) := \exp(-\lambda_0 + 1)$ defines the normalization factor $Z^{-1}(\boldsymbol{x}; \lambda_1, \ldots, \lambda_s)$, where $Z(\boldsymbol{x}; \lambda_1, \ldots, \lambda_s) = \int p(\boldsymbol{\theta}) \exp\left(-\sum_{j=1}^{s} \lambda_j \eta_j(\boldsymbol{\theta})\right) d\boldsymbol{\theta}$ is referred as the partition function of $q^*(\boldsymbol{\theta})$. Substituting (2.8) into the Lagrangian functional (2.7), we have the dual objective function

$$\begin{aligned}
\mathcal{L}(q^*, \boldsymbol{\lambda}) = & \frac{1}{Z(\boldsymbol{x}; \lambda_1, \ldots, \lambda_s)} \int p(\boldsymbol{\theta}) \exp\left(-\sum_{j=1}^{s} \lambda_j \eta_j(\boldsymbol{\theta})\right) \left[-\sum_{j=1}^{s} \lambda_j \eta_j(\boldsymbol{\theta})\right] d\boldsymbol{\theta} \\
& + \frac{1}{Z(\boldsymbol{x}; \lambda_1, \ldots, \lambda_s)} \sum_{j=1}^{s} \lambda_j \int p(\boldsymbol{\theta}) \exp\left(-\sum_{j=1}^{s} \lambda_j \eta_j(\boldsymbol{\theta})\right) \eta_j(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
& - \lambda_0 + 1 \\
= & -\lambda_0 + 1 = -\log Z(\boldsymbol{x}; \lambda_1, \ldots, \lambda_s).
\end{aligned}$$

Therefore, the variables $(\lambda_1, \ldots, \lambda_s)$ are optimal solutions of the dual optimization problem

$$\begin{aligned}
\max_{\lambda_1, \ldots, \lambda_s} \quad & -\log \int p(\boldsymbol{\theta}) \exp\left(-\sum_{j=1}^{s} \lambda_j \eta_j(\boldsymbol{\theta})\right) d\boldsymbol{\theta} \\
\text{s.t.} \quad & \mathbb{E}_{q(\boldsymbol{\theta}; \boldsymbol{\lambda})}\left[\eta_j(\boldsymbol{\theta})\right] = T_j(\boldsymbol{x}), \quad j = 1, \ldots, s.
\end{aligned} \tag{2.9}$$

The final solution of (2.8) has the form

$$q^*(\boldsymbol{\theta}|\boldsymbol{x}) := p(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{s} T_j(\boldsymbol{x}) \eta_j(\boldsymbol{\theta}) - A(\boldsymbol{x})\right) \tag{2.10}$$

where $A(\boldsymbol{x}) := \log Z(\boldsymbol{x})$ is the *log-partition function*. A family of distributions that can be expressed in (2.10) is said to belong to the *exponential family*, denoted as $\mathcal{P}$. The functions $\boldsymbol{\eta} := [\eta_j(\boldsymbol{\theta})] \in \mathbb{R}^s$ correspond to a set of natural parameters (or mean parameters) and $\boldsymbol{T} = (T_j(\boldsymbol{x}))_j$ is a set of sufficient statistics since the conditional distribution q does not depend on the data $\boldsymbol{x}$ given $\boldsymbol{T}$. Note that the formulation (2.6) directly learns a natural parameterization of exponential families, since the data $\boldsymbol{x}$ and the canonical parameters $\boldsymbol{\theta}$ only affect the linear constraints via the sufficient statistics $\boldsymbol{T}$ and the natural parameters $\boldsymbol{\eta}$. The exponential families include many of the most common distributions, such as the normal distribution, exponential distribution, Poisson distribution, Bernoulli distribution, gamma distribution, beta distribution, binomial distribution, multinomial distribution, Dirichlet distribution etc.

To learn discriminative models using maximum entropy principle, T. Jaakkola proposed the Maximum Entropy Discrimination (MED) [Jaakkola et al., 1999]. Let $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$

be i.i.d data from joint distribution $p(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{\theta})$, where $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ and $\boldsymbol{y}_i \in \mathcal{Y} := \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k\}$ for $\boldsymbol{e}_i \in \mathbb{R}^k$ is an all-0's vector with i-th entry as 1. The MED replaces the mapping $\eta_j(\boldsymbol{\theta})$ with the *log-likelihood ratio test function*

$$f(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{y}, \boldsymbol{\theta}) := \log\left(\frac{p(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{\theta})}{p(\boldsymbol{x}_i, \boldsymbol{y}; \boldsymbol{\theta})}\right) = \log\left(\frac{p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{x}_i, \boldsymbol{\theta})}\right), i = 1, \ldots, n.$$

The last equality holds since $p(\boldsymbol{x}_i)$ is fixed. Then the MED solves the following problem:

$$\begin{aligned} \min_{q \geq 0} \quad & \mathbb{KL}\left(q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta})\right) + \sum_i \mathbb{KL}\left(q(\rho_i) \,\|\, p(\rho_i)\right) \\ \text{s.t.} \quad & \mathbb{E}_q\left[f(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{y}, \boldsymbol{\theta}) - \rho_i\right] \geq 0, \;\; \forall\, \boldsymbol{y} \neq \boldsymbol{y}_i, i = 1, \ldots, n. \\ & \int q(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1, \end{aligned} \tag{2.11}$$

where $\{\rho_i\}$ defines a set of margins with the prior distribution $p(\rho_i) \propto \exp(-c(s_\alpha - \rho_i))$ for $\rho_i \leq s_\alpha$, s_α is chosen to be some α-percentile of the margins obtained by standard MAP procedure. The prior of $\boldsymbol{\theta}$ is usually chosen to be the Gaussian process with $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{K})$ and $[\boldsymbol{K}]_{i,j} := K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Similar to above derivations, the optimal solution is

$$q^*(\boldsymbol{\theta}; \boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{x}, \boldsymbol{\lambda})} p(\boldsymbol{\theta}) \exp\left(-\sum_{i=1}^n \sum_{\boldsymbol{y} \in \mathcal{Y}, \boldsymbol{y} \neq \boldsymbol{y}_i} \lambda_{i,\boldsymbol{y}} f(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{y}, \boldsymbol{\theta})\right) \exp\left(-\sum_{i=1}^n \sum_{\boldsymbol{y} \in \mathcal{Y}, \boldsymbol{y} \neq \boldsymbol{y}_i} \rho_i \lambda_{i,\boldsymbol{y}}\right),$$

where the dual variables $\{\lambda_{i,\boldsymbol{y}}\}$ are obtained by solving the dual optimization problem

$$\max_{\boldsymbol{\lambda} \geq 0} -\log \int p(\boldsymbol{\theta}) \exp\left(-\sum_{i=1}^n \sum_{\boldsymbol{y} \in \mathcal{Y}, \boldsymbol{y} \neq \boldsymbol{y}_i} \lambda_{i,\boldsymbol{y}} f(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{y}, \boldsymbol{\theta})\right) \exp\left(-\sum_{i=1}^n \sum_{\boldsymbol{y} \in \mathcal{Y}, \boldsymbol{y} \neq \boldsymbol{y}_i} \rho_i \lambda_{i,\boldsymbol{y}}\right) d\boldsymbol{\theta}.$$

The classifier is defined as $\hat{y} := \arg\max_{\boldsymbol{y} \in \mathcal{Y}} \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\right]$. Compared with (2.6), the solution of MED problem in (2.11) does not necessarily belong to the exponential family $\mathcal{P}$, since the log-likelihood ratio function $f(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{y}, \boldsymbol{\theta})$ usually is not separable as $\eta_i(\boldsymbol{\theta}) T_i(\boldsymbol{x})$. As discussed in [Jaakkola et al., 1999, Jebara, 2011], MED is a Bayesian extension of the support vector machine (SVM), and it can be seen as learning a convex combination of random classifiers as opposed to SVM, which learns a single classifier. From a computational perspective, computing the dual variables is a challenging problem since it involves determining a log-partition function as in (2.9).

2.4 Graphical Models and Exponential Families

Graphical models [Lauritzen, 1996, Wainwright et al., 2008, Koller and Friedman, 2009] bring together the graph theory and probabilistic modeling into a power formalism in multivariate statistical analysis. For a random vector $\boldsymbol{x} := (x_1, \ldots, x_d) \sim p(\boldsymbol{x}; \boldsymbol{\theta})$, $p(\boldsymbol{x}; \boldsymbol{\theta})$ is a graphical model associated with some graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$, if $\mathcal{V} = \{1, \ldots, d\}$ and $x_i \perp\!\!\!\perp x_j | x_{d-\{i,j\}}$ if $(i, j) \notin \mathcal{E}$. The main task in graphical model learning is to infer the parameters $\boldsymbol{\theta}$ from i.i.d. data $\{\boldsymbol{x}_m\}_{m=1}^n$. Various inference algorithms, such as the sum-product and max-product message-passing algorithms [Lauritzen, 1996], the expectation propagation algorithm [Rasmussen and Williams, 2006, Koller and Friedman, 2009] and the Markov chain Monte Carlo methods [Robert and Casella, 1999, Koller and Friedman, 2009].

In [Wainwright et al., 2008], Wainwright et al. introduce the variational inference methods which are based on the maximum entropy learning model in (2.6). As discussed in previous section, the variational form (2.6) provides a natural parameterization of the exponential family $\mathcal{P}$. It is known that a few well-studied graphical models belong to the exponential family, including the Gaussian graphical models (GMM) [Lauritzen, 1996], the Ising model [Ising, 1925], Boltzmann machine [Ackley et al., 1985, Hinton and Salakhutdinov, 2006] and the log-linear models [Lauritzen, 1996]. For instance, the Gaussian graphical model has the form

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\Theta}, \boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta}^T \boldsymbol{x} - \frac{1}{2}\text{tr}\left(\boldsymbol{\Theta}\boldsymbol{x}\boldsymbol{x}^T\right) - A(\boldsymbol{\theta}, \boldsymbol{\Theta})\right\}$$

where $\boldsymbol{\Theta} \succ \mathbf{0}$ is the precision matrix, $\boldsymbol{\theta} := \boldsymbol{\Theta}\boldsymbol{\mu}$ for mean $\boldsymbol{\mu}$ and $A(\boldsymbol{\theta}, \boldsymbol{\Theta}) \propto -\frac{1}{2}\log\det\boldsymbol{\Theta} + \frac{1}{2}\text{tr}\left(\boldsymbol{\Theta}^{-1}\boldsymbol{\theta}\boldsymbol{\theta}^T\right)$ is the log-partition function. For high-dimensional data, the inference of graphical model lack of inefficiency and accuracy. Wainwright et al. then propose to use the marginal polytope, which defines each η_j so that it only involves a few variables. Then the constraint $\mathbb{E}_q\left[\eta_j(\boldsymbol{\theta})\right] = T_j$ in (2.6) only involves the marginal distribution, which can be computed efficiently via message-passing algorithm.

The feasible region $\mathcal{M} := \{q \in \Delta : \mathbb{E}_q\left[\eta_j(\boldsymbol{\theta})\right] = T_j, \quad j = 1, \ldots, s\}$ defines a set of distributions whose mean parameters satisfies the equality constraint $\mathbb{E}_q\left[\boldsymbol{\eta}\right] = \boldsymbol{T}$. Note that for any $p, q \in \mathcal{M}$, $\alpha p + (1-\alpha)q \in \mathcal{M}, \alpha \in [0, 1]$, thus the distributions in $\mathcal{M}$ belong to a *mixture family* [Amari and Nagaoka, 2007]. The mixture family $\mathcal{M}$ is *not* a subset of exponential family $\mathcal{P}$. On the other hand, for any $p, q \in \mathcal{P}$, $p^\alpha q^{1-\alpha}/Z \in \mathcal{P}$, where Z is the partition function. In Chapter 4, we use this fact in multi-view fusion.

2.5 Convex Duality, Information Geometry and Bregman Divergence

According to information geometry [Amari and Nagaoka, 2007], the set of solutions of the maximum entropy problem (2.6), for all possible mean parameters $\boldsymbol{\eta} \in \mathbb{R}^s$, defines a *smooth manifold* $\mathcal{S} := \{q^* \in \Delta : q^* = \arg\min_p \mathbb{KL}(q \,\|\, u) \;\text{ s.t. } \mathbb{E}_q[\boldsymbol{T}(x)] = \boldsymbol{\eta}, \;\; \forall \boldsymbol{\eta} \in \mathbb{R}^s\} \subset \mathcal{P} \cap \mathcal{M}$, where u is the uniform distribution. The reason why $\mathcal{S}$ is a manifold is because there exists a smooth one-to-one mapping $\boldsymbol{\eta} : \mathcal{S} \to \mathbb{R}^s$ so that for each $q \in \mathcal{S}$, there corresponds to a unique mean parameter $\boldsymbol{\eta}(q) := \boldsymbol{\eta} \in \mathbb{R}^s$. The uniqueness comes from the convexity of the maximum entropy problem. $\boldsymbol{\eta}$ is a coordinate system and $\mathcal{S}$ is referred as a statistical manifold. Formally speaking, a *statistical manifold* is a set of distributions equipped with a coordinate system that locally maps from a neighborhood of distributions to a neighborhood in Euclidean space. Both the exponential family $\mathcal{P}$ and the mixture family $\mathcal{M}$ are statistical manifolds. Unlike the Euclidean space, a statistical manifold is non-Euclidean, meaning that a geodesic curve on a statistical manifold is not a straight line in space.

Given the smooth manifold $\mathcal{P}$, the KL-divergence in (2.6) as well as the Hellinger distance can be seen as inducing a geometric structure on $\mathcal{P}$. Moreover, we refer to the operator $q^* = \text{e-proj}_{\mathcal{M}}(p) := \arg\min_{q \in \mathcal{M}} \mathbb{KL}(q \,\|\, p)$ as the *e-projection over* $\mathcal{M}$, since for $p \in \mathcal{P}$, $\text{e-proj}_{\mathcal{M}}(p) \in \mathcal{P}$. Similarly, the operator $q^* = \text{m-proj}_{\mathcal{P}}(p) := \arg\min_{q \in \mathcal{P}} \mathbb{KL}(p \,\|\, q)$ is defined as the *m-projection over* $\mathcal{P}$, since for $p \in \mathcal{M}$, $\text{m-proj}_{\mathcal{P}}(p) \in \mathcal{M}$. As shown in Figure 1.2 in Section 1.3, the maximum entropy learning is seen as e-projection of prior p over $\mathcal{M}$. An important fact from the information geometry is that $\mathcal{P}$ and $\mathcal{M}$ have dual geometric structure to each other [Amari and Nagaoka, 2007], which means that the e-projection on $\mathcal{M}$ is orthogonal.

There is a corresponding convex analysis perspective through the *convex duality* between the maximum entropy learning and the maximum likelihood estimation. Given a convex function $f : \Omega \subset \mathbb{R}^s \to \mathbb{R}$ and an inner product $\langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle$ on $\mathbb{R}^s \times \mathbb{R}^s$, the conjugate dual function of f is defined as

$$f^*(\boldsymbol{\mu}) := \sup_{\boldsymbol{\theta} \in \Omega} \{\langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle - f(\boldsymbol{\theta})\}. \tag{2.12}$$

The variable $\boldsymbol{\mu} \in \mathbb{R}^s$ is referred as the dual variable. Denote $f(\boldsymbol{\theta}) := A(\boldsymbol{\theta}) = \log \int q(\boldsymbol{\theta}) d\boldsymbol{\theta}$ as the log-partition function for $q(\boldsymbol{\theta}) := \exp\{\langle \boldsymbol{T}(\boldsymbol{x}), \boldsymbol{\theta} \rangle - A(\boldsymbol{\theta})\} \in \mathcal{P}$ in exponential family. As seen in Section 2.3, if $\boldsymbol{\mu} = \mathbb{E}_{q(\boldsymbol{\theta})}[T(\boldsymbol{x})]$, the mean parameter of $q(\boldsymbol{\theta})$, then in this case the right-hand side is the optimal value of the maximum log-likelihood estimation of

$\boldsymbol{\theta}$. Following the proof in [Wainwright et al., 2008], the conjugate dual function $f^*(\boldsymbol{\mu}) := A^*(\boldsymbol{\mu}) = -H(q(\boldsymbol{\theta}(\boldsymbol{\mu})))$, the negative entropy of q, where $\boldsymbol{\theta}(\boldsymbol{\mu})$ is the unique canonical parameter satisfying the dual matching condition $\mathbb{E}_{q(\boldsymbol{\theta}(\boldsymbol{\mu}))}[T(\boldsymbol{x})] = \nabla A(\boldsymbol{\theta}(\boldsymbol{\mu})) = \boldsymbol{\mu}$. Also since for a convex closed function f, $f^{**} = f$, the log-partition function $A(\boldsymbol{\theta})$ has a variational representation in turns of its dual

$$A(\boldsymbol{\theta}) = \sup_{\boldsymbol{\mu} \in \mathcal{T}} \{\langle \boldsymbol{\theta}\,,\, \boldsymbol{\mu} \rangle + H(q(\boldsymbol{\theta}'(\boldsymbol{\mu})))\} \tag{2.13}$$

where $\mathcal{T} := \{\boldsymbol{\mu} \in \mathbb{R}^s : \exists q \in \mathcal{P},\ \boldsymbol{\mu} = \mathbb{E}_q[T(\boldsymbol{x})]\}$, $H(q(\boldsymbol{\theta}'(\boldsymbol{\mu}))) = -A^*(\boldsymbol{\mu})$ is the entropy of distribution $q(\boldsymbol{\theta}'(\boldsymbol{\mu}))$. The problem in (2.13) is essentially maximum entropy learning under a linear constraint in $\mathcal{T}$. In other word, the maximum likelihood estimation in (2.12) and the maximum entropy learning in (2.13) are *dual* problems to each other. The mean parameter $\boldsymbol{\mu}$ and the canonical parameter $\boldsymbol{\theta}$ satisfy the dual matching condition $\boldsymbol{\mu} = \mathbb{E}_{q(\boldsymbol{\theta})}[T(\boldsymbol{x})]$.

Finally, a related divergence measure is the *Bregman divergence* [Bregman, 1967]. which is defined as

$$\mathbb{D}^{\phi}(\boldsymbol{\theta} \parallel \boldsymbol{\mu}) := \phi(\boldsymbol{\theta}) - \phi(\boldsymbol{\mu}) - \nabla\phi(\boldsymbol{\mu})^T(\boldsymbol{\theta} - \boldsymbol{\mu}),$$

where $\phi : \Omega \subset \mathbb{R}^s \to \mathbb{R}$ is a real-valued strictly convex function defined over a convex domain $\Omega \subset \mathbb{R}^s$. $\mathbb{D}^{\phi}(\boldsymbol{\theta} \parallel \boldsymbol{\mu}) \geq 0$ for all $(\boldsymbol{\theta}, \boldsymbol{\mu})$, where the equality holds if and only if $\boldsymbol{\theta} = \boldsymbol{\mu}$. For $\phi(\boldsymbol{\theta}) := \|\boldsymbol{\theta}\|_2^2$, the resulting Bregman divergence is the Euclidean distance between $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$. For $\phi(\boldsymbol{\theta}) := \sum_i(\theta_i \log \theta_i - \theta_i)$, the resulting Bregman divergence becomes the unnormalized KL-divergence $\mathbb{D}^{\phi}(\boldsymbol{\theta} \parallel \boldsymbol{\mu}) = \sum_i(\theta_i \log \frac{\theta_i}{\mu_i} - \theta_i + \mu_i)$. Similar to KL-divergence, the Bregman divergence is used as a robust surrogate function in the field of supervised learning [Murata et al., 2004, Nock and Nielsen, 2009, Santos-Rodríguez et al., 2009, Liu and Vemuri, 2011], clustering [Banerjee et al., 2007, Ackermann and Blömer, 2010], matrix factorization [Dhillon and Sra, 2005, Tsuda et al., 2005], low-rank kernel approximation [Kulis et al., 2009], graphical model learning [Friedman et al., 2008] etc.

In Chapter 5, we deal with the matrix Bregman divergence. For instance, the *LogDet divergence* [Kulis et al., 2009, Wang et al., 2010] is defined as

$$\mathbb{D}^{\phi}(\boldsymbol{\Theta}_1 \parallel \boldsymbol{\Theta}_2) = \mathbb{D}^{det}(\boldsymbol{\Theta}_1 \parallel \boldsymbol{\Theta}_2) := \operatorname{tr}\left[\boldsymbol{\Theta}_1(\boldsymbol{\Theta}_2)^{-1}\right] - \log\det\left[\boldsymbol{\Theta}_1(\boldsymbol{\Theta}_2)^{-1}\right] - s, \tag{2.14}$$

for $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \in \mathbb{R}^{s \times s}$ and all positive definite $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \succ \mathbf{0}$. Here the corresponding $\phi(\boldsymbol{\Theta}) := -\sum_i \log \lambda_i$, for $(\lambda_1, \ldots, \lambda_s)$ are eigenvalues of $\boldsymbol{\Theta}$.

CHAPTER 3

Robust Maximum Entropy Training on Approximated Minimal-entropy Set

3.1 Introduction

Large margin classifiers, such as the support vector machine (SVM) [Schölkopf and Smola, 2002] and the MED classifier [Jaakkola et al., 1999], have enjoyed great popularity in the signal processing and machine learning communities due to their broad applicability, robust performance, and the availability of fast software implementations. When the training data is representative of the test data, the performance of MED/SVM has theoretical guarantees that have been validated in practice [Bousquet and Elisseeff, 2002, Schölkopf and Smola, 2002, Bartlett and Mendelson, 2003]. Moreover, since the decision boundary of the MED/SVM is solely defined by a few support vectors, the algorithm can tolerate random feature distortions and perturbations.

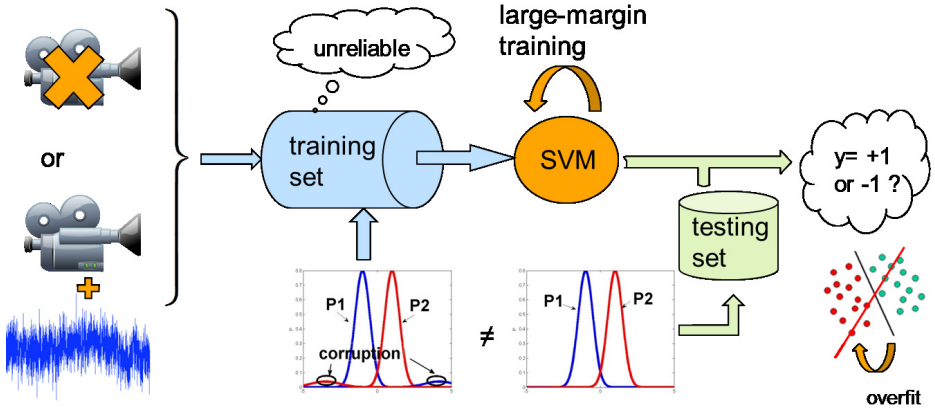


Figure 3.1: Due to corruption in the training data the training and testing sample distributions are different from each other, which introduces errors into the decision boundary.

However, in many real applications, anomalous measurements are inherent to the data

set due to strong environmental noise or possible sensor failures. Such anomalies arise in industrial process monitoring, video surveillance, tactical multi-modal sensing, and, more generally, any application that involves unattended sensors in difficult environments (Fig. 3.1). Anomalous measurements are understood to be observations that have been corrupted, incorrectly measured, mis-recorded, drawn from different environments than those intended, or occurring too rarely to be useful in training a classifier Yang et al. [2010]. If not robustified to anomalous measurements, classification algorithms may suffer from severe degradation of performance. Therefore, when anomalous samples are likely, it is crucial to incorporate outlier detection into the classifier design. This chapter provides a new robust approach to design outlier resistant large margin classifiers.

3.1.1 Problem setting and our contributions

We divide the class of supervised training methods into four categories, according to how anomalies enter into different learning stages.

Table 3.1: Categories for supervised training algorithms via different assumption of anomalies

	Training set (uncorrupted)	Training set (corrupted)
Test set (uncorrupted)	classical learning algorithms (e.g. [Freund and Schapire, 1995, Vapnik and Vapnik, 1998, Jaakkola et al., 1999])	Robust classification & training (e.g. [Bartlett and Mendelson, 2003, Bousquet and Elisseeff, 2002, Krause and Singer, 2004, Xu et al., 2006, Wang et al., 2008, Tyler, 2008, Masnadi-Shirazi and Vasconcelos, 2009, Long and Servedio, 2010, Forero et al., 2012, Ding et al., 2013], **this chapter**)
Test set (corrupted)	anomaly detection (e.g. [Schölkopf et al., 1999, Scott and Nowak, 2006, Hero, 2006, Sricharan and Hero, 2011])	Domain adaptation & transfer learning (e.g. [Blitzer et al., 2006, Dai et al., 2007, Pan and Yang, 2010])

As shown in Table 3.1, a majority of learning algorithms assume that the training and test samples follow the same nominal distribution and neither are corrupted by anomalies. Under this assumption, an empirical error minimization algorithm can achieve consistent performance on the test set. In the case that anomalies exist only in the test data, one can apply anomaly detection algorithms, e.g. [Scott and Nowak, 2006, Hero, 2006, Chandola et al., 2009, Sricharan and Hero, 2011], to separate the anomalous samples from nominal ones. Under additional assumptions on the nominal set, these algorithms can effectively

identify an anomalous sample under given false alarm rate and miss rate. Furthermore, in the case that both training and test set are corrupted, possibly with different corruption rate, domain adaptation or transfer learning methods may be applied [Blitzer et al., 2006, Daume III and Marcu, 2006, Pan and Yang, 2010].

This chapter falls into the category of *robust classification & training* in which possibly anomalous samples occur in the training set. Such a problem is relevant, for example, when high quality clean training data is too expensive or too difficult to obtain. In [Bousquet and Elisseeff, 2002, Bartlett and Mendelson, 2003, Krause and Singer, 2004], the test set is assumed to be uncorrupted so that the test error provides an unbiased estimate of the generalization error on the *nominal data set*, which is a standard measure of performance for robust classifiers. We adopt this assumption, although we also evaluate the proposed robust classifier when the test set is also corrupted with limited corruption rate. Our *goal* is to train a classifier that minimizes the generalization error with respect to the nominal data distribution when the *training* set may be corrupted.

The area of robust classification has been thoroughly investigated in both theory [Bartlett and Mendelson, 2003, Bousquet and Elisseeff, 2002, Krause and Singer, 2004, Xu et al., 2006, Wu and Liu, 2007, Tyler, 2008, Masnadi-Shirazi and Vasconcelos, 2009] and applications [Wang et al., 2008, Long and Servedio, 2010, Forero et al., 2012, Ding et al., 2013]. Tractable robust classifiers that identify and remove outliers, called the Ramp-Loss based learning methods, have been studied in [Song et al., 2002, Bartlett and Mendelson, 2003, Xu et al., 2006, Wang et al., 2008]. Among these methods, Xu et al. [Xu et al., 2006] proposed the *Robust-Outlier-Detection (ROD)* method as an outlier detection and removal algorithm using the soft margin framework. Training the ROD algorithm involves solving an optimization problem, for which dual solution is obtained via semi-definite programming (SDP). Like all the Ramp-Loss based learning models, this optimization is non-convex requiring random restarts to ensure a globally optimal solution [Long and Servedio, 2010, Yang et al., 2010]. In this chapter, in contrast to the models above, a *convex* framework for robust classification is proposed and a tractable algorithm is presented that finds the unique optimal solution of a penalized entropy-based objective function.

Our proposed algorithm is motivated by the basic principle underlying the so-called *minimal volume (MV) /minimal entropy (ME) set anomaly detection method* [Schölkopf et al., 1999, Scott and Nowak, 2006, Hero, 2006, Sricharan and Hero, 2011]. Such methods are expressly designed to detect anomalies in order to attain the lowest possible false alarm and miss probabilities. In machine learning, nonparametric algorithms are often preferred since they make fewer assumptions on the underlying distribution. Among these methods, we focus on the *Geometric Entropy Minimization (GEM)* algorithm [Hero, 2006,

Sricharan and Hero, 2011]. This algorithm estimates the ME set based on the k-nearest neighbor graph (k-NNG), which is shown to be the Uniformly Most Powerful Test at given level when the anomalies are drawn from an unknown mixture of known nominal density and uniform anomalous density[Hero, 2006]. A *key contribution* of this chapter is the incorporation of the non-parametric GEM anomaly detection into a binary classifier under a non-parametric corrupt-data model.

The proposed framework, called the *GEM-MED*, follows a *Bayesian* perspective. It is an extension of the well-established MED approach proposed by Jaakkola et al. [Jaakkola et al., 1999]. MED performs Bayesian large margin classification via the maximum entropy principle and it subsumes SVM as a special case. The MED model can also solve the parametric anomaly detection [Jaakkola et al., 1999] problem and has been extended to multitask classification [Jebara, 2011]. A naive application of MED to robust classification might use a two-stage approach that implements an anomaly detector on the training set prior to training the MED classifier, which is sub-optimal. In this chapter, we propose GEM-MED as a unified approach that jointly solves an anomaly detection and classification problem via the MED framework. The GEM-MED explicitly incorporates the anomaly detection false-alarm constraint and the mis-classification rate constraint into a maximum entropy learning framework. Unlike the two-stage approach, GEM-MED finds anomalies by investigating both the underlying sample distribution and the sample-label relationship, allowing anomalies in support vectors to be more effectively suppressed. As a Bayesian approach, GEM-MED requires no tuning parameter as compared to other anomaly-resistant classification approaches, such as ROD [Xu et al., 2006]. We demonstrate the superior performance of the GEM-MED anomaly-resistant classification approach over other robust learning methods on simulated data and on a real data set combining sensor failure. The real data set contains human-alone and human-leading-animal footsteps, collected in the field by an acoustic sensor array [Damarla et al., 2011, Damarla, 2012, Huang et al., 2011].

What follows is a brief outline of the chapter. In Section **II**, we review MED as a general framework to perform classification and other inference tasks. The proposed combined GEM-MED approach is presented in Section **III**. A variational implementation of GEM-MED is introduced in Section **IV**. Experimental results based on synthetic data and real data are presented in Section **V**. Our conclusions are discussed in Section **3.6**.

3.2 From MED to GEM-MED: A General Routine

Denote the training data set as $\mathcal{D}_t := \{(y_n, \mathbf{x}_n)\}_{n \in T}$, where each sample-pair $(y_n, \boldsymbol{x}_n) \in \mathcal{Y} \times \mathcal{X} = \mathcal{D}$ are independent. Denote the feature set $\mathcal{X} \subset \mathcal{R}^p$ and the label set as $\mathcal{Y}$. For

simplicity, let $\mathcal{Y} = \{-1, 1\}$. The test data set is denoted as $\mathcal{D}_s := \{\mathbf{x}_m\}_{m \in S}$. We assume that $\{(y_n, \mathbf{x}_n)\}_{n \in T}$ are i.i.d. realizations of random variable (Y, X) with distribution $\mathcal{P}_t$, conditional probability density $p(X|Y = y, \Theta)$ and prior $p(Y = y), y \in \mathcal{Y}$, where Θ is the set of unknown model parameters. We denote by $p(Y = y, X; \Theta) = p(X|Y = y, \Theta)p(Y = y)$ the parameterized joint distribution of (Y, X). The distribution of test data, denoted as $\mathcal{P}_s$, is defined similarly. $\mathcal{P}_{nom}$ denotes the nominal distribution. Finally, we define the probability simplex $\Delta_{\mathcal{Y} \times \mathcal{X}}$ over the space $\mathcal{Y} \times \mathcal{X}$.

3.2.1 MED for Classification and Parametric Anomaly Detection

The Maximum entropy discrimination (MED) approach to learning a classifier was proposed by Jaakkola et al [Jaakkola et al., 1999]. The MED approach is a Bayesian maximum entropy learning framework that can either perform conventional classification, when $\mathcal{P}_t = \mathcal{P}_s = \mathcal{P}_{nom}$, or anomaly detection, when $\mathcal{P}_t \neq \mathcal{P}_s$, and $\mathcal{P}_t = \mathcal{P}_{nom}$. In particular, assume that all parameters in Θ are random with prior distribution $p_0(\Theta)$. Then MED is formulated as finding the posterior distribution $q(\Theta)$ that minimizes the relative entropy

$$\mathbb{KL}\left(q(\Theta) \parallel p_0(\Theta)\right) := \int \log\left(\frac{q(\Theta)}{p_0(\Theta)}\right) q(d\Theta) \tag{3.1}$$

subject to a set of P constraints on the risk or loss:

$$\int \mathcal{L}_i\left(p, (y_n, \boldsymbol{x}_n); \Theta\right) q(d\Theta) \leq 0, \ \forall n \in T, 1 \leq i \leq P. \tag{3.2}$$

The constraint functions $\{\mathcal{L}_i\}_{i=1}^{P}$ can correspond to losses associated with different type of errors, e.g. misdetection, false alarm or misclassification. For example, the classification task defines a parametric discriminant function $\mathcal{F}_C : \Delta_{\mathcal{Y} \times \mathcal{X}} \times \mathcal{D} \to \mathbb{R}_+$ as

$$\mathcal{F}_C\left(p, (y_n, \boldsymbol{x}_n); \Theta\right) := \log p(Y = y_n|\boldsymbol{x}_n; \Theta)/p(Y \neq y_n|\boldsymbol{x}_n; \Theta).$$

In the case of the SVM classification, the loss function is defined as

$$\mathcal{L}_i = \mathcal{L}_C\left(p, (y_n, \boldsymbol{x}_n); \Theta\right) := \left[\xi_n - \mathcal{F}_C\left(p, (y_n, \boldsymbol{x}_n); \Theta\right)\right]. \tag{3.3}$$

Other definitions of discriminant functions are also possible [Jaakkola et al., 1999].

An example of an anomaly detection test function $\mathcal{L}_i = \mathcal{L}_D : \Delta_{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}$, is

$$\mathcal{L}_D\left(p, \boldsymbol{x}_n; \Theta\right) := -\left[\log p(\boldsymbol{x}_n; \Theta) - \beta\right], \tag{3.4}$$

where $p(\boldsymbol{x}_n;\Theta)$ is the marginal likelihood $p(\boldsymbol{x}_n;\Theta)=\sum_{y_n\in\mathcal{Y}}p(X=\boldsymbol{x}_n|Y=y_n,\Theta)p(Y=y_n)$. The constraint function (3.4) has the interpretation as local entropy of X in the neighborhood of $X=\boldsymbol{x}_n$. Minimization of the *average* constraint function yields the minimal entropy anomaly detector [Hero, 2006, Sricharan and Hero, 2011]. The solution to the minimization (3.2) yields a posterior distribution $p(Y=y|\boldsymbol{x}_n,\overline{\Theta})$ where $\overline{\Theta}:=\Theta\cup\{\xi_n\}\cup\{\beta\}$. This lead to a discrimination rule

$$y^*=\text{argmin}_y\left\{-\int\log p(y,\boldsymbol{x}_m\,;\,\overline{\Theta})q(d\overline{\Theta})\right\},\ \boldsymbol{x}_m\in\mathcal{D}_s. \tag{3.5}$$

when applied to the test data $\mathcal{D}_s$.

The decision region $\{\boldsymbol{x}\in\mathcal{X}|Y=y\}$ of MED can have various interpretations depending on the form of the constraint function (3.3) and (3.4). For the anomaly detection constraint (3.4), it is easily seen that the decision region is a *β-level-set region* for the marginal $p(\boldsymbol{x};\overline{\Theta})$, denoted as $\Psi_\beta:=\{\boldsymbol{x}_n\in\mathcal{X}\mid\log p(\boldsymbol{x}_n;\overline{\Theta})\geq\beta\}$. Here Ψ_β is the *rejection region* associated with the test: declare $\boldsymbol{x}_m\in\mathcal{D}_s$ as anomalous whenever $\boldsymbol{x}_m\notin\Psi_\beta$; and declare it as nominal if $\boldsymbol{x}_m\in\Psi_\beta$. With a properly-constructed decision region, the MED model, as a projection of prior distribution $p_0(\Theta)$ into this region, can provide performance guarantees in terms of the error rate or the false alarm rate and can result in improved accuracy [Jebara, 2011, Zhu et al., 2011].

Similar to the SVM, the MED model readily handles nonparametric classifiers. For example, the discriminant function $\mathcal{F}_C\,(p,(y,\boldsymbol{x});\Theta)$ can take the form $y[\Theta(\boldsymbol{x})]$ where $\Theta=f$ is a random function, and $f:\mathcal{X}\to\mathcal{Y}$ can be specified by a Gaussian process with Gaussian covariance kernel $K(\cdot,\cdot)$. More specifically, $f\in\mathcal{H}$, where $\mathcal{H}$ is a Reproducing Kernel Hilbert Space (RKHS) associated with kernel function $K:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$. See [Jebara, 2011] for more detailed discussion.

MED utilizes a weighted ensemble strategy that can improve the classifier stability [Jaakkola et al., 1999]. However, like SVM, MED is sensitive to anomalies in the training set.

3.2.2 Robustified MED with Anomaly Detection Oracle

Assume an *oracle* exists that identifies anomalies in the training set. Using this oracle, partition the training set as $\mathcal{D}_t=\mathcal{D}_t^{nom}\cup\mathcal{D}_t^{anm}$, where $(\boldsymbol{x}_n,y_n)\sim\mathcal{P}_{nom}$ if $(\boldsymbol{x}_n,y_n)\in\mathcal{D}_t^{nom}$ and $(\boldsymbol{x}_n,y_n)\not\sim\mathcal{P}_{nom}$, if $(\boldsymbol{x}_n,y_n)\in\mathcal{D}_t^{anm}$. Given the oracle, one can achieve robust classification simply by constructing a classifier and an anomaly detector simultaneously

on $\mathcal{D}_t^{nom}$. This results in a naive implementation of robustified MED as

$$\min_{q(\overline{\Theta})\in\Delta_{\overline{\Theta}}} \quad \mathbb{KL}\left(q(\overline{\Theta}) \parallel p_0(\overline{\Theta})\right) \tag{3.6}$$

$$\text{s.t.} \int \mathcal{L}_C\left(p,(y_n,\boldsymbol{x}_n);\overline{\Theta}\right) q(d\overline{\Theta}) \leq 0,\ (\boldsymbol{x}_n,y_n)\in\mathcal{D}_t^{nom}, \tag{3.7}$$

$$\int \mathcal{L}_D\left(p,\boldsymbol{x}_n;\overline{\Theta}\right) q(d\overline{\Theta}) \leq 0,\ (\boldsymbol{x}_n,y_n)\in\mathcal{D}_t^{nom}, \tag{3.8}$$

where $\overline{\Theta}=\Theta\cup\{\beta\}\cup\{\xi_n\}_{n\in T}$, the large-margin error function $\mathcal{L}_C$ is defined in (3.3) and the test function $\mathcal{L}_D$ is defined in (3.4). The prior is defined as $p_0(\overline{\Theta})=p_0(\Theta)p_0(\beta)\prod_{n\in T}p_0(\xi_n)$.

Of course, the oracle partition $\mathcal{D}_t = \mathcal{D}_t^{nom}\cup\mathcal{D}_t^{anm}$ is not available *a priori*. The parametric estimator $\widehat{\Psi}_\beta$ of Ψ_β can be introduced in place of $\mathcal{D}_t^{nom}$ in (3.6). However, the estimator $\widehat{\Psi}_\beta$ is difficult to implement and can be severely biased if there is model mismatch.

Below, we propose an alternative nonparametric estimate of the decision region Ψ_β that learns the oracle partition.

3.3 The GEM-MED: Model Formulation

3.3.1 Anomaly Detection using Minimal-entropy Set

As an alternative to a parametric estimator of the level-set $\Psi_\beta := \{\boldsymbol{x}_m\in\mathcal{X} \mid \log p(\boldsymbol{x}_m;\overline{\Theta})\geq\beta\}$, we propose to use a non-parametric estimator [Wasserman, 2010] based on the *minimal-entropy (ME) set* $\Omega_{1-\beta}$. The ME set $\Omega_{1-\beta} := \arg\min_A\{H(A)|\int_A p(\mathbf{x})d\mathbf{x}\geq\beta\}$ is referred as the *minimal-entropy-set of false alarm level* $1-\beta$, where $H(A)=-\int_A \log p(\mathbf{x})\,p(\mathbf{x})d\mathbf{x}$ is the Shannon entropy of the density $p(\mathbf{x})$ over the region A. This minimal-entropy-set is equivalent to the *epigraph-set* $\{A:\ \int_A p(\mathbf{x})d\mathbf{x}\geq\beta\}$ as illustrated in Fig. 3.2.

Given $\Omega_{1-\beta}$, the ME anomaly test is as follows: a sample $\mathbf{x}_n$ is declared anomalous if $\mathbf{x}_n\notin\Omega_{1-\beta}$; and it is declared nominal, when $\mathbf{x}_n\in\Omega_{1-\beta}$. It is established in [Hero, 2006] that when $p(x)$ is a known density, this test is a Uniformly Most Powerful Test (UMPT) [Scharf, 1991] at level β of the hypothesis $H_0: x\sim p(x)$ vs. $H_1: x\sim p(x)+\epsilon U(x)$, where $U(x)$ is the uniform density and $\epsilon\in[0,1]$ is an unknown mixture coefficient.

3.3.2 The BP-kNNG Implementation of GEM

Several methods have been proposed to empirically approximate the ME set $\Omega_{1-\beta}$ including: kernel density estimation [Scott and Nowak, 2006]; the k-point minimal spanning tree

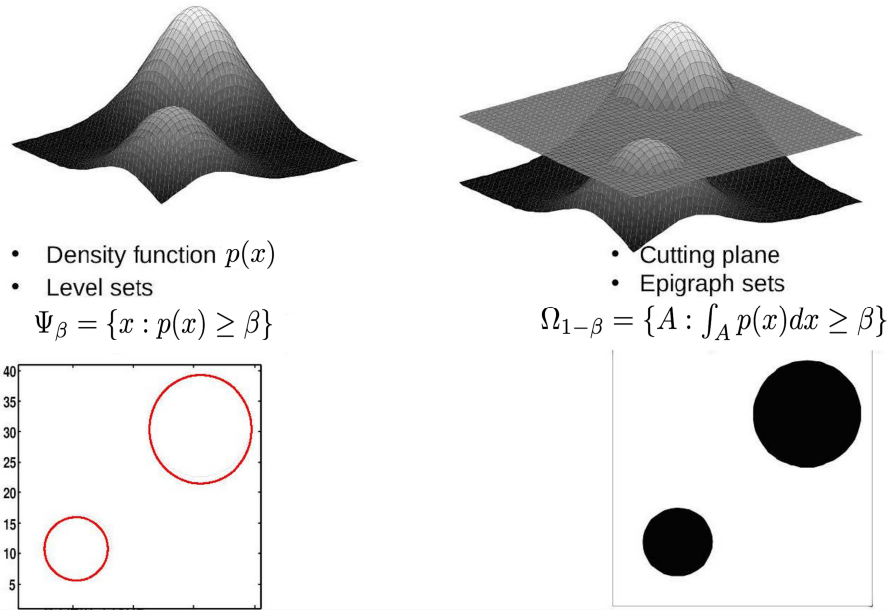


Figure 3.2: The comparison of level-set (left) and the epigraph-set (right) w.r.t. two continuous density function $p(x)$. The minimum-entropy-set is computed based on the epigraph-set.

[Hero and Michel, 1999]; the leave-one-out k-nearest-neighbor graph [Sricharan and Hero, 2011]; and the average k-nearest-neighbor distance [Root et al., 2015]. In [Sricharan and Hero, 2011], the Bipartite k-Nearest-Neighbor (BP-kNN) based algorithm was proposed as an alternative approximation. The BP-kNN solves the following discrete optimization problem:

$$A_c^* \in \arg \min_{A_c \subset \mathcal{D}_t^{N,c}} L(A_c, \mathcal{D}_t^{M,c}),$$
$$\text{where } L(A_c, \mathcal{D}_t^{M,c}) := \sum_{\boldsymbol{x}_n \in A_c} d_k(\boldsymbol{x}_n, \mathcal{D}_t^{M,c}),$$

and where A_c is a set of distinct $K = |T|(1-\beta)$ points in $\mathcal{D}_t^{N,c}$ (see Fig. 3.3 for illustration). It is shown in [Sricharan and Hero, 2011] that $A_c^* = \widehat{\Omega}_{1-\beta}$ is an asymptotically consistent estimator of the ME set. Equivalently, let $\eta_n \in \{0,1\}$ be the indicator function of the event $\mathbf{x}_n \in A_c$ and define $d_n := d_k(\boldsymbol{x}_n, \mathcal{D}_t^{M,c})$. Then it can easily be shown that the algorithm in [Sricharan and Hero, 2011] finds the optimal binary variables $\left\{\eta_n \in \{0,1\} \,|\boldsymbol{x}_n \in \mathcal{D}_t^{N,c}\right\}, n = 1, \ldots, N$, that minimize

$$\sum_{\boldsymbol{x}_n \in \mathcal{D}_t^{N,c}} \eta_n d_n \quad \text{subject to} \quad \sum_{\boldsymbol{x}_n \in \mathcal{D}_t^{N,c}} \eta_n \geq K. \tag{3.9}$$

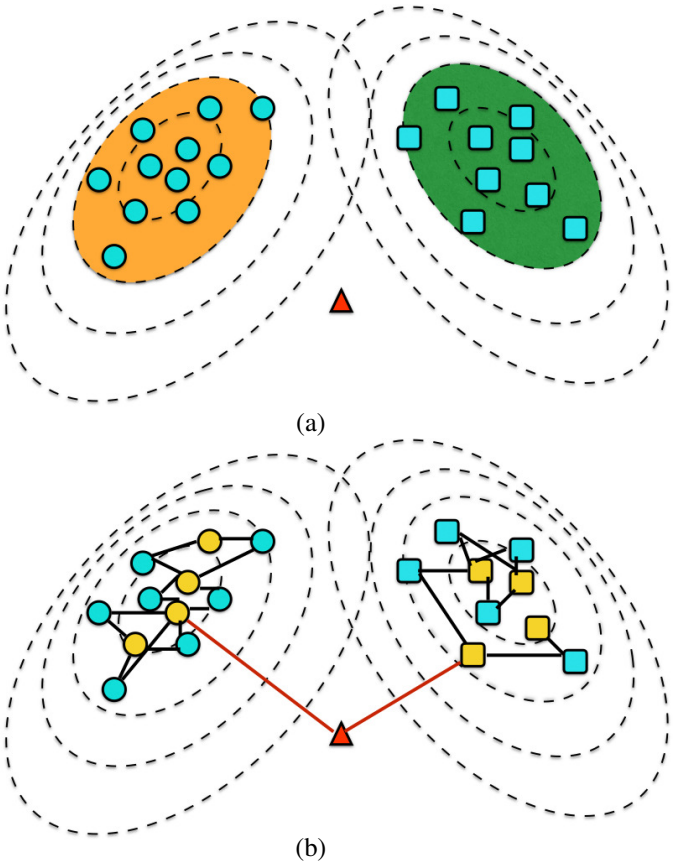


Figure 3.3: Figure (a) illustrates ellipsoidal minimum entropy (ME) sets for two dimensional Gaussian features in the training set for class 1 (orange region) and class 2 (green region). These ME sets have coverage probabilities $1-\beta$ under each class distribution and correspond to the regions of maximal concentration of the densities. The blue disks and blue squares inside these regions correspond to the nominal training samples under class 1 and class 2, respectively. An outlier (in red triangle) falls outside of both of these regions. Figure (b) illustrates the bipartite 2-NN graph approach to identify the anomalous point, where the yellow disks and squares are reference samples in each class that are randomly selected from the training set. Note that the average 2-NN distance for anomalies should be significantly larger than that for the nominal samples.

This representation makes the BP-kNN implementation of GEM naturally adaptable to our GEM-MED framework. Specifically, the binary weights $\eta_n \in \{0,1\}$ are relaxed to continuous weights in the unit interval $[0,1]$ for all $n \in T$. After relaxation, the constraint in (3.9) becomes $\sum_n \eta_n / |T| \geq \hat{\beta}$, where $\hat{\beta} = K/|T| = (1-\beta) > 0$ is set so that the optimal solution $\{\eta_n | \boldsymbol{x}_n \in A_c^*\}$ is feasible and the all-zero solution is infeasible. With the set of weights $\{\eta_n\}_{n\in T}$, the GEM problem in (3.9) can be transformed into a set of nonparametric constraints that fit the framework (3.6). This is discussed below.

3.3.3 The GEM-MED as Non-parametric Robustified MED

Now we can implement the framework in (3.6). Denote $\overline{\Theta} := \Theta \cup \left\{\hat{\beta}\right\} \cup \{\xi_n\}_{n\in T} \cup \{\eta_n\}_{n\in T} \cup \{\gamma_z\}_{z\in\{\pm 1\}}$, where $\Theta, \{\xi_n\}_{n\in T}$ are parameters as defined in (3.6), $\{\eta_n\}_{n\in T}$ are weights in Sec. 3.3.2 and $\hat{\beta}, \{\gamma_z\}_{z\in\{\pm 1\}}$ are variables to be defined later.

According to the objective function in (3.9), we specify the test function $\tilde{\mathcal{L}}_D$ as

$$\begin{aligned}&\tilde{\mathcal{L}}_D(\overline{\Theta}, \boldsymbol{y}; z, \boldsymbol{d}) := \tilde{\mathcal{L}}_D(\{\eta_n\}, \{\gamma_z\}, \boldsymbol{y}; z, \boldsymbol{d})\\ &= \left(\sum_n \mathbb{1}\left\{y_n = z\right\} \eta_n d_n / |T| - \gamma_z\right), \quad z \in \{\pm 1\},\end{aligned}$$

where $\gamma_z \geq 0, z \in \{\pm 1\}$ is the threshold associated with d_n on $\mathcal{D}_t \cap \{\boldsymbol{x}_n | y_n = z\}$. Compared with (3.9), if $\gamma_z = L_z^* + \epsilon$, where L_z^* is the optimal value in (3.9) and $\epsilon > 0$ is small enough, then for $\{\eta_n\}_{n\in T}$ satisfying $\tilde{\mathcal{L}}_D \leq 0$, the region $\left\{\boldsymbol{x}_n : \eta_n > \frac{1}{2}\right\}$ is concentrated on $\widehat{\Omega}_{1-\beta} \cap \{\boldsymbol{x}_n | y_n = z\}, z \in \{\pm 1\}$.

As discussed in 3.3.2, the constraint in (3.9) becomes the inequality constraint $\sum_{n|y_n=z} \eta_n / |T| \geq \hat{\beta}$.

Assuming that $\overline{\Theta}$ is random with unknown distribution $q(\overline{\Theta})$, the above expected constraints becomes

$$\int \tilde{\mathcal{L}}_D(\overline{\Theta}, \boldsymbol{y}; z, \boldsymbol{d}) q(d\overline{\Theta}) \leq 0, z \in \{\pm 1\}, \tag{3.10}$$

$$\int \left[\sum_{n:y_n=z} \eta_n / |T|\right] q(d\overline{\Theta}) \geq \hat{\beta}, z \in \{\pm 1\}. \tag{3.11}$$

The constraint (3.10) is referred as *the entropy constraint* and constraint (3.11) is the *epigraph constraint*. As discussed above, the region $\left\{\boldsymbol{x}_n | \eta_n > \frac{1}{2}\right\}$ for $q(\overline{\Theta})$ satisfying (3.10) and (3.11) is concentrated on $\widehat{\Omega}_{1-\beta} \cap \{\boldsymbol{x}_n | y_n = z\}$ in each class $z \in \{\pm 1\}$ on average. With $\tilde{\mathcal{L}}_D$, the test constraint in (3.6) is replaced by (3.10) and (3.11).

For the classification part in (3.6), given η_n associated with each sample, the error constraints in (3.6) is replaced by *reweighted* error constraints

$$\int \left[\eta_n \mathcal{L}_C\left(p, (y_n, \boldsymbol{x}_n); \overline{\Theta}\right)\right] q(d\overline{\Theta}) \leq 0, \ n \in T,$$

with $\mathcal{L}_C$ defined as in (3.3). Note that these constraints are applied to the entire training set. Summarizing, we have the following:

Definition The *Geometric-Entropy-Minimization Maximum-Entropy-Discrimination (GEM-MED)*

method solves

$$\begin{aligned}\min_{q(\overline{\Theta})\in\Delta_{\overline{\Theta}}} \quad & \mathbb{KL}\left(q(\overline{\Theta}) \parallel p_0(\overline{\Theta})\right) \\ \text{s.t.} \quad & \int \left[\eta_n \mathcal{L}_C\left(p,(y_n,\boldsymbol{x}_n);\overline{\Theta}\right)\right] q(d\overline{\Theta}) \leq 0,\ n\in T, \\ & \int \tilde{\mathcal{L}}_D(\overline{\Theta},\boldsymbol{y};z,\boldsymbol{d}) q(d\overline{\Theta}) \leq 0,\ z\in\{\pm 1\}, \\ & \int \left[\sum_{n:y_n=z} \eta_n / |T|\right] q(d\overline{\Theta}) \geq \hat{\beta},\ z\in\{\pm 1\}\end{aligned} \tag{3.12}$$

where $\overline{\Theta}$, $\mathcal{L}_C$ and $\tilde{\mathcal{L}}_D$ are defined as before.

3.4 Implementation

3.4.1 Projected Stochastic Gradient Descent Algorithm

Note that (3.12) is a convex optimization w.r.t. the unknown distribution $q(\overline{\Theta})$ [Jaakkola et al., 1999, Cover and Thomas, 2012]. Therefore, it can be solved using the Karush-Kuhn-Tucker (KKT) conditions, which will result in a unique solution. We make the following simplifying assumptions under which our a computational algorithm is derived to solve (3.12).

1. Assume that a kernelized SVM is used for the classifier discriminant $\mathcal{F}_C$ function. Following [Zhu et al., 2014, Jebara, 2011], we assume that the decision function f follows a Gaussian random process on $\mathcal{X}$, i.e., a positive definite covariance kernel $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ is defined for all $\boldsymbol{x}_i,\boldsymbol{x}_j\in\mathcal{X}$ and all finite dimensional distributions, i.e., distributions of samples $(f(\boldsymbol{x}_i))_{i\in T}$, follow the multivariate normal distribution

$$(f(\boldsymbol{x}_i))_{i\in T} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{K}), \tag{3.13}$$

   where $\boldsymbol{K}=[K(\boldsymbol{x}_i,\boldsymbol{x}_j)]_{i,j\in T}$ is a specified covariance matrix. For example, $K(\boldsymbol{x}_i,\boldsymbol{x}_j) := \exp(-\gamma\|\mathbf{x}_i-\mathbf{x}_j\|_2^2)$ for Gaussian RBF kernel covariance function.

2. Assume a separable prior, as commonly used in Bayesian inference [Jaakkola et al., 1999, Blei et al., 2003, Zhu et al., 2014]

$$p_0(\overline{\Theta}) = p_0(\Theta)\prod_{n\in T} p_0(\xi_n)\prod_{n\in T} p_0(\eta_n)\prod_{z\in\{\pm 1\}} p_0(\gamma_z). \tag{3.14}$$

3. Assume that the hyperparameters $\{\xi_n\}$ are exponential random variables and the indicator variables $\{\eta_n\}$ are independent Bernoulli random variables,

$$\begin{aligned} p_0(\xi_n) &\propto \exp(-c_\xi(1-\xi_n)),\ \xi_n \in (-\infty, 1],\ n \in T; \\ p_0(\eta_n) &= Ber(p_\eta) \\ &\text{with } p_\eta = \frac{1}{1+\exp(-(a_\eta - \eta_n))} \\ &\qquad := \sigma(a_\eta - \eta_n),\ \eta_n \in \{0,1\},\ n \in T; \\ p_0(\gamma_z) &= \delta_{\hat{\gamma}_z}(\gamma_z);\ z \in \{\pm 1\}, \end{aligned} \tag{3.15}$$

where (a_η, c_ξ) are parameters and $\hat{\gamma}_z$ is the upper bound estimate for minimal-entropy in each class $z = \pm 1$ given by GEM algorithm. $\sigma(x) = 1/(1+\exp(-x))$ is the sigmoid function.

Now by solving the primal version of optimization problem (3.12), we have

Theorem 3.4.1 *The GEM-MED problem in* (3.12) *is convex with respect to the unknown distribution* $q(\overline{\Theta})$ *and the unique optimal solution is a* generalized Gibbs distribution *with the density:*

$$q(d\overline{\Theta}) = \frac{1}{Z(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})} p_0(d\overline{\Theta}) \exp\left(-E(\overline{\Theta}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})\right), \tag{3.16}$$

where

$$\begin{aligned} E(\overline{\Theta}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa}) &:= E(\Theta, \hat{\beta}, \{\xi_n\}, \{\eta_n\}, \{\gamma_z\}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa}) \\ &= \sum_{n \in T} \lambda_n \eta_n \mathcal{L}_{C,\Theta,\xi_n} - \sum_{z \in \{\pm 1\}} \mu_z \tilde{\mathcal{L}}_{D,z} \\ &\quad - \sum_{z \in \{\pm 1\}} \kappa_z \sum_{n: y_n = z} \eta_n / |T| + \sum_{z \in \{\pm 1\}} \kappa_z \hat{\beta} \end{aligned}$$

with $\overline{\Theta} = \Theta \cup \left\{\hat{\beta}\right\} \cup \{\xi_n\}_{n \in T} \cup \{\eta_n\}_{n \in T} \cup \{\gamma_{+1}, \gamma_{-1}\}$*and where the dual variables* $\boldsymbol{\lambda} = \{\lambda_n, n \in T\}$, $\boldsymbol{\mu} = (\mu_z, z \in \pm 1)$ *and* $\boldsymbol{\kappa} = (\kappa_z, z \in \pm 1)$ *are all nonnegative.* $Z(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})$ *is* the partition function*, which is given as*

$$Z(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa}) = \int \exp\left(-E(\overline{\Theta}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})\right) p_0(d\overline{\Theta}). \tag{3.17}$$

The factor $\mathcal{L}_{C,\Theta,\xi_n} := \mathcal{L}_C(\cdot; \Theta, \xi_n)$ *is defined as in* (3.3)*,* $\tilde{\mathcal{L}}_{D,z} := \tilde{\mathcal{L}}_D(\cdot; z, \cdot)$ *is defined as in* (3.10)*. See the Appendix Sec. 3.7.1 for a detailed derivation.*

Algorithm 1 The (kernel) GEM-MED algorithm

Require: The training set $\mathcal{D}_t \subset \mathcal{X} \times \{\pm 1\}$ and the test set $\mathcal{D}_s$. The projection gradient step parameter $\varphi, \psi, \tau > 0$. Prior distribution and assumptions given as (13)-(15). The kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is specified.

1: **Initialize:** Set $\boldsymbol{\mu}_0 = \mathbf{0}, \boldsymbol{\kappa}_0 = \mathbf{0}$. $\boldsymbol{\lambda}_0$ is set by applying conventional MED on $\mathcal{D}$

2: **for** $t = 1, \ldots, T$ or until converge **do**

3: Compute the gradient of log-partition function w.r.t $\boldsymbol{\lambda}_t$, $\boldsymbol{\mu}_t$ and $\boldsymbol{\kappa}_t$, respectively, i.e. $\frac{\partial -\log Z(\boldsymbol{\lambda}_t, \boldsymbol{\mu}_t, \boldsymbol{\kappa}_t)}{\partial \lambda_n}$, $\frac{\partial -\log Z(\boldsymbol{\lambda}_t, \boldsymbol{\mu}_t, \boldsymbol{\kappa}_t)}{\partial \mu_z}$ and $\frac{\partial -\log Z(\boldsymbol{\lambda}_t, \boldsymbol{\mu}_t, \boldsymbol{\kappa}_t)}{\partial \kappa_z}$ according to the formula (23)-(25) where the expectation is approximated via Gibbs sampling described as above.

4: Update λ_n, μ_z and κ_z via projected gradient descent, i.e.

$$\lambda_{n,(t+1)} = \text{proj}_{\{\lambda:\, 0 \le \lambda \le C_1\}} \left\{ \lambda_{n,t} - \varphi \frac{\partial \log Z((\boldsymbol{\mu}_t, \boldsymbol{\lambda}_t, \boldsymbol{\kappa}_t))}{\partial \lambda_n} \right\} \quad n \in T,$$

$$\mu_{z,(t+1)} = \text{proj}_{\{\mu:\, \mu \ge 0\}} \left\{ \mu_{z,t} - \psi \frac{\partial \log Z(\boldsymbol{\mu}_t, \boldsymbol{\lambda}_t, \boldsymbol{\kappa}_t)}{\partial \mu_z} \right\} \quad z \in \{-1, +1\},$$

$$\kappa_{z,(t+1)} = \text{proj}_{\{\kappa:\, \kappa \ge 0\}} \left\{ \kappa_{z,t} - \tau \frac{\partial \log Z(\boldsymbol{\mu}_t, \boldsymbol{\lambda}_t, \boldsymbol{\kappa}_t)}{\partial \kappa_z} \right\} \quad z \in \{-1, +1\},$$

where $\text{proj}_{\{x:\, 0 \le x \le C\}}\{w\} \equiv \min\left(\max(x, 0), C\right)$ defines the projection of x on the feasible set $\{z :\, 0 \le z \le C\}$.

5: **end for**

Ensure: Assign label for test sample $\mathbf{x}_m \in \mathcal{D}_s$ as

$$y^* = \text{sign}\left\{ \sum_{n \in T} \hat{\eta}_n \lambda_n^* y_n K(\boldsymbol{x}_m, \boldsymbol{x}_n) \right\}, \quad \boldsymbol{x}_m \in \mathcal{D}_s$$

where $\hat{\eta}_n = \mathbb{E}\left[\eta_n | f\right]$ at the final iteration of step 4.

Moreover, we specify the error function as

$$\mathcal{L}_C\left(p, (y_n, \boldsymbol{x}_n); \Theta, \xi_n\right) := \xi_n - y_n f(\boldsymbol{x}_n), \tag{3.18}$$

where $\Theta := f : \mathcal{X} \to \mathcal{Y}$ is a decision function associated with a nonparametric classifier as defined in Sec 3.2.1.

Theorem 3.4.2 *Assume that* (3.13)*,* (3.14)*,* (3.15) *hold, the dual optimization problem is given as*

$$\begin{aligned} &\max_{\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa} \ge 0} -\log Z(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa}) \\ &= -\log \int \exp\left(-E(\overline{\Theta}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})\right) p_0(d\overline{\Theta}) \end{aligned} \tag{3.19}$$

$$= \sum_{n\in T} (\lambda_n + \log(1-\lambda_n/c)) - \sum_{z\in\{\pm 1\}} \mu_z \hat{\gamma}_z + \hat{\beta} \sum_{z\in\{\pm 1\}} \kappa_z$$
$$- \log \int \exp\left(\frac{1}{2} Q(\boldsymbol{K} \odot (\boldsymbol{y}\boldsymbol{y}^T), (\boldsymbol{\lambda} \odot \boldsymbol{\eta}))\right)$$
$$\times\, p_0(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^T (-\boldsymbol{\mu} \otimes \boldsymbol{d} + \boldsymbol{\kappa} \otimes \boldsymbol{e})\right) d\boldsymbol{\eta} \tag{3.20}$$

where $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})$ *are nonnegative dual variables as defined in* (3.16)*,* $\boldsymbol{e}$ *is the all* 1*'s vector,* $\odot$ *is Hadamard product,* $\otimes$ *is the Kronecker product, respectively, and*

$$Q(\boldsymbol{K}, \boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{K} \boldsymbol{x}$$

is the quadratic form associated with the kernel K*.*

See Appendix Sec. 3.7.2 for derivations of this result.

It is seen from (3.19) that the dual objective function is concave w.r.t. dual variables $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})$. However, the integral in (3.20) is not closed form, so an explicit form as a quadratic optimization in SVM is not available. Nevertheless, the only coupling in (3.20) comes from the joint distribution $q(f, \boldsymbol{\eta})$. In particular, under the prior assumption (3.13), (3.14), (3.15), the optimal solution (3.16) satisfies

1. $q(\overline{\Theta}) = q(f, \boldsymbol{\eta}) \prod_n q(\xi_n) q(\gamma_{+1}) q(\gamma_{-1})$ is factorized.

2. $q(\boldsymbol{\eta}|f) = \prod_{n\in T} q(\eta_n|f)$, i.e. the $\{\eta_n, n \in T\}$ are conditional independent given the decision boundary function f. Moreover,

$$q(\eta_n|f) = Ber(q_\eta), \quad \text{with } q_\eta = \sigma(\rho_n \mathcal{F}_n(f)) \tag{3.21}$$

   where $\rho_n := \log \frac{1-p_0(\eta_n=1)}{p_0(\eta_n=1)}$, $\mathcal{F}_n(f) := \lambda_n [y_n f(\mathbf{x}_n) - 1] - \mu_{y_n} h_n + \kappa_{y_n} / |T|$, $\sigma(\cdot)$ is the sigmoid function as (3.15).

3. $f|\boldsymbol{\eta} \sim \mathcal{N}(f|\hat{f}_{\boldsymbol{\eta},\boldsymbol{\lambda}}(\cdot), \boldsymbol{K})$, where

$$\hat{f}_{\boldsymbol{\eta},\boldsymbol{\lambda}}(\cdot) = \sum_{n\in T} \lambda_n \eta_n y_n K(\cdot, \boldsymbol{x}_n) \in \mathcal{H} \tag{3.22}$$

See Appendix Sec. 3.7.3 for details.

Given above results, we propose to use the *projected stochastic gradient descent (PSGD)* [Bertsekas, 1999, Murphy, 2012] algorithm to solve the dual optimization problem in

(3.20). The gradient vectors of the dual objective function in (3.20) w.r.t. $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, $\boldsymbol{\kappa}$, respectively, are computed as

$$\begin{aligned}
&\frac{\partial}{\partial \lambda_n}\left[-\log Z(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa})\right] \\
&= 1 - \mathbb{E}_{q(f,\boldsymbol{\eta})}\left[\eta_n y_n f(\mathbf{x}_n)\right] + \frac{c}{c-\lambda_n},\ n \in T; \qquad (3.23)\\
&\frac{\partial}{\partial \mu_z}\left[-\log Z(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa})\right] \\
&= \mathbb{E}_{q(f,\boldsymbol{\eta})}\left\{\sum_{n:y_n=z}\eta_n d_n\right\} - \hat{\gamma}_z,\ \ z \in \{\pm 1\}; \qquad (3.24)\\
&\frac{\partial}{\partial \kappa_z}\left[-\log Z(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa})\right] \\
&= \hat{\beta} - \frac{1}{|T|}\mathbb{E}_{q(f,\boldsymbol{\eta})}\left[\sum_{n:y_n=z}\eta_n\right],\ \ z \in \{\pm 1\}. \qquad (3.25)
\end{aligned}$$

Note that the expectation w.r.t. $q(f,\boldsymbol{\eta})$ are approximated by Gibbs sampling with each conditional distribution given by (3.21), (3.22). For a detailed implementation of the Gibbs sampler, see the Appendix Sec. 3.7.4.

A complete description of algorithm is presented in **Algorithm 1**. It is remarked that in (3.21) the probability of $\{\eta_n = 0\}$ is proportional to the sum of margin of classification and negative local entropy value. The role of the dual variables (η_n, μ_c) in (3.21) and (3.22) is to balance the classification margin $y\,f(\cdot)$ and local entropy h in determining the anomalies.

3.4.2 Prediction and Detection on Test Samples

The GEM-MED classifier is similar to the standard MED classifier in (3.5):

$$\begin{aligned}
y^* &= \text{argmax}_y\left\{\int y f(\boldsymbol{x}_m) q(f|\hat{\boldsymbol{\eta}},\mathcal{D}_t) df\right\}, \\
&= \text{sign}\left\{\sum_{n\in T}\hat{\eta}_n\lambda_n^* y_n K(\boldsymbol{x}_m,\boldsymbol{x}_n)\right\} \quad \boldsymbol{x}_m \in \mathcal{D}_s. \qquad (3.26)
\end{aligned}$$

where $\hat{\boldsymbol{\eta}}$ is the conditional mean estimator of $\boldsymbol{\eta}$ given by Algorithm 1.

The GEM-MED was optimized on the training set to detect and mitigate anomaly corrupted training samples. When there are also anomalies in the test sample, an anomaly detection method can be applied independently to screen out these samples (at a given false positive rate) before applying GEM-MED to classify them. Such a two-stage approach to handling anomalies in the test sample is obviously not optimal. An optimal joint approach

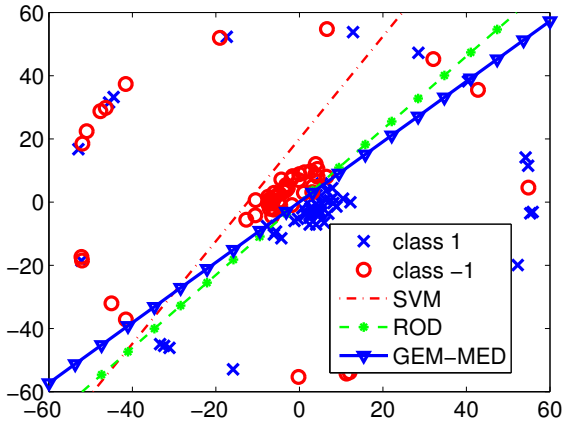


Figure 3.4: The classification decision boundary for SVM, ROD and GEM-MED on the simulated data set with two bivariate Gaussian distribution $\mathcal{N}(\boldsymbol{m}_{+1}, \Sigma)$, $\mathcal{N}(\boldsymbol{m}_{-1}, \Sigma)$ in the center and a set of anomalous samples for both classes distributed in a ring. Note that SVM is biased toward the anomalies (within outer ring support) and ROD and GEM-MED are insensitive to the anomalies.

to handling anomalies in the training and test samples is worthwhile future direction which will not be investigated here.

3.5 Experiments

We illustrate the performance of the proposed GEM-MED algorithm on simulated data as well as on a real data collected in a field experiment. We compare the proposed GEM-MED with the SVM implemented by *LibSVM* [Chang and Lin, 2011] and the ROD algorithm implemented with code obtained from the authors of [Xu et al., 2006]. For the simulated data experiment, a linear kernel SVM is implemented, and for the real data, a Gaussian RBF kernel SVM with kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|_2^2)$ is implemented and the kernel parameter $\gamma > 0$ is tuned via 5-fold-cross validation.

3.5.1 Simulated Experiment

For each class $c \in \{\pm 1\}$, we generate samples from the bivariate Gaussian distribution $\mathcal{N}(\boldsymbol{m}_{+1}, \Sigma)$ and $\mathcal{N}(\boldsymbol{m}_{-1}, \Sigma)$, with mean $\boldsymbol{m}_{-1} = (3, 3)$ and $\boldsymbol{m}_{+1} = -\boldsymbol{m}_{-1}$ and common covariance $\Sigma = \begin{bmatrix} 20 & 16 \\ 16 & 20 \end{bmatrix}$. The sample follows the log-linear model $\log p(y, \boldsymbol{x}; \overline{\Theta}) \propto 1/2\, y(\mathbf{w}^T \mathbf{x} + b)$ where $\overline{\Theta} = (\mathbf{w}, b)$. A Gaussian prior was used as $p_0(\overline{\Theta}) = \mathcal{N}(\mathbf{w};\, \mathbf{0}, \sigma_w^2 \mathbf{I}) \mathcal{N}(b; 0, \sigma_b^2)$.

We followed the same models as in [Xu et al., 2006]. In particular, the anomalies in the training set were drawn uniformly from a ring with an inner radius of R and outer radius $R + 1$, where R was assigned as one of the values $[15, 35, 55, 75]$. Define R to be

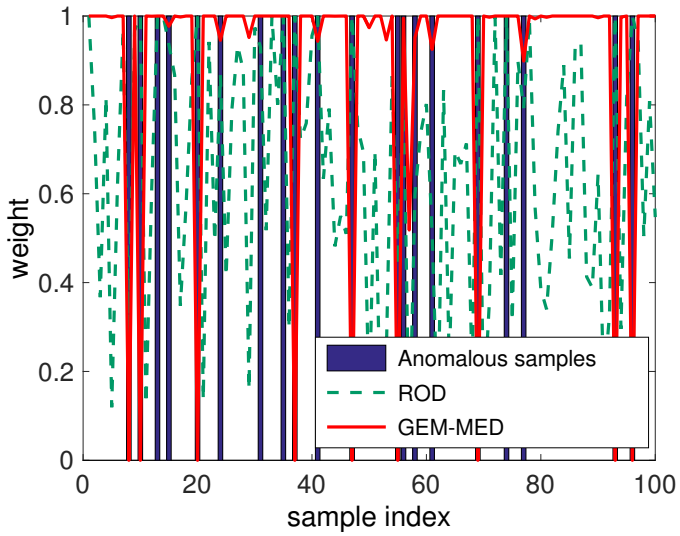


Figure 3.5: The Illustration of anomaly score $\hat{\eta}_n$ for GEM-MED and ROD. The GEM-MED is more accurate than ROD in term of anomaly detection.

the *noise level* of the data set, since the larger R the higher the discrepancy between the nominal distribution and the anomalous distribution. The samples then were labeled as $\{0, 1\}$ with equal probability. The size of the training set was 100 for each class, and the ratio of anomaly samples was r_a. The test set contained 2000 uncorrupted samples from each class. See Fig. 3.4 for a realization of the data set and the classifiers.

We first compare the classification accuracy of SVM, Robust-Outlier-Detection (ROD) with outlier parameter ρ and GEM-MED, under noise level R and a range of corruption rates $r_a \in \{0.2, 0.3, 0.4, 0.5\}$. We used the BP-kNNG implementation of GEM, where the k-nearest neighbor parameter $k = 5$. In the update of the GEM-MED dual variables $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})$, the learning rate (φ, ψ, τ) is chosen based on a comparison of classification performance of the GEM-MED under a range of noise levels R and corruption rates r_a, as shown in Fig. 3.7 (a)-(c). Note that when $\varphi \in [1, 4] \times 10^{-3}, \psi \in [1, 4] \times 10^{-2}, \tau \in [1, 5] \times 10^{-2}$, the performance of the GEM-MED is stable in terms of the averaged missclassification error and the variance. We fix (φ, ψ, τ) in the stable range in the following experiments. For the ROD, we investigated a range of algorithm parameters, in particular outlier parameter $\rho \in \{0.02, 0.2, 0.6\}$ for comparison, and we observed that the value $\rho = 0.02$ gives the best classification performance regardless of the setting of $R \in \{15, 35, 55, 75\}$ or $r_a \in \{0.2, 0.3, 0.4, 0.5\}$. Recall that the ROD parameter ρ is a fixed threshold that determines the proportion of anomalies, i.e., the proportion of nonzero η_n [Xu et al., 2006]. Compared to the ROD, the GEM-MED as a Bayesian method requires no tuning parameter to control the proportion of anomalies. In the experiments below, we compare the ROD for a range of outlier parameters ρ with GEM-MED for a single choice of (φ, ψ, τ), which

were tuned via 5-fold-cross-validation of misclassification rate over 50 trial runs.

We first compare the classification accuracy of SVM, Robust-Outlier-Detection (ROD) with outlier parameter ρ and GEM-MED, under noise level R and a range of corruption rates $r_a \in \{0.2, 0.3, 0.4, 0.5\}$. We used the BP-kNNG implementation of GEM, where the k-nearest neighbor parameter $k = 5$. In the update of the GEM-MED dual variables $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})$, the learning rate (φ, ψ, τ) is chosen based on a comparison of classification performance of the GEM-MED under a range of noise levels R and corruption rates r_a, as shown in Fig. 3.7 (a)-(c). Note that when $\varphi \in [1, 4] \times 10^{-3}, \psi \in [1, 4] \times 10^{-2}, \tau \in [1, 5] \times 10^{-2}$, the performance of the GEM-MED is stable in terms of the averaged missclassification error and the variance. We fix (φ, ψ, τ) in the stable range in the following experiments. For the ROD, we investigated a range of algorithm parameters, in particular outlier parameter $\rho \in \{0.02, 0.2, 0.6\}$ for comparison, and we observed that the value $\rho = 0.02$ gives the best classification performance regardless of the setting of $R \in \{15, 35, 55, 75\}$ or $r_a \in \{0.2, 0.3, 0.4, 0.5\}$. Recall that the ROD parameter ρ is a fixed threshold that determines the proportion of anomalies, i.e., the proportion of nonzero η_n [Xu et al., 2006]. Compared to the ROD, the GEM-MED as a Bayesian method requires no tuning parameter to control the proportion of anomalies. In the experiments below, we compare the ROD for a range of outlier parameters ρ with GEM-MED for a single choice of (φ, ψ, τ), which were tuned via 5-fold-cross-validation of misclassification rate over 50 trial runs.

Fig. 3.6(a) shows the miss-classification error $(\%)$ versus various noise level R (with $r_a = 0.2$), and Fig. 3.6(b) shows the miss-classification error under different corruption rate settings (with $R = 55$). In both experiments, GEM-MED outperforms ROD and SVM in terms of classification accuracy. Note that when the noise level or the corruption rate increases, the training data become less representative of the test data and the difference between their distributions increases. This causes a significant performance deterioration for the SVM/MED method, which is demonstrated in Fig. 3.6(a) and Fig. 3.6(b). Fig. 3.5 also shows the bias of the SVM classifier towards the anomalies that lie in the ring. Comparing to GEM-MED and ROD in Fig 3.6(a) and Fig. 3.6(b), the former method is less sensitive to the anomalies. Moreover, since the GEM-MED model takes into account the marginal distribution for the training sample, it is more adaptive to anomalies in the training set, as compared to ROD, which does not use any prior knowledge about the nominal distribution but only relies on the predefined outlier parameter ρ to limit the training loss.

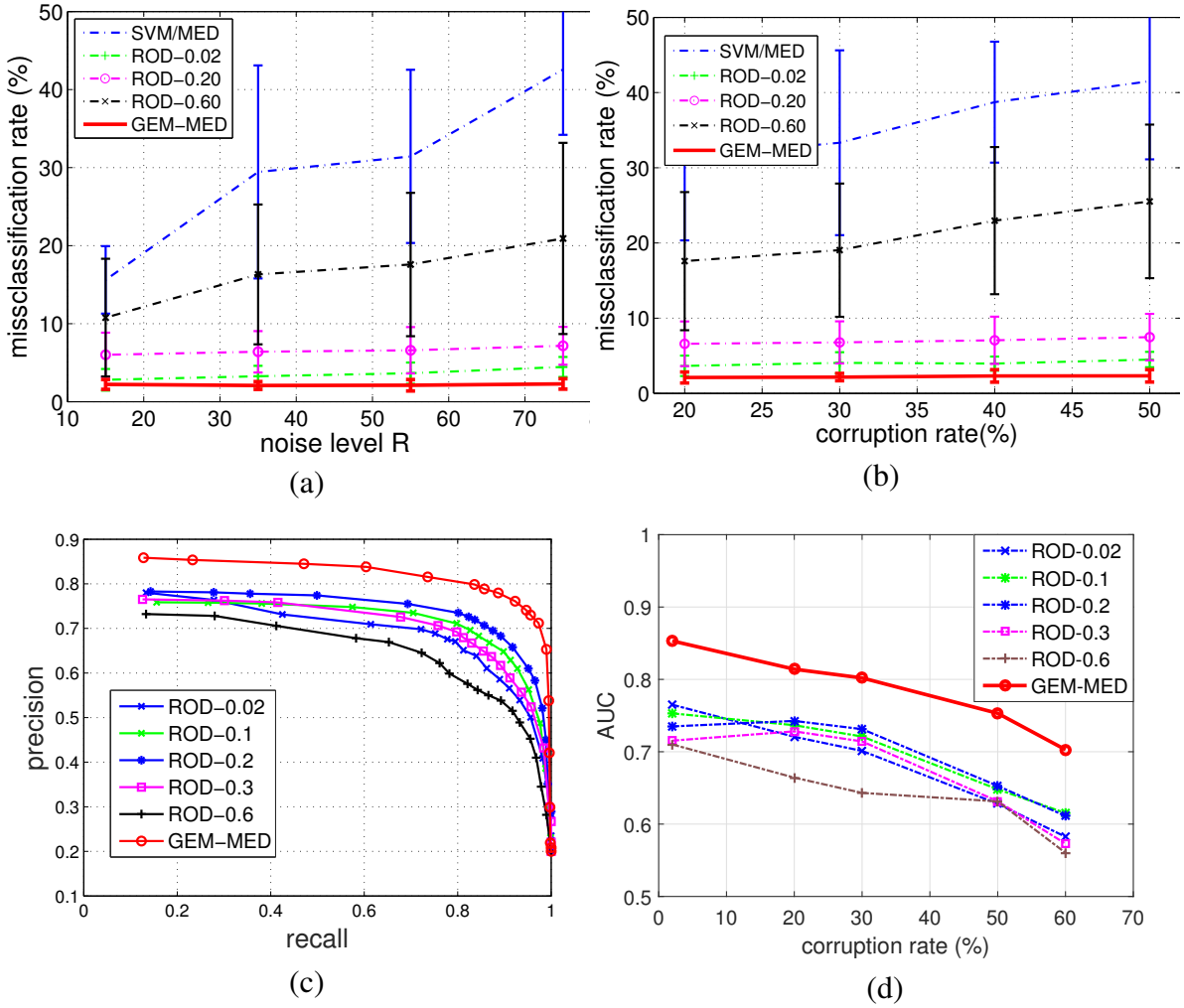


Figure 3.6: (a) Miss-classification error (%) vs. noise level R for corruption rate $r_a = 0.2$. (b) Miss-classification error (%) vs. corruption rate $\mathbb{E}[\eta]$ for ring-structureed anomaly distribution having ring $R = 55$. (c) Recall-precision curve for GEM-MED and RODs on simulated data for corruption rate $= 0.2$. (d) The AUC vs. corruption rate r_a for GEM-MED and ROD with a range of outlier parameters ρ. From (a) and (b), GEM-MED outperforms both SVM/MED and ROD for various ρ in classification accuracy. From (c), under the same corruption rate, we see that GEM-MED outperforms ROD in terms of the precision-recall behavior. This due to the superiority of GEM constraints in enforcing anomaly penalties into the classifier. From (d), The GEM-MED outperforms RODs in terms of AUC for the range of investigated corruption rates.

In Fig. 3.6(c) we compare the performance of GEM-MED and ROD in terms of precision vs recall for the same corruption rate as in Fig. 3.6(a) and 3.6(b). In ROD and GEM-MED, the estimated weights $\eta_n \in [0, 1]$ for each sample can be used to infer the likelihood of anomalies. In particular, in GEM-MED the corresponding latent variable estimate $\hat{\eta}_n$ is obtained at the final iteration of the Gibbs sampling procedure, as described in Appendix Sec. 3.7.4. Following the anomaly ranking procedure in [Xu et al., 2006], these anomaly scores are placed in ascending order. We compute the precision and recall using this ordering by averaging over 50 runs. Precision and recall are measures that are commonly used in data mining [Japkowicz and Shah, 2011]:

$$\begin{aligned} \text{Precision} &= \frac{|\{n : \eta_n \leq \rho_c\} \cap \{n : (\mathbf{x}_n, y_n) \text{ are anomalous}\}|}{|\{n : \eta_n \leq \rho_c\}|} \\ \text{Recall} \quad &= \frac{|\{n : \eta_n \leq \rho_c\} \cap \{n : (\mathbf{x}_n, y_n) \text{ are anomalous}\}|}{|\{n : (\mathbf{x}_n, y_n) \text{ are anomalous}\}|}, \end{aligned}$$

where the threshold ρ_c is a cut-off threshold that is swept over the interval $[0, 1]$ to trace out the precision-recall curves in Fig. 3.6(c). It is evident from the figure that the proposed GEM-MED outlier resistant classifier has better precision-recall performance than ROD. Other corruption rates r_a lead to similar results. In Fig. 3.6(d), we compare the performance of GEM-MED, RODs under different corruption rates in terms of the Area Under the Curve (AUC), a commonly used measure in data mining [Japkowicz and Shah, 2011]. Similar to Fig.3.6(c), the GEM-MED outperforms RODs in terms of AUC for the range of investigated corruption rates.

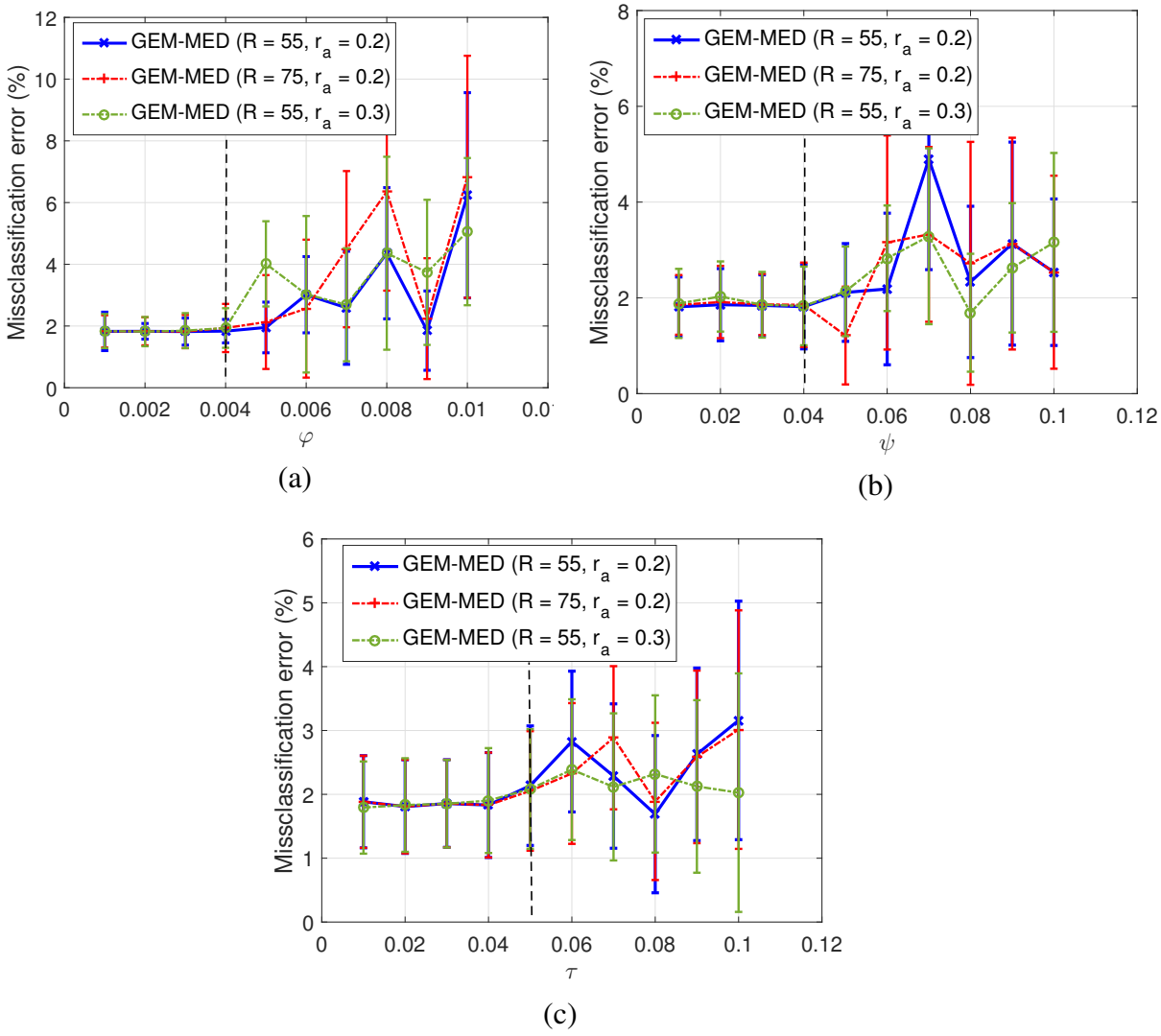


Figure 3.7: The classification error of GEM-MED vs. (a) learning rates φ, when $(\psi = 0.01, \tau = 0.02)$; (b) vs. ψ when $(\varphi = 0.001, \tau = 0.02)$ and (c) vs. τ when $(\varphi = 0.001, \psi = 0.01)$. The vertical dotted line in each plot separates the breakdown region (to the right) and the stable region of misclassification performance. These threshold values do not vary significantly as the noise level R and corruption rate r_a vary over the ranges investigated.

3.5.2 Footstep Classification

The proposed GEM-MED method was evaluated on experiments on a real data set collected by the U.S. Army Research Laboratory [Huang et al., 2011, Damarla, 2012, Nguyen et al., 2011]. This data set contains footstep signals recorded by a multisensor system, which includes four acoustic sensors and three seismic sensors. All the sensors are well-synchronized and operate in a natural environment, where the acoustic signal recordings are corrupted by environmental noise and intermittent sensor failures. The task is to discriminate between human-alone footsteps and human-leading-animal footsteps. We use the signals collected via four acoustic sensors (labeled sensor 1,2,3,4) to perform the classification. See Fig. 3.8. Note that the fourth acoustic sensor suffers from sensor failure, as evidenced by its very noisy signal record (bottom panel of Fig. 3.8). The data set involves 84 human-alone subjects and 66 human-leading-animal subjects. Each subject contains 24 75%-overlapping sample segments to capture temporal localized signal information. We randomly selected 25 subjects with 600 segments from each class as the training set. The test set contains the rest of the subjects. In particular, it contains 1416 segments from human-alone subjects and 984 segments from human-leading-animal subjects. A more detailed description of the dataset is given in [Huang et al., 2011, Damarla, 2012].

In a preprocessing step, for each segment, the time interval with strongest signal response is identified and signals within a fixed size of window (1.5 second) are extracted from the background. Fig. 3.9 shows the spectrogram (dB) of human-alone footsteps and human-leading-animal footsteps using the short-time Fourier transform [Sejdić et al., 2009], as a function of time (second) and frequency (Hz). The majority of the energy is concentrated in the low frequency band and the footstep periods differ between these two classes of signals. For features, we extract a mel-frequency cepstral coefficient (MFCC, [Mermelstein, 1976]) vector using a 50 msec. window. Only the first 13 MFCC coefficients were retained, which were experimentally determined to capture an average 90% of the power in the associated cepstra. There are in total 150 windows for each segment, resulting in a matrix of MFCC coefficients of size 13×150. We reshaped the matrix of MFCC features to obtain a 1950 dimensional feature vector for each segment. We then apply PCA to reduce the dimensionality from 1950 to 50, while preserving 85% of the total power. The above procedures for preprocessing follows exactly from [Nguyen et al., 2011].

We compare the performance of kernel SVM, kernel MED, ROD for outlier parameter $\rho \in [0.01, 1]$, and GEM-MED by training on the four sensors individually as well as in combination. For the combined sensors we used an augmented feature vector of dimension 200 via feature concatenation. We used a Gaussian RBF kernel function for the matrix $\mathbf{K}$ in the Gaussian process prior for the SVM decision function f. For the optimization

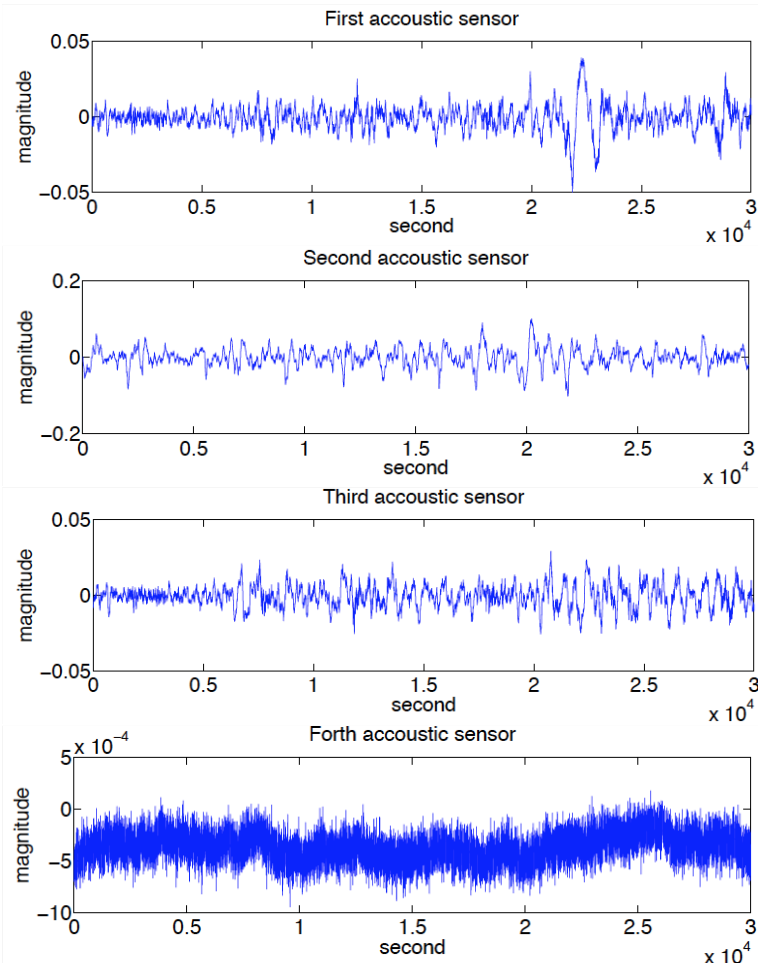


Figure 3.8: A snapshot of human-alone footstep collected by four acoustic sensors.

of GEM-MED we used a separable prior and exponentially distributed hyperparameters, as indicated by (3.14) and (3.15). Finally, the BP-kNNG implementation of GEM was applied on the training samples in the MFCC feature space with $k = 10$ nearest neighbors. The threshold ϑ is set using the Leave-One-Out resampling strategy [Hero, 2006], where each holdout sample corresponds to an entire segment.

Note that all classifiers were learned from a corrupted training set. Since the test set is also corrupted we used an anomaly detection algorithm (GEM with 5% false alarm rate) to produce a test set with few anomalies, called the nominal test set. This allows us to report the performance of the various algorithms on both the clean test data and on the corrupted test data. Table 3.2 shows the classification accuracy of the methods (trained on the training set alone) applied to nominal test set and Table 3.3 shows the result on the entire corrupted test set. For ROD only $\rho = 0.02$ and $\rho = 0.20$ are shown; it was determined that $\rho = 0.20$ achieves the best performance in the range $\rho \in [0.01, 1]$. In Table

Table 3.2: Classification accuracy **on nominal (clean) test set** for footstep experiment with different sensor combinations, with the best performance shown in **bold**.

Classification Accuracy (%) mean ± standard error						
sensor no.	kernel SVM	kernel MED	ROD-0.02	ROD-0.2	GEM + SVM	GEM-MED
1	71.2 ± 8.2	71.1 ± 5.3	73.7 ± 3.7	76.0 ± 2.5	72.5 ± 4.2	$\mathbf{78.4 \pm 3.3}$
2	60.8 ± 12.5	62.3 ± 10.2	71.5 ± 7.3	76.5 ± 5.3	70.3 ± 2.5	$\mathbf{82.1 \pm 3.1}$
3	60.5 ± 14.2	60.0 ± 13.1	63.2 ± 5.4	$\mathbf{67.6 \pm 4.2}$	56.5 ± 3.5	66.8 ± 4.5
4	59.6 ± 10.1	58.4 ± 8.2	71.8 ± 7.2	73.2 ± 4.2	76.5 ± 2.7	$\mathbf{80.1 \pm 3.1}$
1,2,3,4	75.9 ± 7.5	78.6 ± 5.1	79.2 ± 3.7	79.8 ± 2.5	75.2 ± 3.3	$\mathbf{84.0 \pm 2.3}$

Table 3.3: Classification accuracy **on the entire (corrupted) test set** for footstep experiment with different sensor combinations, with the best performance shown in **bold**.

Classification Accuracy (%) mean ± standard error						
sensor no.	kernel SVM	kernel MED	ROD-0.02	ROD-0.2	GEM + SVM	GEM-MED
1	65.2 ± 10.6	65.8 ± 10.2	68.5 ± 8.3	70.0 ± 6.8	70.2 ± 5.5	$\mathbf{72.5 \pm 4.8}$
2	54.9 ± 11.8	55.2 ± 11.0	63.2 ± 9.8	68.1 ± 7.5	68.5 ± 7.8	$\mathbf{76.3 \pm 3.9}$
3	50.7 ± 10.0	52.0 ± 10.5	56.8 ± 8.5	56.9 ± 7.3	56.5 ± 3.5	$\mathbf{60.1 \pm 5.3}$
4	57.0 ± 12.3	57.5 ± 12.1	69.6 ± 9.2	69.8 ± 5.1	70.2 ± 4.2	$\mathbf{75.0 \pm 4.0}$
1,2,3,4	70.8 ± 8.8	71.0 ± 8.5	73.6 ± 7.2	74.8 ± 6.9	75.1 ± 3.3	$\mathbf{76.8 \pm 2.5}$

3.2, it is seen that the GEM-MED method outperforms the ROD-ρ algorithms for all values of ρ as a function of classification accuracy when individual sensors 1,2,4 are used. Notice that when used alone neither kernel MED nor kernel SVM is resistant to the sensor failures in the training set, which explains their poor accuracy in sensor 3 and sensor 4. Also in the column *GEM+MED* of Table 3.2, we first trained a GEM anomaly detector to screen out

Table 3.4: Anomaly detection accuracy with different sensors, with the best performance shown in **bold**.

Anomaly Detection Accuracy (%) mean ± standard error			
sensor no.	ROD-0.02	ROD-0.2	GEM-MED
1	30.2 ± 1.3	59.0 ± 3.5	$\mathbf{70.5 \pm 1.3}$
2	23.5 ± 2.6	$\mathbf{63.5 \pm 2.8}$	63.4 ± 2.5
3	5.3 ± 1.4	48.1 ± 3.3	$\mathbf{72.8 \pm 1.5}$
4	22.8 ± 3.2	65.2 ± 4.2	$\mathbf{88.1 \pm 2.1}$
1, 2, 3, 4	38.5 ± 6.3	63.3 ± 5.5	$\mathbf{88.5 \pm 4.1}$

5% of the noisy training set, then trained a MED classifier on the rest of the training data. Note that GEM-MED learns both the detector and the classifier jointly on noisy training data. Table 3.2 shows that the two stage training approach has poor performance in highly corrupted sensors 3 and 4. This is due to the fact that when the GEM detector is learned without inferring the classification margin, it cannot effectively limit the negative influence of those corrupted samples that are close to the class boundary. In Table 3.3, we show the classification accuracy when both the nominal and anomalous test samples are involved in evaluation. We observe a performance degradation for all methods due to the irregularity of the outliers in the test set. In spite of this, the GEM-MED maintains a superior performance over all other methods. This reflects the superiority of the proposed joint classification and detection approach of GEM-MED as compared with *GEM + MED* approach.

Table 3.4 compares the anomaly detection accuracies on both *training and test data* for ROD and GEM-MED, where the accuracy is computed relative to ground truth anomalies. Note that GEM-MED has significant improvement in accuracy over ROD when trained individually on sensors 1,3,4, respectively, and when trained on all of the combined sensors. When trained on sensor 2 alone, the accuracies of GEM-MED and ROD-0.2 are essentially equivalent. In sensor 2 the anomalies appear to occur in concentrated bursts and we conjecture that that a GEM-MED model that accounts for clustered and dependent anomalies may be able to do better. Such an extension is left to future work.

3.6 Conclusion

In this chapter we proposed a unified GEM-MED approach for anomaly-resistant classification. We demonstrated its performance advantages in terms of both classification accuracy and detection rate on a simulated data set and on a real footstep data set, as compared to an anomaly-blind Ramp-Loss-based classification method (ROD). Further work could include generalization to the setting of multiple sensor types where anomalies exist in both training and test sets.

3.6.1 Acknowledgment

This work was supported in part by the U.S. Army Research Lab under ARO grant WA11NF-11-1-103A1. We also thanks Xu LinLi and Kumar Sricharan for their inputs on this work.

3.7 Appendices

3.7.1 Derivation of theorem 3.4.1

Proof: The proof of the convexity of the problem can be seen in chapter 12 of the standard textbook [Jaakkola et al., 1999], since the problem is with respect to the distribution q. The uniqueness of the solution follows directly from the fact that the problem is convex.

The Lagrangian function is given as

$$\begin{aligned}
&\mathcal{L}(q, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}) \\
&= \mathbb{E}_q \left[\log q - \log p_0\right] + \sum_{n \in T} \lambda_n \mathbb{E}_q \left[\eta_n \mathcal{L}_C\right] - \sum_{z \in \{\pm 1\}} \mu_z \mathbb{E}_q \left[\tilde{\mathcal{L}}_{D,z}\right] \\
&\quad - \sum_{z \in \{\pm 1\}} \kappa_z \mathbb{E}_q \left[\sum_{n: y_n = z} \eta_n / \left|T\right| - \hat{\beta}\right]
\end{aligned}$$

with dual variables $\boldsymbol{\lambda} = \{\lambda_n, n \in T\} \succeq \mathbf{0}$, $\boldsymbol{\mu} = (\mu_z, z \in \pm 1) \succeq \mathbf{0}$ and $\nu \geq 0$.

Then the result follows directly from solving a system of equations according to the KKT condition.

3.7.2 Derivation of theorem 3.4.2

Proof: According to [Jaakkola et al., 1999], the dual optimization is given as

$$
\begin{aligned}
&\max_{\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa}\geq 0} -\log Z(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\kappa}) \\
&= -\log \prod_{n\in T} \int \exp\left(-c(1-\xi_n)-\lambda_n\xi_n\right) d\xi_n \\
&\quad \times \int\int \exp\left(-\frac{1}{2}\boldsymbol{f}^T\boldsymbol{K}^{-1}\boldsymbol{f}+\sum_n \lambda_n\eta_n y_n f_n\right) d\boldsymbol{f} \\
&\quad \times p_0(\boldsymbol{\eta}) \exp\left(-\sum_{z\in\{\pm 1\}}\mu_z\sum_{n:z}\eta_n d_n+\sum_{z\in\{\pm 1\}}\mu_z\hat{\gamma}_z\right. \\
&\qquad\qquad \left.+\sum_{z\in\{\pm 1\}}\kappa_z\sum_{n:z}\eta_n+\sum_{z\in\{\pm 1\}}\kappa_z\hat{\beta}\right) d\boldsymbol{\eta} \\
&= \sum_{n\in T}\left(\lambda_n+\log\left(1-\lambda_n/c\right)\right)-\sum_{z\in\{\pm 1\}}\mu_z\hat{\gamma}_z-\left(\sum\kappa_z\right)\hat{\beta} \\
&\quad -\log\int \exp\left(\frac{1}{2}Q(\boldsymbol{K},(\boldsymbol{\lambda}\odot\boldsymbol{\eta}\odot\boldsymbol{y}))+\boldsymbol{\eta}^T\left(-\boldsymbol{\mu}\otimes\boldsymbol{d}+\boldsymbol{\kappa}\otimes\boldsymbol{e}\right)\right) \\
&\qquad\qquad \times p_0(\boldsymbol{\eta}) d\boldsymbol{\eta}
\end{aligned}
$$

where

$$
\begin{aligned}
Q(\boldsymbol{K},\boldsymbol{x}) &= \boldsymbol{x}^T\boldsymbol{K}\boldsymbol{x} \\
Q(\boldsymbol{K},(\boldsymbol{\lambda}\odot\boldsymbol{\eta})) &:= (\boldsymbol{\lambda}\odot\boldsymbol{\eta})^T\boldsymbol{K}(\boldsymbol{\lambda}\odot\boldsymbol{\eta}) \\
&= \boldsymbol{\lambda}^T\left(\boldsymbol{K}\odot(\boldsymbol{\eta}\boldsymbol{\eta}^T)\right)\boldsymbol{\lambda} \\
&= Q(\boldsymbol{K}(\boldsymbol{\eta}),\boldsymbol{\lambda}).
\end{aligned}
$$

3.7.3 Derivation of (3.21), (3.22)

Proof: The expression for $q(\overline{\Theta})$ is given as

$$
\begin{aligned}
q(\overline{\Theta}) \propto &\exp\left(-\frac{1}{2}\boldsymbol{f}^T\boldsymbol{K}^{-1}\boldsymbol{f}+\sum_n\lambda_n\eta_n y_n f_n\right) \\
&\times p_0(\boldsymbol{\eta})\exp\left(-\sum_{z\in\{\pm 1\}}\mu_z\sum_{n:z}\eta_n d_n+\sum_{z\in\{\pm 1\}}\kappa_z\sum_{n:z}\eta_n\right)
\end{aligned}
$$

$$
\begin{aligned}
&\times \prod_{n\in T} \exp\left(-c + (c-\lambda_n)\xi_n\right) \\
&= q(f,\boldsymbol{\eta}) \prod_n q(\xi_n)
\end{aligned}
$$

Given all $\eta_n, n \in T$,

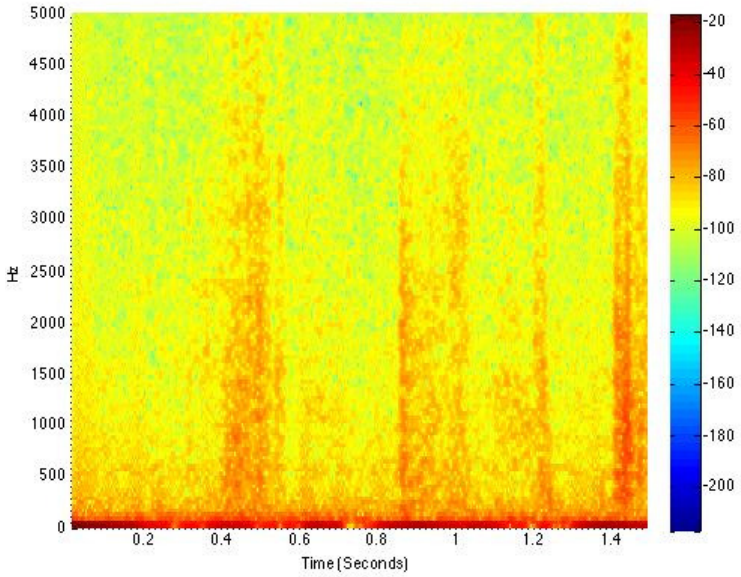
$$
\begin{aligned}
q(f|\boldsymbol{\eta}) &\propto \exp\left(-\frac{1}{2}\boldsymbol{f}^T\boldsymbol{K}^{-1}\boldsymbol{f} + \sum_n (\lambda_n\eta_n) f_n\right) \\
&= \exp\left(-\frac{1}{2}\left(\boldsymbol{f} - \boldsymbol{K}\left(\boldsymbol{\lambda}\odot\boldsymbol{\eta}\odot\boldsymbol{y}\right)\right)^T \boldsymbol{K}^{-1}\left(\boldsymbol{f} - \boldsymbol{K}\left(\boldsymbol{\lambda}\odot\boldsymbol{\eta}\odot\boldsymbol{y}\right)\right)\right) \\
&= \mathcal{N}\left(\boldsymbol{K}\left(\boldsymbol{\lambda}\odot\boldsymbol{\eta}\odot\boldsymbol{y}\right), \boldsymbol{K}\right).
\end{aligned}
$$

On the other hand, given f, $\boldsymbol{\eta} = (\eta_n, n \in T)$ are fully separated in above formula, therefore $q(\boldsymbol{\eta}|f) = \prod_n q(\eta_n|f)$.

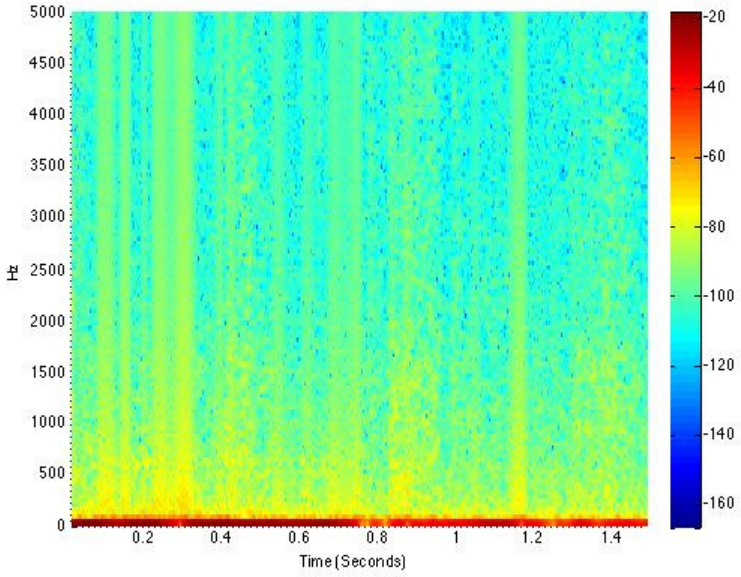
3.7.4 Implementation of Gibbs sampler

We implement a Gibbs sampler [Robert and Casella, 2013] to estimate $\mathbb{E}_{q(f,\boldsymbol{\eta})}\left[G(f,\boldsymbol{\eta})\right]$, where G is a general function of f and $\boldsymbol{\eta}$, as expressed in (3.23), (3.24), (3.25). The following procedure is applied iteratively

- Initialization: Set $\hat{\boldsymbol{\eta}}_0 = [1,\ldots,1]^T$ and set a fixed dual parameter $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})$. Let $G_0 = 0$.
- For each $t = 1, 2, \ldots, T_G$ or until convergence
  1. Given $\hat{\boldsymbol{\eta}}_{t-1} = (\hat{\eta}_{n,t-1})$, generate decision value $f_t(\mathbf{x}_n), n = 1,\ldots,N$ according to the Gaussian process with mean function $\hat{f}_t(\cdot) = \sum_{n\in T} \lambda_n \hat{\eta}_{n,t-1} y_n K(\cdot, \boldsymbol{x}_n)$.
  2. Given $\{f_t(\mathbf{x}_n)\}_{1\le n\le N}$, for $r = 1,\ldots,N_r$,
     (a) generate latent variables $\eta_{n,t}^{(r)} \in \{0,1\}$ according to the Bernoulli distribution with parameter as in (3.21) for each n independently.
  3. Compute the sample mean of $\hat{\eta}_{n,t} = \frac{1}{N_r}\sum_{r=1}^{N_r} \eta_{n,t}^{(r)} \in [0,1], n = 1,\ldots,N$. Let $\hat{\boldsymbol{\eta}}_t = (\hat{\eta}_{n,t})_{1\le n\le N}$.
  4. Evaluate G_t via $G_t = \frac{t-1}{t} G_{t-1} + \frac{1}{t} G(\hat{f}_t, \hat{\boldsymbol{\eta}}_t)$
- Output the approximate expectation $\hat{\mathbb{E}}_{q(f,\boldsymbol{\eta})}\left[G(f,\boldsymbol{\eta})\right] = G_T$ as well as the mean estimate $\hat{\boldsymbol{\eta}}_T$ and $\hat{f}_T(\boldsymbol{x}_n), 1 \le n \le N$ when the Gibbs chain process becomes stationary.



(a)



(b)

Figure 3.9: The power spectrogram (dB) vs. time (sec.) and frequency (Hz.) for a human-alone footstep (a) and a human-leading-animal footstep (b). Observe that the period of periodic footstep is a discriminative feature that separates these two signals.

CHAPTER 4

Multi-view Learning on Statistical Manifold via Stochastic Consensus Constraints

4.1 Introduction

In many applications, data are available from multiple sources (views) for which to train an object multiclass classifier. However, multi-view samples are often not fully annotated leading to degradation of classifier performance. For example, crowdsourcing [Whitla, 2009] has been used to annotate data in applications ranging from network analysis [Xiong and Svensson, 2002], video surveillance [Snoek and Worring, 2005] and multimedia retrieval [Snoek et al., 2010] For other applications, data are collected from less controlled environments like mobile devices [Satyanarayanan, 2011], open-source databases [Kushmerick, 1999] or public webpages [Craven et al., 2000]. Such approaches can lead to problems of missing labels [Mann and McCallum, 2010], of noise corruptions [Xie et al., 2014] and of *multi-view inconsistency* [Christoudias et al., 2008, Yoon et al., 2014] These problems can be formulated as semi-supervised multi-view learning problems with *weakly-labeled* data [Ivanov et al., 2001, Bockhorst and Craven, 2002, Bergamo and Torresani, 2010]. Data are called *weakly-labeled* when their class labels are provided according to conditional probability distributions. In other words, true label instances are only probabilistically linked to their associated multi-view sample instances. In this chapter, our *goal* is to perform weakly-labeled multi-view classification in the presence of such multi-view inconsistency.

Conventional multi-view learning methods fall into two categories: feature fusion (*early fusion*) and decision fusion (*late fusion*). In the former case, the goal of the algorithm is to find a joint feature representation of multiple views for classifications [Hardoon et al., 2004, Kakade and Foster, 2007, Ngiam et al., 2011], and in the latter case, multiple models are learned independently within each view and their outputs are combined to form a final classification result [Yager, 1987, Collins and Singer, 1999, Klein, 2004]. When noise

corruption and multi-view inconsistency exist, neither of these two schemes works well. Feature fusion approaches are known to be sensitive to the noise in a single view [Hardoon et al., 2004, Pan and Yang, 2010, Ngiam et al., 2011]. Decision fusion approaches, on the other hand, have difficulty handling the multi-view inconsistency problem, since the partial observations from single views may reveal inconsistent or even contradictory information, causing unreliable final decisions [Christoudias et al., 2008]. This chapter provides an alternative *model fusion* approach, where a *consensus view* is learned by fusing probabilistic models from different views. The predictions of all single view models are enforced to agree in the consensus view. Fig 4.1 illustrates the difference between the two conventional methods and our proposed method. Note that for the case of Gaussian features with unknown means and known covariances, all three cases give the identical result, because the posterior distribution of a Gaussian random variable X given another Gaussian random variable Y is itself Gaussian with mean parameter equal to $\mathbb{E}[X|Y]$, which is linear in Y.

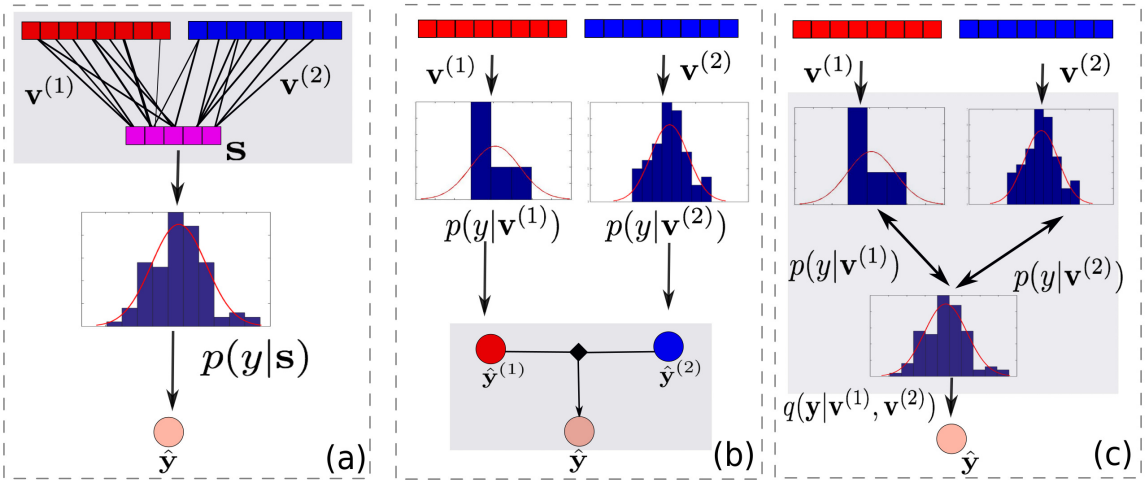


Figure 4.1: Illustration of multi-view learning approaches for classifying the multi-class label y given multi-view data $v^{(1)}$, $v^{(2)}$ with two views. Early multi-view fusion (a) combines the views into a composite view s, e.g., using algebraic combining rules, from which a posterior probability $p(y|s)$ is determined and a MAP estimator $\hat{y} = \text{argmax}_y p(y|s)$ is derived. High level multi-view fusion (b) fuses single view MAP estimates $\hat{y}^{(1)}$ and $\hat{y}^{(2)}$, obtained by maximizing the respective single view posteriors $p(y|v^{(1)})$ and $p(y|v^{(1)})$. The proposed consensus-based multi-view maximum entropy discrimination (COM-MED) method (c) forms a consensus estimate $q(y|v^{(1)}, v^{(2)})$ of the posterior distribution given pairs of multi-views from which a multi-view MAP estimator $\hat{y}$ is derived.

In this chapter, we assume that data in each view come in the form of probability distributions or histograms. Such data representations are widely used in database indexing [Agarwal et al., 2009], image and action recognition [Lowe, 2004, Scovanner et al., 2007], gene micro-array expression [Yang and Speed, 2002], manifold learning [Carter et al., 2009] and classification [Muandet et al., 2012]. In these tasks, a histogram or an em-

Table 4.1: The comparison of multi-view learning methods (**Bold** for the proposed method, $\surd$ for yes and $\times$ for no.)

	fusion category	parsimony[1]	partially-labeled	noise tolerance	multi-view inconsis-tency	# of views
CCA [Hardoon et al., 2004]	feature	$\times$	$\surd$	$\times$	$\times$	2
Bi-DAE [Ngiam et al., 2011]	feature	$\times$	$\surd$	$\times$	$\times$	2
Bayes-Fusion	decision	$\surd$	$\times$	$\surd$	$\times$	≥ 2
Co-Boosting [Collins and Singer, 1999]	decision	$\surd$	$\surd$	$\times$	$\times$	≥ 2
Co-training [Blum and Mitchell, 1998]	consens.	$\surd$	$\surd$	$\surd$	$\times$	2
Bayes Co-trn [Yu et al., 2007]	consens.	$\times$	$\surd$	$\surd$	$\surd$	≥ 2
SVM-2K [Farquhar et al., 2005]	consens.	$\times$	$\surd$	$\times$	$\times$	2
MV-MED [Sun and Chao, 2013]	consens.	$\surd$	$\surd$	$\times$	$\times$	2
COM-MED	consens.	$\surd$	$\surd$	$\surd$	$\surd$	≥ 2

pirical probability distribution function (p.d.f.) provides an efficient low-dimensional *non-Euclidean* representation as compared to the original vectorial representation. For instance, in [Ivanov et al., 2001], a p.d.f. is provided by crowd-sourcing, indicating the reliability of each sample. In this paper, we propose to learn a set of parametric conditional p.d.f. representations for multi-view data simultaneously as well as a consensus-view model on the space of all parametric conditional p.d.f.s, i.e., a *statistical manifold* [Amari and Nagaoka, 2007]. These learned p.d.f.s can further be used to construct local classifiers, while the consensus-view model maintains the shared information among multiple views.

A *key* contribution of this chapter is a multi-view learning framework on statistical manifolds, namely the *COM-MED*, using *stochastic consensus-based regularization*. This

[1] *Parsimony* means that there are few tuning parameters.

framework extends to the conventional co-regularlization method [Sindhwani et al., 2005, Farquhar et al., 2005, Kakade and Foster, 2007, Sindhwani and Rosenberg, 2008, Sun and Chao, 2013] on Euclidean space to non-Euclidean statistical manifolds. The proposed stochastic consensus measure is defined using information-theoretic divergences, such as the Kullback-Leibler divergence (KL-divergence), the Bhattacharyya distance [Kailath, 1967] or the α-divergence [Hero et al., 2001]. According to [Hero et al., 2001, Carter et al., 2009], these divergence measures take into account the intrinsic non-Euclidean geometry of the statistical manifold [Amari and Nagaoka, 2007] and are robust to noise corruption in single views. The COM-MED is based on the well-established Maximum Entropy Discrimination (MED) approach proposed by Jaakkola et al [Jaakkola et al., 1999]. MED performs Bayesian large-margin classification via the maximum entropy principle and it subsumes the support vector machine (SVM) as a special case. Note that Sun et al. [Sun and Chao, 2013] have proposed a multi-view version of MED recently, referred as MV-MED, where they imposed a shared-margin constraint over all view-specific discriminate functions. Our method does not rely on this heuristic constraint, but directly projects onto the underlying statistical manifold.

4.1.1 A Comparison of Multi-view Learning Methods

The proposed Consensus-based Multi-view Maximum Entropy Discrimination (COM-MED) method can be compared qualitatively with several popular multi-view learning methods [Table 4.1]. These include early fusion methods such as Canonical Correlation Analysis (CCA [Hardoon et al., 2004, Kakade and Foster, 2007]) and Bi-modal Deep Autoencoder (Bi-DAE) [Ngiam et al., 2011] methods, late fusion methods such as Bayesian fusion [Klein, 2004] and Co-Boosting [Collins and Singer, 1999], and the co-regularization methods [Sindhwani et al., 2005, Sindhwani and Rosenberg, 2008] (abbreviated as *consens.*), such as Co-training [Blum and Mitchell, 1998], Bayesian Co-training [Yu et al., 2007] and SVM-2K [Farquhar et al., 2005] algorithms. In Table 4.1 the comparison criteria include parsimony (few model parameters), the capability to handle noise corruption and multi-view inconsistency, the accommodation of partially-labeled data and the applicability to more than two views. Feature fusion methods such as the CCA and Bi-DAE are both sensitive to noise in each view and suffer from the problem of noise propagation across different views. The Bayesian methods have high sample complexity due to the high dimensionality of joint distribution. The consensus-based methods such as Co-training, Bayesian Co-training, SVM-2K and our proposed method are less sensitive to local noise and requires less parameters in modeling since we learn a predictive distribution on each single view independently. As shown in Table 4.1, the COM-MED method enjoys all of these

advantages. As will seen in Section 4.4, COM-MED scales well when the number of views increases, since it can learn each single view predictive distribution in parallel given a consensus-view distribution.

We quantitatively demonstrate the superior performance of the COM-MED over these methods in Section **4.5** on a collection of simulated data sets and two publicly available real data sets, the WebKB dataset for web-page classification [Craven et al., 2000] and the Internet Ads dataset in the UCI Machine Learning Repository [Lichman, 2013]. We also demonstrate COM-MED on a multi-sensor data set containing human-alone and human-leading-animal footsteps, collected in the field by an acoustic sensor array [Damarla et al., 2011, Nguyen et al., 2011, Damarla, 2012].

What follows is a brief outline of the chapter: In Section **4.2.1**, we review the co-regularization method in the Euclidean space. The proposed COM-MED method as a co-regularization method on statistical manifold is presented in Section **4.2.3**. In Section **4.3**, we analyze the robustness of the proposed stochastic consensus measure under noise perturbations. A variational Expectation-Maximization (EM) based algorithm is introduced in Section **4.4**. Experimental results based on synthetic data and real data are presented in Section **4.5**. Our conclusions are given in Section **4.6**.

4.2 Problem formulation

Consider a multi-view domain $\mathcal{X}^1 \times \ldots \mathcal{X}^V \times \mathcal{Y} := \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}^i \subset \mathbb{R}^{d_i}$ is the sample domain of view i, $\mathcal{X}$ is the joint sample domain and $\mathcal{Y}$ is the shared multiclass label domain and V is the number of views. Let $\mathbf{x}_n := [\mathbf{x}_n^1, \ldots, \mathbf{x}_n^V]$ be a multi-view sample in $\mathcal{X}$. Assume that $\mathbf{x}_n$ is associated with a label y_n for $n \in L$, the set of indices of labeled samples, but that there is no label for $\mathbf{x}_n$ for $n \in U$, the set of indices of unlabeled samples. For each view i, there is a mapping $\boldsymbol{x}^i \mapsto p_i(\cdot|\boldsymbol{x}^i)$ that associates the sample $\boldsymbol{x}^i$ with a conditional p.d.f. $p_i(\cdot|\boldsymbol{x}^i) : \mathcal{Y} \to \mathbb{R}$, i.e., the associated posterior probability. Denote $\mathcal{D}_{[V]} := \{\boldsymbol{x}_n\}_{n \in L \cup U}$ as a set of the N independent multi-view sample points, including $|L|$ labeled points and $|U|$ unlabeled points. Note that for labeled points, the posterior probability is a point mass at y_n, i.e., $p_i(y|\boldsymbol{x}_n^i) := \delta_{y=y_n}, n \in L$. A slice of $\mathcal{D}_{[V]}$ within view i is denoted as $\mathcal{D}_i := \{\mathbf{x}_n^i\}$.

Consider a parametric family of p.d.f.s p, referred as $\mathcal{M}$, i.e.,

$$\mathcal{M} := \{p_{\boldsymbol{\theta}} := p(y; \boldsymbol{\theta}),\ y \in \mathcal{Y} \mid \boldsymbol{\theta} \in \Theta\},$$

where $\boldsymbol{\theta}$ is a parameterization of the p.d.f. $p_{\boldsymbol{\theta}} \in \mathcal{M}$, and $\Theta \subset \mathbb{R}^k$ is the parameter set. The

set $\mathcal{M}$ is called a *statistical manifold* on domain $\mathcal{Y}$ [Amari and Nagaoka, 2007]. Denote $p_{\boldsymbol{\theta}_n^i} := p_i(y|\boldsymbol{\theta}_i(\boldsymbol{x}_n^i)) \in \mathcal{M}$, for $n \in L \cup U$, $1 \leq i \leq V$ as the parameterized conditional p.d.f. with parameter $\boldsymbol{\theta}_n^i := \boldsymbol{\theta}_i(\boldsymbol{x}_n^i) \in \Theta$. The notation emphasizes the dependence of a parameter $\boldsymbol{\theta}$ on the associated point $\boldsymbol{x}_n^i$.

In the following, we first consider a semi-supervised multi-view learning model on the Euclidean feature space $\mathcal{X}$.

4.2.1 Co-regularization on Euclidean space

The *co-regularization* methods for semi-supervised multi-view learning were proposed in [Farquhar et al., 2005, Sindhwani et al., 2005, Sindhwani and Rosenberg, 2008]. These methods learn multiple view-specific discriminate functions $f_i : \mathcal{X}^i \to \mathbb{R}$ jointly, where each $f_i \in \mathcal{H}_i$, the Reproducing Kernel Hilbert Space (RKHS) associated with a kernel $K_i : \mathcal{X}_i \times \mathcal{X}_i \to \mathbb{R}_+$ for $1 \leq i \leq V$. In the case of $V = 2$, they find an optimal pair of $\{f_1, f_2\}$ that minimizes the sum of the empirical loss functions over all views

$$\sum_{i=1}^{2} \left\{ \widehat{\mathbb{E}}_{n \in L} \left[\mathcal{L}_i(y_n, \boldsymbol{x}_n^i, f_i) \right] + \|f_i\|_{\mathcal{H}_i}^2 \right\}$$

under a *consensus constraint*

$$\widehat{\mathbb{E}}_{m \in U} \left[\left\| f_1(\boldsymbol{x}_m^1) - f_2(\boldsymbol{x}_m^2) \right\|_2^2 \right] \leq \rho, \tag{4.1}$$

where $\widehat{\mathbb{E}}_{n \in L}$ is the empirical expectation over L, $\|\cdot\|_{\mathcal{H}_i}$ is a norm defined in $\mathcal{H}_i$, $\rho > 0$ is a threshold and $\mathcal{L}_i : \mathcal{Y} \times \mathcal{X}^i \times \mathcal{H}_i \to \mathbb{R}_+ \cup \{0\}$ defines a classification loss function within view i, for $i = 1, 2$. The prediction for a two-view point $\boldsymbol{x} = [\boldsymbol{x}^1, \boldsymbol{x}^2]$ is then made from the output of an averaged discriminate function

$$g(\boldsymbol{x}) = \frac{1}{2} \sum_{i=1}^{2} f_i(\boldsymbol{x}^i).$$

In [Dasgupta et al., 2002], it is shown that a high probability of agreement between outputs of f_1 and f_2 guarantees that there exists a polynomial time algorithm to learn $g(\cdot)$ with small generalization error. The underling principle is referred as the *consensus principle* [Xu et al., 2013]. In other words, even if each function f_i is biased, they could learn from each other in order to reach a general agreement and the consensus of opinion will reveal the ground truth.

In [Yu et al., 2007], the consensus constraint in (4.1) is extended to cover the multi-view

case $V > 2$, where each function $f_i(\cdot)$ is compared with a common function $g(\cdot)$ so that

$$g(\cdot) := \arg \min_{h:\mathcal{X}\to\mathbb{R}} \widehat{\mathbb{E}}_{m\in U} \left[\sum_{i=1}^{V} \left\| h(\boldsymbol{x}_m) - f_i(\boldsymbol{x}_m^i) \right\|_2^2 \right],$$

which gives $g(\cdot) = \frac{1}{V}\sum_{i=1}^{V} f_i(\cdot)$ if U is sufficiently large.

4.2.2 Measure Label Inconsistency on Statistical Manifold via Stochastic Consensus Constraint

In the presence of noise corruption and multi-view label inconsistency, however, the use of a ℓ_2-distance-based consensus measure becomes unsatisfactory. First, by comparing their absolute difference, it does not take into account the reliability of the prediction of each f_i. In other word, insisting upon an agreement between a high-confidence classifier and a low-confidence, one will increase the bias for the overall system. Moreover, as demonstrated in [Christoudias et al., 2008], the co-regularization method using constraint (4.1) is sensitive to inconsistency among views.

In this chapter, we take into account of the effect of noise corruption and multi-view label inconsistency as a perturbation of posterior distribution $p_i(y|\boldsymbol{\theta}_i(\boldsymbol{x}_n^i))$ over the statistical manifold $\mathcal{M}$. A stochastic consensus constraint using an information-theoretic divergence $\mathbb{D}\left(\cdot \,\|\, \cdot\right) : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+ \cup \{0\}$ is proposed to address this noise effect [Amari and Nagaoka, 2007]. Examples of $\mathbb{D}\left(\cdot \,\|\, \cdot\right)$ include the KL-divergence [Kullback and Leibler, 1951],

$$\begin{aligned}\mathbb{D}\left(p_{\boldsymbol{\theta}_n^i} \,\|\, p_{\boldsymbol{\theta}_n^j}\right) &:= \mathbb{KL}\left(p_{\boldsymbol{\theta}_n^i} \,\|\, p_{\boldsymbol{\theta}_n^j}\right) \\ &= \sum_y p_i(y|\boldsymbol{\theta}_n^i) \log\left(\frac{p_i(y|\boldsymbol{\theta}_n^i)}{p_j(y|\boldsymbol{\theta}_n^j)}\right),\end{aligned}$$

and the Bhattacharyya distance [Bhattachayya, 1943]

$$B\left(p_{\boldsymbol{\theta}_n^i} \,\|\, p_{\boldsymbol{\theta}_n^j}\right) = -2\log \sum_y \sqrt{p_i(y|\boldsymbol{\theta}_n^i) p_j(y|\boldsymbol{\theta}_n^j)}$$

The stochastic consensus constraint is then defined as

$$\widehat{\mathbb{E}}_{m\in U} \left[\sum_{i=1}^{V} \mathbb{D}\left(q(y|\boldsymbol{x}_m) \,\|\, p_i(y|\boldsymbol{\theta}_m^i)\right) \right] \leq \rho, \tag{4.2}$$

where $q(y|\boldsymbol{x}_m) \in \mathcal{M}$ is a common p.d.f. shared among all views, also referred as the *consensus-view p.d.f.* Here $\rho > 0$ is a fixed threshold.

Similar to the role of ℓ_2-distance in Euclidean space, the divergence $\mathbb{D}\left(p_{\boldsymbol{\theta}'} \parallel p_{\boldsymbol{\theta}}\right)$ defines a *Riemannian metric* [Amari and Nagaoka, 2007] on the manifold $\mathcal{M}$ in a local neighborhood of $p_{\boldsymbol{\theta}}$ using the *Fisher information matrix* $\boldsymbol{J}(\boldsymbol{\theta}) := \left[-\mathbb{E}\left[\frac{\partial^2}{\partial\theta_s \partial\theta_t} \log p(y;\boldsymbol{\theta})\right]\right]$. Indeed, $J(\boldsymbol{\theta})$ is the Riemannian metric tensor associated with $\mathcal{M}$, and any f-divergence is locally equivalent to the it in the sense that:

$$\mathbb{KL}\left(p_{\boldsymbol{\theta}'} \parallel p_{\boldsymbol{\theta}}\right) = \frac{1}{2}\Delta\boldsymbol{\theta}^T \boldsymbol{J}(\boldsymbol{\theta})\Delta\boldsymbol{\theta} + o(\|\Delta\boldsymbol{\theta}\|^2),$$

where $\Delta\boldsymbol{\theta} := \boldsymbol{\theta}' - \boldsymbol{\theta}$ [Carter et al., 2011]. It is this local equivalence to a Riemannian metric over the space of posterior distributions that is the primary motivation for the proposed consensus constraint (4.2).

The consensus-view p.d.f. $q(y|\boldsymbol{x}_m)$ is defined as

$$q(y|\cdot) := \arg\min_{h(y|\cdot)\in\mathcal{M}} \widehat{\mathbb{E}}_{m\in U}\left[\sum_{i=1}^{V} \mathbb{D}\left(h(y|\boldsymbol{x}_m) \parallel p_i(y|\boldsymbol{\theta}_m^i)\right)\right]. \tag{4.3}$$

In the case of KL-divergence in (4.2), according to Appendix 4.8.1, $\log q(y|\boldsymbol{x}_n) = \frac{1}{V}\sum_{i=1}^{V} \log p_i(y|\boldsymbol{\theta}_n^i) + c(\boldsymbol{x})$, where $c(.)$ is a function of the multiview data $\boldsymbol{x}_n$ that makes $q(y|\boldsymbol{x})$ a properly normalized distribution over y. Fig. 4.2 illustrates a region defined by the consensus constraint in (4.2), when $\mathcal{M}$ is the space of all finite dimensional histograms, which is the upper hemisphere shown in Fig 4.2, and there are 5 views. In this case the consensus constraint (4.2) is a hyperspherical simplex (shown in yellow) and the consensus view is the centroid denoted by q.

4.2.3 Co-regularization on Statistical Manifold via COM-MED

Besides the stochastic constraint in (4.2), we need to reformulate the learning task in each single view. Note that each p.d.f. $p_i(y|\boldsymbol{\theta}_i(\boldsymbol{x}_n^i)), n \in L \cup U, 1 \le i \le V$, is indexed by the parameter function $\boldsymbol{\theta}_i(\boldsymbol{x}_n^i)$. Our goal for each view is to learn the parameters $\boldsymbol{\theta}_n^i := \boldsymbol{\theta}_i(\boldsymbol{x}_n^i)$, for $n \in L \cup U, 1 \le i \le V$. This is equivalent to learn parameterizations on the manifold $\mathcal{M}$.

The Maximum Entropy Discrimination (MED), proposed by Jaakkola et al. [Jaakkola et al., 1999], provides a flexible way to achieve such a goal in each view. Let Ψ_i be the set of all unknown model parameters in view i, including $\left\{\boldsymbol{\theta}_n^i, n \in L \cup U\right\}$. Assume that all these parameters in Ψ_i are random with a prior distribution $p_0(\Psi_i)$. The MED learns a

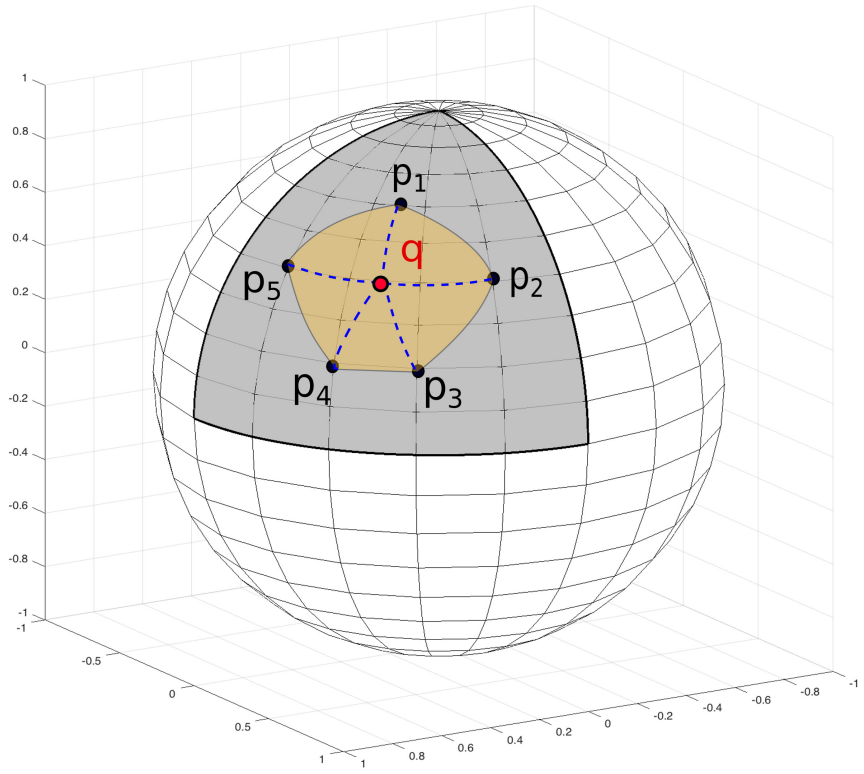


Figure 4.2: The illustration of a region defined by the consensus constraint in (4.2), when $\mathcal{M}$ is the space of all finite dimensional histograms, which is the upper hemisphere shown above, and there are 5 views. In this case the consensus constraint (4.2) is a hyperspherical simplex (shown in yellow) and the consensus view is the centroid denoted by q.

posterior distribution $q(\Psi_i) := q(\Psi_i|\mathcal{D}_i)$ that minimizes the KL-divergence

$$\mathbb{KL}\left(q(\Psi_i|\mathcal{D}_i) \parallel p_0(\Psi_i)\right) = \int \log\left(\frac{q(\Psi_i|\mathcal{D}_i)}{p_0(\Psi_i)}\right) q(\Psi_i|\mathcal{D}_i) d\Psi_i \tag{4.4}$$

subject to a set of constraints on the classification loss

$$\int \mathcal{L}_i(y_n, p_{\boldsymbol{\theta}_n^i}; \Psi_i) q(\Psi_i|\mathcal{D}_i) d\Psi_i \leq 0, \quad n \in L, \tag{4.5}$$

where the loss function $\mathcal{L}_i : \mathcal{Y} \times \mathcal{M} \to \mathbb{R}_+ \cup \{0\}$, for instance, is defined to be a large-margin loss in the SVM binary classification, i.e.,

$$\mathcal{L}_i(y_n, p_{\boldsymbol{\theta}_n^i}; \Psi_i) := \xi_n^i - \log\left(\frac{p_i(y = y_n|\,\boldsymbol{\theta}_n^i)}{p_i(y \neq y_n|\,\boldsymbol{\theta}_n^i)}\right), \tag{4.6}$$

with a set of additional non-negative slack variables $\{\xi_n^i\}_{n \in L}$ defined in Ψ_i.

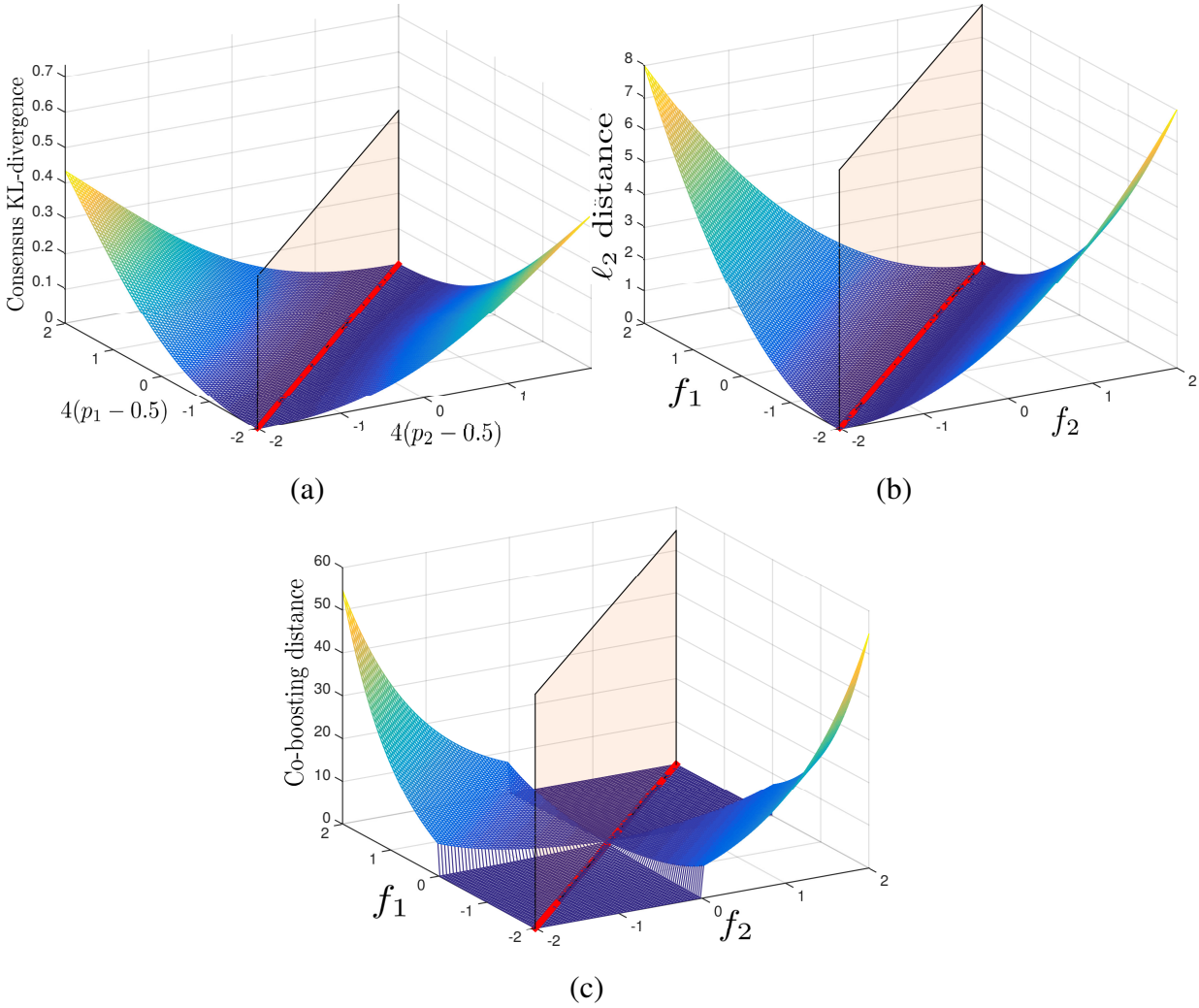


Figure 4.3: The comparison for two-view-consensus measures. (a) corresponds to the proposed stochastic consensus measure in (4.2); (b) corresponds to the ℓ_2-distance measure in the co-regularization in RKHS (4.1); (c) corresponds to the exp-distance measure in the Co-Boosting [Collins and Singer., 1999]. The red dash-line in the diagonal for (p_1, p_2) is the consensus line, when $p_1 = p_2$. Note that the curvature of the stochastic consensus measure around the consensus line is smaller than the rest of two measures, indicating its robustness in the presence of noise perturbation and multi-view inconsistency.

Combing (4.2), (4.4) and (4.5), we introduce the *Consensus-based Multi-view Maximum Entropy Discrimination (COM-MED)* method as a co-regularization method on $\mathcal{M}$, which solves

$$\min_{\substack{q(\Psi_1),\ldots,\\ q(\Psi_V),q(\rho)\in\Delta}} \sum_{i=1}^{V} \mathbb{KL}\left(q(\Psi_i) \parallel p_0(\Psi_i)\right) + \mathbb{KL}\left(q(\rho) \parallel p_0(\rho)\right) \tag{4.7}$$

$$\text{s.t.} \quad \int \mathcal{L}_i(y_n, p_{\boldsymbol{\theta}_n^i}; \Psi_i) q(\Psi_i) d\Psi_i \leq 0, \quad n \in L, 1 \leq i \leq V,$$

$$\widehat{\mathbb{E}}_{m\in U}\left[\sum_{i=1}^{V} \int \left[\mathbb{D}\left(q(y|\boldsymbol{x}_m) \parallel p_{\boldsymbol{\theta}_m^i}\right) - \rho\right] q(\Psi_i) q(\rho) d\Psi_i d\rho\right] \leq 0, \tag{4.8}$$

where the threshold $\rho > 0$ is random with prior distribution $p_0(\rho)$, Δ is the probability simplex and the consensus-view p.d.f. $q(y|\cdot)$ is given from (4.3). For a test sample $\boldsymbol{x} = [\boldsymbol{x}^1, \ldots, \boldsymbol{x}^V]$, the predicted label is given according to the Maximum a posteriori (MAP) rule

$$\hat{y} = \arg\max_{y\in\mathcal{Y}} \log q(y|\boldsymbol{x}).$$

From (4.7), we see that the COM-MED method solves V learning tasks via MED jointly, using the expectation of the stochastic consensus constraint in (4.2). Note that the optimization problem (4.7) can be decoupled into V independent subproblems, given a fixed consensus-view p.d.f. $q(y|\boldsymbol{x})$ on the unlabeled samples. As seen in (4.3), q is the centroid of $\{p_{\boldsymbol{\theta}_i}\}_{i=1}^{V}$, which is in turn determined by the results of all subproblems in each view. This motivates a solution method based on the *variational Expectation-Maximization (EM)* [Ganchev et al., 2010, Zhu et al., 2011], which will be discussed in Section 4.4. In the following section, we analyze the behavior of the stochastic consensus constraint (4.2) under small perturbations of p_i.

4.3 Analysis of Consensus Constraints

For simplicity, let $V = 2$ and we use the KL-divergence $\mathbb{KL}\left(q \parallel p_{\boldsymbol{\theta}_m^i}\right)$ in (4.2). Consider a perturbation of $p_{\boldsymbol{\theta}^i}$ on $\mathcal{M}$ due to the noise corruption in $\boldsymbol{x}^i$, resulting in $p_{\boldsymbol{\theta}^i+\Delta\boldsymbol{\theta}^i} \in \mathcal{M}$. As discussed in Section 4.2.2, since the consensus-view p.d.f. between $p_{\boldsymbol{\theta}^1}$ and $p_{\boldsymbol{\theta}^2}$ is proportional to the average of two p.d.f.s in the log-space, we denote it as $q_{(\boldsymbol{\theta}^1,\boldsymbol{\theta}^2)}$.

Substituting $\log q_{(\boldsymbol{\theta}^1,\boldsymbol{\theta}^2)} \propto \frac{1}{2}(\log p_{\boldsymbol{\theta}^1} + \log p_{\boldsymbol{\theta}^2})$ to the stochastic constraint (4.2), we have

$$\sum_{i=1}^{2} \mathbb{KL}\left(q_{(\boldsymbol{\theta}^1,\boldsymbol{\theta}^2)} \parallel p_{\boldsymbol{\theta}^i}\right) = -2\log \sum_{y} \sqrt{p_1(y|\boldsymbol{\theta}^1)p_2(y|\boldsymbol{\theta}^2)} + C(\boldsymbol{\theta}_1, \boldsymbol{\theta_2})$$
$$= B\left(p_{\boldsymbol{\theta}^1} \parallel p_{\boldsymbol{\theta}^2}\right) + C(\boldsymbol{\theta}_1, \boldsymbol{\theta_2}),$$

which is proportional to the Bhattacharyya distance between p.d.f.s $p_{\boldsymbol{\theta}^1}$ and $p_{\boldsymbol{\theta}^2}$. Note that $B\left(p_{\boldsymbol{\theta}^1} \parallel p_{\boldsymbol{\theta}^2}\right) = 0$, when a consensus is reached, i.e., $\boldsymbol{\theta}^2 = \boldsymbol{\theta}^1$.

Fig 4.3 shows the consensus measure as a function of the two single-view prediction values. We compare the stochastic consensus measure in Fig 4.3 (a) with the ℓ_2-distance measure (4.1) in Fig 4.3 (b) and the exp-distance measure for Co-Boosting [Collins and Singer, 1999] in Fig 4.3 (c). Note that the stochastic consensus measure grows slowly as compared to the other two measures when $\Delta\boldsymbol{\theta}^1$ is large. This shows its robustness with respect to anomalies and noise corruptions. Moreover, in the neighborhood of the consensus line $\left\{(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2)|\boldsymbol{\theta}^1 = \boldsymbol{\theta}^2\right\}$, the curvature of the stochastic consensus measure is determined *adaptively* via the Fisher information matrix at $\boldsymbol{\theta}^1$, as compared to a constant curvature for both the ℓ_2-distance and the exp-distance measures. In terms of this, for a poor estimator of $\boldsymbol{\theta}^1$ with small $\boldsymbol{J}(\boldsymbol{\theta}^1)$, the consensus constraint (4.2) becomes loose, reducing the negative impact of the inconsistency upon the whole system.

4.4 Algorithm

Note that given $q \in \mathcal{M}$, the primal problem in (4.7) is convex in each view. We can solve them view-by-view using the Karush-Kuhn-Tucker (KKT) conditions. On the other hand, given all $\{p_{\theta_i}\}_{i=1}^{V} \in \mathcal{M}$, $q \in \mathcal{M}$ lies in the centroid of the region spanned by $\{p_{\theta_i}\}_{i=1}^{V}$ on $\mathcal{M}$. A *variational EM-based* algorithm [Ganchev et al., 2010, Zhu et al., 2011] can be derived to solve (4.7) under the following model assumptions:

1. Assume binary classifications, i.e., $\mathcal{Y} = 2$, and the logistic posterior distribution $p_i(y|\theta_n^i)$ is

$$p_i(y = 1|\theta_n^i) = \frac{1}{1 + \exp\left(-\theta_n^i\right)},$$
$$p_i(y = -1|\theta_n^i) = \frac{1}{1 + \exp\left(\theta_n^i\right)} \tag{4.9}$$

   for $i = 1, \ldots, V$. $\theta_n^i \in \Theta \subset \mathbb{R}$.

2. In each view i, assume that $(\theta_n^i)_{n=1,\ldots}$ follows a Gaussian random process [Rasmussen and Williams, 2006] on $\mathcal{X}^i$, i.e., a positive-definite covariance kernel $K_i(\boldsymbol{x}_m^i, \boldsymbol{x}_n^i)$

is defined for all $\boldsymbol{x}_m^i, \boldsymbol{x}_n^i \in \mathcal{X}^i$ and for any $N \geq 1$,

$$(\theta_n^i)_{n=1}^N \sim \mathcal{N}(\mathbf{0}, \boldsymbol{K}_i), \tag{4.10}$$

where $\boldsymbol{K}_i = [K_i(\boldsymbol{x}_m^i, \boldsymbol{x}_n^i)]_{m,n=1}^N$ is a covariance matrix. An example is the Gaussian RBF kernel coveriance function $K_i(\boldsymbol{x}_m^i, \boldsymbol{x}_n^i) := \exp\left(-\gamma_i \left\|\boldsymbol{x}_m^i - \boldsymbol{x}_n^i\right\|_2^2\right)$.

3. Assume a separable prior, as commonly used in Bayesian inference [Jaakkola et al., 1999, Zhu et al., 2014]

$$p_0(\Psi_i) = p_0((\theta_n^i)_{n=1}^N) \prod_{n \in L} p_0(\xi_n^i), i = 1, \ldots, V. \tag{4.11}$$

4. Assume that the hyperparameters $\{\xi_n^i\}$ and ρ are exponential random variables. For $1 \leq i \leq V$,

$$\begin{aligned} p_0(\xi_n^i) &\propto \exp(-c_\xi^{(i)}(1 - \xi_n^i)),\ \xi_n^i \in (-\infty, 1],\ n \in L; \\ p_0(\rho) &\propto \exp(-c_\rho \rho),\ \rho \in [0, \infty), \end{aligned} \tag{4.12}$$

where $\{c_\xi^{(i)}\}_{i=1}^V$ and c_ρ are parameters.

4.4.1 Solving the Subproblem in Each View, given $q \in \mathcal{M}$

Given $q \in \mathcal{M}$, the stochastic consensus constraint function in (4.8) can be decoupled into V sub-constraint functions. For view i, the sub-constraint function is

$$\begin{aligned} &\int \widehat{\mathbb{E}}_{m \in U} \left[\mathbb{KL}\left(q(y|\boldsymbol{x}_m) \parallel p_i(y|\theta_m^i)\right)\right] q(\Psi_i) d\Psi_i \\ \propto &- \int \widehat{\mathbb{E}}_{m \in U} \left[\mathbb{E}_{q(y|\boldsymbol{x}_m)} \left[\log p_i(y|\theta_m^i)\right]\right] q(\Psi_i) d\Psi_i, \end{aligned} \tag{4.13}$$

where the integrand is the *cross-entropy loss* [De Boer et al., 2005]. From (4.22) in Appendix 4.8.2, we see that given a reference $\overline{\theta_m^i}$ from the previous iteration, the cross-entropy loss has second-order approximation

$$\begin{aligned} (4.13) = H(\overline{\theta_m^i}) - \frac{1}{2} d(q, \overline{\theta_m^i}) \theta_m^i + \frac{1}{8} R_m^i \left(\theta_m^i - \overline{\theta_m^i}\right)^2 \\ + o(\left\|\theta_m^i - \overline{\theta_m^i}\right\|^3), \end{aligned} \tag{4.14}$$

where $H(\overline{\theta_m^i})$ is the entropy of $p_{\overline{\theta}_m^i}$, $d_m := d(q, \overline{\theta_m^i}) = (\mathbb{E}_q[y] - \mathbb{E}_{p_{\overline{\theta}_m^i}}[y]))$ is the averaged difference between the prediction by the consensus-view p.d.f $q(y|\boldsymbol{x}_m)$ and that by the

single view p.d.f. $p_i(y|\overline{\theta_m^i})$. And $R_m^i := [1 - |\mathbb{E}_{p_{\overline{\theta}_m^i}}[y]|^2] \to 0$, for $|\mathbb{E}_{p_{\overline{\theta}_m^i}}[y]| \to 1$. Using this approximation, we can solve the primal problem in (4.7) for each view i. The result is given below.

Theorem 4.4.1 *Given $q \in \mathcal{M}$, the threshold $\rho > 0$ and the reference parameters $\{\overline{\theta_m^i}\}_{m\in U}$, the primal problem in* (4.7) *in each view i is convex with respect to the unknown distribution $q(\Psi_i)$ and the unique optimal solution is a generalized Gibbs distribution with the density:*

$$q(d\Psi_i) = \frac{1}{Z(\boldsymbol{\lambda}^i, \mu^i)} p_0(d\Psi_i) \exp\left(-E(\Psi_i; \boldsymbol{\lambda}^i, \mu^i)\right), \tag{4.15}$$

where

$$\begin{aligned} E(\Psi_i; \boldsymbol{\lambda}^i, \mu^i) &:= E\left(\left\{\theta_n^i\right\}_{n=1}^N, \left\{\xi_n^i\right\}_{n\in L}; \left\{\lambda_n^i\right\}_{n\in L}, \mu^i\right) \\ &= \sum_{n\in L} \lambda_n^i \mathcal{L}_i(y_n, \xi_n^i, \theta_n^i) \\ &\quad - \mu^i \sum_{m\in U} \left(\frac{1}{2} d(q, \overline{\theta_m^i})\theta_m^i - \frac{1}{8} R_m^i((\theta_m^i)^2 - 2\overline{\theta_m^i}\theta_m^i)\right), \end{aligned}$$

with $\Psi_i = \{\theta_n^i\}_{n=1}^N \cup \{\xi_n^i\}_{n\in L}$ and where the dual variables $\boldsymbol{\lambda}^i = (\lambda_n^i)_{n\in L}$ and μ^i are all nonnegative. $Z(\boldsymbol{\lambda}^i, \mu^i)$ is the partition function, which is given as

$$Z(\boldsymbol{\lambda}^i, \mu^i) = \int p_0(d\Psi_i) \exp\left(-E(\Psi_i; \boldsymbol{\lambda}^i, \mu^i)\right). \tag{4.16}$$

We use the large-margin classification loss (4.6) *in $\mathcal{L}_i(y_n, \xi_n^i, \theta_n^i)$. See the Appendix 4.8.3 for a detailed derivation.*

Since the subproblem in each view i is convex, we can equivalently solve a dual version of the optimization problem (4.7). In fact, we have the following result:

Theorem 4.4.2 *Under assumptions* (4.9)-(4.12)*, the dual optimization problem in view i is given as*

$$\begin{aligned} \max_{\boldsymbol{\lambda}^i \succeq \mathbf{0}, \mu^i \geq 0} \; & \sum_{n\in L} \left(\lambda_n^i + \log\left(1 - \lambda_n^i / c_\xi^{(i)}\right)\right) + (\mu^i + \log(1 - \mu^i / c_\rho)) \\ & - \frac{1}{2} \left[(\boldsymbol{\lambda}^i)^T, \mu^i\right] \boldsymbol{A}_i^T \widetilde{\boldsymbol{K}}_i \boldsymbol{A}_i \left[(\boldsymbol{\lambda}^i)^T, \mu^i\right]^T, \end{aligned} \tag{4.17}$$

where $\widetilde{\boldsymbol{K}}_i := \boldsymbol{K}_i - \boldsymbol{K}_i \boldsymbol{Q} \left(\alpha \boldsymbol{D}_{\boldsymbol{R}^i}^{-1} + (\boldsymbol{K}_i)_{uu}\right)^{-1} \boldsymbol{Q}^T \boldsymbol{K}_i$, $\left[(\boldsymbol{\lambda}^i)^T, \mu^i\right]^T \in \mathbb{R}^{|L|+1}$ are the nonnegative dual variables. $\boldsymbol{d}^i := (d_m^i)_{m\in U}, \boldsymbol{R}^i := (R_m^i)_{m\in U}$ are defined in (4.14)*, $\Delta \boldsymbol{y}^i := (\boldsymbol{d}^i/2 + \boldsymbol{R}^i \odot \overline{\boldsymbol{\theta}_i}/4) \in \mathbb{R}^{|U|}$, and $\odot$ is the point-wise matrix product.*

$$\boldsymbol{A}_i := \begin{bmatrix} diag\,(\boldsymbol{y}) & \boldsymbol{0} \\ \boldsymbol{0} & \Delta \boldsymbol{y}^i \end{bmatrix} \in \mathbb{R}^{N\times(|L|+1)}, \quad \boldsymbol{Q} := \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_{|U|} \end{bmatrix} \in \mathbb{R}^{N\times|U|}, \tag{4.18}$$

$\boldsymbol{D}_{\boldsymbol{R}^i} := \frac{1}{4} diag\,(R^i_m)_{m\in U} = diag\left(\left[\boldsymbol{J}(\overline{\boldsymbol{\theta}}^i)\right]\right)$ *is the diagonal of the Fisher information matrix at* $\overline{\boldsymbol{\theta}}^i$ *and* $\alpha > 0$ *is a constant shrinkage factor.*

See Appendix 4.8.4 for detailed derivations. Note that the dual optimization problem (4.17) can be solved by any SVM solver, such as the *LibSVM* [Chang and Lin, 2011]. We summarize the result as follows,

Theorem 4.4.3 *The primal solution* $q(\Psi_i), i = 1, \ldots, V$ *satisfies the following::*

1. *The posterior distribution* $q(\Psi_i)$ *is factorized as* $q(\Psi_i) = q(\theta^i_1, \ldots, \theta^i_N) \prod_n q(\xi^i_n) q(\rho)$.

2. *The conditional mean* $\mathbb{E}_{q(\Psi_i)}\left[\theta^i_s | \boldsymbol{x}^i_s, \overline{\boldsymbol{\theta}}^i, \mathcal{D}^i\right]$ *is given as*

$$\sum_{n\in L} y_n \lambda^i_n \widetilde{K}_i(\boldsymbol{x}^i_n, \boldsymbol{x}^i_s) + \mu^i \sum_{m\in U} \Delta\bar{y}^i_m \widetilde{K}_i(\boldsymbol{x}^i_m, \boldsymbol{x}^i_s),$$

   where $\widetilde{K}_i(\cdot,\cdot)$ *is the modified kernel function in* (4.17) *and* $\Delta\bar{y}^i_m := \frac{1}{2} d^i_m + \frac{1}{4} R^i_m \overline{\theta}^i_m, m \in U$ *with* d^i_m, R^i_m *defined in Appendix 4.8.2.*

3. *The conditional variance* $\mathbb{V}_{q(\Psi_i)}[\boldsymbol{\theta}^i | \overline{\boldsymbol{\theta}}^i, \mathcal{D}^i]$ *is given as*

$$\widetilde{\boldsymbol{K}}_i = \boldsymbol{K}_i - \boldsymbol{K}_i \boldsymbol{Q} \left(\alpha \boldsymbol{D}^{-1}_{\boldsymbol{R}^i} + (\boldsymbol{K}_i)_{uu}\right)^{-1} \boldsymbol{Q}^T \boldsymbol{K}_i$$

   where $\boldsymbol{D}_{\boldsymbol{R}^i}$ *is the diagonal of Fisher information matrix evaluated at* $\overline{\boldsymbol{\theta}}^i$.

Note that the conditional variance above depends on $R^i_m = [\boldsymbol{J}(\overline{\boldsymbol{\theta}}^i)]_{m,m}$. When $R^i_m \to 0$ for all $m \in U$, the conditional variance becomes $\boldsymbol{K}_i$. From a geometric point of view, $p_i(y|\overline{\theta}^i_m)$ has reached the vertex of the manifold $\mathcal{M}$ and thus the algorithm should stop, since no further information gain is expected given the previous predictions.

From Appendix 4.8.1, we see that the consensus-view p.d.f. $q(y|\boldsymbol{x})$ is proportional to the average of $\{p_{\theta^i}\}_{i=1}^V$ in log-space, i.e.,

$$\log q(y|\boldsymbol{x}_s) \propto \frac{1}{V} \sum_{i=1}^{V} \log p_i(y|\widehat{\theta}^i_s), \quad s \in L \cup U.$$

where $\widehat{\theta}_s^i = \mathbb{E}_{q(\Psi_i)}[\boldsymbol{\theta}_s^i | \boldsymbol{x}_s^i, \overline{\boldsymbol{\theta}}^i, \mathcal{D}^i]$ is provided by each single-view classifier. Finally, the variational EM-based algorithm is summarized in Algorithm 2.

Algorithm 2 COM-MED via variational EM

Require: Training samples $\mathcal{D}_{[V]}$ from V views, with the fully labeled part $\{(\mathbf{x}_n, y_n), n \in L\}$ and the unlabeled part $\{\mathbf{x}_m, m \in U\}$. The kernel function K^i defined on $\mathcal{X}_i \times \mathcal{X}_i$. The nonnegative hyperparameter $\{c_\xi^{(i)}\}_{i=1}^V$ and c_ρ for $\boldsymbol{\xi}^i, 1 \le i \le V$ and ρ. The shrinkage factor $\alpha > 0$.

1: **Initialize:** Choose an random subset of labeled data $\widehat{\mathcal{D}}_{[V]}^L$. Learn an initial p.d.f. $p_{i,0}$ from $\widehat{\mathcal{D}}_i^L$ independently for each view i via SVM. Find an initial estimate of the model parameter $\hat{\theta}_{m,0}^i \equiv \mathbb{E}_{p_{i,0}}\left[\theta_m^i | \, \boldsymbol{x}_m^i, \widehat{\mathcal{D}}_i^L\right]$, $m \in \mathcal{D}_i / \widehat{\mathcal{D}}_i^L$ for each i.

2: **for** $t = 0, \ldots, T$ or until converge **do**

3: (**E-step**) Find the consensus-view distribution $q_{t+1} \in \mathcal{M}$.

$$\log q_{t+1}(y|\boldsymbol{x}_s) \propto \frac{1}{V} \sum_{i=1}^{V} \log p_i(y|\overline{\theta}_{s,t}^i), \; s \in L \cup U$$

4: Make predictions via the consensus-view p.d.f. $\bar{y}_{m,(t+1)} = \mathbb{E}_{q_{t+1}(y)}[y | \, \boldsymbol{x}_m]$, $m \in U$

5: **for** $i = 1, \ldots, V$ **do**

6: Make single-view predictions via p.d.f. $p_{\overline{\theta}_{m,t}^i}$ as $\hat{y}_{m,t}^i = \mathbb{E}_{\overline{\theta}_{m,t}^i}[y | \, \boldsymbol{x}_m^i]$ and $R_{m,t}^i = (1 - |\hat{y}_{m,t}^i|^2)$.

7: Compute the difference $d_{m,(t+1)}^i = \bar{y}_{m,(t+1)} - \hat{y}_{m,t}^i$ and then compute $\Delta\bar{y}_{m,(t+1)}^i := \frac{1}{2} d_{m,(t+1)}^i + \frac{1}{4} R_{m,t}^i \overline{\theta}_{m,t}^i$ for $m \in U$.

8: (**M-step**) Solve for the optimal dual variables $(\boldsymbol{\lambda}^i, \mu^i)$ via (4.17) for view i.

9: Update the p.d.f. $p_{\overline{\theta}_{s,(t+1)}^i}$ using the posterior mean

$$\begin{aligned} \overline{\theta}_{s,(t+1)}^i &= \mathbb{E}_{q_i}\left[\theta_s^i | \, \boldsymbol{x}_s^i, \, \overline{\theta}_{s,t}^i, \, \mathcal{D}^i\right] \\ &= \sum_{n \in L} y_n \lambda_n^i K^i(\boldsymbol{x}_s^i, \boldsymbol{x}_n^i) + \mu^i \sum_{m \in U} \Delta\bar{y}_m^i K^i(\boldsymbol{x}_s^i, \boldsymbol{x}_m^i), \\ & \qquad\qquad s \in L \cup U, \end{aligned}$$

10: **end for**

11: **end for**

Ensure: Assign label for test sample $\boldsymbol{x}_s$ as

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \log q(y|\boldsymbol{x}_s).$$

4.4.2 Implementation Complexity

For each iteration, our COM-MED algorithm has $O((|L| + k)^3)$ time complexity for the dual optimization (4.17) along with an update of a $(|L| + k) \times (|L| + k)$ matrix in each view, where $|L|$ is the size of labeled samples and k is the number of clusters in unlabeled samples. Also it has $O(VN^2)$ memory complexity, which is prohibited for $N > 1000$. The *major advantage* of the COM-MED is that the time complexity does not grow with respect to $|U|$, the size of the unlabeled samples, since the information measure in (4.8) provides an efficient summary of the unlabeled data.

4.5 Experiments

We compare the proposed COM-MED model with the SVM-2K method proposed by Farquhar et al. [Farquhar et al., 2005], the Co-Laplacian SVM (CoLapSVM) by Sindahwani et al. [Sindhwani et al., 2005], the MV-MED by Sun et al. [Sun and Chao, 2013] as well as the standard MED for each view. For a fair comparison, we focus on two-view learning, i.e. $V = 2$ since the SVM-2K is for two-views only. For all MED-based algorithms, we use the Gaussian Process as a prior with the radial basis kernel function $K_i(\mathbf{x}_n^i, \mathbf{x}_m^i) = \exp(-\gamma_i \|\mathbf{x}_n^i - \mathbf{x}_m^i\|^2), \forall m, n,\ i \leq V$, where c_i is obtained by 5-fold cross-validation.

4.5.1 Footstep Classification

In the first experiment, we use the **ARL-Footstep** [Damarla et al., 2011, Nguyen et al., 2011] data. This data set contains footstep signals recorded by a multisensor system, which includes four acoustic sensors (, labeled as 1-4, respectively,) and two seismic sensors (, labeled as 5,6). All the sensors are well-synchronized and operates in a natural environment, where the environmental noises may exist in the either view.

The task is to discriminate between human footsteps and human-leading-animal footsteps. The data set involves 84 human-alone subjects and 66 human-leading-animal subjects. Each subject contains 24 75%-overlapping sample segments to capture temporal localized signal information. We randomly selected 25 subjects with 600 segments from each class as the training set. The test set contains the rest of the subjects. In particular, it contains 1416 segments from human-alone subjects and 984 segments from human-leading-animal subjects. A more detailed description of the dataset is given in [Huang et al., 2011, Damarla, 2012, Xie et al., 2014].

In the preprocessing step, the time periods containing strongest signal response are identified and signals within a fixed size of window (i.e. 23 sec.) are extracted from the

Table 4.2: Classification accuracy with different data set, with the best performance shown in **bold**.

Classification Accuracy (%) mean ± standard error						
Dataset.	MED (single views)		SVM-2K	CoLapSVM	MV-MED	COM-MED
ARL Footstep (Sensor 1,2, $\|L\| = 50$)	71.1 ± 5.3	69.1 ± 7.5	73.3 ± 5.2	75.2 ± 6.0	77.5 ± 6.5	$\mathbf{85.5} \pm 6.1$
WebKB4 ($\|L\| = 15$)	76.6 ± 10.2	77.1 ± 10.1	79.0 ± 10.0	83.6 ± 9.0	85.9 ± 8.7	$\mathbf{91.7} \pm 5.8$
Internet Ads ($\|L\| = 50$)	87.3 ± 0.9	86.2 ± 1.4	82.5 ± 4.3	85.9 ± 3.2	88.8 ± 2.3	$\mathbf{92.7} \pm 0.7$

Table 4.3: The classification accuracy (%) for two *homoegenous* views in **ARL-Footstep** dataset (The best one is in **bold**.)

	MED view i	MED view j	SVM-2K	CoLapSVM	MV-MED	COM-MED
Sensor $1, 2$	71.1 ± 5.3	69.1 ± 7.5	73.3 ± 5.2	75.2 ± 2.6	77.5 ± 6.5	$\mathbf{85.5} \pm 6.1$
Sensor $1, 3$	74.4 ± 9.7	52.8 ± 18.5	56.1 ± 7.8	73.5 ± 3.5	72.8 ± 3.7	$\mathbf{80.2} \pm 3.1$
Sensor $1, 4$	72.6 ± 7.3	63.7 ± 15.6	58.1 ± 8.5	70.5 ± 6.5	73.2 ± 1.5	$\mathbf{77.0} \pm 5.3$
Sensor $2, 3$	66.0 ± 9.3	57.2 ± 12.3	60.2 ± 7.1	75.2 ± 8.1	73.1 ± 4.2	$\mathbf{81.3} \pm 5.7$
Sensor $2, 4$	70.8 ± 7.8	65.6 ± 12.3	73.7 ± 7.0	73.5 ± 5.2	72.7 ± 4.8	$\mathbf{75.6} \pm 6.5$
Sensor $5, 6$	89.3 ± 1.5	86.1 ± 2.2	90.3 ± 2.1	90.5 ± 0.8	93.1 ± 0.5	$\mathbf{95.5} \pm 3.2$

background noises. We extract mel-frequency cepstral coefficients (MFCC, [Rabiner and Juang, 1993]) for acoustic signals using a 50 msec. window. Only the first 13 MFCC coefficients were retained, which were experimentally determined to capture an average 90% of the power in the associated cepstra. There are in total 150 windows for each segment, resulting in a matrix of MFCC coefficients of size 13×150. We reshaped the matrix of MFCC features to obtain a 1950 dimensional feature vector for each segment. We then apply PCA to reduce the dimensionality from 1950 to 50, while preserving 85% of the total power. For each seismic signal, we use the multilevel discrete wavelet transform (DWT) [Mallat, 1999] with 3 levels of the Daubechies wavelets [Daubechies, 1992, Mahmoodabadi et al., 2005, Sinha et al., 2005]. Then we apply PCA to reduce the dimensionality of wavelet coefficients to 200, while preserving 85% of the total power. The above procedures for preprocessing follows exactly from [Nguyen et al., 2011].

In Table 4.3 and 4.4, we compare the classification accuracy of our COM-MED methods with SVM-2K, CoLapSVM, MV-MED and single-view MED as baseline. Specifically, two out of six sensors are chosen to build two-view models. Table 4.3 compares the classifi-

Table 4.4: The classification accuracy (%) for two *heterogenous* views in **ARL-Footstep** dataset (The best one is in **bold**.)

	MED view i	MED view j	SVM-2K	CoLapSVM	MV-MED	COM-MED
Sensor $1,5$	71.4 ± 3.6	89.5 ± 2.4	86.5 ± 1.1	90.1 ± 1.5	91.5 ± 2.1	$\mathbf{96.2} \pm 2.1$
Sensor $1,6$	75.6 ± 3.5	85.1 ± 3.2	83.2 ± 2.0	91.0 ± 2.3	90.3 ± 2.4	$\mathbf{93.2} \pm 4.1$
Sensor $2,5$	73.4 ± 4.1	89.2 ± 5.3	90.1 ± 2.5	90.5 ± 3.6	90.8 ± 4.2	$\mathbf{94.5} \pm 3.8$
Sensor $2,6$	72.8 ± 6.1	87.2 ± 3.8	88.7 ± 2.3	91.1 ± 2.5	92.5 ± 2.7	$\mathbf{94.3} \pm 4.2$
Sensor $3,5$	56.5 ± 12.1	89.2 ± 2.8	77.6 ± 6.8	78.6 ± 1.5	79.8 ± 6.8	$\mathbf{81.5} \pm 2.0$
Sensor $3,6$	54.1 ± 15.2	87.5 ± 3.1	78.7 ± 7.5	79.3 ± 1.2	80.2 ± 5.6	$\mathbf{83.1} \pm 6.5$
Sensor $4,5$	52.9 ± 10.8	89.0 ± 3.0	72.1 ± 9.8	75.3 ± 5.3	76.2 ± 6.5	$\mathbf{84.5} \pm 1.8$
Sensor $4,6$	52.7 ± 12.5	87.0 ± 2.6	70.9 ± 8.5	72.8 ± 6.5	73.8 ± 5.6	$\mathbf{85.6} \pm 2.5$

cation accuracy for SVM-2K, CoLapSVM, MV-MED and COM-MED using *homogeneous* views, i.e., sensors are of the same type, while in Table 4.4, we compare these models using *heterogeneous* views, i.e. sensors are of different types. Note that it is also known from the data source that sensors $[3, 4]$ have more corruption than the other acoustic sensors $[1, 2]$ and the seismic sensors [5-6] are of high quality. Therefore, we drop the case $[3, 4]$, since there is no good sensor in this combination.

We see that our COM-MED outperforms SVM-2K, CoLapSVM and MV-MED in terms of the accuracy of classification and it improves over the single-view MED. This is partially due to the incorporation of both the decision and the confidence level of classification in the COM-MED model. The SVM-2K does not consider the confidence level explicitly and the model only optimizes the average decision of two classifiers by enforcing a common decision made on both views. The CoLapSVM learns multiple view-specific classifiers by assuming smoothness of classifiers on the underlying data manifold. Similar to SVM-2K, CoLapSVM still relies on the Euclidean distance as the measure of disagreement between view-specific classifiers, therefore it suffers from the view-corruption and multi-view inconsistency problem as well.

On the other hand, for homogeneous-view in Table 4.3 such as sensors $[1, 2]$, it is seen that the SVM-2K, CoLapSVM and MV-MED all improve over the single view method. This is because the corresponding samples from homogenous sensors can be viewed as following the same distribution. Co-regularization methods such as SVM-2K and CoLapSVM can then take advantages of extra information from the alternative view to improve the performance of single-view classifier. The proposed COM-MED method does account for this heterogeneity of distributions, thus improves over the other methods.

Table 4.3 and 4.4 also show the robustness of our COM-MED method, compared with

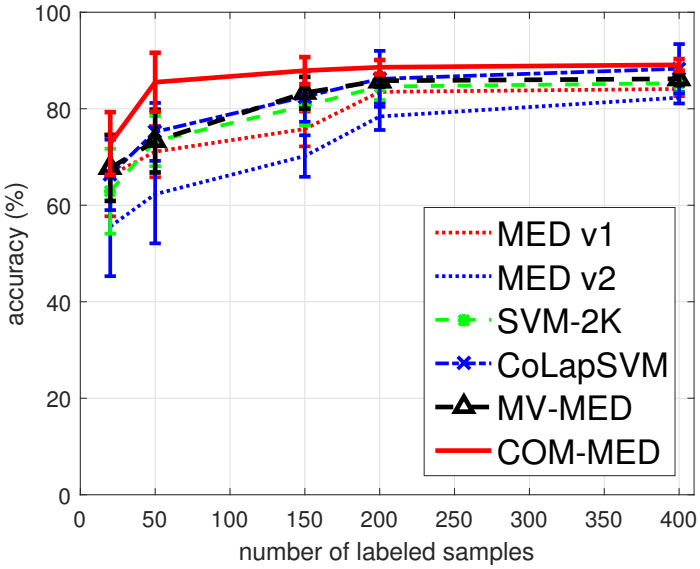


Figure 4.4: The classification accuracy vs. the size of labeled set for **ARL-Footstep** data set (Sensor 1,2). The proposed COM-MED outperforms MV-MED, CoLapSVM, SVM-2K and two single-view MEDs (view 1 and 2) and it has good stability when the number of labeled samples is small.

SVM-2K, CoLapSVM, MV-MED and single-view MED. For single-view MED, it is a large-margin classifier, which is known to be sensitive to corruption in the training sets [Xie et al., 2014]. In other words, including more sensors does not ensure better performance due to the intermittency in failed sensors and increased variance. For instance, compare the case of sensors $[1, 2]$ with sensors $[1, 4]$ and sensors $[2, 4]$, it is seen that if any sensor is of poor quality, neither SVM-2K, CoLapSVM nor MV-MED provides guarantee to improve over the best one-view classifier. This is due to the lack of robustness of the ℓ_2 consensus-measure used in both methods. Note that noisy measurement and the outliers may cause perturbations in output of the decisions, thus both the decision regularization and margin regularization are unreliable under this situation. On the other hand, COM-MED uses the stochastic consensus constraints that are insensitive to the data perturbations and outliers. Therefore it achieves superior performance in both accuracy and variance compared to single view classifiers and conventional co-regularization methods such as SVM-2K, CoLapSVM.

Fig. 4.4 shows the accuracy and the standard deviation for the four methods as the size of the labeled set increases. As more ground truth labels are used, the performances of all training methods increases, while COM-MED shows its superior performance consistently. On the other hand, it is seen from the plot that the relative performance gain of COM-MED is larger when the size of the weakly-labeled set are much larger than the fully-labeled

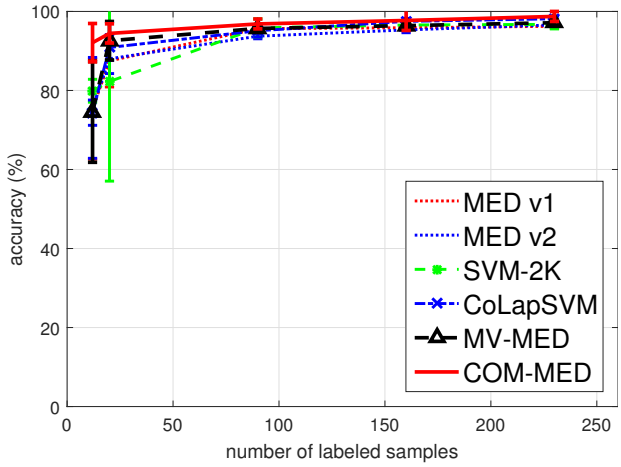


Figure 4.5: The classification accuracy vs. the size of labeled set for **WebKB4** data set. Unlike previous example, for this dataset, all multi-view learning algorithms are performing similarly well, although COM-MED still outperforms the rest. Note that **WebKB4** is the first dataset used by Co-training to demonstrate its success. It is a easy dataset for our task.

one. COM-MED is more suitable for large multi-view dataset with few annotations, since in this case, neither the conventional co-regularization such as SVM-2K and CoLapSVM or the heuristic-based MV-MED yields reliable and reasonably good initial estimate of the classifier.

4.5.2 Web-Page Classification

The **WebKB4** [Craven et al., 2000] data set is widely-used in multi-view learning literature [Blum and Mitchell, 1998, Sindhwani and Rosenberg, 2008]. It consists of 1051 two-view web pages collected from computer science department web sites at four universities. There are 230 course pages and 821 non-course pages. The two natural views are words in a web page and words appearing in the links pointing to that page. We follow the preprocessing step in [Sindhwani and Rosenberg, 2008], and extract a 3000-dimensional feature vector via the bag-of-words representation in the page view and a 1840-dimensional feature vector in the link view. Then we compute the term frequency-inverse document frequency weights (TF-IDF) features from the document word matrix. The feature vector is length normalized.

In Table 4.2, we see that our COM-MED has significantly better classification performance as compared to SVM-2K, CoLapSVM and MV-MED, when the labeled set is small, i.e., $|L| = 15$. Also, according to Fig. 4.5, when more labeled samples are included, all four methods have similarly good performance, even for the single-view MED. The COM-MED performs better with a few labeled samples because its stability relies on a good estimate of confidence on the unlabeled training samples, which is less affected by the amount of

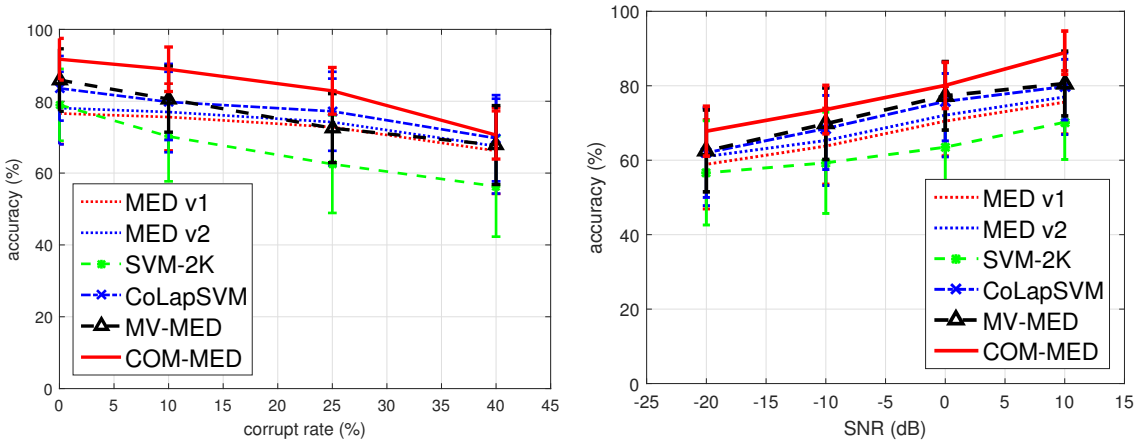


Figure 4.6: The classification accuracy vs. the corruption rate (%) for (a) **WebKB4** data set and (b) **Internet Ads** data set, where i.i.d. Gaussian random noise $\mathcal{N}(0, \sigma^2)$ is added in either of the two views. Here we choose the signal-to-noise ratio $SNR := \mathbb{E}\left[\|X\|^2\right]/\sigma^2 = 10$. Also, the classification accuracy vs. the $SNR(dB)$ (i.e. $10\log_{10}(SNR)$) with corruption rate 10% for (c) **WebKB4** data set and (d) **Internet Ads** data set. The proposed COM-MED outperforms MV-MED, CoLapSVM, SVM-2K and two single-view MEDs (view 1 and 2) and it is robust when both corrupt rate increases and SNR decreases.

the labeled training samples.

To evaluate the robustness of the single view MED, SVM-2K, CoLapSVM, MV-MED and COM-MED, we randomly add i.i.d. Gaussian noise $\mathcal{N}(0, \sigma^2)$ to a random subset of data in either of the two views. Each view is selected with equal probability. Note that since the underlying distribution of data are changed by noise corruption, there exists inconsistency among different views. Experiments are conducted under different *corrupt rate* and *signal-to-noise ratios (SNR)*. The former is defined as the percentage of corrupted data in dataset and the latter $SNR := \mathbb{E}\left[\|X\|^2\right]/\sigma^2$. Fig. 4.6 (a) and (c) show the classification accuracy vs. the corrupt rate (%) and SNR, respectively. It is seen that as the noise level and the corrupt rate increases, SVM-2K has the worst performance since it relies on the CCA, which is sensitive to the noise in each single view. By using the stochastic consensus constraint, the COM-MED outperforms CoLapSVM, SVM-2K and MV-MED in terms of reliability of its performance.

4.5.3 Internet Advertisement Classification

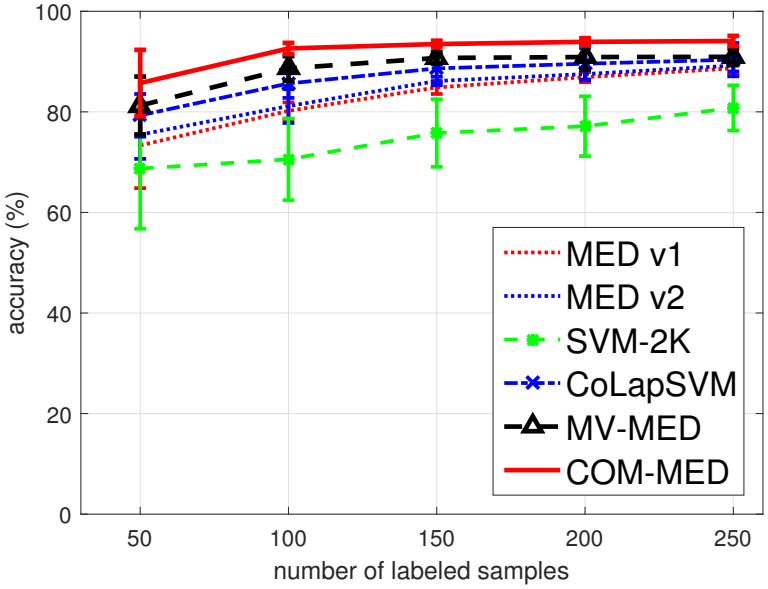
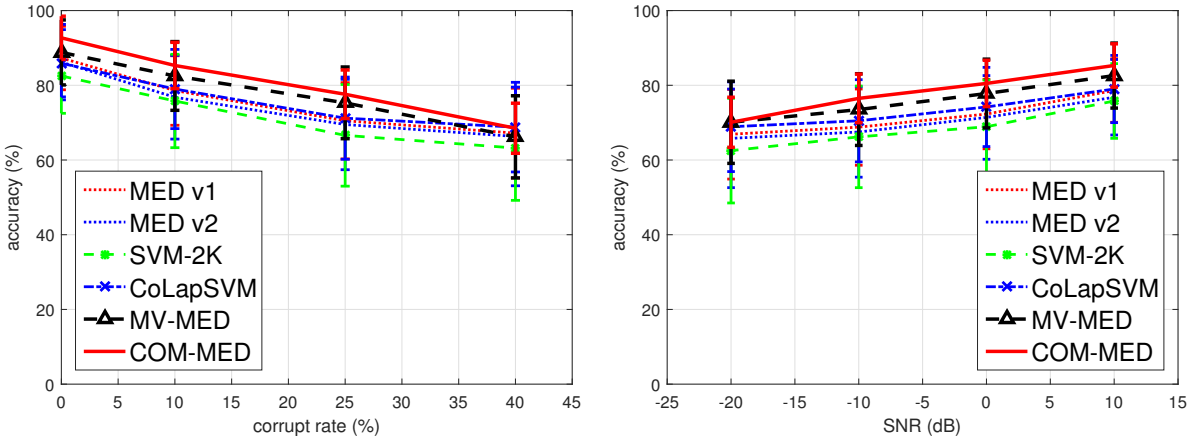


Figure 4.7: The classification accuracy vs. the size of labeled set for **Internet Ads** data set. Similar to above results, the proposed COM-MED outperforms MV-MED, CoLapSVM, SVM-2K and two single-view MEDs (view 1 and 2) and it has good stability when the number of labeled samples is small.



(a) (b)

Figure 4.8: The classification accuracy vs. the corruption rate (%) for (a) **WebKB4** data set and (b) **Internet Ads** data set, where i.i.d. Gaussian random noise $\mathcal{N}(0, \sigma^2)$ is added in either of the two views. Here we choose the signal-to-noise ratio $SNR := \mathbb{E}\left[\|X\|^2\right]/\sigma^2 = 10$. Also, the classification accuracy vs. the $SNR(dB)$ (i.e. $10\log_{10}(SNR)$) with corruption rate 10% for (c) **WebKB4** data set and (d) **Internet Ads** data set. The proposed COM-MED outperforms MV-MED, CoLapSVM, SVM-2K and two single-view MEDs (view 1 and 2) and it is robust when both corrupt rate increases and SNR decreases.

The **Internet Ads** [Kushmerick, 1999] data set consists of 3279 instances including 458 ads images and 2820 non-ads images. The first view describes the image itself, i.e., words in images' URL and caption, while the other view contains all other features, i.e., words from URLs of pages that contain the image and pages which the image points to. For each view, we extract the bag-of-words representations, which results in a $587-$dimensional vector in view 1 and a $967-$dimension vector in view 2. We set the size of training set as 600 and $|L| = 50$.

From Table 4.2 and Fig. 4.7, we see that our COM-MED still performs better than SVM-2K, MV-MED and single-view MED. It is seen that COM-MED is more stable as the size of the labeled training set increases, while SVM-2K has much worse stability performance. Also, similar to the experiment in **WebKB** dataset, we manually contaminated the **Internet Ads** datasets by randomly adding i.i.d. Gaussian noise. Adopting the same setting as the **WebKB** experiments, we see in Fig. 4.8 (a) and (b) that the COM-MED is more robust in the presence of noise corruption in single view, compared to single-view MED, CoLapSVM, SVM-2K and MV-MED.

4.6 Conclusion

In this chapter, we proposed a multi-view maximum entropy learning model on statistical manifolds via stochastic consensus constraints. In particular, the Kullback-Liebler divergence is used to measure the dissimilarity of information contents in different views. Experiments show that the proposed COM-MED method is robust in the presence of corruption and outliers and it achieves superior classification performance over other multi-view learning methods. A further improvement of COM-MED might be achieved by introducing a nonparametric Bayesian framework such as the Dirichlet process [Blei et al., 2006] to handle clusters in sample domain, e.g., [Zhu et al., 2014].

4.7 Acknowledge

This research was partially supported by US Army Research Office (ARO) grants W911NF-11-1-0391, WA11NF-11-1-103A1. The authors are grateful for the support of U.S. Army Research Lab. for providing the experimental platform to collect and process a collection multi-sensor data used in this paper.

4.8 Appendices

4.8.1 Result for consensus-view p.d.f. in (4.3)

Let $q(y|\boldsymbol{x}) \in \mathcal{M}$ be the consensus-view p.d.f.. From (4.3), for the KL-divergence,

$$q(y|\boldsymbol{x}) := \arg\min_{h(y|\boldsymbol{x})\in\mathcal{M}} \sum_{i=1}^{V} \mathbb{KL}\left(h(y|\boldsymbol{x}_m) \parallel p_i(y|\theta_m^i)\right).$$

Given $\{p_{\theta^i}\}$ where $p_{\theta^i} := p_i(y|\theta^i)$, we take the derivative with respect to h and let it to be zero, i.e.,

$$\begin{aligned}
&\frac{\partial}{\partial h(y|\boldsymbol{x})} \sum_{i=1}^{V} \sum_{y} h(y|\boldsymbol{x}) \log\left(\frac{h(y|\boldsymbol{x})}{p_i(y|\theta^i)}\right) + \lambda\left(\sum_{y} h(y|\boldsymbol{x}) - 1\right) \\
&= \sum_{i=1}^{V} h(y|\boldsymbol{x}) \left(\frac{h(y|\boldsymbol{x})}{p_i(y|\theta^i)}\right)^{-1} \frac{1}{p_i(y|\theta^i)} + \sum_{i=1}^{V} \log\left(\frac{h(y|\boldsymbol{x})}{p_i(y|\theta^i)}\right) \\
&\quad + \lambda = 0,
\end{aligned}$$

where $\lambda \geq 0$ is the dual variable for the normalization constraint.The optimal consensus posterior $q(y|x)$ is the function $h(y|x)$ that is the solution to the above equation. This yields:

$$\log q(y|\boldsymbol{x}) = \frac{1}{V} \sum_{i=1}^{V} \log p_i(y|\theta^i) - c(\boldsymbol{x}),$$

where $c(\boldsymbol{x})$ corresponds to the log-normalization factor. ∎

4.8.2 Approximation of the cross-entropy loss in (4.13)

The cross-entropy loss in (4.13) is

$$-\int \widehat{\mathbb{E}}_{m\in U}\left[\mathbb{E}_{q(y|\boldsymbol{x}_m)}\left[\log p_i(y|\theta_m^i)\right]\right] q(\Psi_i) d\Psi_i,$$

where

$$\begin{aligned}
\log p_i(y|\theta_m^i) &= \frac{1}{2} y\theta_m^i - \log\left(\exp\left(\frac{1}{2}\theta_m^i\right) + \exp\left(-\frac{1}{2}\theta_m^i\right)\right) \\
&= \frac{1}{2} y\theta_m^i - \log\left(2\cosh\left(\frac{1}{2}\theta_m^i\right)\right) \qquad (4.19) \\
\mathbb{E}_{q(y|\boldsymbol{x}_m)}\left[\log p_i(y|\theta_m^i)\right] &= \frac{1}{2} \bar{y}\theta_m^i - \log\left(2\cosh\left(\frac{1}{2}\theta_m^i\right)\right) \qquad (4.20)
\end{aligned}$$

for $\bar{y} = \mathbb{E}_{q(y|\boldsymbol{x}_m)}[y]$, $\cosh(x) = \frac{\exp(x)+\exp(-x)}{2}$. For the nonlinear part, the Taylor expansion at $\theta_m^i = \overline{\theta_m^i}$ gives

$$\begin{aligned}
&\log\left(2\cosh\left(\frac{1}{2}\theta_m^i\right)\right) = \log\left(2\cosh\left(\frac{1}{2}\overline{\theta_m^i}\right)\right) \\
&\quad + \frac{1}{2}\tanh(0.5\overline{\theta_m^i})\left(\theta_m^i - \overline{\theta_m^i}\right) + \frac{1}{8}R(\overline{\theta_m^i})\left(\theta_m^i - \overline{\theta_m^i}\right)^2 \\
&\quad + o(\left\|\theta_m^i - \overline{\theta_m^i}\right\|^3), \\
&= \mathbb{E}_{p_{\overline{\theta}_m^i}}[y]\,\theta_m^i + H(\overline{\theta}_m^i) + \frac{1}{8}R(\overline{\theta_m^i})\left(\theta_m^i - \overline{\theta_m^i}\right)^2 \\
&+ o(\left\|\theta_m^i - \overline{\theta_m^i}\right\|^3),
\end{aligned} \tag{4.21}$$

where $\overline{\theta_m^i} \in \Theta$ is a reference point, $H(\overline{\theta}_m^i) = -\mathbb{E}_{\overline{\theta}_m^i}[\log p_{\overline{\theta}_m^i}]$ is the entropy of $p_{\overline{\theta}_m^i}$, $R(\overline{\theta_m^i}) := (1 - |\mathbb{E}_{p_{\overline{\theta}_m^i}}[y]|^2) \in [0,1]$. Note that $\tanh(0.5\overline{\theta_m^i}) = \mathbb{E}_{p_{\overline{\theta}_m^i}}[y]$ is a sigmoid function as shown in Fig. 4.9. Substituting (4.21) into (4.20), we have

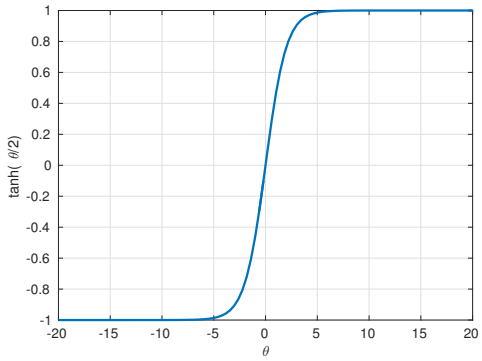


Figure 4.9: The function $\tanh(\theta/2)$

$$\begin{aligned}
(4.20) &= -H(\overline{\theta_m^i}) + \frac{1}{2}d(q,\overline{\theta_m^i})\theta_m^i - \frac{1}{8}R(\overline{\theta_m^i})\left(\theta_m^i - \overline{\theta_m^i}\right)^2 \\
&+ o(\left\|\theta_m^i - \overline{\theta_m^i}\right\|^3),
\end{aligned} \tag{4.22}$$

where $d(q,\overline{\theta_m^i}) = (\bar{y} - \mathbb{E}_{p_i}\left[y|\overline{\theta_m^i}\right])$. See that $\frac{1}{4}R(\overline{\theta_m^i}) \equiv -\nabla^2 \log p(y|\overline{\theta}_m^i) \approx 0$ for large $\left\|\overline{\theta_m^i}\right\|_2$, which makes (4.22) linear. ∎

4.8.3 Proof of theorem 4.4.1

For given $q \in \mathcal{M}$, the Lagrangian in view i is

$$\mathcal{F}\left(\{\theta_n^i\}_{n=1}^N, \{\xi_n^i\}_{n\in L}; \{\lambda_n^i\}_{n\in L}, \mu^i, \kappa\right)$$

$$
\begin{aligned}
&= \int q(\Psi_i) \log\left(\frac{q(\Psi_i)}{p_0(\Psi_i)}\right) + \int q(\Psi_i) \sum_{n\in L} \lambda_n^i \mathcal{L}_{i,n,\Psi_i} \\
&\quad + \int q(\Psi_i)\mu^i \left[\sum_{m\in U}\left(-\frac{1}{2}d_m\theta_m^i + \frac{1}{8}R_m((\theta_m^i)^2 - 2\overline{\theta_m^i}\theta_m^i)\right)\right. \\
&\quad \left. -\rho\right] + \kappa\left(\int q(\Psi_i) - 1\right)
\end{aligned}
$$

Taking the derivative with respect to $q(\Psi_i)$ and letting it to be zero, we have

$$
\begin{aligned}
\log q(\Psi_i) &= \log p_0(\Psi_i) - \sum_{n\in L}\lambda_n^i \mathcal{L}_{i,n,\Psi_i} \\
&\quad + \sum_{m\in U}\mu^i\left(\frac{1}{2}d_m\theta_m^i - \frac{1}{8}R_m((\theta_m^i)^2 - 2\overline{\theta_m^i}\theta_m^i) + \rho\right) \\
&\quad - \kappa - 1,
\end{aligned}
$$

which gives the Result 4.4.1. ∎

4.8.4 Proof of theorem 4.4.2

Proof: Following [Jaakkola et al., 1999], by strong duality, the optimal dual variables can be computed by

$$
\max_{\boldsymbol{\lambda}^i \succeq \mathbf{0}, \mu^i \geq 0} -\log Z(\boldsymbol{\mu}^i, \lambda^i)
$$

where Z is defined as (4.16). Under assumption (4.9)-(4.12), the objective function is computed as

$$
\begin{aligned}
&-\log\int_{\Psi_i} p_0(\Psi_i)\exp\left(-\sum_{n\in L}\lambda_n^i\xi_n^i + \sum_{n\in L}\lambda_n^i y_n\theta_n^i + \right. \\
&\left.\sum_{m\in U}\mu^i\left[\frac{1}{2}d_m + \frac{1}{4}R_m\overline{\theta_m^i}\right]\theta_m^i - \frac{1}{8}\mu^i\sum_{m\in U}R_m(\theta_m^i)^2\right) \\
&= \sum_{n\in L}\left(\lambda_n^i + \log\left(1 - \lambda_n^i/c_\xi^{(i)}\right)\right) \\
&\quad -\log\int_{\boldsymbol{\theta}_i}\exp\left(-\frac{1}{2}\boldsymbol{\theta}_i^T\boldsymbol{K}_i^{-1}\boldsymbol{\theta}_i + \boldsymbol{\theta}_i^T[\boldsymbol{\nu}(\boldsymbol{\lambda}^i,\mu^i)] - \frac{\mu^i}{8}\boldsymbol{\theta}_i^T\boldsymbol{Q}\boldsymbol{D}_{\boldsymbol{R}^i}\boldsymbol{Q}^T\boldsymbol{\theta}_i\right) \\
&= \sum_{n\in L}\left(\lambda_n^i + \log\left(1 - \lambda_n^i/c_\xi^{(i)}\right)\right) \\
&\quad -\frac{1}{2}\boldsymbol{\nu}(\boldsymbol{\lambda}^i,\mu^i)^T\left[\boldsymbol{K}_i^{-1} + \frac{\mu^i}{4}\boldsymbol{Q}\boldsymbol{D}_{\boldsymbol{R}^i}\boldsymbol{Q}^T\right]^{-1}\boldsymbol{\nu}(\boldsymbol{\lambda}^i,\mu^i)
\end{aligned}
$$

$$= \sum_{n \in L} \left(\lambda_n^i + \log\left(1 - \lambda_n^i / c_\xi^{(i)}\right)\right)$$
$$- \frac{1}{2} \boldsymbol{\nu}(\boldsymbol{\lambda}^i, \mu^i)^T \boldsymbol{A}_i^T \left[\boldsymbol{K}_i^{-1} + \frac{\mu^i}{4} \boldsymbol{Q} \boldsymbol{D}_{\boldsymbol{R}^i} \boldsymbol{Q}^T \right]^{-1} \boldsymbol{A}_i \boldsymbol{\nu}(\boldsymbol{\lambda}^i, \mu^i) \tag{4.23}$$

where $\boldsymbol{d}^i := (d_m^i)_{m \in U}$, $\boldsymbol{R}^i := (R_m^i)_{m \in U}$, and $\Delta \boldsymbol{y}^i := (\frac{1}{2}\boldsymbol{d}^i + \frac{1}{4}\boldsymbol{R}^i \odot \overline{\boldsymbol{\theta}_i})$, $\boldsymbol{\nu}(\boldsymbol{\lambda}^i, \mu^i) = \left[(\boldsymbol{\lambda}^i \odot \boldsymbol{y})^T, \mu^i\right]^T \in \mathbb{R}^{|L|+1}$,

$$\boldsymbol{A}_i := \begin{bmatrix} \boldsymbol{I}_{|L|} & \boldsymbol{0} \\ \boldsymbol{0} & \Delta \boldsymbol{y}^i \end{bmatrix} \in \mathbb{R}^{N \times (|L|+1)}, \ \boldsymbol{Q} := \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_{|U|} \end{bmatrix} \in \mathbb{R}^{N \times |U|},$$
$$\boldsymbol{D}_{\boldsymbol{R}^i} = \frac{1}{4} \begin{bmatrix} R_1^i & & \\ & \ddots & \\ & & R_{|U|}^i \end{bmatrix}$$

and $\odot$ is the point-wise matrix product. Note that by the Matrix inversion lemma,

$$\left[\boldsymbol{K}_i^{-1} + \frac{\mu^i}{4} \boldsymbol{Q} \boldsymbol{D}_{\boldsymbol{R}^i} \boldsymbol{Q}^T \right]^{-1}$$
$$= \boldsymbol{K}_i - \boldsymbol{K}_i \boldsymbol{Q} \left(\frac{4}{\mu^i} \boldsymbol{D}_{\boldsymbol{R}^i}^{-1} + \boldsymbol{Q}^T \boldsymbol{K}_i \boldsymbol{Q} \right)^{-1} \boldsymbol{Q}^T \boldsymbol{K}_i$$
$$= \boldsymbol{K}_i - \boldsymbol{K}_i \boldsymbol{Q} \left(\frac{4}{\mu^i} \boldsymbol{D}_{\boldsymbol{R}^i}^{-1} + (\boldsymbol{K}_i)_{uu} \right)^{-1} \boldsymbol{Q}^T \boldsymbol{K}_i$$

where

$$\boldsymbol{K}_i \boldsymbol{Q} = \begin{bmatrix} (\boldsymbol{K}_i)_{lu} \\ (\boldsymbol{K}_i)_{uu} \end{bmatrix}, \ \boldsymbol{Q}^T \boldsymbol{K}_i \boldsymbol{Q} = (\boldsymbol{K}_i)_{uu}$$
$$\boldsymbol{A}_i^T \boldsymbol{K}_i \boldsymbol{Q} = \begin{bmatrix} (\boldsymbol{K}_i)_{lu} \\ (\Delta \boldsymbol{y}^i)^T (\boldsymbol{K}_i)_{uu} \end{bmatrix}$$

As discussed in Appendix 4.8.2, when $\boldsymbol{\theta}_i$ deviates from the reference $\overline{\boldsymbol{\theta}_i}$, then the perturbation factor $\boldsymbol{R} \to \boldsymbol{0}$ and the kernel matrix becomes $\boldsymbol{K}_i$, which implies that the unlabeled samples are of no use since the classifier is already good enough. The dual variable μ^i has a shrinkage effect on the perturbation term when μ^i is small and when μ^i is large, we can drop it. In the formulation (4.17), we replace $(\mu^i)^{-1}$ by a constant shrinkage factor $\alpha > 0$.

Finally, given $(\boldsymbol{\lambda}^i, \mu^i)$, the mean $\mathbb{E}_{q(\Psi_i)}[\theta_m^i]$ is computed using the property of the Gaus-

sian process, i.e.

$$\sum_{n\in L} y_n \lambda_n^i \widetilde{K}_i(\boldsymbol{x}_n^i, \boldsymbol{x}_s^i) + \mu^i \sum_{m\in U} \Delta\bar{y}_m^i \widetilde{K}_i(\boldsymbol{x}_m^i, \boldsymbol{x}_s^i),$$

where $\widetilde{K}_i(\cdot,\cdot)$ is the modified kernel function in (4.17) and $\Delta\bar{y}_m^i := \frac{1}{2}d_m^i + \frac{1}{4}R_m^i\overline{\theta}_m^i, m \in U$ with d_m^i, R_m^i as above. This completes the proof.

CHAPTER 5

Collaborative Network Topology Learning from Partially Observed Relational Data

5.1 Introduction

Learning a dependency graph $\mathcal{G}$ given relational data $\boldsymbol{x}$ is an important task for natural language processing [Jurafsky and Martin, 2014]; sensor networks [Hall and Llinas, 1997]; recommendation systems [Aggarwal et al., 1999] and artificial intelligence [Ferber, 1999]. In many situations, however, a learner only has access to a limited amount of data, while the rest of data are either missing or protected by the system due to the privacy or security concerns. For instance, recommendation system of a company has proprietary information regarding its own registered customers, who may be influenced by a number of agents, including other customers of the company but also including other people whose information is not accessible to the company. Without such external information confounding marginal correlations may exist between customers who are conditionally uncorrelated. In this case, the conditional dependencies, specifically, partial correlations, between the customers may be more accurately estimated if information about the partial correlations of the non-customers is available. Similarly, in a sensor network with limited power budget, a subset of sensors that were actively collecting data in the recent past may have gone into sleeping mode. A processor thus only has measurements from the active sensors at the current time as well as a possible information summary (e.g., spatial correlation) of the sleeping nodes based on their recent past data. The spatial partial correlations of the sleeping network may be used to better estimate the partial correlations between the awake sensors.

In each of these scenarios, the inaccessible (*latent* [1]) data have influence on the dependency structure underling the accessible (*observed*) ones. It may therefore be advantageous

[1] In this chapter, the words *latent* and *inaccessible* are used interchangeably. Similar for the words *observed* and *accessible*.

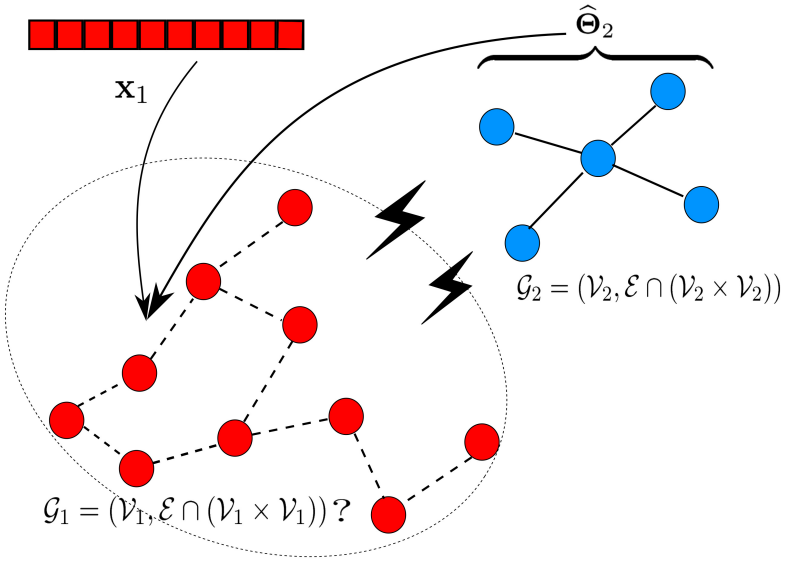


Figure 5.1: An illustration of the problem of *learning sub-network toplogy from partially observed data*. The red rectanges represent observed data $\boldsymbol{x}_1$, which is a subset of full data $\boldsymbol{x}$. The red vertices are affected by the blue vertices through some unknown links. Data on the blue vertices are not observed directly but a noisy summary $\widehat{\mathbf{\Theta}}_2$ regarding their relationship graph $\mathcal{G}_2$ is given. The task is to infer the unknown edges of subnetwork $\mathcal{G}_1$ from partially observed data $\boldsymbol{x}_1$ in $\mathcal{G}_1$ and a summary $\widehat{\mathbf{\Theta}}_2$ of $\mathcal{G}_2$.

to collaborate with external sources in order to improve the performance of a graph learning system on the observed dataset. In this chapter, we consider a situation where in addition to an observed subset of data $\boldsymbol{x}_1$, the learner may receive a noisy summary of the partial correlations of latent data from external sources. These partial correlations are, up to a constant, specified by the inverse covariance matrix, denoted by $\widehat{\mathbf{\Theta}}_2$. The task is to learn the dependency sub-network $\mathcal{G}_1$ among $\boldsymbol{x}_1$ collaboratively given asymmetric information from two sources.

Learning graph topology from data may be an ill-conditioned problem in the sense that there may be insufficient data to accurately determine the topology - there is high sensitivity of the topology estimate to small variations in the data. Regularization is commonly used to address ill-conditioned problems. For example, in the graph signal processing GSP [Zhu and Rabbat, 2012, Narang et al., 2013a, Sandryhaila and Moura, 2013, 2014b, Shuman et al., 2013], it is assumed that the data $\boldsymbol{x}$ is smooth over the underlying graph $\mathcal{G}$ in the sense that its graph Fourier transform is band-limited. That is, for each $v \in \mathcal{V}$, the datum $\boldsymbol{x}_v$ is similar to its nearest neighbors in $\mathcal{G}$. However, in the case that the signals obey a Gaussian graphical model the marginal distribution of observed variables in a sub-graph may not

be smooth relative to the graph Fourier transform that is based on the eigenspectrum of the graph Laplacian over the entire network. As a result, the task of learning sub-network topology from partially observed data is challenging, especially for GSP. In this chapter, we focus on a random graph signal that follows a Gaussian graphical model [Lauritzen, 1996, Koller and Friedman, 2009, Zhang et al., 2015]. Specifically, the joint distribution is multivariate Gaussian that factorizes according to an unknown network $\mathcal{G}$. The network $\mathcal{G}$ is partitioned into the target network $\mathcal{G}_1$ and the external network $\mathcal{G}_2$, and the observed data $\boldsymbol{x}_1$ follow a marginal Gaussian distribution. See Figure 5.3 for an illustration.

Given complete observations over all nodes of the graph, learning the Gaussian graphical model can be solved efficiently via sparse inverse covariance estimation [Lauritzen, 1996, Rue and Held, 2005, Banerjee et al., 2008, Friedman et al., 2008, Rothman et al., 2008, Wainwright et al., 2008, Yuan, 2010, Chen et al., 2011, Pavez and Ortega, 2016]. However, these methods have difficulties dealing with partially observed data. This is due to the effect of marginalization [Koller and Friedman, 2009], which introduces phantom dependency edges between vertices that connects to common latent variables. In other words, due to the existence of latent factors, the marginal precision matrix (or, inverse of marginal covariance matrix) defines a dependency structure that is equivalent to the corresponding subgraph of the Gaussian graphical model. To take into account of this effect, the LV-GGM was introduced by Chandrasekaran et al. [Chandrasekaran et al., 2011, 2012] and extended to learning latent variable precision matrices by [Meng et al., 2014]. It summarizes the effect of marginalization as a dense low-rank matrix. The LV-GGM then effectively separates out this low-rank matrix from the marginal precision matrix by solving a semi-definite programming (SDP) problem, resulting in a sparse matrix whose support set coincides with the edge set of the the underling sub-network. Despite its success, LV-GGM has two limitations: First, it assumes that the effect of latent variables is *global* and there exist edges between each latent vertex and the observed vertex set. Second, LV-GGM does not fully utilize partial information that may be available on the latent variables. In many applications, a noisy summary $\widehat{\boldsymbol{\Theta}}_2$ of correlation structure between the latent variables is available. Therefore it is plausible for us to improve over LV-GGM given this additional relevant information.

In this chapter, we consider the situation where the influence of each latent variable decayed outside a neighborhood so that only a small potion of them have influence on topology of the target network $\mathcal{G}_1$. This occurs in applications such as the spatial correlation analysis of sensor networks [Jindal and Psounis, 2004, Vuran et al., 2004, Jindal and Psounis, 2006, Dai and Akyildiz, 2009], field estimation [Nowak et al., 2004], image clustering [Kumar and Hebert, 2003] and geographical data analysis [Mai and Beroza, 2002].

Specifically, the proposed *Decayed-influence Latent variable Gaussian Graphical Model (**DiLat-GGM**)* generalizes the LV-GGM by incorporating the dependence structure between latent variables. In company customer example above, the decayed influence model assumes that the behavior of each individual customer is affected only by people who have direct friendship relationship with him/her. Similarly, in sensor network examples above, it is assumed that the spatial correlation of measurements between two sensors decays drastically with respect to their relative distance. Besides the decayed-influence assumption, the proposed DiLat-GGM also includes a latent feature selection procedure by introducing additional row sparsity structure on the conditional cross-covariance matrix. From a network perspective, it induces a topology with sparse inter-connection between the target network and the external network. The full network $\mathcal{G}$ thus resembles a network with block structure, which is common in social networks, sensor networks and other distributed systems [Barthelemy, 2004, Newman, 2005, Brandes, 2008, Jackson, 2010]. From a multi-view learning perspective, DiLat-GGM can be seen as a *two-view learning system* given asymmetric information flow from both the internal view and the external view.

What follows is an outline of the chapter: In Section **5.2**, we review and discuss the network topology inference with a Gaussian graphical model. In Section **5.2.2**, we deal with the case when full information from a single source is available and in Section **5.2.3**, we consider the situation where only a subset of data from a single source is accessible. These lead to solutions based on the graphical Lasso and the LV-GGM, respectively. The DiLat-GGM method as a generalization of LV-GGM is proposed in Section **5.2.4**. In Section **5.3**, an sparsity constrained precision matrix estimation algorithm based on convex-concave programming is introduced. Experimental results based on synthetic data and real data are presented in Section **5.4**. Our conclusions are given in Section **5.5**.

5.2 Problem Formulation

5.2.1 Notation and Preliminaries

Consider an undirected network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the vertex set and $\mathcal{E}$ is the edge set. $|\mathcal{V}| = n$. Define the $\boldsymbol{A} = [a_{i,j}] \in \mathbb{R}^{n\times n}$ as the adjacency matrix. For simplicity, assume that $\mathcal{G}$ is unweighted. The normalized graph Laplacian matrix $\boldsymbol{L} := \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$, where $\boldsymbol{D}$ is a diagonal matrix with diagonal entries $d_{i,i} = \sum_{j\in\mathcal{V}} a_{i,j}$. Each vertex of the graph is associated with m i.i.d samples $\boldsymbol{x}_v \in \mathbb{R}^m$. Denote $\boldsymbol{X} = [\boldsymbol{x}_v]_{v\in\mathcal{V}} \in \mathbb{R}^{n\times m}$ with $\boldsymbol{x}_v$ as its v-th row. Let $\boldsymbol{x} = [x_1, \ldots, x_n]$ be the random vector whose i.i.d realizations correspond to columns $\{\boldsymbol{X}^{(j)}\}_{j=1}^m$ of $\boldsymbol{X}$. We assume that each row $\boldsymbol{x}$ of $\boldsymbol{X}$ obeys a Gaussian graphical

model faithful to $\mathcal{G}$ [Lauritzen, 1996]. Specifically, $\boldsymbol{x}$ is n-variate Gaussian distributed with zero mean and covariance $\boldsymbol{\Sigma}$ where the inverse covariance, or precision, matrix $\boldsymbol{\Theta}$ is sparse with non-zero entries corresponding to the location of edges in $\mathcal{G}$. We use the shorthand notation $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}^{-1})$, where $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1} \in \mathbb{R}^{n \times n}$ denotes the inverse of covariance matrix $\boldsymbol{\Sigma}$, or, the precision matrix. Assume that the Gaussian graphical model (GGM) $\mathcal{N}(\mathbf{0}, \boldsymbol{\Theta})$ factorizes according to $\mathcal{G}$, i.e., in the condition independence notation of [Lauritzen, 1996], $x_i \perp\!\!\!\perp x_j | \boldsymbol{x} - \{x_i, x_j\} \Leftrightarrow (i,j) \notin \mathcal{E}$. Thus the support of off-diagonal entries $\{(i,j) \in \mathcal{V} \times \mathcal{V} : \boldsymbol{\Theta}_{i,j} \neq 0, i \neq j\}$ is equal to $\mathcal{E}$.

Let the n vertices of $\mathcal{G}$ be partitioned into two sets $\mathcal{V}_1$ and $\mathcal{V}_2$ of cardinalities $|\mathcal{V}_1| = n_1$ and $|\mathcal{V}_2| = n_2$, respectively, i.e., $\mathcal{V}_1 \cap \mathcal{V}_2 = \emptyset$ and $\mathcal{V}_1 \cup \mathcal{V}_2 = \mathcal{V}$. Define the corresponding $n_1 \times n_1$ and $n_2 \times n_2$ sub-matrices $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_{\mathcal{V}_1}$ and $\boldsymbol{\Theta}_2 = \boldsymbol{\Theta}_{\mathcal{V}_2}$ of the precision matrix $\boldsymbol{\Theta}$. Likewise, let $\mathcal{E}_1$ and $\mathcal{E}_2$ denote the edges in $\mathcal{G}$ associated with the partition, i.e., $\mathcal{E}_i$ are the edges in $\mathcal{E}$ that connect vertices in $\mathcal{V}_i$, i=1,2. The sub-graph $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ is called the *target* sub-network while $\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2)$ is called the *external* network.

The learning problem we consider here is to infer $\boldsymbol{\Theta}_1$, and in particular the edges in the graph $\mathcal{G}_1$, given measurements of only the subset$\mathcal{V}_1$ of the vertices of the complete graph $\mathcal{G}$, i.e., estimate $\boldsymbol{\Theta}_1$ when only measurements from the target network $\boldsymbol{x}_{\mathcal{V}_1}$ are available. It is well known that, unless there are no edges in $\mathcal{G}$ connecting $\mathcal{V}_1$ and $\mathcal{V}_2$, in which case $\boldsymbol{\Theta}$ is block diagonal, unbiased estimation of $\boldsymbol{\Theta}_1$ from partial measurements $\boldsymbol{x}_{\mathcal{V}_1}$ is not possible [Lauritzen, 1996, Koller and Friedman, 2009, Wiesel et al., 2010]. This is because the precision matrix of $\boldsymbol{x}_{\mathcal{V}_1}$ is not necessarily equal to $\boldsymbol{\Theta}_1$, unless $\boldsymbol{\Theta}$ is block diagonal. Indeed, the marginal distribution of target variables $\boldsymbol{x}_{\mathcal{V}_1}$ contains *phantom* edges not in subgraph $\mathcal{G}_1$ that are due to marginalization over unobserved external variables $\boldsymbol{x}_{\mathcal{V}_2}$ in the larger graph $\mathcal{G}$. Marginalization has in effect converted these unobserved variables into *latent variables* creating phantom edges. The latent data $\boldsymbol{x}_2 := [x_v]_{v \in \mathcal{V}_2}$ are inaccessible to the learner, but a noisy summary $\widehat{\boldsymbol{\Theta}}_2 \in \mathbb{R}^{n_2 \times n_2}$ of their dependency structure $\mathcal{G}_2$ is provided by an external source. The task of the learner is to learn $\mathcal{G}_1$ given $\boldsymbol{x}_{\mathcal{V}_1}$ and $\widehat{\boldsymbol{\Theta}}_2$, where $\widehat{\boldsymbol{\Theta}}_2 \succ \mathbf{0}$ has the representation

$$\widehat{\boldsymbol{\Theta}}_2 = \widehat{\boldsymbol{L}}_2 + \sigma_L^2 \boldsymbol{G} \tag{5.1}$$

where $\widehat{\boldsymbol{L}}_2$ is an estimate of inverse covariance matrix over $\boldsymbol{x}_2$, $\sigma_L > 0$ and $\boldsymbol{G} = \frac{1}{n_2^2} \boldsymbol{H}\boldsymbol{H}^T$ is a Gram matrix generated by Gaussian random matrix $\boldsymbol{H} \in \mathbb{R}^{n_2 \times n_2}$ with $\boldsymbol{H}_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. It will be assumed that the $\widehat{\boldsymbol{\Theta}}_2$ is statistically independent of the target data $\boldsymbol{x}_{\mathcal{V}_1}$.

The representation (5.1) can be motivated by the following social networking example. In a certain market there are $n = n_1 + n_2$ active customers connected by friendship

network $\mathcal{G}$ and whose behaviors are random variables $\boldsymbol{x}$ that follow a GGM that is faithful to the friendship graph $\mathcal{G}$. Company A and company B have exclusive access to the behaviors of customers $\mathcal{V}_1$ and $\mathcal{V}_2$, respectively. Each company is trying to infer their respective customer's friendship networks $\mathcal{G}_1$ and $\mathcal{G}_2$ from observed customer behaviors $\boldsymbol{x}_{\mathcal{V}_1}$ and $\boldsymbol{x}_{\mathcal{V}_2}$, respectively. As the companies are in competition they do not want to share their raw behavior data but they do want to improve their ability to learn about their own customer's friendship networks by sharing summary information. In particular, Company A is interested in learning $\mathcal{G}_1$ from $\boldsymbol{x}_{\mathcal{V}_1}$ and Company B is willing to share (or sell) a noisy summary of its own customer behaviors in the form of (5.1). In this setting, $\widehat{\boldsymbol{L}}_2$ corresponds to Company B's empirical estimate of its customer's precision matrix. $\boldsymbol{G}$ corresponds to a Wishart noise of level σ_L^2 that is added to $\widehat{\boldsymbol{L}}_2$ in order to preserve the IP of Company B or the privacy of its customers. If the empirical precision matrix estimate $\boldsymbol{G}$ is constructed from Company B's archival data, collected in the distant past, then $\widehat{\boldsymbol{\Theta}}_2$ will be statistically independent of Company A's current data.

We use the standard O-notation and Ω-notation [Cormen, 2009]: $f(n) = O(g(n))$ if $f(n) \leq cg(n)$ for some constant $c < \infty$; $f(n) = \Omega(g(n))$ if $f(n) \geq c'g(n)$ for some constant $c' > 0$.

5.2.2 Inference Network Topology with Full Data

When the learner has access to all the vertices in the network there have been many approaches proposed for estimating the precision matrix $\boldsymbol{\Theta}$ of the GGM. In [Meinshausen and Bühlmann, 2006] a lasso regression approach was proposed where the measurements at each node are regressed onto the measurements of all other nodes, with a sparsity constraint to determine nearest neighbors. In [Marjanovic and Hero, 2015] an ℓ_0-penalized maximum likelihood approach was taken using coordinate ascent optimization. This method extended the ℓ_1-penalized maximum likelihood approach of [Yuan and Lin, 2007, d'Aspremont et al., 2008, Friedman et al., 2008, Yuan, 2010], which maximizes the ℓ_1-regularized log-likelihood function. Note that maximizing the likelihood function is equivalent to minimizing the KL-divergence between the model distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}^{-1})$ and the empirical distribution. In particular, the maximum penalized likelihood estimator has the representation

$$\begin{aligned}\widehat{\boldsymbol{\Theta}} &= \arg\min_{\boldsymbol{\Theta}\succeq\mathbf{0}} \ \mathbb{KL}\left(\widehat{p}(\boldsymbol{x}) \,\|\, \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}^{-1})\right) + \alpha_m \|\boldsymbol{\Theta}\|_1 \\ &= \arg\min_{\boldsymbol{\Theta}\succeq\mathbf{0}} \ -\log\det\boldsymbol{\Theta} + \mathrm{tr}\left(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}\right) + \alpha_m \|\boldsymbol{\Theta}\|_1 , \end{aligned} \tag{5.2}$$

where $\mathbb{KL}\left(p \,\|\, q\right) = \int p \log \frac{p}{q}$ is the KL-divergence, $\widehat{p}(\boldsymbol{x}) = \sum_m \delta_{\boldsymbol{x}_m}(\boldsymbol{x})$ is the empirical distribution from data set $\boldsymbol{X} := \{\boldsymbol{x}\}_m$, $\widehat{\boldsymbol{\Sigma}} := \boldsymbol{X}\boldsymbol{X}^T/m$ is the $n \times n$ sample covariance matrix of $\boldsymbol{X}$, $\alpha_m = O(\sqrt{\frac{\log(n)}{m}}) > 0$ is a regularization parameter that depends on the number of vertices n and the number of samples m. The support set of $\widehat{\boldsymbol{\Theta}}$ then corresponds to an estimate of the edge set $\mathcal{E}$. Problem (5.2) is convex and many efficient algorithms have been proposed to solve it [Wang et al., 2010, d'Aspremont et al., 2008, Duchi et al., 2008, Friedman et al., 2008, Yuan, 2010, Hsieh et al., 2011, 2013, 2014, Treister and Turek, 2014]. For instance, the graphical Lasso and its variants [Friedman et al., 2008, Mazumder and Hastie, 2012] are popular algorithms. In [Pavez and Ortega, 2016] , the authors considered an extended framework to learn a generalized Laplaican matrix [Biyikoğu et al., 2007], which is suitable for graph signal analysis. In [Ravikumar et al., 2008], an ℓ_0-penalized version of (5.2) was proposed. It is shown that, if $m = \Omega(d^2 \log(n))$ with maximal vertex degree $d > 0$, high-dimensional consistency of edge recovery using $\widehat{\boldsymbol{\Theta}}$ can be achieved, under incoherence conditions.

If only a marginal covariance matrix $\widehat{\boldsymbol{\Sigma}}_1 := \boldsymbol{X}_1\boldsymbol{X}_1^T/m$ is available for a subset $\mathcal{V}_1$ of the vertices, inverse covariance estimation using (5.2) does not guarantee recovery of the underlying sub-network $\mathcal{G}_1$, due to the inclusion of phantom edges in the network after marginalization, as explained above. If there exist edges between $\mathcal{V}_1$ and the remaining vertices $\mathcal{V}_2 = \mathcal{V} - \mathcal{V}_1$, then $\mathcal{V}_2$ introduce hidden factors that globally affect the observed data. In this situation, the true marginal precision matrix may not even be sparse, and using a sparse GGM to represent the marginal distribution $p(\boldsymbol{x}_1)$ may lead to severe bias.

5.2.3 Sub-network Inference via Latent Variable Gaussian Graphical Model

To interpret for the global effect of latent variables explicitly, one can use a partitioned matrix inverse identity [Petersen and Pedersen, 2012] that has been previously used to elucidate local vs global views of GGM's [Meng et al., 2014]

$$\widetilde{\boldsymbol{\Theta}}_1 := \left(\boldsymbol{\Sigma}_1\right)^{-1} = \boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_{12}\left(\boldsymbol{\Theta}_2\right)^{-1}\boldsymbol{\Theta}_{21}, \tag{5.3}$$

where $\widetilde{\boldsymbol{\Theta}}_1$ is the marginal precision matrix over $\boldsymbol{x}_1$, $\boldsymbol{\Theta}_1$ is the principal submatrix of the global precision matrix $\boldsymbol{\Theta}$ over $\mathcal{V}_1$. Similarly, $\boldsymbol{\Theta}_{12}$ and $\boldsymbol{\Theta}_2$ are sub-blocks of full precision matrix $\boldsymbol{\Theta}$ for $\mathcal{V}_1 \times \mathcal{V}_2$ and $\mathcal{V}_2 \times \mathcal{V}_2$. It is seen from (5.3) that the marginal precision matrix consists of two terms: the first term is the inverse of conditional covariance matrix $\boldsymbol{\Theta}_1 = \left(\boldsymbol{\Sigma}_{1|2}\right)^{-1}$ and it is sparse. The second term characterizes the effect of marginalization and

it is low-rank for $|\mathcal{V}_2| = n_2 < n_1$. Therefore, according to (5.3), there exists a sparse plus low-rank separation for the marginal precision matrix. Also it is seen that the support of $\boldsymbol{\Theta}_1$ coincides with the edge set of $\mathcal{G}_1$, since $P(\boldsymbol{x}_1|\boldsymbol{x}_2;\boldsymbol{\Theta}_1)$ factorizes over $\mathcal{G}_1$.

Chandrasekaran et al. [Chandrasekaran et al., 2011, 2012] introduced the latent variable Gaussian graphical model (LV-GGM), which explicitly finds such separation $(\widehat{\boldsymbol{C}}, \widehat{\boldsymbol{M}})$ while maximizing the regularized marginal log-likelihood

$$\begin{aligned} \min_{\boldsymbol{C},\boldsymbol{M}} \quad & -\log\det(\boldsymbol{C}-\boldsymbol{M}) + \operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}_1(\boldsymbol{C}-\boldsymbol{M})\right) + \alpha_m \|\boldsymbol{C}\|_1 + \beta_m \|\boldsymbol{M}\|_* \\ \text{s.t.} \quad & \boldsymbol{C}-\boldsymbol{M} \succeq \boldsymbol{0} \\ & \boldsymbol{M} \succeq \boldsymbol{0}, \end{aligned} \tag{5.4}$$

where $\widehat{\boldsymbol{\Sigma}}_1$ is the sample marginal covariance, $\alpha_m = O(\sqrt{\frac{\log(n)}{m}}) > 0$ and $\beta_m = O(\|\boldsymbol{\Sigma}_1\|_2 \sqrt{\frac{n}{m}})$ are regularization parameters for the ℓ_1-norm and the nuclear-norm, respectively. By solving (5.4), LV-GGM finds an estimate of marginal precision matrix $\boldsymbol{R}$ where $\boldsymbol{R} := \widehat{\boldsymbol{C}} - \widehat{\boldsymbol{M}}$, for which $\widehat{\boldsymbol{C}}$ is sparse due to the ℓ_1-norm regularizer and $\widehat{\boldsymbol{M}}$ is low-rank due to the nuclear-norm regularizer. In [Chandrasekaran et al., 2012], it is proved that, given that some identifiability conditions hold and $m = \Omega(d^4\, n)$ for maximal vertex degree d, the support of estimated sparse matrix $\widehat{\boldsymbol{C}}$ equals to the support of conditional precision matrix $\boldsymbol{\Theta}_1$ with high probability. It then provides a theoretical guarantee for edge recovery of $\mathcal{G}_1$ using marginalized data. Note that the identifiability condition requires that the low-rank matrix is *dense and incoherent*. In other word, the effect of latent variables due to marginalization must not be confused with the dependency structure in $p(\boldsymbol{x}_1|\boldsymbol{x}_2;\boldsymbol{\Theta}_1)$. In [Meng et al., 2014], the authors proved Frobenius norm error bounds for estimating the precision matrix of an LV-GGM under weaker conditions than [Chandrasekaran et al., 2011, 2012].

Solving (5.4) requires solving a semi-definite program, which is slow. Efficient implementations based on LogdetPPA [Wang et al., 2010], the ADMM [Ma et al., 2013] and AltGD [Xu et al., 2017] have been proposed.

The disadvantages of the LV-GGM are two-fold: First, it cannot be used to infer the latent variables. In fact, LV-GGM has no knowledge on the latent variables directly except that their size is smaller than the observed ones. Second, the incoherence of low-rank matrix $\boldsymbol{M}$ implies that the influence of each latent variable is uniform and global. In network perspective, it means that every latent variable has direct influence on the observed data, regardless its network geodesic distance to observed vertices. This is overcomplicated, especially when the size of latent vertex set is greater than that of the observed vertex set but only a small potion of them are effective in inference of $\mathcal{G}_1$. In Figure 5.2 (b), it is seen than

LV-GGM implicitly assumes that there exist dense interactions between the observed vertices and the latent vertices, while there is no interaction between latent vertices. This does no fit the ground truth network in Figure 5.2 (a), which has sparser interactions between the observed vertices and the latent vertices, and has interactions between latent vertices. This motivates us to find a localized latent variable model in Section 5.2.4.

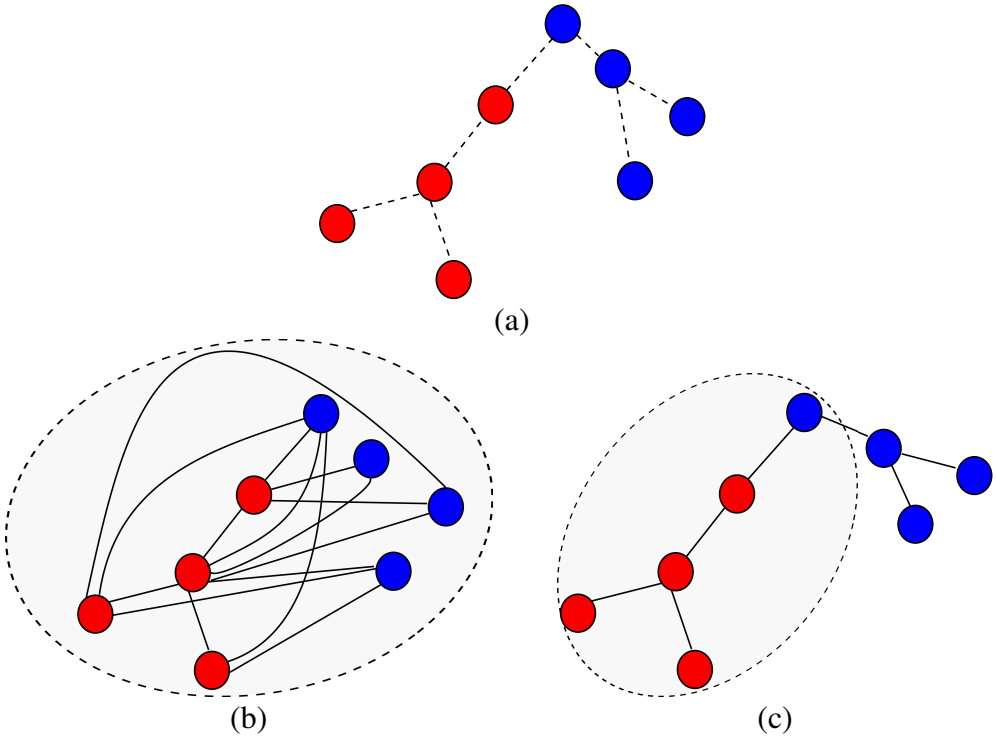


Figure 5.2: (a) A network-structured dataset. Data on red vertice are observed and data on the blue vertex are not. The dashed edges represent the underlying unknown network. (b) The global influence model for the LV-GGM. Note that every latent variable has at least one direct link to the observed dataset and there is no direct interactions between latent variables. The shaded region is the neighborhood $\mathcal{N}(\mathcal{V}_1)$ of observed vertices, which indicates that all latent variables have an effect in inference of the sub-network. (c) The decayed influence model. Only latent vertices within a local neighborhood (shaded region) have influence in inference of the sub-network.

5.2.4 Sub-network Inference under Decayed Influence

The LV-GGM induces a *global influence* model for latent variables, which has limitations in real applications. A better model is a *decayed influence model*: the influence of each variable decays drastically outside its neighborhood regions. Such neighborhood can be defined according to, for example, k-nearest neighbors, or the graph geodesic distances between two variables [Tenenbaum et al., 2000, Costa and Hero, 2004, Cormen, 2009]. The comparison between the global influence model and the decayed influence model is

illustrated in Figure 5.2.

Specifically, let $\mathcal{N}(v)$ be a neighborhood of v in $\mathcal{G}$. Denote $\mathcal{N}(\mathcal{V}_1) := \cup_{u\in\mathcal{V}_1}\mathcal{N}(u)$. For latent vertex set $\mathcal{V}_2$, we can partition it into two groups: the boundary set $\delta\mathcal{V}_2 := \mathcal{V}_2 \cap \mathcal{N}(\mathcal{V}_1) = \{v \in \mathcal{V}_2 : v \in \mathcal{N}(u) \text{ for some } u \in \mathcal{V}_1\}$ and the interior set $\mathring{\mathcal{V}}_2 := \mathcal{V}_2 - \mathcal{N}(\mathcal{V}_1)$. According to the decayed influence model, the influence of latent variables in $\mathring{\mathcal{V}}_2$ over $\mathcal{V}_1$ is negligible.

To better characterize the local effect, define $\boldsymbol{B} := \boldsymbol{\Theta}_{12}\boldsymbol{\Theta}_2^{-1} \in \mathbb{R}^{n_1\times n_2}$ so that $\boldsymbol{\Theta}_{12} = \boldsymbol{B}\boldsymbol{\Theta}_2$. With the partition $\mathcal{V}_2 = \mathring{\mathcal{V}}_2 \cup \delta\mathcal{V}_2$, we have

$$\boldsymbol{\Theta}_2\boldsymbol{B}^T = \boldsymbol{\Theta}_{12}^T = \boldsymbol{\Theta}_{21} = \begin{bmatrix} \boldsymbol{\Theta}_{\mathring{\mathcal{V}}_2,1} \\ \boldsymbol{\Theta}_{\delta\mathcal{V}_2,1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{\Theta}_{\delta\mathcal{V}_2,1} \end{bmatrix}, \tag{5.5}$$

which has row-sparse structure. With $\boldsymbol{B} := \boldsymbol{\Theta}_{12}\boldsymbol{\Theta}_2^{-1}$, we can reparameterize the low rank matrix in (5.3) as

$$\begin{aligned} \boldsymbol{M} &:= \boldsymbol{\Theta}_{12}\boldsymbol{\Theta}_2^{-1}\boldsymbol{\Theta}_{21} \\ &= \left[\boldsymbol{\Theta}_{12}\left(\boldsymbol{\Theta}_2\right)^{-1}\right]\boldsymbol{\Theta}_2\left[\left(\boldsymbol{\Theta}_2\right)^{-1}\boldsymbol{\Theta}_{21}\right] = \boldsymbol{B}\boldsymbol{\Theta}_2\boldsymbol{B}^T. \end{aligned} \tag{5.6}$$

Combining (5.6) and (5.5) with the objective function in (5.4), we obtained a ***Decayed-influence Latent variable Gaussian Graphical Model (DiLat-GGM)*** as

$$\min_{\boldsymbol{C},\boldsymbol{B}} \quad -\log\det\left(\boldsymbol{C} - \boldsymbol{B}\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T\right) + \operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}_1\left(\boldsymbol{C} - \boldsymbol{B}\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T\right)\right) + \alpha_m\|\boldsymbol{C}\|_1 + \beta_m\left\|\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T\right\|_{2,1} \tag{5.7}$$

$$\text{s.t.} \quad \boldsymbol{C} - \boldsymbol{B}\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T \succeq \boldsymbol{0},$$

where $\alpha_m, \beta_m > 0$ are regularization parameters and $\left\|\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T\right\|_{2,1} := \left\|\widehat{\boldsymbol{\Theta}}_{21}\right\|_{2,1} = \sum_i \left\|\left(\widehat{\boldsymbol{\Theta}}_{21}\right)_i\right\|_2$ is the mixed-$\ell_{2,1}$ norm that induces row sparsity on $\left\{\left(\widehat{\boldsymbol{\Theta}}_{21}\right)_i\right\}_{i=1}^{n_2}$ [Bach et al., 2012]. The matrix $\widehat{\boldsymbol{\Theta}}_2 \in \mathbb{R}^{n_2\times n_2}$ is a fixed pre-defined positive definite matrix.

Remark

- Similar to graphical Lasso and LV-GGM, the choice of regularization parameter α and β depends on the size of the graph n (or n_1), the number of measurements on each node m and the relative rank of the marginal covariance $\widehat{\boldsymbol{\Sigma}}_1$ [Ravikumar et al., 2008, Chandrasekaran et al., 2012, Meng et al., 2014]. In Section 5.4, we choose $\alpha = \varphi\sqrt{\frac{\log(n)}{m}}$ and $\beta = r\sqrt{\frac{n}{m}}$ for $\varphi \in [0.1, 0.5]$ and $r \in [0.5, 2]$, which results in a good performance.

- The matrix $\widehat{\boldsymbol{\Theta}}_2$ summarizes the inter-dependency between latent variables, which

comes from an external sources. The *key* for this setting is that, due to the privacy and security concerns, the external source cannot share with the proprietary information of its own data $\boldsymbol{x}_2$ directly except for a noisy summary of the past correlation of $\boldsymbol{x}_2$ or a dependency structure $\widehat{\mathcal{G}}_2$ that specify the relationships among $\boldsymbol{x}_2$. In practice, the matrix $\widehat{\boldsymbol{\Theta}}_2$ can be obtained in external source via

$$\widehat{\boldsymbol{\Theta}}_2 = \widehat{\boldsymbol{L}}_2 + \sigma_L^2 \boldsymbol{G},$$

where $\widehat{\boldsymbol{L}}_2$ is an estimate of inverse covariance matrix over $\boldsymbol{x}_2$, $\sigma_L > 0$ and $\boldsymbol{G} = \frac{1}{n_2^2}\boldsymbol{H}\boldsymbol{H}^T$ is a Gram matrix generated by Gaussian random matrix $\boldsymbol{H} \in \mathbb{R}^{n_2 \times n_2}$ with $\boldsymbol{H}_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. $\widehat{\boldsymbol{L}}_2$ can also be chosen as the generalized Laplacian matrix [Pavez and Ortega, 2016] of the true graph $\mathcal{G}_2$. If the external source is unwilling to share any data-related information with the learner, it can simply choose $\widehat{\boldsymbol{\Theta}}_2 = \boldsymbol{I}$ or $\widehat{\boldsymbol{\Theta}}_2 = \text{diag}\,(\boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_2 := (\theta_{21}, \dots \theta_{2,n_2}) \in \mathbb{R}^{n_2}$. In the latter case, the ℓ_{21} norm in (5.7) imposes an weighted regularization on different columns of $\boldsymbol{B}$. From Bayesian perspective, the external source can generate $\widehat{\boldsymbol{\Theta}}_2 \sim \text{Wishart}_{n_2}^{-1}(\boldsymbol{\Psi}, r)$, where $\text{Wishart}_{n_2}^{-1}(\boldsymbol{\Psi}, r)$ is a n_2-variate inverse Wishart distribution with scale matrix $\boldsymbol{\Psi} \in \mathbb{R}^{n_2 \times n_2}$ and degree of freedom r.

- Given $\boldsymbol{B}$, we can infer the effective (non-zero) latent variables via the conditional mean $\boldsymbol{\mu}_{2|1} := \boldsymbol{B}^T \boldsymbol{x}_1$. This is another benefit for using the proposed DiLat-GGM.

The *main advantages* of the DiLat-GGM are 1) it takes into account the matrix $\widehat{\boldsymbol{\Theta}}_2$, which summarizes the inter-dependency among latent variables. As compared to LV-GGM, which does not exploit knowledge regarding the latent variables, DiLat-GGM exploits external network structure of latent variables and their influence on the target network. 2) DiLat-GGM explicitly learns the linear mapping $\boldsymbol{B}$, which enables estimation of the hidden variables via the conditional mean $\boldsymbol{\mu}_{2|1} := \boldsymbol{B}^T \boldsymbol{x}_1$. Thus it can be used as a graph signal interpolation method on $\mathcal{V}_2$, given that we have prior knowledge of the network $\mathcal{G}_2$. This benefit comes at the expense of losing the convexity of the problem. As a result, DiLat-GGM can be seen as a *non-convex* generalization of LV-GGM with the row-sparsity penalty controlling the rank of low-rank term $\boldsymbol{M}$. 3) DiLat-GGM learns the sub-network by combining both the reliable proprietary data from an internal source and the unreliable summary of data from an external source. In the next section, we propose to learn DiLat-GGM using the convex-concave algorithm.

5.3 Efficient Optimization Solver for DiLat-GGM

In this section, we propose an efficient algorithm to solve (5.7).

5.3.1 A Difference-of-Convex Programming Reformulation

We first reformuate the problem (5.7). Note that, since $\widehat{\boldsymbol{\Theta}}_2 \succ \mathbf{0}$, by the Schur complement theorem [Boyd and Vandenberghe, 2004], the constraint

$$\boldsymbol{C} - \boldsymbol{B}\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T \succeq \mathbf{0} \Leftrightarrow \begin{bmatrix} \boldsymbol{C} & \boldsymbol{B} \\ \boldsymbol{B}^T & \widehat{\boldsymbol{\Theta}}_2^{-1} \end{bmatrix} \succeq \mathbf{0}$$

$$\text{and } -\log\det\left(\boldsymbol{C} - \boldsymbol{B}\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T\right) + \log\det\widehat{\boldsymbol{\Theta}}_{\mathbf{2}} = -\log\det\begin{bmatrix} \boldsymbol{C} & \boldsymbol{B} \\ \boldsymbol{B}^T & \widehat{\boldsymbol{\Theta}}_2^{-1} \end{bmatrix}.$$

Thus the problem (5.7) can be reformulated as

$$\min_{\boldsymbol{C},\boldsymbol{B}} \quad -\log\det\begin{bmatrix} \boldsymbol{C} & \boldsymbol{B} \\ \boldsymbol{B}^T & \widehat{\boldsymbol{\Theta}}_2^{-1} \end{bmatrix} + \text{tr}\left(\widehat{\boldsymbol{\Sigma}}_1\left(\boldsymbol{C} - \boldsymbol{B}\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T\right)\right) + \alpha_m\|\boldsymbol{C}\|_1 + \beta_m\left\|\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T\right\|_{2,1} \tag{5.8}$$

$$\text{s.t.} \quad \begin{bmatrix} \boldsymbol{C} & \boldsymbol{B} \\ \boldsymbol{B}^T & \widehat{\boldsymbol{\Theta}}_2^{-1} \end{bmatrix} \succeq \mathbf{0}$$

It is seen that the feasible region is convex in terms of $(\boldsymbol{C}, \boldsymbol{B})$. However, the problem (5.8) is non-convex, since $\text{tr}\left(\widehat{\boldsymbol{\Sigma}}_1\left(\boldsymbol{C} - \boldsymbol{B}\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T\right)\right) = \text{tr}\left(\widehat{\boldsymbol{\Sigma}}_1\boldsymbol{C}\right) - \text{tr}\left(\widehat{\boldsymbol{\Sigma}}_1\boldsymbol{B}\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T\right)$ is non-convex with respect to $(\boldsymbol{C}, \boldsymbol{B})$. Let

$$f(\boldsymbol{C}, \boldsymbol{B}) := -\log\det\begin{bmatrix} \boldsymbol{C} & \boldsymbol{B} \\ \boldsymbol{B}^T & \widehat{\boldsymbol{\Theta}}_2^{-1} \end{bmatrix} + \text{tr}\left(\widehat{\boldsymbol{\Sigma}}_1\boldsymbol{C}\right) + \alpha_m\|\boldsymbol{C}\|_1 + \beta_m\left\|\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T\right\|_{2,1} \tag{5.9}$$

$$g(\boldsymbol{B}) := \text{tr}\left(\widehat{\boldsymbol{\Sigma}}_1\boldsymbol{B}\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T\right). \tag{5.10}$$

See that the $f(\boldsymbol{C}, \boldsymbol{B})$ is convex in $(\boldsymbol{C}, \boldsymbol{B})$. For $g(\boldsymbol{B})$, we have the following proposition:

Proposition 5.3.1 *Given that $\widehat{\boldsymbol{\Sigma}}_1 \succ \mathbf{0}$ and $\widehat{\boldsymbol{\Theta}}_2 \succ \mathbf{0}$ are positive definite matrices, the function $g(\boldsymbol{B})$ is convex in $\boldsymbol{B}$. Moreover, the Hessian of $g(\boldsymbol{B})$ is $\widehat{\boldsymbol{\Sigma}}_1 \otimes \widehat{\boldsymbol{\Theta}}_2 \succ \mathbf{0}$.*

Proof: We can vectorize the matrix $\boldsymbol{B} \in \mathbb{R}^{n_1 \times n_2}$ as $\text{vec}(\boldsymbol{B}) \in \mathbb{R}^{n_1 n_2 \times 1}$ by stacking

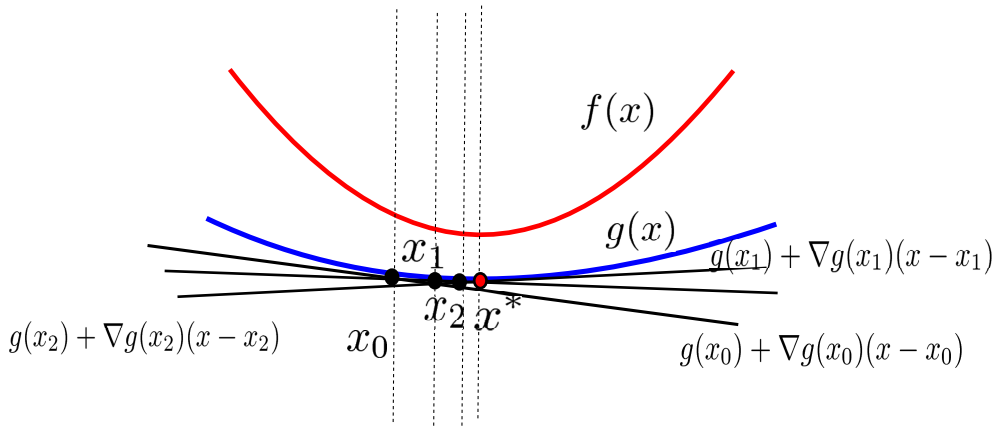


Figure 5.3: An illustration of convex-concave procedure. $f(x)$ and $g(x)$ are both convex functions and we want to find the $x^* = \operatorname{argmin}(f(x) - g(x))$ (red point). We begin by x_0 and iteratively find $x_t := \operatorname{argmin}(f(x) - g(x_{t-1}) - \nabla g(x_{t-1})(x - x_{t-1}))$, where $g(x_t) + \nabla g(x_t)(x - x_t)$ is the tagent plane of $g(x)$ at x_t. Also see that the convergence rate is determined by the difference between curvatures of f and curvatures of g.

columns of $\boldsymbol{B}$. Then

$$\begin{aligned} g'(\text{vec}\,(\boldsymbol{B})) := g(\boldsymbol{B}) &= \text{tr}\left(\left(\boldsymbol{B}^T\widehat{\boldsymbol{\Sigma}}_1\right)^T\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T\right) \\ &= \text{vec}\left(\boldsymbol{B}^T\widehat{\boldsymbol{\Sigma}}_1\right)^T\text{vec}\left(\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T\right). \end{aligned}$$

Since $\text{vec}\,(\boldsymbol{AB}) = (\boldsymbol{I}\otimes\boldsymbol{A})\,\text{vec}\,(\boldsymbol{B}) = \left(\boldsymbol{B}^T\otimes\boldsymbol{I}\right)\text{vec}\,(\boldsymbol{A})$, we have

$$\begin{aligned} g'(\text{vec}\,(\boldsymbol{B})) &= \left(\left(\widehat{\boldsymbol{\Sigma}}_1\otimes\boldsymbol{I}\right)\text{vec}\left(\boldsymbol{B}^T\right)\right)^T\left(\left(\boldsymbol{I}\otimes\widehat{\boldsymbol{\Theta}}_2\right)\text{vec}\left(\boldsymbol{B}^T\right)\right) \\ &= \text{vec}\left(\boldsymbol{B}^T\right)^T\left(\left(\widehat{\boldsymbol{\Sigma}}_1\otimes\boldsymbol{I}\right)\left(\boldsymbol{I}\otimes\widehat{\boldsymbol{\Theta}}_2\right)\right)\text{vec}\left(\boldsymbol{B}^T\right) \\ &= \text{vec}\left(\boldsymbol{B}^T\right)^T\left(\widehat{\boldsymbol{\Sigma}}_1\otimes\widehat{\boldsymbol{\Theta}}_2\right)\text{vec}\left(\boldsymbol{B}^T\right) \end{aligned}$$

Thus the Hessian is

$$\frac{\partial^2 g'(\text{vec}\,(\boldsymbol{B}))}{\partial\text{vec}\left(\boldsymbol{B}^T\right)\partial\text{vec}\left(\boldsymbol{B}^T\right)^T} = \left(\widehat{\boldsymbol{\Sigma}}_1\otimes\widehat{\boldsymbol{\Theta}}_2\right).$$

Since $\widehat{\boldsymbol{\Sigma}}_1 \succ \boldsymbol{0}$ and $\widehat{\boldsymbol{\Theta}}_2 \succ \boldsymbol{0}$, and the eigenvalue of $\left(\widehat{\boldsymbol{\Sigma}}_1\otimes\widehat{\boldsymbol{\Theta}}_2\right)$ is given as $\lambda_i(\widehat{\boldsymbol{\Sigma}}_1)\lambda_j(\widehat{\boldsymbol{\Theta}}_2) > 0, i = 1,\ldots,n_1, j = 1,\ldots,n_2$, where $\lambda_i(\widehat{\boldsymbol{\Sigma}}_1)$ is i-th eigenvalue of $\widehat{\boldsymbol{\Sigma}}_1$ and $\lambda_j(\widehat{\boldsymbol{\Theta}}_2)$ is j-th eigenvalue of $\widehat{\boldsymbol{\Theta}}_2$. Since the Hessian is positive definite, it follows that $g(\boldsymbol{B})$ is a convex function. □

From the above proposition, the objective function of $f(\boldsymbol{C}, \boldsymbol{B}) - g(\boldsymbol{B})$ is a difference of two convex functions, which implies that (5.8) is a difference of convex (DC) programming problem. The convex-concave procedure (CCP) introdued in [Yuille et al., 2002, Yuille and Rangarajan, 2003, Lipp and Boyd, 2016] provides a powerful heuristic method that finds the local solution of DC problem. Specifically, it convexifies the concave function $-g(\boldsymbol{B})$ by linearization as

$$\begin{aligned}
\tilde{g}(\boldsymbol{B}; \boldsymbol{B}_t) &= g(\boldsymbol{B}_t) + \operatorname{tr}\left(\nabla_{\boldsymbol{B}} g(\boldsymbol{B}_t)^T (\boldsymbol{B} - \boldsymbol{B}_t)\right) \\
\text{where } \nabla_{\boldsymbol{B}} g(\boldsymbol{B}_t) &= \nabla_{\boldsymbol{B}} \operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}_1 \boldsymbol{B} \widehat{\boldsymbol{\Theta}}_2 \boldsymbol{B}^T\right) \Big|_{\boldsymbol{B}_t} \\
&= 2\widehat{\boldsymbol{\Sigma}}_1 \boldsymbol{B}_t \widehat{\boldsymbol{\Theta}}_2
\end{aligned}$$

Then the CCP iteratively solves a *convex programming problem* [Wang et al., 2010] given $\boldsymbol{B}_t \in \mathbb{R}^{n_1 \times n_2}$,

$$\begin{aligned}
&(\boldsymbol{C}_{t+1}, \boldsymbol{B}_{t+1}) \\
:= \arg\min_{\boldsymbol{C}, \boldsymbol{B}} &- \log\det \begin{bmatrix} \boldsymbol{C} & \boldsymbol{B} \\ \boldsymbol{B}^T & \widehat{\boldsymbol{\Theta}}_2^{-1} \end{bmatrix} + \operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}_1 \boldsymbol{C}\right) - g(\boldsymbol{B}_t) - \operatorname{tr}\left(\nabla_{\boldsymbol{B}} g(\boldsymbol{B}_t)^T (\boldsymbol{B} - \boldsymbol{B}_t)\right) \\
&+ \alpha_m \|\boldsymbol{C}\|_1 + \beta_m \left\|\widehat{\boldsymbol{\Theta}}_2 \boldsymbol{B}^T\right\|_{2,1} \\
= \arg\min_{\boldsymbol{C}, \boldsymbol{B}} &- \log\det \begin{bmatrix} \boldsymbol{C} & \boldsymbol{B} \\ \boldsymbol{B}^T & \widehat{\boldsymbol{\Theta}}_2^{-1} \end{bmatrix} + \operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}_1 \left(\boldsymbol{C} - 2\boldsymbol{B}\boldsymbol{D}_t^T\right)\right) + \alpha_m \|\boldsymbol{C}\|_1 + \beta_m \left\|\widehat{\boldsymbol{\Theta}}_2 \boldsymbol{B}^T\right\|_{2,1} \\
\text{s.t.} \quad & \begin{bmatrix} \boldsymbol{C} & \boldsymbol{B} \\ \boldsymbol{B}^T & \widehat{\boldsymbol{\Theta}}_2^{-1} \end{bmatrix} \succeq \boldsymbol{0},
\end{aligned} \tag{5.11}$$

where $\boldsymbol{D}_t := \boldsymbol{B}_t \widehat{\boldsymbol{\Theta}}_2 := \boldsymbol{\Theta}_{12}^{(t)}$ since by definition $\boldsymbol{B} := \boldsymbol{\Theta}_{12} \widehat{\boldsymbol{\Theta}}_2^{-1}$. Software packages such as CVX [Grant et al., 2012] and CVXPY [Diamond and Boyd, 2016] can be used to solve above problem. The CCP is also known as a majorization minimization (MM) algorithm [Ortega and Rheinboldt, 2000, Hunter and Lange, 2004], which is a generalization of the EM algortihm [Dempster et al., 1977]. In [Lange et al., 2000, Naghsh et al., 2013], DC problems are solved using MM approaches.

5.3.2 Solving Convex Subproblems

The subproblem (5.11) is a convex programming problem, so it can be solved via a general solver. However, due to its special structure, we can find a fast implementation.

First, we reformulate the objective function of (5.11) as

$$\begin{aligned}\min_{\boldsymbol{C},\boldsymbol{B}} \quad & -\log\det\begin{bmatrix}\boldsymbol{C} & \boldsymbol{B}\\ \boldsymbol{B}^T & \widehat{\boldsymbol{\Theta}}_2^{-1}\end{bmatrix} + \operatorname{tr}\left(\begin{bmatrix}\widehat{\boldsymbol{\Sigma}}_1 & -\widehat{\boldsymbol{\Sigma}}_1\boldsymbol{D}_t\\ -\boldsymbol{D}_t^T\widehat{\boldsymbol{\Sigma}}_1 & \gamma_t\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{C} & \boldsymbol{B}\\ \boldsymbol{B}^T & \widehat{\boldsymbol{\Theta}}_2^{-1}\end{bmatrix}\right)\\ & + \alpha_m\left\|\begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{C} & \boldsymbol{B}\\ \boldsymbol{B}^T & \widehat{\boldsymbol{\Theta}}_2^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{I}\\ \boldsymbol{0}\end{bmatrix}\right\|_1 + \beta_m\left\|\begin{bmatrix}\boldsymbol{0} & \widehat{\boldsymbol{\Theta}}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{C} & \boldsymbol{B}\\ \boldsymbol{B}^T & \widehat{\boldsymbol{\Theta}}_2^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{I}\\ \boldsymbol{0}\end{bmatrix}\right\|_{2,1}\\ \text{s.t.} \quad & \begin{bmatrix}\boldsymbol{C} & \boldsymbol{B}\\ \boldsymbol{B}^T & \widehat{\boldsymbol{\Theta}}_2^{-1}\end{bmatrix} \succeq \boldsymbol{0}.\end{aligned} \tag{5.12}$$

The parameter $\gamma_t > \left\|\boldsymbol{D}_t^T\widehat{\boldsymbol{\Sigma}}_1\boldsymbol{D}_t\right\|_2$, the ℓ_2 norm, so that $\gamma_t\boldsymbol{I} - \boldsymbol{D}_t^T\widehat{\boldsymbol{\Sigma}}_1\boldsymbol{D}_t \succ \boldsymbol{0}$. Therefore

$$\boldsymbol{S}_{\gamma_t}(\boldsymbol{D}_t) := \begin{bmatrix}\widehat{\boldsymbol{\Sigma}}_1 & -\widehat{\boldsymbol{\Sigma}}_1\boldsymbol{D}_t\\ -\boldsymbol{D}_t^T\widehat{\boldsymbol{\Sigma}}_1 & \gamma_t\boldsymbol{I}\end{bmatrix} \succ \boldsymbol{0} \tag{5.13}$$

is positive definite. We can compare the CCP that iteratively solves (5.12) with the EM algorithm developed in Appendix 5.6.1. Note that for the EM algorithm, $\widehat{\boldsymbol{\Theta}}_2$ is not fixed but rather is updated during the iterations. Also the EM algorithm has no additional sparsity regularization on the off-diagonal terms. As a result, the CCP in (5.11) yields a different solution than the EM algorithm.

Define a new variable $\boldsymbol{R} := \begin{bmatrix}\boldsymbol{C} & \boldsymbol{B}\\ \boldsymbol{B}^T & \widehat{\boldsymbol{\Theta}}_2^{-1}\end{bmatrix} \in \mathbb{R}^{n\times n}$ and denote $\boldsymbol{J}_1 := \begin{bmatrix}\boldsymbol{I}_{n_1} & \boldsymbol{0}\end{bmatrix}^T \in \mathbb{R}^{n\times n_1}$, $\boldsymbol{J}_2 := \begin{bmatrix}\boldsymbol{0} & \boldsymbol{I}_{n_2}\end{bmatrix}^T \in \mathbb{R}^{n\times n_2}$, $\boldsymbol{Q} := \begin{bmatrix}\boldsymbol{0} & \widehat{\boldsymbol{\Theta}}_2\end{bmatrix}^T \in \mathbb{R}^{n\times n_2}$. The problem (5.12) becomes

$$\begin{aligned}\min_{\boldsymbol{R}} \quad & -\log\det\boldsymbol{R} + \operatorname{tr}\left(\boldsymbol{S}_{\gamma_t}(\boldsymbol{D}_t)\,\boldsymbol{R}\right) + \alpha_m\left\|\boldsymbol{J}_1^T\boldsymbol{R}\,\boldsymbol{J}_1\right\|_1 + \beta_m\left\|\boldsymbol{Q}^T\boldsymbol{R}\boldsymbol{J}_1\right\|_{2,1}\\ \text{s.t.} \quad & \boldsymbol{J}_2^T\boldsymbol{R}\,\boldsymbol{J}_2 = \widehat{\boldsymbol{\Theta}}_2^{-1}\\ & \boldsymbol{R} \succeq \boldsymbol{0}.\end{aligned} \tag{5.14}$$

We can find the optimal $\boldsymbol{C}$ and $\boldsymbol{B}$ from the $\boldsymbol{R}_1$ and $\boldsymbol{R}_{12}$ blocks of the optimal $\boldsymbol{R}$. Note that the function $\mathbb{D}^{log}\left(\boldsymbol{R}\,\|\,\boldsymbol{S}^{-1}\right) := \operatorname{tr}(\boldsymbol{R}\boldsymbol{S}) - \log\det(\boldsymbol{R}\boldsymbol{S}) - n$ is referred as the LogDet divergence [Kulis et al., 2009]. It belongs to the family of Bregman matrix divergences, which is widely used in machine learning [Murata et al., 2004, Dhillon and Sra, 2005, Banerjee et al., 2007, Kulis et al., 2009, Santos-Rodríguez et al., 2009, Ackermann and Blömer, 2010]. Therefore the CCP that iteratively optimizes (5.14) can be also seen as a

Algorithm 3 DiLat-GGM initialization based on heuristic rule

Require: Sample covariance on observed data $\widehat{\boldsymbol{\Sigma}}_1 \succ \mathbf{0} \in \mathbb{R}^{n_1 \times n_1}$. The nonnegative regularization parameter $\alpha, \beta > 0$. The pre-defined nonegative definite matrix $\widehat{\boldsymbol{\Theta}}_2 \succ \mathbf{0} \in \mathbb{R}^{n_2 \times n_2}$.

1: Find the marginal precision matrix $\widehat{\boldsymbol{\Sigma}}_1^{-1}$.

2: Compute the sparse part $\boldsymbol{C}_0 = \text{soft-threshold}(\widehat{\boldsymbol{\Sigma}}_1^{-1}, \alpha)$;

3: Compute the low-rank part $\boldsymbol{M}_0 = \text{Prox}_{\boldsymbol{M}}(\boldsymbol{C}_0 - \widehat{\boldsymbol{\Sigma}}_1^{-1}, \beta^{'})$, where $\text{Prox}_{\boldsymbol{M}}(\boldsymbol{Z}, \beta^{'})$ is defined in (5.15).

4: Find $\boldsymbol{B}_0$ from $\boldsymbol{C}_0, \boldsymbol{M}_0$ according to (5.16);

Ensure: Output $(\boldsymbol{C}_0, \boldsymbol{B}_0, \boldsymbol{M}_0)$.

sequential LogDet divergence minimization problem.

Remark The subproblem (5.14) can be solved using interior point method [Boyd and Vandenberghe, 2004] implemented via CVX [Grant et al., 2012] and CVXPY [Diamond and Boyd, 2016]. The drawback is that its time complexity can be as high as $O(a^2b^{2.5} + ab^{3.5}) = O(n^{6.5})$, where $a = O(n^2)$ is the number of optimization parameters and $b = O(n)$ is the size of semi-definite matrices [Nemirovskii, 2004]. In Appendix 5.6.3, we have developed a faster implementation using Alternating direction methods of multipliers (ADMM) [Bertsekas, 2015], summarized in Algorithm 5. The proposed **Algorithm 5** has time complexity $O(n^3)$, much faster compared to the interior point method.

5.3.3 Initialization and Stopping Criterion

Learning DiLat-GGM model involves solving a non-convex optimization problem, whose performance relies on the choice of initial feasible solution $(\boldsymbol{C}_0, \boldsymbol{B}_0)$. Following the idea from [Xu et al., 2017], we can choose the matrix $(\boldsymbol{C}_0, \boldsymbol{B}_0)$ via a heuristic-based rule. Specifically, $\boldsymbol{C}_0 := \text{soft-threshold}(\widehat{\boldsymbol{\Sigma}}_1^{-1}, \alpha)$ where $\text{soft-threshold}(x, \alpha) := \text{sign}(x)$ $(|x| - \alpha)_+$ is acted on each entry of the matrix. Note that $\boldsymbol{M}_0 := \boldsymbol{B}_0 \widehat{\boldsymbol{\Theta}}_2 \boldsymbol{B}_0^T$ can be obtained by $\boldsymbol{M}_0 = \text{Prox}_{\boldsymbol{M}}(\boldsymbol{C}_0 - \widehat{\boldsymbol{\Sigma}}_1^{-1}, \beta^{'})$. The operator $\text{Prox}_{\boldsymbol{M}}(\boldsymbol{Z}, \beta^{'})$ is defined as

$$\text{Prox}_{\boldsymbol{M}}(\boldsymbol{Z}, \beta^{'}) := \min_{\boldsymbol{M}} \frac{1}{2} \|\boldsymbol{M} - \boldsymbol{Z}\|_F^2 + \beta^{'} \|\boldsymbol{M}\|_* + \mathbb{1}\{\boldsymbol{M} \succeq \mathbf{0}\}. \tag{5.15}$$

The optimal solution $\boldsymbol{L}^*$ has the eigen-decomposition

$$\boldsymbol{M}^* = \boldsymbol{U} \text{diag}(\boldsymbol{\zeta}) \boldsymbol{U}^T$$

where the eigenvalues $\zeta_i = \max\{\sigma_i - \beta^{'}, 0\}$ for $\boldsymbol{Z} := \boldsymbol{U} \text{diag}([\sigma_i]) \boldsymbol{U}^T$.

Algorithm 4 DiLat-GGM via Convex-concave procedure

Require: Sample covariance on observed data $\widehat{\boldsymbol{\Sigma}}_1 \succ \mathbf{0} \in \mathbb{R}^{n_1 \times n_1}$. The nonnegative regularization parameter $\alpha, \beta > 0$. The index set of observed data $\mathcal{V}_1$ and the index set of the latent data $\mathcal{V}_2$. The pre-defined nonegative definite matrix $\widehat{\boldsymbol{\Theta}}_2 \succ \mathbf{0} \in \mathbb{R}^{n_2 \times n_2}$.

1: **Initialize:** Use heuristic-based rule as in **Algorithm 3** or random initialization with best result reported. Return $(\boldsymbol{C}_0, \boldsymbol{B}_0, \boldsymbol{M}_0)$.

2: **for** $t = 1, \ldots, T$ or until converge **do**

3: Construct matrix $\boldsymbol{S}_{t-1} := \begin{bmatrix} \widehat{\boldsymbol{\Sigma}}_1 & -\widehat{\boldsymbol{\Sigma}}_1 \boldsymbol{D}_{t-1} \\ -\boldsymbol{D}_{t-1}^T \widehat{\boldsymbol{\Sigma}}_1 & \gamma_{t-1} \boldsymbol{I} \end{bmatrix} \succ \mathbf{0}$, where $\boldsymbol{D}_{t-1} := \boldsymbol{B}_{t-1} \widehat{\boldsymbol{\Theta}}_2$, $\gamma_{t-1} > \left\| \boldsymbol{D}_{t-1}^T \widehat{\boldsymbol{\Sigma}}_1 \boldsymbol{D}_{t-1} \right\|_2$, the ℓ_2 norm.

4: Solve the convex subproblem (5.14). See **Algorithm 5** in Appendix 5.6.2. Return $\boldsymbol{R}_t = \begin{bmatrix} \boldsymbol{C}_t & \boldsymbol{B}_t \\ \boldsymbol{B}_t^T & \widehat{\boldsymbol{\Theta}}_2^{-1} \end{bmatrix}$.

5: **end for**

Ensure: Output $\boldsymbol{C}_T := [\boldsymbol{R}_T]_{\mathcal{V}_1 \times \mathcal{V}_1}$ and $\boldsymbol{B}_T := [\boldsymbol{R}_T]_{\mathcal{V}_1 \times \mathcal{V}_2}$.

Finally, we obtain $\boldsymbol{B}_0 \in \mathbb{R}^{n_1 \times n_2}$ from $\boldsymbol{M}_0 \in \mathbb{R}^{n_1 \times n_1}$ and $\widehat{\boldsymbol{\Theta}}_2 \in \mathbb{R}^{n_2 \times n_2}$. Let $\widehat{\boldsymbol{\Theta}}_2 = \boldsymbol{V} \boldsymbol{\Lambda} \boldsymbol{V}^T$ be the eigen-decompostion of $\widehat{\boldsymbol{\Theta}}_2$, where the eigenvalue $0 < \lambda_0 \leq \lambda_1 \leq \ldots \leq \lambda_{n_2}$ in increasing order and $\boldsymbol{M}_0 = \boldsymbol{U} \operatorname{diag}(\boldsymbol{\zeta}) \boldsymbol{U}^T$ with eigenvalues $\zeta_0 \geq \zeta_1 \geq \ldots \geq \zeta_r = 0 = \ldots \zeta_{n_1}$ in decreasing order, where $r = \operatorname{rank}(\boldsymbol{M}_0)$. We want to find $\boldsymbol{B}_0$ satisfying the quadratic equation

$$\boldsymbol{M}_0 = \boldsymbol{B}_0 \widehat{\boldsymbol{\Theta}}_2 \boldsymbol{B}_0.$$

It is seen that

$$\boldsymbol{B}_0 = \begin{cases} \boldsymbol{U} \begin{bmatrix} \boldsymbol{\Psi} \\ \mathbf{0} \end{bmatrix} \boldsymbol{V}^T, & \text{if } n_1 \geq n_2 \\ \boldsymbol{U} \begin{bmatrix} \boldsymbol{\Psi} & \mathbf{0} \end{bmatrix} \boldsymbol{V}^T, & \text{if } n_2 > n_1 \end{cases} \tag{5.16}$$

is a feasible solution. Here $\boldsymbol{\Psi} := \operatorname{diag}(\psi_0, \psi_1, \ldots, \psi_k, 0, \ldots, 0)$ where $\psi_i = \sqrt{\frac{\zeta_i}{\lambda_i}}$ for $0 \leq i \leq k$ and $k = \min\{r, n_2\}$ if $n_1 \geq n_2$ or $k = r$ for $n_2 > n_1$. Note that $\psi_i \geq \psi_{i+1}$. In [Yuille and Rangarajan, 2003, Lanckriet and Sriperumbudur, 2009], it is suggested to start with random initialization and choose the best solution that has the minimal objective value. We can replace $\widehat{\boldsymbol{\Sigma}}_1$ by a random positive definite matrix to achieve this. In our experiments in Section 5.4, we choose the heuristic-based initialization with inverse of empirical covariance matrix.

A reasonable stopping criterion is that the improvement of objective values being less than a fixed threshold δ:

$$h(\boldsymbol{C}_{t+1}, \boldsymbol{B}_{t+1}) - h(\boldsymbol{C}_t, \boldsymbol{B}_t) \leq \delta,$$

where $h(\boldsymbol{C}, \boldsymbol{B}) := f(\boldsymbol{C}, \boldsymbol{B}) - g(\boldsymbol{B})$ be the objective function of problem (5.7), where $f(\boldsymbol{C}, \boldsymbol{B})$ is defined in (5.9) and $g(\boldsymbol{B})$ is defined in (5.10).

Finally, the CCP-based algorithm to solve DiLat-GGM is summarized in **Algorithm 4**.

5.3.4 Local Convergence Analysis

It is shown in [Lanckriet and Sriperumbudur, 2009] that a general CCP is a special form of the MM algorithm [Ortega and Rheinboldt, 2000, Hunter and Lange, 2004]. It is seen that for the proposed Algorithm 4, the following theorem holds:

Theorem 5.3.2 *([Yuille and Rangarajan, 2003, Lanckriet and Sriperumbudur, 2009, Lipp and Boyd, 2016]) Let $h(\boldsymbol{C}, \boldsymbol{B}) := f(\boldsymbol{C}, \boldsymbol{B}) - g(\boldsymbol{B})$ be the objective function of* (5.7)*, where both $f(\boldsymbol{C}, \boldsymbol{B})$ as defined in* (5.9) *and $g(\boldsymbol{B})$ as defined in* (5.10) *are convex in $(\boldsymbol{C}, \boldsymbol{B})$. Assumes $\{(\boldsymbol{C}_t, \boldsymbol{B}_t)\}_{t=0}^{\infty}$ is a sequence of solutions of sub-problems* (5.14) *in Algorithm 4. Then at each iteration, the value of objective function monotonically decreases, i.e. $h(\boldsymbol{C}_{t+1}, \boldsymbol{B}_{t+1}) \leq h(\boldsymbol{C}_t, \boldsymbol{B}_t)$. Thus $\{h(\boldsymbol{C}_t, \boldsymbol{B}_t)\}_{t=0}^{\infty}$ converges.*

Proof: Assume that $(\boldsymbol{C}_t, \boldsymbol{B}_t)$ is a feasible solution of (5.7). So $(\boldsymbol{C}_t, \boldsymbol{B}_t)$ is also a feasible solution of the convex subproblem (5.14). Let the objective value be $v_t := h(\boldsymbol{C}_t, \boldsymbol{B}_t)$. Then

$$v_t = f(\boldsymbol{C}_t, \boldsymbol{B}_t) - g(\boldsymbol{B}_t) = f(\boldsymbol{C}_t, \boldsymbol{B}_t) - \tilde{g}(\boldsymbol{B}_t; \boldsymbol{B}_t) \geq f(\boldsymbol{C}_{t+1}, \boldsymbol{B}_{t+1}) - \tilde{g}(\boldsymbol{B}_{t+1}; \boldsymbol{B}_t),$$

where the last inequality follows since at iteration t, $(\boldsymbol{C}_{t+1}, \boldsymbol{B}_{t+1})$ minimizes the objective function $f(\boldsymbol{C}, \boldsymbol{B}) - \tilde{g}(\boldsymbol{B}; \boldsymbol{B}_t)$. Finally, since $-g(\boldsymbol{B})$ is concave in $\boldsymbol{B}$, $-\tilde{g}(\boldsymbol{B}; \boldsymbol{B}_t) \geq -g(\boldsymbol{B})$ for all $\boldsymbol{B}$ with equality holds if and only if $\boldsymbol{B} = \boldsymbol{B}_t$. So we have

$$v_t \geq f(\boldsymbol{C}_{t+1}, \boldsymbol{B}_{t+1}) - \tilde{g}(\boldsymbol{B}_{t+1}; \boldsymbol{B}_t) \geq v_{t+1}.$$

Since the sequence $\{v_t\}_{t=0}^{\infty}$ is nonincreasing and $v_t \geq 0$, the iterations will converge. □

Although the objective value converges, there is no guarantee for general CCP problems that $\{(\boldsymbol{C}_t, \boldsymbol{B}_t)\}_{t=0}^{\infty}$ converges, even to a local minima [Lipp and Boyd, 2016]. However, according to the Theorem 4 in [Lanckriet and Sriperumbudur, 2009], if both f and g are

real-valued differentiable functions and ∇g is continuous, and if some additional conditions are satisfied, then all limit points of the solution sequence $\{(\boldsymbol{C}_t, \boldsymbol{B}_t)\}_{t=0}^{\infty}$ are stationary points of the original DC-problem (5.7). Specifically, we have the following theorem:

Theorem 5.3.3 *Let $f(\boldsymbol{C}, \boldsymbol{B})$ and $g(\boldsymbol{B})$ be the convex functions defined in* (5.9) *and* (5.10)*, respectively. Denote the point-to-set map $\mathcal{S}(\boldsymbol{Z}) := \arg\min_{\boldsymbol{C},\boldsymbol{B}}\{f(\boldsymbol{C}, \boldsymbol{B}) - \tilde{g}(\boldsymbol{B}; \boldsymbol{Z}) : (\boldsymbol{C}, \boldsymbol{B}) \in \Omega\}$, where $f(\boldsymbol{C}, \boldsymbol{B}) - \tilde{g}(\boldsymbol{B}; \boldsymbol{Z})$ is the objective function of the subproblem* (5.14)*, and $\Omega := \left\{(\boldsymbol{C}, \boldsymbol{B}) : \boldsymbol{C} - \boldsymbol{B}\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T \succeq \boldsymbol{0}\right\}$ is the feasible region. Then all limit points of $\{(\boldsymbol{C}_t, \boldsymbol{B}_t)\}_{t=0}^{\infty}$ are stationary points of the original DC-problem* (5.7)*. Moreover, the limit of the objective value $\lim_{t\to\infty} f(\boldsymbol{C}_t, \boldsymbol{B}_t) - g(\boldsymbol{B}_t) = f(\boldsymbol{C}^*, \boldsymbol{B}^*) - g(\boldsymbol{B}^*)$, where $(\boldsymbol{C}^*, \boldsymbol{B}^*)$ is some stationary point of* (5.7)*.*

Proof: We establish the theorem by confirming the condition of Theorem 4 in [Lanckriet and Sriperumbudur, 2009]. First observe that both f and g are differentiable function in $(\boldsymbol{C}, \boldsymbol{B})$ and $\nabla g(\boldsymbol{B}) = 2\widehat{\boldsymbol{\Sigma}}_1\boldsymbol{B}_t\widehat{\boldsymbol{\Theta}}_2$ is continuous in $\boldsymbol{B}$. Next note that at each iteration t, the new subproblem $\mathcal{S}(\boldsymbol{B}_t)$ does not depends on previous result $\boldsymbol{C}_t$. Let $\boldsymbol{C}_t$ be fixed and the feasible region Ω is reparameterized in terms of $\boldsymbol{B}$ as $\Omega_{\boldsymbol{B}} := \left\{\boldsymbol{B} : \boldsymbol{B}\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T \preceq \boldsymbol{C}\right\}$. Since the subproblem (5.14) is convex, there exists a unique global minimizer, thus the set $\mathcal{S}(\boldsymbol{B}_t) \neq \emptyset$ for any $\boldsymbol{B}_t \in \Omega_{\boldsymbol{B}}$. Finally, note that $\|\boldsymbol{C}_t\|_2 \leq \rho$, since $\boldsymbol{C}_t$ is an optimal solution for the subproblem (5.14) in previous iteration. We see that $\Omega_{\boldsymbol{B}} \subset \left\{\boldsymbol{B} : \boldsymbol{B}\widehat{\boldsymbol{\Theta}}_2\boldsymbol{B}^T \preceq \rho\boldsymbol{I}\right\}$, which is a compact subset in $\mathbb{R}^{n_1 \times n_1}$. As a result, $\mathcal{S}(\boldsymbol{B}_t)$ is uniformly compact over $\Omega_{\boldsymbol{B}}$. Hence, according to Theorem 4 in [Lanckriet and Sriperumbudur, 2009], the above results hold. □

Note that the above convergence result holds for any random initialization. However, since the DiLat-GGM is non-convex, it may have multiple stationary points. Thus the final limit point does depend on the initialization.

5.4 Experiments

In this section, we compare the performance of the **DiLat-GGM** [2] on synthetic datasets with two graph topology learning algorithms: the graphical Lasso (**GLasso**) [Friedman et al., 2008], the latent variable Gaussian graphical model (**LV-GGM**) [Chandrasekaran et al., 2012] and the EM version of LV-GGM (**EM-GLasso**) [Yuan, 2012] described in Appendix 5.6.1. The gLasso is implemented using scikit-learn Python package [Pedregosa et al., 2011]. The LV-GGM is implemented via ADMM algorithm as in [Ma et al., 2013].

[2] The code is available in https://github.com/TianpeiLuke/LatNet

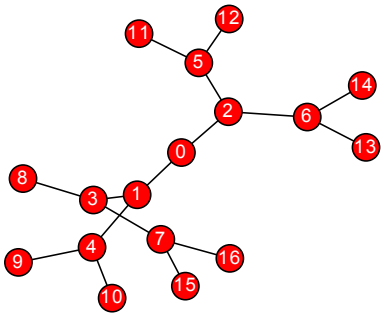
Table 5.1: Edge selection error for different graphs, with the best performance shown in **bold**.

Mean Jaccard distance error ($\times 100\%$)					
Network	GLasso	EM-GLasso	GenLap	LV-GGM	DiLat-GGM
complete binary tree ($h = 3, n_1 = 10$)	55.7	65.2	**12.8**	36.4	18.8
complete binary tree ($h = 4, n_1 = 17$)	11.3	32.1	22.4	3.5	**2.2**
complete binary tree ($h = 5, n_1 = 36$)	15.0	26.6	50.9	3.3	**2.5**
grid ($w = 5, h = 5, n_1 = 15$)	39.3	40.7	**5.7**	23.3	12.8
grid ($w = 7, h = 7, n_1 = 30$)	10.4	18.0	20.8	7.7	**4.6**
grid ($w = 9, h = 9, n_1 = 49$)	10.3	25.1	32.7	7.8	**5.4**
Erdős-Rényi ($n = 15, p = 0.05, n_1 = 10$)	19.6	25.4	**7.9**	15.0	13.9
Erdős-Rényi ($n = 30, p = 0.05, n_1 = 20$)	9.6	22.3	23.0	6.2	**4.5**
Erdős-Rényi ($n = 60, p = 0.05, n_1 = 40$)	10.8	32.5	61.1	8.1	**6.5**
Erdős-Rényi ($n = 60, p = 0.1, n_1 = 40$)	39.3	43.5	63.4	34.1	**27.2**
Erdős-Rényi ($n = 60, p = 0.15, n_1 = 40$)	54.9	56.2	62.1	52.2	**50.2**

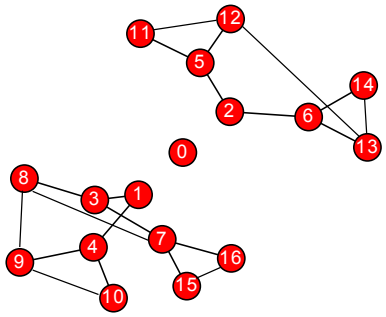
See Appendix 5.6.2. We also includes the generalized Laplacian learning (**GenLap**) [Pavez and Ortega, 2016] which is a variant of dual gLasso.

To illustrate performance of the proposed algorithm 5, we use a synthesis data set $(\boldsymbol{X}, \mathcal{G})$. First, a full network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $n = |\mathcal{V}|$ is generated. Let $\boldsymbol{L}$ be the normalized Laplacian matrix. The random graph signal $\boldsymbol{x} \in \mathbb{R}^n$ is generated from the Laplacian matrix $\boldsymbol{L}$ according to the distribution $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where the covariance matrix $\boldsymbol{\Sigma} = (\boldsymbol{L} + \epsilon \boldsymbol{I})^{-1}$, for a small $\epsilon > 0$. Note that the true precision matrix for the full data $\boldsymbol{\Theta} = \boldsymbol{L} + \epsilon \boldsymbol{I}$, whose support set is equal to the edge set $\mathcal{E}$ of the graph $\mathcal{G}$. A set of m i.i.d realizations of $\boldsymbol{x}$ is generated, which is referred to as $\boldsymbol{X} = [\boldsymbol{x}_v]_{v \in \mathcal{V}} \in \mathbb{R}^{n \times m}$ with $\boldsymbol{x}_v$ as its v-th row. Let $m > n$. In all experiments below, we choose $m = 500$, which is sufficiently large for well-conditioned estimates, given the size of graph under consideration.

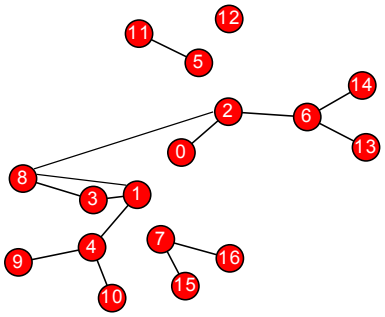
A local sub-network $\mathcal{G}_1$ is sampled via the breadth-first search strategy [Cormen, 2009]: we randomly choose an initial point $u_0 \in \mathcal{V}$, and set $\mathcal{V}_1^{(0)} = \{u_0\}$. At each iteration $t = 1, \ldots,$, for each $v \in \mathcal{V}_1$, we find all neighbors of v as $\mathcal{N}_v = \{w | (w, v) \in \mathcal{E}\}$. Then $\mathcal{V}_1^{(t)} := \mathcal{V}_1^{(t-1)} \cup \bigcup_{v \in \mathcal{V}_1^{(t-1)}} \mathcal{N}_v$. The sampling procedure stops when $|\mathcal{V}_1| > n_0^{'}$. Let $|\mathcal{V}_1| = n_1$



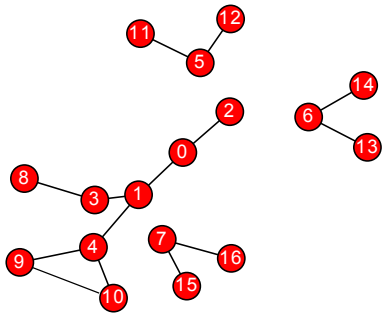
(a) Ground truth



(b) GLasso



(c) LV-GGM



(d) **DiLat-GGM**

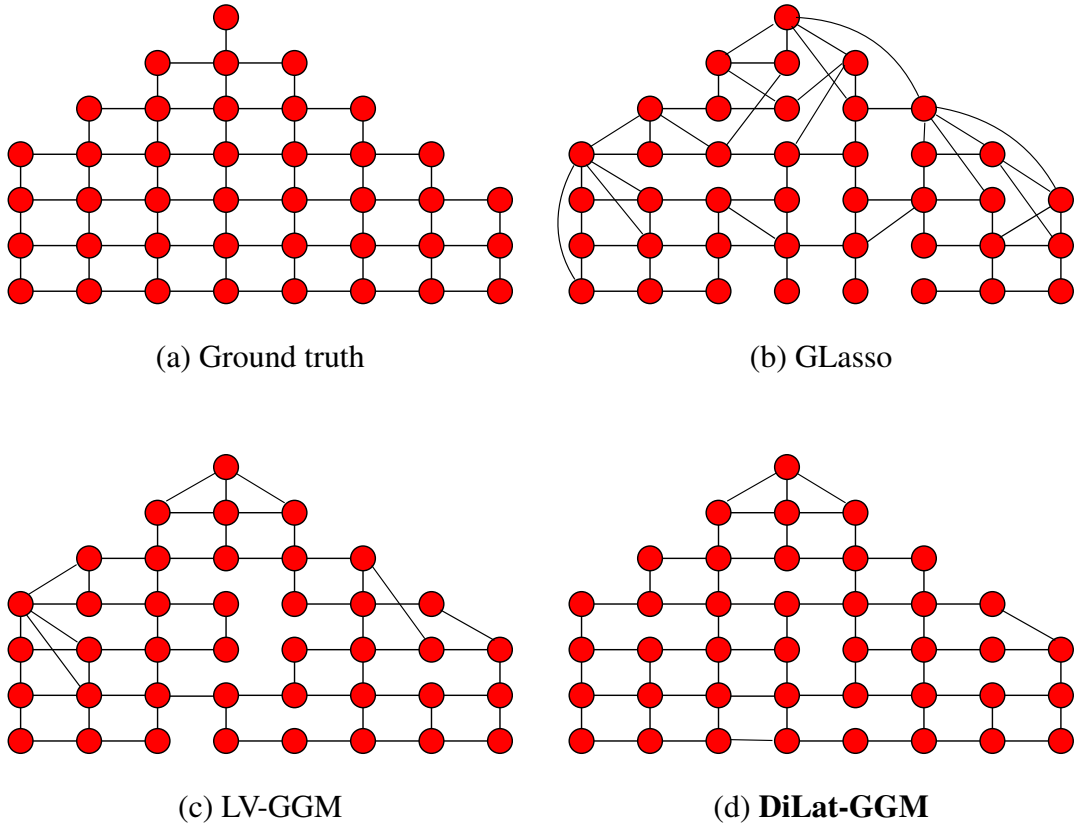
Figure 5.4: (a) The ground truth is a balanced binary tree with height $h = 3$. (b) The graph learned by **GLasso** with optimal $\alpha = 0.6$ (c) The graph learned by **LV-GGM** with optimal $\alpha = 0.1, \beta = 0.15$ (d) The graph learned by **DiLat-GGM** with optimal $\alpha = 0.15, \beta = 1$. It is seen that GLasso has high false positives (cross-edges between leaves) due to the marginalization effect. Compare to LV-GGM, the DiLat-GGM has fewer missing edges and less false positives.

and define the sub-network $\mathcal{G}_1 := (\mathcal{V}_1, \mathcal{E}_1)$ with $\mathcal{E}_1 := \mathcal{E} \cap (\mathcal{V}_1 \times \mathcal{V}_1)$. The remaining vertex set is $\mathcal{V}_2$ and it forms a network $\mathcal{G}_2$. Let $\boldsymbol{L}_1$ be the normalized Laplacian matrix for $\mathcal{G}_1$.

Given $\mathcal{V}_1$, we choose the corresponding data $\boldsymbol{X}_1 := [\boldsymbol{x}_v]_{v \in \mathcal{V}_1}$. The task is to find the sub-network topology $\mathcal{G}_1$ given partially observed data $\boldsymbol{X}_1$. To measure the accuracy of the edge selection, we use the Jaccard distance [Jaccard, 1901, Choi et al., 2010b] between two sets A, B as

$$dist_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} \in [0, 1].$$

The Jaccard distance is a widely used similarity measure in structure estimation [Toldo



(a) Ground truth (b) GLasso

(c) LV-GGM (d) **DiLat-GGM**

Figure 5.5: (a) The ground truth of size $n_1 = 40$ with a grid structure. (b) The graph learned by **GLasso** with optimal $\alpha = 0.4$ (c) The graph learned by **LV-GGM** with optimal $\alpha = 0.1, \beta = 0.15$ (d) The graph learned by **DiLat-GGM** with optimal $\alpha = 0.2, \beta = 1$. It is seen that GLasso has high false positives (cross-edges between leaves) due to the marginalization effect. Compare to LV-GGM, the DiLat-GGM has fewer missing edges and less false positives.

and Fusiello, 2008], clustering [Ferdous et al., 2009] and information retrieval [Manning et al., 2008]. In the experiment, the set $A := \left\{(i,j) \mid \widehat{\boldsymbol{C}}_{i,j} \neq 0, i > j\right\}$ is the support set of estimated sparse precision matrix $\widehat{\boldsymbol{C}}$ is chosen to compare with the true edge set $B := \{(i,j) \mid \boldsymbol{L}_1 \neq 0, i > j\}$. In most of the experiments in this section, we choose the best performance after a grid search of regularization parameter $\alpha \in [10^{-2}, 0.7]$, $\beta \in [0.01, 5]$. For DiLat-GGM, we choose the matrix $\widehat{\boldsymbol{\Theta}}_2$ as

$$\widehat{\boldsymbol{\Theta}}_2 = \widehat{\boldsymbol{L}}_2 + \sigma_L^2 \boldsymbol{G},$$

where $\widehat{\boldsymbol{L}}_2$ is an estimate of inverse covariance matrix over $\boldsymbol{x}_2$, $\sigma_L > 0$ and $\boldsymbol{G} = \frac{1}{n_2^2}\boldsymbol{H}\boldsymbol{H}^T$ is a Gram matrix generated by Gaussian random matrix $\boldsymbol{H} \in \mathbb{R}^{n_2 \times n_2}$ with $\boldsymbol{H}_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.

We first compare the performance of edge selection for GLasso, GenLap, LV-GGM

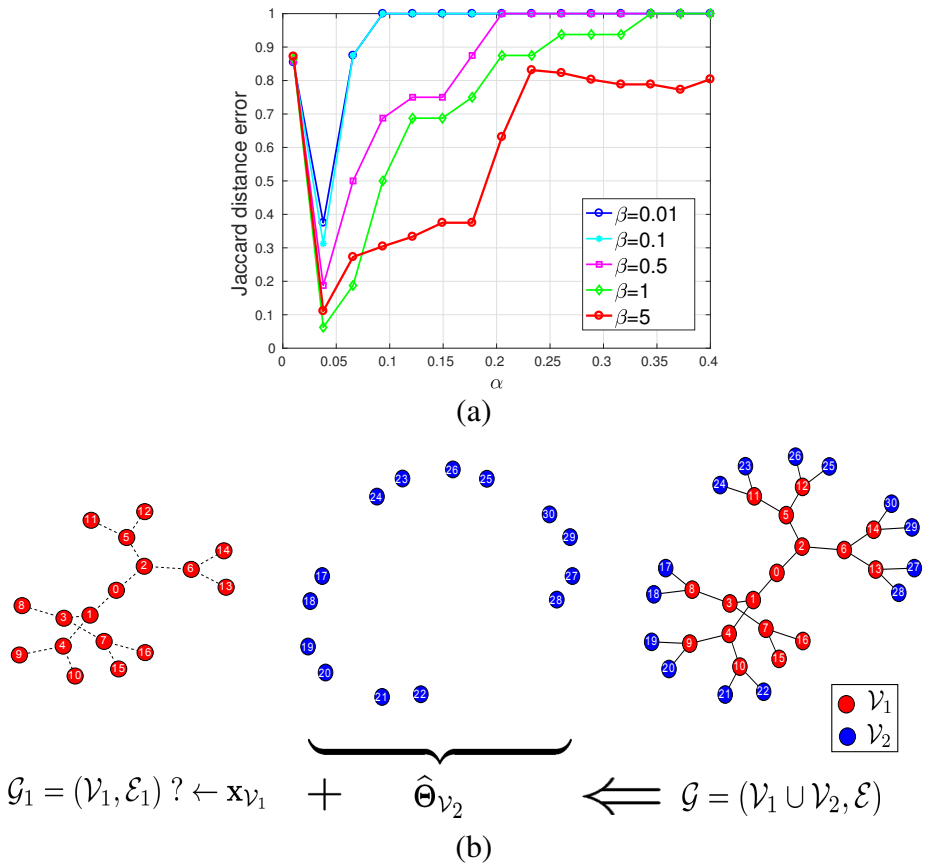


Figure 5.6: (a) The sensitivity of DiLat-GGM for a fixed complete binary tree graph ($h = 4$) under the different choice of regularization parameter α and β. The network is illustrated as $\mathcal{G}$ in (b). The performance is measured in terms of Jaccard distance error. (b) Illustration of experiments in (a). The ground truth network $\mathcal{G}$ on the right is a complete binary tree graph ($h = 4$) with observed variables on red vertices. The task is to infer the marginal network $\mathcal{G}_1$ for red vertices (left) given data $\boldsymbol{x}_{\mathcal{V}_1}$ on its nodes (red) and a summary of latent network (center) $\widehat{\mathbf{\Theta}}_{\mathcal{V}_2} = \widehat{\boldsymbol{L}}_2$, where $\widehat{\boldsymbol{L}}_2$ is an estimate of inverse covariance matrix over $\boldsymbol{x}_2$ (blue vertices). See that all the latent variables are conditional independent given the observed data $\boldsymbol{x}_{\mathcal{V}_1}$.

and the proposed DiLat-GGM. In Figure 5.4 and Figure 5.5, we compare these methods qualitatively by showing their learned network under the choice of optimal parameters. The ground truth is a balanced binary tree with $h = 3$ in Figure 5.4 and a neighborhood in a 8×8 grid network in Figure 5.5, respectively. It is seen that the learned network by DiLat-GGM has fewer missing rate and false positive rate in edge detection compared to GLasso and LV-GGM in both networks. The GLasso, however, has a higher false positive rate in boundary vertices due to the effect of marginalization. Table 5.1 shows the mean edge selection error under different graphs for GLasso, GenLap, LV-GGM and DiLat-GGM in terms of the Jaccard distance. All results are based on an average of 50 runs and for each run we choose the best performance after a grid search of regularization parameter $\alpha \in [10^{-2}, 0.7]$, $\beta \in [0.01, 5]$. For DiLat-GGM, $\widehat{\mathbf{\Theta}}_2 = \widehat{\boldsymbol{L}}_2 + \sigma_L^2 \boldsymbol{G}$ as defined above with

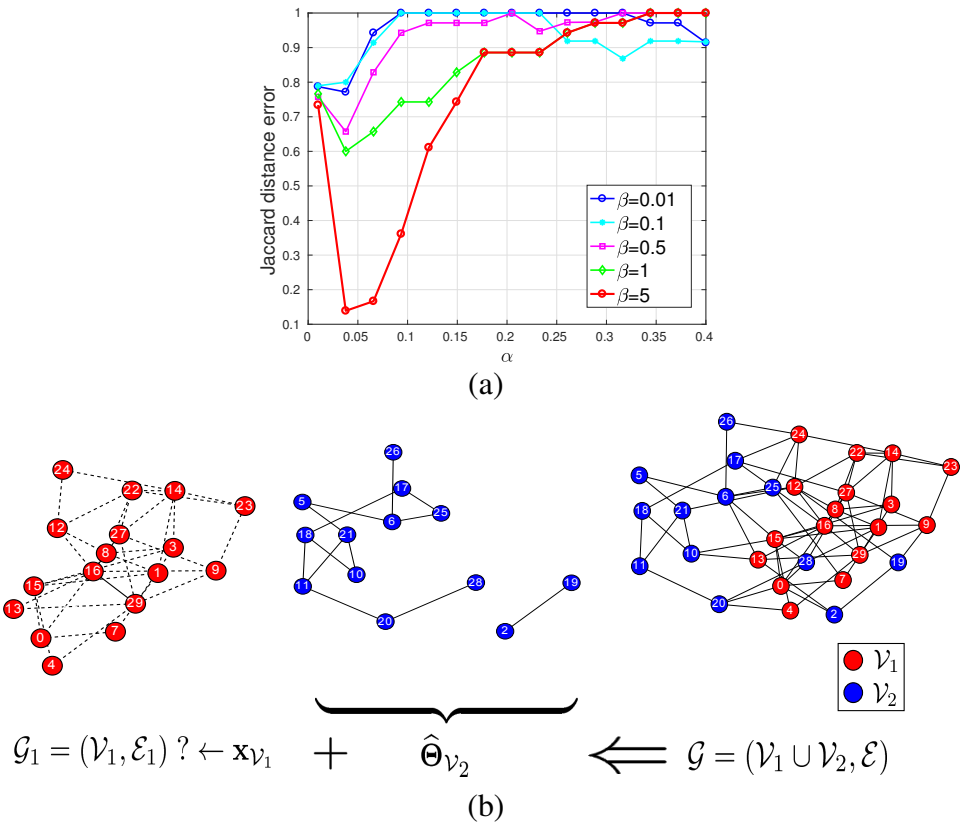


Figure 5.7: (a) The sensitivity of DiLat-GGM for a Erdős-Rényi graph model with $n = 30, p = 0.16$ in (b) under the different choice of regularization parameter α and β. The performance is measured in terms of Jaccard distance error. (b) Illustration of experiments in (a). The underling network is a realization of a Erdős-Rényi graph model with observed variables on red vertices. The task is to infer the marginal network $\mathcal{G}_1$ for red vertices (left) given data $\boldsymbol{x}_{\mathcal{V}_1}$ on its nodes (red) and a summary of latent network (center) $\widehat{\mathbf{\Theta}}_{\mathcal{V}_2} = \widehat{\boldsymbol{L}}_2$, where $\widehat{\boldsymbol{L}}_2$ is an estimate of inverse covariance matrix over $\boldsymbol{x}_2$ (blue vertices). Compared with Figure 5.6, latent variables are conditional depedent on each other given the observed data $\boldsymbol{x}_{\mathcal{V}_1}$.

$\sigma_L = 0.1$. We compare the result with different graph topology, including the complete binary tree with height h, the grid network with width w and height h and the Erdős-Rényi graph with size n and edge probability p. As shown in the table, our proposed DiLat-GGM reaches superior performance compared to GLasso, LV-GGM, EM-GLasso for all investigated networks. When the size of full network is very small, the GenLap algorithm reach the best performance. This is due to the strong interaction between latent variables and the observed variables, which causes a decrease in performance for GLasso, LV-GGM and DiLat-GGM. These three algorithms rely on the soft-regularizer such as the ℓ_1 norm to learn a sparse graph, which is not as strong as the non-negative constraint in GenLap. The performance of the GenLap algorithm decreases drastically when the size of the network getting large due to the bias induced by the strong non-negative constraint. For the Erdős-

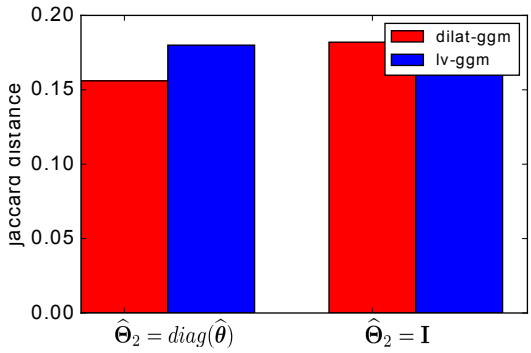


Figure 5.8: A comparison between DiLat-GGM and LV-GGM when $\widehat{\boldsymbol{\Theta}}_2 = \text{diag}(\widehat{\boldsymbol{\theta}})$ where $\widehat{\boldsymbol{\theta}} = [\hat{\theta}_1, \ldots, \hat{\theta}_{n_2}]$ is an estimate of conditional variance over $\boldsymbol{x}_2$ and when $\widehat{\boldsymbol{\Theta}}_2 = \boldsymbol{I}$. We use the same balanced binary tree as in Figure 5.6 but with non-identical conditional variances over $\boldsymbol{x}_2$. See that for DiLat-GGM, $\widehat{\boldsymbol{\Theta}}_2 = \text{diag}(\widehat{\boldsymbol{\theta}})$ performs better than $\widehat{\boldsymbol{\Theta}}_2 = \boldsymbol{I}$, since $\widehat{\boldsymbol{\theta}}$ accounts for the actual conditional variances in $\boldsymbol{x}_2$.

Rényi graph, the performance for all algorithms decreases as the edge probability increases. This reflects the increase of the bias induced by the sparsity penalty.

We also compare the EM-GLasso and DiLat-GGM in Table 5.1. As mentioned in Section 5.3, both algorithms share some similarities and DiLat-GGM can be seen as a generalization of EM-Glasso. However, the ability of conditional feature selection imposed by the row sparsity regularization in DiLat-GGM permits a better to fit to the underlying structures of the graph. As a result, DiLat-GGM outperforms EM-GLasso in all cases investigated.

In Figure 5.6, we demonstrate the sensitivity of the DiLat-GGM model under different choices of regularization parameter α and β. The results are based on an average of 20 runs with fixed graph topology and fixed choice of observed sub-network. We use the $\widehat{\boldsymbol{\Theta}}_2$ as above. The results in Figure 5.6 (a) is based on smooth data over a complete binary tree in Figure 5.6 (b). Figure 5.6 (c) shows the external network $\mathcal{G}_2$ corresponding to $\widehat{\boldsymbol{\Theta}}_2$. Note that the latent variables in $\mathcal{G}_2$ are all independent. Similarly, the results in Figure 5.7 (a) is over a realization of Erdős-Rényi graph $(n = 30, p = 0.16)$ in Figure 5.7 (b). From both Figure 5.6 (a) and Figure 5.7 (a) we see that when α increases the learned graph becomes too sparse so the Jaccard distance error increases. The choice of β controls the row sparsity of the conditional cross precision $\boldsymbol{\Theta}_{21}$, if it is too small, the DiLat-GGM cannot capture the local effect of the latent variables, which decreases its performance in sub-network learning. Both plots show that an optimal choice of (α, β) exists for the DiLat-GGM, which is the same as the other graphical model selection methods such as the GLasso and LV-GGM. In practice, it is observed that $\alpha = \varphi\sqrt{\frac{\log(n)}{m}}$ and $\beta = r\sqrt{\frac{n}{m}}$ for $\varphi \in [0.1, 0.5]$ and $r \in [0.5, 2]$ will result in a good performance.

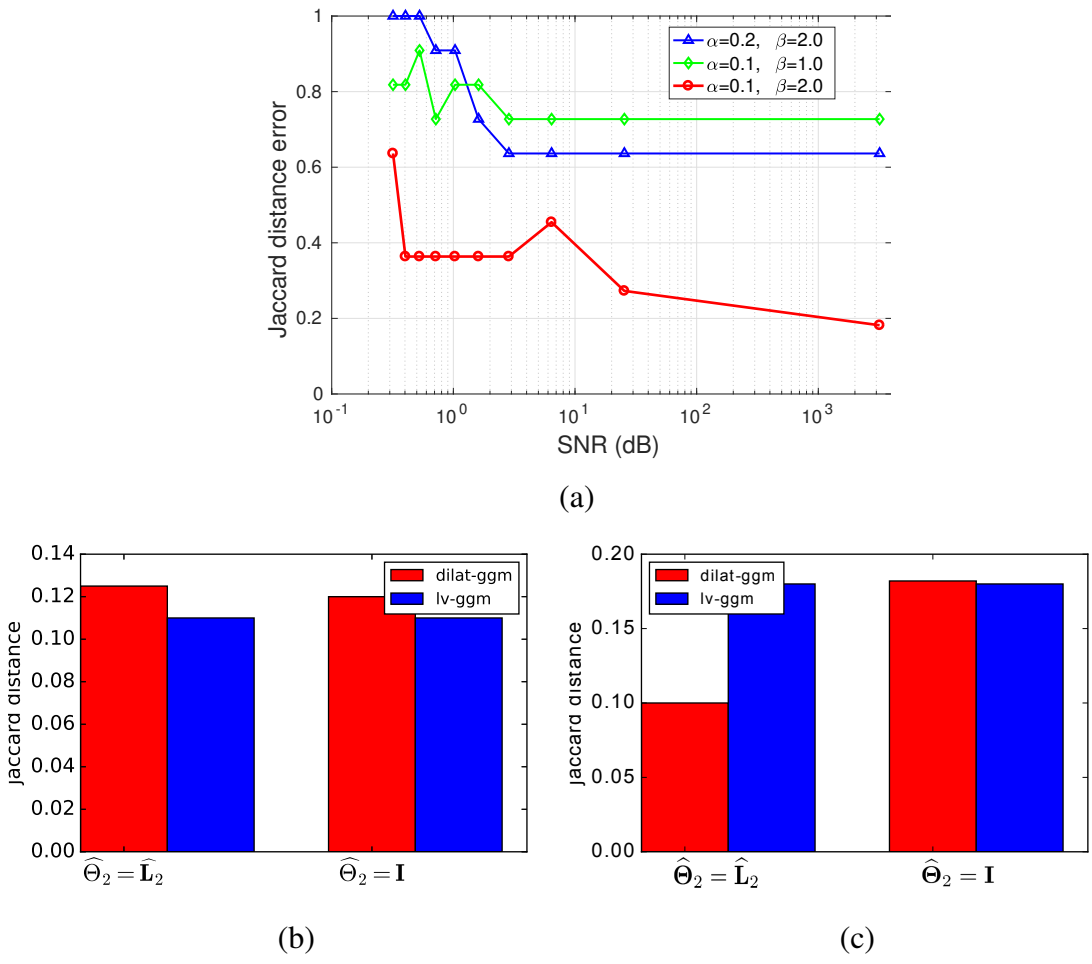


Figure 5.9: (a) The robustness of DiLat-GGM under different (α, β) when Θ_2 is corrupted. The underlying network is the same as Figure 5.7. Note that when the Signal-to-Noise Ratio (SNR) decreases, the performance of DiLat-GGM decreases. (b)-(c) A comparison between DiLat-GGM and LV-GGM when $\widehat{\mathbf{\Theta}}_2 = \widehat{\boldsymbol{L}}_2$ for the inverse covariance of $\boldsymbol{x}_2$ and when $\widehat{\mathbf{\Theta}}_2 = \boldsymbol{I}$. In (b), we use the same graph as in Figure 5.6 with equal conditional variance over $\boldsymbol{x}_2$. In (c), we use the same graph as in Figure 5.7. Note that when the non-informative prior $\widehat{\mathbf{\Theta}}_2 = \boldsymbol{I}$ is chosen, the performance of DiLat-GGM is slightly worse than that of LV-GGM due to its non-convexity. The performance of DiLat-GGM improves for a great amount when $\widehat{\mathbf{\Theta}}_2$ is known to fit the latent network $\mathcal{G}_2$. Also see that when the latent variables are all conditional independent with equal conditional variance, the identity matrix $\widehat{\mathbf{\Theta}}_2 = \boldsymbol{I}$ is optimal. In this case, the LV-GGM has better performance than DiLat-GGM.

In Figure 5.9 (a), we evaluate the robustness of DiLat-GGM when the pre-defined matrix $\widehat{\mathbf{\Theta}}_2$ is corrupted by noise. In specific, we define $\widehat{\mathbf{\Theta}}_2 = \widehat{\boldsymbol{L}}_2 + \sigma_L^2 \boldsymbol{G}$, where $\sigma_L \in [10^{-2}, 5]$, $\widehat{\boldsymbol{L}}_2$ is the inverse covariance of $\boldsymbol{x}_2$ evaluated using the ground truth data $\boldsymbol{x}_2$. The Signal-to-Noise Ratio (SNR) is defined as $\left(\log \frac{\|\widehat{\boldsymbol{L}}_2\|_F^2}{\|\sigma_L^2 \boldsymbol{G}\|_F^2} \right)$ (dB). It is seen that when the SNR

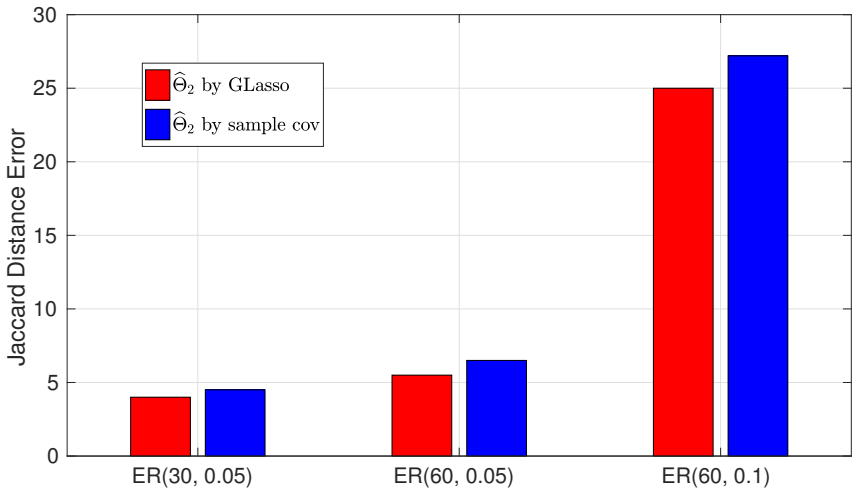


Figure 5.10: A comparison of the Jaccard distance error of DiLat-GGM when $\widehat{\boldsymbol{\Theta}}_2$ is estimated by the GLasso or the inverse of sample covariance matrix $\widehat{\boldsymbol{L}}_2$. The underling network is generated from the Erdős-Rényi (ER) graph model with different (n, p). See that using the GLasso as a precision matrix estimator, the DiLat-GGM has better performance compared to the case of the inverse of sample covariance matrix. This is because the GLasso esimator has lower variance compared with the inverse of sample covariance matrix.

decreases, the performance of DiLat-GGM deteriorates. However, it is also seen that when the noise level is within a range of relatively small values, the performance of DiLat-GGM is stable, indicating its robustness under the uncertainty in $\widehat{\boldsymbol{\Theta}}_2$. In Figure 5.9 (b) and (c), we compare the performance of DiLat-GGM and LV-GGM when $\widehat{\boldsymbol{\Theta}}_2 = \widehat{\boldsymbol{L}}_2$, inverse covariance of ground truth $\boldsymbol{x}_2$ and when $\widehat{\boldsymbol{\Theta}}_2 = \boldsymbol{I}$, the uniform prior. In (b), we choose the network as illustrated in Figure 5.6 with equal conditional variance over $\boldsymbol{x}_2$. In (c), we use the same graph as in Figure 5.7. It is seen that when $\widehat{\boldsymbol{\Theta}}_2 = \boldsymbol{I}$, no prior knowledge of latent variables is given, the performance of DiLat-GGM is slightly worse than that of LV-GGM, since DiLat-GGM is seen as a non-convex reformulation of LV-GGM and its performance is affected by the choice of initialization. However, when the prior knowledge regarding the dependency structure of latent variables is given, the performance of DiLat-GGM improves a lot since it utilizes the inner structure of latent variables to effectively reduce the number of latent variables in concern. This is equivalent to a feature selection procedure in latent space. Also, a comparison between Figure 5.9 (b) and (c) shows that when the latent variables are all conditional independent with equal conditional variance, the identity matrix $\widehat{\boldsymbol{\Theta}}_2 = \boldsymbol{I}$ is optimal. In this case, the LV-GGM has better performance than DiLat-GGM. In Figure 5.8, we compare the performance of DiLat-GGM and LV-GGM over the balanced tree network as in Figure 5.6 but with non-identical conditional variances over $\boldsymbol{x}_2$. It is seen that for DiLat-GGM, $\widehat{\boldsymbol{\Theta}}_2 = \text{diag}(\widehat{\boldsymbol{\theta}})$ performs better than $\widehat{\boldsymbol{\Theta}}_2 = \boldsymbol{I}$, since $\widehat{\boldsymbol{\theta}}$ accounts for the actual conditional variances in $\boldsymbol{x}_2$.

In Figure 5.10, we compare the performance of DiLat-GGM when the inverse of sample covariance matrix $\widehat{\boldsymbol{L}}_2$ is replaced by the GLasso precision matrix estimator. Note that since the GLasso estimator has lower variance compared to the inverse of sample covariance estimator, i.e., the DiLat-GGM has improved topology estimation accuracy.

5.5 Conclusion

We proposed the DiLat-GGM, a delayed-influence latent variable Gaussian graphical model to learn the conditional connectivity of a sub-graph of a GMM with partially observed data. By incorporating a row-sparsity regularization, DiLat-GGM performs the feature selection during the model selection and learns a linear mapping to estimate the latent variables. The problem involves solving a DC-programming and an efficient solver based on CCP and ADMM has been proposed. Theoretical analysis shows that the proposed algorithm guarantees to converge to a local stationary point. Experiments on synthetic dataset show its superior performance compared to the conventional Gaussian graphical model selection methods such as the Glasso and the LV-GGM. Future research directions including development of fast optimization method for large-scale datasets and extension of DiLat-GGM to learn both sub-networks $\mathcal{G}_1$ and $\mathcal{G}_2$ simultaneously.

5.6 Appendix

5.6.1 The EM algorithm to solve LV-GGM

A great advantage of LV-GGM is that it is a convex problem and the global optimal solution is guaranteed. Before Chandrasekaran et al. [Chandrasekaran et al., 2011, 2012], a natural way to solve a latent variable inference problem such as LV-GGM is via Expectation Maximization (EM) algorithm. EM algorithm is a heuristic based algorithm that minimizes the upper bound of the negative log-marginal likelihood functions (w.r.t. the latent variables)

$$\begin{aligned}\min_{\boldsymbol{\Theta}_1 \succeq \mathbf{0}} & -\log \int p(\boldsymbol{x}_1, \boldsymbol{x}_2|\boldsymbol{\Theta}_1) d\boldsymbol{x}_2 + \alpha \left\|\boldsymbol{\Theta}_1\right\|_1 \\ & \leq \mathbb{E}_{\boldsymbol{x}_2|\boldsymbol{x}_1, \widehat{\boldsymbol{\Theta}}^{(t)}} \left[-\log p(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{\Theta}_1)\right] - H(\boldsymbol{x}_2|\boldsymbol{x}_1, \widehat{\boldsymbol{\Theta}}^{(t)}) + \alpha \left\|\boldsymbol{\Theta}_1\right\|_1 . \end{aligned} \tag{5.17}$$

where the first term is the joint log-likelihood conditioned on the observed variables and $H(\boldsymbol{x}_2|\boldsymbol{x}_1, \widehat{\boldsymbol{\Theta}}^{(t)}) = -\mathbb{E}_{p_{\boldsymbol{x_2}|\boldsymbol{x}_1}} \left[\log p_{\boldsymbol{x_2}|\boldsymbol{x}_1}\right]$ is the Shannon entropy. We find the upper bound in (5.17)

$$\begin{aligned} Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(t)}) &:= \mathbb{E}_{\boldsymbol{x}_2|\boldsymbol{x}_1, \widehat{\boldsymbol{\Theta}}^{(t)}} \left[-\log p(\boldsymbol{x}_1, \boldsymbol{x}_2|\boldsymbol{\Theta}) + \alpha \left\|\boldsymbol{\Theta}_1\right\|_1\right] \\ &= -\log\det \boldsymbol{\Theta} + \operatorname{tr}\left(\mathbb{E}_{\boldsymbol{x}_2|\boldsymbol{x}_1, \widehat{\boldsymbol{\Theta}}^{(t)}} \left[\widehat{\boldsymbol{\Sigma}}\right] \boldsymbol{\Theta}\right) + \alpha \left\|\boldsymbol{\Theta}_1\right\|_1 \end{aligned}$$

where $\widehat{\boldsymbol{\Sigma}}$ is the covariance matrix for full data. Then using the fact that $P(\boldsymbol{x}_2|\boldsymbol{x}_1, \widehat{\boldsymbol{\Theta}}^{(t)})$ is also a Gaussian distribution with mean and covariance

$$\begin{aligned} \mathbb{E}\left[\boldsymbol{x}_2|\boldsymbol{x}_1, \widehat{\boldsymbol{\Theta}}^{(t)}\right] &= \widehat{\boldsymbol{\Sigma}}_{21}^{(t)} \left(\widehat{\boldsymbol{\Sigma}}_1^{(t)}\right)^{-1} \boldsymbol{x}_1 \\ \operatorname{Cov}(\boldsymbol{x}_2|\boldsymbol{x}_1, \widehat{\boldsymbol{\Theta}}^{(t)}) &= \widehat{\boldsymbol{\Sigma}}_2^{(t)} - \widehat{\boldsymbol{\Sigma}}_{21}^{(t)} \left(\widehat{\boldsymbol{\Sigma}}_1^{(t)}\right)^{-1} \widehat{\boldsymbol{\Sigma}}_{12}^{(t)} \end{aligned}$$

for $\widehat{\boldsymbol{\Sigma}}^{(t)} = \left(\widehat{\boldsymbol{\Theta}}^{(t)}\right)^{-1}$, the EM algorithm is described as below:

For $t = 1, \ldots,$ until convergence:

1. **M-step**: Find the estimate of joint inverse covariance matrix $\widehat{\boldsymbol{\Theta}}^{(t)}$ via graphical Lasso. That is solve the following problem

$$\widehat{\boldsymbol{\Theta}}^{(t)} = \arg\min_{\boldsymbol{\Theta} \succeq \mathbf{0}} -\log\det \boldsymbol{\Theta} + \operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}^{(t-1)} \boldsymbol{\Theta}\right) + \alpha \left\|\boldsymbol{J}_1^T \boldsymbol{\Theta} \boldsymbol{J}_1\right\|_1 \tag{5.18}$$

where $\boldsymbol{J}_1 := \begin{bmatrix} \boldsymbol{I}_{n_1} & \mathbf{0} \end{bmatrix}^T \in \mathbb{R}^{n \times n_1}$.

2. **E-step**: Find the conditional expectation of the full covariance $\widehat{\boldsymbol{\Sigma}}^{(t)} := \mathbb{E}_{\boldsymbol{x}_2|\boldsymbol{x}_1, \widehat{\boldsymbol{\Theta}}^{(t)}} \left[\widehat{\boldsymbol{\Sigma}}\right]$ given the observed data via imputation

$$\widehat{\boldsymbol{\Sigma}}^{(t)} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & -\boldsymbol{\Sigma}_1 \widehat{\boldsymbol{\Theta}}_{12}^{(t)} \left(\widehat{\boldsymbol{\Theta}}_2^{(t)}\right)^{-1} \\ -\left(\widehat{\boldsymbol{\Theta}}_2^{(t)}\right)^{-1} \widehat{\boldsymbol{\Theta}}_{21}^{(t)} \boldsymbol{\Sigma}_1 & \left(\widehat{\boldsymbol{\Theta}}_2^{(t)}\right)^{-1} + \widehat{\boldsymbol{\Theta}}_{21}^{(t)} \boldsymbol{\Sigma}_1 \widehat{\boldsymbol{\Theta}}_{12}^{(t)} \end{bmatrix} \tag{5.19}$$

where $\widehat{\boldsymbol{\Theta}}_2^{(t)} = \boldsymbol{I}$ if all hidden variables are conditional independent. Here $\boldsymbol{\Sigma}_1$ is the empirical covariance matrix on the observed node.

Note that compared to (5.13), the matrix in (5.19) is equal to $\begin{bmatrix} \boldsymbol{I} & \widehat{\boldsymbol{\Theta}}_2 \end{bmatrix} \boldsymbol{S}_1 (\boldsymbol{D}_t) \begin{bmatrix} \boldsymbol{I} \\ \widehat{\boldsymbol{\Theta}}_2 \end{bmatrix}$ except for the principal submatrix corresponding to the latent variables. Also in the CCP in (5.14), the conditional covariance $\widehat{\boldsymbol{\Theta}}_2$ is fixed, but the EM algorithm also learns $\widehat{\boldsymbol{\Theta}}_2$ in M-step.

5.6.2 Solving the latent variable Gaussian graphical model via ADMM

From the formulation of LV-GGM,

$$\begin{aligned} (\mathbf{S}^*, \mathbf{L}^*) = \arg\min_{\mathbf{L},\ \mathbf{S}} & -\frac{m}{2} \log\det(\mathbf{S} - \mathbf{L}) + \frac{m}{2} \text{tr}\left(\widehat{\boldsymbol{\Sigma}}_o (\mathbf{S} - \mathbf{L})\right) + \alpha_m \left(\lambda \|\mathbf{S}\|_1 + \|\mathbf{L}\|_*\right) \\ \text{s.t.} \quad & \mathbf{S} - \mathbf{L} \succeq \mathbf{0} \\ & \mathbf{L} \succeq \mathbf{0} \end{aligned}$$

we can instead separate the constraints and the log-likelihood function with additional copy of $\boldsymbol{R} := \boldsymbol{S} - \boldsymbol{L}$. Then the above problem becomes

$$\begin{aligned} (\mathbf{R}^*, \mathbf{S}^*, \mathbf{L}^*) = \arg\min_{\mathbf{L},\ \mathbf{R},\ \mathbf{S}} & -\frac{m}{2} \log\det(\mathbf{R}) + \frac{m}{2} \text{tr}\left(\widehat{\boldsymbol{\Sigma}}_o \mathbf{R}\right) + \alpha_m \|\mathbf{S}\|_1 + \gamma_m \|\mathbf{L}\|_* \\ \text{s.t.} \quad & \mathbf{R} - \mathbf{S} + \mathbf{L} = \mathbf{0} \\ & \mathbf{R} \succ \mathbf{0} \\ & \mathbf{L} \succeq \mathbf{0}. \end{aligned}$$

Denote $\boldsymbol{Z} := (\boldsymbol{R}, \boldsymbol{S}, \boldsymbol{L})$. We can again separate out the linear equality constraint with another copy $\boldsymbol{Z}' := (\boldsymbol{R}', \boldsymbol{S}', \boldsymbol{L}')$. Therefore, we solve the ADMM with concensus constraint

$$\begin{aligned}\min_{\mathbf{L},\,\mathbf{R},\,\mathbf{S},\,\mathbf{L}',\,\mathbf{R}',\,\mathbf{S}'} \quad & -\frac{m}{2}\log\det(\mathbf{R}) + \frac{m}{2}\mathrm{tr}\left(\widehat{\boldsymbol{\Sigma}}_o\mathbf{R}\right) + \alpha_m\|\mathbf{S}\|_1 + \gamma_m\|\mathbf{L}\|_* + \mathbb{1}\left\{\mathbf{R}' - \mathbf{S}' + \mathbf{L}' = 0\right\} \\ \text{s.t.} \quad & \mathbf{R} - \mathbf{S} + \mathbf{L} = \mathbf{R}' - \mathbf{S}' + \mathbf{L}' \\ & \mathbf{R} \succ \mathbf{0} \\ & \mathbf{L} \succeq \mathbf{0}. \end{aligned} \tag{5.20}$$

ADMM relies on the easy computation of the proximal projection

$$\mathrm{Prox}(\boldsymbol{Z}, \xi) := \min_{\boldsymbol{R}} \frac{1}{2\xi}\|\boldsymbol{R} - \boldsymbol{Z}\|_F^2 + \mathcal{R}(\boldsymbol{R})$$

1. For $\boldsymbol{R} \succ \mathbf{0}$,

$$\mathrm{Prox}_{\boldsymbol{R}}(\boldsymbol{Z}, \xi) := \min_{\boldsymbol{R}} \frac{1}{2\xi}\|\boldsymbol{R} - \boldsymbol{Z}\|_F^2 - \log\det(\mathbf{R}) + \mathrm{tr}\left(\widehat{\boldsymbol{\Sigma}}_o\mathbf{R}\right)$$

which has optimal solution

$$\begin{aligned} \boldsymbol{R} &= \boldsymbol{U}\mathrm{diag}(\boldsymbol{\gamma})\boldsymbol{U}^T \\ \text{for } \xi\widehat{\boldsymbol{\Sigma}}_o - \boldsymbol{Z} &:= \boldsymbol{U}\mathrm{diag}(\boldsymbol{\sigma})\boldsymbol{U}^T \\ \gamma_i &= \frac{-\sigma_i + \sqrt{\sigma_i^2 + 4\,\xi}}{2} \end{aligned}$$

2. For $\boldsymbol{S}$ sparse,

$$\mathrm{Prox}_{\boldsymbol{S}}(\boldsymbol{Z}, \xi) := \min_{\boldsymbol{S}} \frac{1}{2\xi}\|\boldsymbol{S} - \boldsymbol{Z}\|_F^2 + \alpha\|\boldsymbol{S}\|_1$$

which has optimal solution

$$\boldsymbol{S} = \text{soft-threshold}(\boldsymbol{Z}, \xi\,\alpha)$$

3. For $\boldsymbol{L} \succeq \mathbf{0}$ low-rank,

$$\mathrm{Prox}_{\boldsymbol{L}}(\boldsymbol{Z}, \xi) := \min_{\boldsymbol{L}} \frac{1}{2\xi}\|\boldsymbol{L} - \boldsymbol{Z}\|_F^2 + \gamma\|\boldsymbol{L}\|_* + \mathbb{1}\left\{\boldsymbol{L} \succeq \mathbf{0}\right\}$$

which has optimal solution

$$\begin{aligned} \boldsymbol{L} &= \boldsymbol{U}\text{diag}\,(\boldsymbol{\zeta})\,\boldsymbol{U}^T \\ \text{for } \boldsymbol{Z} &:= \boldsymbol{U}\text{diag}\,(\boldsymbol{\sigma})\,\boldsymbol{U}^T \\ \zeta_i &= \max\{\sigma_i - \xi\,\gamma, 0\} \end{aligned}$$

Finally we need to adjust $\boldsymbol{Z} := (\boldsymbol{R}, \boldsymbol{S}, \boldsymbol{L})$, according to the dual variables $\boldsymbol{\Lambda} := (\boldsymbol{\Lambda}_R.\boldsymbol{\Lambda}_S, \boldsymbol{\Lambda}_L)$. The ADMM solution is as below:

1. Update $\boldsymbol{W} := \boldsymbol{Z} + \mu\boldsymbol{\Lambda}$;
2. Find new $\boldsymbol{Z} := (\boldsymbol{R}, \boldsymbol{S}, \boldsymbol{L})$, given $\boldsymbol{W} := (\boldsymbol{W}_R, \boldsymbol{W}_S, \boldsymbol{W}_L)$ and $\xi = \mu$, via proximal projection as above;
3. Update $\boldsymbol{T} := \boldsymbol{Z} - \mu\boldsymbol{\Lambda}$;
4. Update

$$\begin{aligned} \boldsymbol{R}' &= \boldsymbol{T}_R - (\boldsymbol{T}_R - \boldsymbol{T}_S + \boldsymbol{T}_L)/3 \\ \boldsymbol{S}' &= \boldsymbol{T}_S + (\boldsymbol{T}_R - \boldsymbol{T}_S + \boldsymbol{T}_L)/3 \\ \boldsymbol{L}' &= \boldsymbol{T}_L - (\boldsymbol{T}_R - \boldsymbol{T}_S + \boldsymbol{T}_L)/3 \end{aligned}$$

5. Update the dual variables via direction of multipliers

$$\boldsymbol{\Lambda} \leftarrow \boldsymbol{\Lambda} - \frac{1}{\mu}(\boldsymbol{Z} - \boldsymbol{Z}')$$

Note that the original problem is convex, so the ADMM (5.20) guarantee to converge to the global minimal. See [Ma et al., 2013] for details.

5.6.3 Solving subproblem (5.14) using ADMM

Denote $\widehat{\boldsymbol{\Theta}}_2^{-1} = \boldsymbol{T}$. By introducing an auxiliary variable $\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}_1 & \boldsymbol{P}_{12} \\ \boldsymbol{P}_{21} & \boldsymbol{P}_2 \end{bmatrix} = \boldsymbol{R}$ and $\boldsymbol{W} = \widehat{\boldsymbol{\Theta}}_2\boldsymbol{P}_{21}$, problem (5.14) becomes

$$\begin{aligned} \min_{\boldsymbol{R},\boldsymbol{P},\boldsymbol{W}} \quad & -\log\det\boldsymbol{R} + \text{tr}\,(\boldsymbol{S}\boldsymbol{R}) + \alpha_m \left\|\boldsymbol{J}_1^T\,\boldsymbol{P}\,\boldsymbol{J}_1\right\|_1 + \beta_m \|\boldsymbol{W}\|_{2,1} \\ \text{s.t.} \quad & \boldsymbol{J}_2^T\,\boldsymbol{P}\,\boldsymbol{J}_2 = \boldsymbol{T} \end{aligned}$$

$$
\begin{aligned}
&\boldsymbol{R} = \boldsymbol{P} \\
&\boldsymbol{W} = \boldsymbol{Q}^T \boldsymbol{P} \boldsymbol{J}_1 \\
&\boldsymbol{R} \succeq \mathbf{0}.
\end{aligned}
$$

where $\boldsymbol{J}_1 := \begin{bmatrix} \boldsymbol{I}_{n_1} & \mathbf{0} \end{bmatrix}^T \in \mathbb{R}^{n \times n_1}$, $\boldsymbol{J}_2 := \begin{bmatrix} \mathbf{0} & \boldsymbol{I}_{n_2} \end{bmatrix}^T \in \mathbb{R}^{n \times n_2}$, $\boldsymbol{Q} := \begin{bmatrix} \mathbf{0} & \widehat{\boldsymbol{\Theta}}_2 \end{bmatrix}^T \in \mathbb{R}^{n \times n_2}$. Following the ADMM procedure, we form an augmented Lagrangian as

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{R}, \boldsymbol{P}) = &- \log\det \boldsymbol{R} + \operatorname{tr}(\boldsymbol{S}\boldsymbol{R}) + \alpha_m \left\| \boldsymbol{J}_1^T \boldsymbol{P} \boldsymbol{J}_1 \right\|_1 + \beta_m \|\boldsymbol{W}\|_{2,1} \\
&+ \mathbb{1}\{\boldsymbol{R} \succeq \mathbf{0}\} + \mathbb{1}\left\{\boldsymbol{J}_2^T \boldsymbol{P} \boldsymbol{J}_2 - \boldsymbol{T} = \mathbf{0}\right\} + \operatorname{tr}\left(\boldsymbol{\Lambda}^T(\boldsymbol{R} - \boldsymbol{P})\right) + \frac{\rho}{2}\|\boldsymbol{R} - \boldsymbol{P}\|_F^2 \\
&+ \operatorname{tr}\left(\boldsymbol{\Lambda}_w^T\left(\boldsymbol{W} - \widehat{\boldsymbol{\Theta}}_2 \boldsymbol{P}_{21}\right)\right) + \frac{\rho_w}{2}\left\|\boldsymbol{W} - \widehat{\boldsymbol{\Theta}}_2 \boldsymbol{P}_{21}\right\|_F^2,
\end{aligned}
$$

where $\boldsymbol{\Lambda} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{\Lambda}_w$ form dual matrices. ADMM minimizes the augmented Lagrangian via block coordinate descent. In specific, it solves two separable problems:

$$
\begin{aligned}
\min_{\boldsymbol{R}} \quad & - \log\det \boldsymbol{R} + \operatorname{tr}(\boldsymbol{S}\boldsymbol{R}) + \operatorname{tr}\left(\boldsymbol{\Lambda}^T(\boldsymbol{R} - \boldsymbol{P})\right) + \frac{\rho}{2}\|\boldsymbol{R} - \boldsymbol{P}\|_F^2 \\
& = - \log\det \boldsymbol{R} + \operatorname{tr}(\boldsymbol{S}\boldsymbol{R}) + \frac{\rho}{2}\left\|\boldsymbol{R} - \boldsymbol{P} + \frac{1}{\rho}\boldsymbol{\Lambda}\right\|_F^2 \\
\text{s.t.} \quad & \boldsymbol{R} \succeq \mathbf{0},
\end{aligned}
\tag{5.21}
$$

and

$$
\begin{aligned}
\min_{\boldsymbol{P}, \boldsymbol{W}} \quad & \alpha_m \left\| \boldsymbol{J}_1^T \boldsymbol{P} \boldsymbol{J}_1 \right\|_1 + \beta_m \|\boldsymbol{W}\|_{2,1} + \operatorname{tr}\left(\boldsymbol{\Lambda}^T(\boldsymbol{R} - \boldsymbol{P})\right) + \frac{\rho}{2}\|\boldsymbol{R} - \boldsymbol{P}\|_F^2 \\
& \quad + \operatorname{tr}\left(\boldsymbol{\Lambda}_w^T\left(\boldsymbol{W} - \widehat{\boldsymbol{\Theta}}_2 \boldsymbol{P}_{21}\right)\right) + \frac{\rho_w}{2}\left\|\boldsymbol{W} - \widehat{\boldsymbol{\Theta}}_2 \boldsymbol{P}_{21}\right\|_F^2 \\
& = \alpha_m \left\| \boldsymbol{J}_1^T \boldsymbol{P} \boldsymbol{J}_1 \right\|_1 + \beta_m \|\boldsymbol{W}\|_{2,1} + \frac{\rho}{2}\left\|\boldsymbol{P} - \boldsymbol{R} - \frac{1}{\rho}\boldsymbol{\Lambda}\right\|_F^2 \\
& \quad + \frac{\rho_w}{2}\left\|\boldsymbol{W} - \widehat{\boldsymbol{\Theta}}_2 \boldsymbol{P}_{21} - \frac{1}{\rho_w}\boldsymbol{\Lambda}_w\right\|_F^2 \\
\text{s.t.} \quad & \boldsymbol{J}_2^T \boldsymbol{P} \boldsymbol{J}_2 = \boldsymbol{T}.
\end{aligned}
\tag{5.22}
$$

From Section 5.6.2, we see that (5.21) corresponds to a proximal operator

$$
\operatorname{Prox}_{\boldsymbol{R}}(\boldsymbol{Z}, \xi) := \min_{\boldsymbol{R} \succ \mathbf{0}} \frac{1}{2\xi}\|\boldsymbol{R} - \boldsymbol{Z}\|_F^2 - \log\det(\mathbf{R}) + \operatorname{tr}(\boldsymbol{S}\mathbf{R}). \tag{5.23}
$$

The optimal solution of above satisfies that the gradient of the objective function

$$\frac{1}{\xi}(\boldsymbol{R}-\boldsymbol{Z})-\boldsymbol{R}^{-1}+\boldsymbol{S}=0.$$

Let the eigen-decomposition of $\xi\boldsymbol{S}-\boldsymbol{Z}:=\boldsymbol{U}\mathrm{diag}\,(\boldsymbol{\sigma})\,\boldsymbol{U}^T$, where $\boldsymbol{\sigma}:=(\sigma_i)$. Then the optimal solution

$$\begin{aligned}\boldsymbol{R}&=\boldsymbol{U}\mathrm{diag}\,(\boldsymbol{\gamma})\,\boldsymbol{U}^T\\ \text{where }\gamma_i&=\frac{-\sigma_i+\sqrt{\sigma_i^2+4\,\xi}}{2}>0.\end{aligned}$$

To solve (5.22), we see that the objective of (5.22) is separable as well. Problem (5.22) is equivalent to

$$\begin{aligned}\min_{\boldsymbol{P}_1,\boldsymbol{P}_{21},\boldsymbol{W}}\ \ &\alpha_m\left\|\boldsymbol{P}_1\right\|_1+\frac{\rho}{2}\left\|\boldsymbol{P}_1-\boldsymbol{R}_1-\frac{1}{\rho}\boldsymbol{\Lambda}_1\right\|_F^2\\ &+\beta_m\left\|\boldsymbol{W}\right\|_{2,1}+\frac{\rho}{2}\left\|\boldsymbol{P}_{21}-\boldsymbol{R}_{21}-\frac{1}{\rho}\boldsymbol{\Lambda}_{21}\right\|_F^2+\frac{\rho_w}{2}\left\|\boldsymbol{W}-\widehat{\boldsymbol{\Theta}}_2\boldsymbol{P}_{21}-\frac{1}{\rho_w}\boldsymbol{\Lambda}_w\right\|_F^2\end{aligned}\tag{5.24}$$

and $\boldsymbol{P}_2=\boldsymbol{T}$. It involves three proximal operators: first,

$$\mathrm{Prox}_{\boldsymbol{P}_1,\alpha}(\boldsymbol{Z},\xi):=\min_{\boldsymbol{P}_1}\frac{1}{2\xi}\left\|\boldsymbol{P}_1-\boldsymbol{Z}\right\|_F^2+\alpha\left\|\boldsymbol{P}_1\right\|_1$$

which is equivalent to

$$\mathrm{Prox}_{\boldsymbol{P}_1,\alpha}(\boldsymbol{Z},\xi)=\text{soft-threshold}(\boldsymbol{Z},\xi\,\alpha).\tag{5.25}$$

Then,

$$\mathrm{Prox}_{\boldsymbol{P}_{21}}(\boldsymbol{Z},\boldsymbol{Z}^{'},\xi,\xi_w):=\min_{\boldsymbol{P}_{21}}\ \frac{1}{2\xi}\left\|\boldsymbol{P}_{21}-\boldsymbol{Z}\right\|_F^2+\frac{1}{2\xi_w}\left\|\widehat{\boldsymbol{\Theta}}_2\boldsymbol{P}_{21}-\boldsymbol{Z}^{'}\right\|_F^2$$

This is a linear transformation

$$\begin{aligned}&\mathrm{Prox}_{\boldsymbol{P}_{21}}(\boldsymbol{Z},\boldsymbol{Z}^{'},\xi,\xi_w)\\ &=\left(\xi_w\boldsymbol{I}+\xi\widehat{\boldsymbol{\Theta}}_2^2\right)^{-1}\left(\xi_w\boldsymbol{Z}+\xi\widehat{\boldsymbol{\Theta}}_2\boldsymbol{Z}^{'}\right)\\ &=\boldsymbol{U}\mathrm{diag}\left[\frac{\xi_w}{\xi_w+\xi\lambda_i^2}\right]_{i,i}\boldsymbol{U}^T\boldsymbol{Z}+\boldsymbol{U}\mathrm{diag}\left[\frac{\xi\lambda_i}{\xi_w+\xi\lambda_i^2}\right]_{i,i}\boldsymbol{U}^T\boldsymbol{Z}^{'}\end{aligned}\tag{5.26}$$

Algorithm 5 DiLat-GGM subproblem via ADMM

Require: Positive definite matrix $\boldsymbol{S} \succ \mathbf{0}$ and $\boldsymbol{S} \in \mathbb{R}^{n\times n}$. The nonnegative regularization parameter $\alpha, \beta > 0$. The pre-defined nonegative definite matrix $\widehat{\boldsymbol{\Theta}}_2 \succeq \mathbf{0}$ and $\widehat{\boldsymbol{\Theta}}_2 \in \mathbb{R}^{n_2\times n_2}$. Let $\boldsymbol{T} = \widehat{\boldsymbol{\Theta}}_2^{-1}$. Let $n_1 = n - n_2$. Dual update parameter $\mu, \mu_w > 0$.

1: **Initialize:** Choose an random matrix $\boldsymbol{R}^{(0)} = \begin{bmatrix} \boldsymbol{R}_1^{(0)} & \boldsymbol{R}_{12}^{(0)} \\ \boldsymbol{R}_{21}^{(0)} & \boldsymbol{R}_2^{(0)} \end{bmatrix} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{R}^{(0)} \succ \mathbf{0}$. $\boldsymbol{\Lambda}^{(0)} = \mathbf{0} \in \mathbb{R}^{n\times n} = \begin{bmatrix} \boldsymbol{\Lambda}_1^{(0)} & \boldsymbol{\Lambda}_{12}^{(0)} \\ \boldsymbol{\Lambda}_{21}^{(0)} & \boldsymbol{\Lambda}_2^{(0)} \end{bmatrix}$. $\boldsymbol{\Lambda}_W^{(0)} = \mathbf{0} \in \mathbb{R}^{n_2\times n_1}$. Let $\boldsymbol{P}^{(0)} = \begin{bmatrix} \boldsymbol{P}_1^{(0)} & \boldsymbol{P}_{12}^{(0)} \\ \boldsymbol{P}_{21}^{(0)} & \boldsymbol{P}_2^{(0)} \end{bmatrix} = \boldsymbol{R} \in \mathbb{R}^{n\times n}$. Choose $\boldsymbol{W}^{(0)} = \widehat{\boldsymbol{\Theta}}_2 \boldsymbol{P}_{21}^{(0)}$.

2: **for** $t = 1, \ldots, T$ or until converge **do**

3: Find $\boldsymbol{P}_1^{(t)} \in \mathbb{R}^{n_1\times n_1}$ via $\boldsymbol{P}_1^{(t)} = \text{Prox}_{\boldsymbol{P}_1,\alpha}(\boldsymbol{R}_1^{(t-1)} + \mu\boldsymbol{\Lambda}_1^{(t-1)}, \mu)$ as in (5.25);

4: **if** $\widehat{\boldsymbol{\Theta}}_2 := \text{diag}\left(\widehat{\boldsymbol{\Theta}}_2\right)$ **then**

5: Find $\boldsymbol{P}_{21}^{(t)} \in \mathbb{R}^{n_2\times n_1}$ via $\boldsymbol{P}_{21}^{(t)} = \text{Prox}'_{\boldsymbol{P}_{21},\beta}(\boldsymbol{R}_{21}^{(t-1)} + \mu\boldsymbol{\Lambda}_{21}^{(t-1)}, \mu)$ as in (5.29)

6: **else**

7: Find $\boldsymbol{W}^{(t)} \in \mathbb{R}^{n_2\times n_1}$ via $\boldsymbol{W}^{(t)} = \text{Prox}_{\boldsymbol{W},\beta}(\widehat{\boldsymbol{\Theta}}_2\boldsymbol{P}_{21}^{(t-1)} - \mu_w\boldsymbol{\Lambda}_W^{(t-1)}, \mu_w)$ as in (5.27);

8: Find $\boldsymbol{P}_{21}^{(t)} = \text{Prox}_{\boldsymbol{P}_{21}}(\boldsymbol{R}_{21}^{(t-1)} + \mu\boldsymbol{\Lambda}_{21}^{(t-1)}, \boldsymbol{W}^{(t)} + \mu_w\boldsymbol{\Lambda}_W^{(t-1)}, \mu, \mu_w)$ as in (5.26);

9: Update dual variables $\boldsymbol{\Lambda}_W$.

$$\boldsymbol{\Lambda}_W^{(t)} = \boldsymbol{\Lambda}_W^{(t-1)} + \frac{1}{\mu_w}\left(\boldsymbol{W}^{(t)} - \widehat{\boldsymbol{\Theta}}_2\boldsymbol{P}_{21}^{(t)}\right)$$

10: **end if**

11: Set $\boldsymbol{P}_2^{(t)} = \boldsymbol{T}$ and $\boldsymbol{P}_{12}^{(t)} = \left(\boldsymbol{P}_{21}^{(t)}\right)^T$. Construct $\boldsymbol{P}^{(t)}$.

12: Find $\boldsymbol{R}^{(t)} \in \mathbb{R}^{n\times n}$ via $\boldsymbol{R}^{(t)} = \text{Prox}_{\boldsymbol{R},\alpha}(\boldsymbol{P}^{(t)} - \mu\boldsymbol{\Lambda}^{(t-1)}, \mu)$ as in (5.23).

13: Update dual variables $\boldsymbol{\Lambda}$

$$\boldsymbol{\Lambda}^{(t)} = \boldsymbol{\Lambda}^{(t-1)} + \frac{1}{\mu}\left(\boldsymbol{R}^{(t)} - \boldsymbol{P}^{(t)}\right).$$

14: **end for**

Ensure: Output $(\boldsymbol{R}^{(T)}, \boldsymbol{P}^{(T)})$ if $\widehat{\boldsymbol{\Theta}}_2$ is diagonal and $(\boldsymbol{R}^{(T)}, \boldsymbol{P}^{(T)}, \boldsymbol{W}^{(T)})$ otherwise.

where $\widehat{\boldsymbol{\Theta}}_2 = \boldsymbol{U}\text{diag}\left[\lambda_i\right]_{i,i}\boldsymbol{U}^T$ is the eigen-decomposition. And the proximal operator

$$\text{Prox}_{\boldsymbol{W},\beta}(\boldsymbol{Z}^{'}, \xi) := \min_{\boldsymbol{W}\in\mathbb{R}^{n_2\times n_1}} \quad \frac{1}{2\xi}\left\|\boldsymbol{W} - \boldsymbol{Z}^{'}\right\|_F^2 + \beta\left\|\boldsymbol{W}\right\|_{2,1},$$

which has optimal solution $\boldsymbol{W}$ with i-th row

$$\boldsymbol{W}_i = \left(1 - \frac{\beta\xi}{\left\|\boldsymbol{Z}_i^{'}\right\|_2}\right)_+ \boldsymbol{Z}_i^{'}, \quad i = 1, \ldots, n_2 \tag{5.27}$$

Note that if the matrix $\widehat{\boldsymbol{\Theta}}_2$ is a diagonal matrix $\widehat{\boldsymbol{\Theta}}_2 = \text{diag}\left(\widehat{\boldsymbol{\theta}}_2\right)$, where $\boldsymbol{\theta}_2 := (\theta_{21}, \ldots \theta_{2,n_2}) \in$

$\mathbb{R}^{n_2}, \theta_{2,i} > 0, i = 1, \ldots, n_2$, we can compute $\boldsymbol{P}_{21}$ directly without introducing $\boldsymbol{W}$ and $\boldsymbol{\Lambda}_w$ using the proximal operator

$$\text{Prox}'_{\boldsymbol{P}_{21},\beta}(\boldsymbol{Z}, \xi) := \min_{\boldsymbol{P}_{21} \in \mathbb{R}^{n_2 \times n_1}} \frac{1}{2\xi} \|\boldsymbol{P}_{21} - \boldsymbol{Z}\|_F^2 + \beta \left\| \widehat{\boldsymbol{\Theta}}_2 \boldsymbol{P}_{21} \right\|_{2,1} \tag{5.28}$$

The optimal solution $\boldsymbol{P}_{21}$ of problem (5.28) has its i-th row

$$(\boldsymbol{P}_{21})_i = \left(1 - \frac{\theta_{2,i}\beta\xi}{\|\boldsymbol{Z}_i\|_2}\right)_+ \boldsymbol{Z}_i, \quad i = 1, \ldots, n_2 \tag{5.29}$$

where $(x)_+ := \max\{x, 0\}$.

Finally, we have the dual updates

$$\begin{aligned} \boldsymbol{\Lambda}^{(t)} &:= \boldsymbol{\Lambda}^{(t-1)} + \rho(\boldsymbol{R} - \boldsymbol{P}) \\ \boldsymbol{\Lambda}_w^{(t)} &:= \boldsymbol{\Lambda}_w^{(t-1)} + \rho_w \left(\boldsymbol{W} - \widehat{\boldsymbol{\Theta}}_2 \boldsymbol{P}_{21}\right) \end{aligned}$$

The algorithm of ADMM is summarized in Algorithm 5.

CHAPTER 6

Conclusion, Discussion and Future Research Directions

6.1 Conclusion and Discussion

In many practical machine learning problems, information comes from multiple sources. Modern information processing systems, such as recommendation systems, vision and audio processing systems, control systems and automated vehicle systems, often need to handle large-scale multi-view data. There is therefore a need for data analyst to develop methodologies that naturally accommodate large-scale data from multiple sources. This thesis proposes approaches that are motivated by information theory and robust multi-view learning. Of principal concern is the robustness of the learning algorithm in the presence of noisy corruption, multi-view inconsistency, intervention of external sources and various uncertainties within the processing system. As seen in Chapter 2, information theory provides a great variety of measures and divergences, which quantify the amount of information and uncertainties shared among multiple systems. This thesis focused on a small subset of information theoretic measures, including KL-divergence in robust learning, multi-view learning and graphical model inference. Use of other divergence measures such as the α-divergence [Hero et al., 2001, Cichocki et al., 2007, Póczos and Schneider, 2011], f-divergence [Moon and Hero, 2014b] and Hellinger distance [Hellinger, 1909, Beran, 1977, Lindsay, 1994, Cutler and Cordero-Brana, 1996, Cieslak et al., 2012] may be worthwhile extensions of this work. Like KL-divergence, these measures and divergences have been shown to be robust to the presence of noise and mixture components in some settings. This make them natural candidates to study in the context of extending beyond the role of KL-divergence in Chapter 3 and Chapter 4.

The success and popularity of KL-divergence in robust multi-view learning lies in its close association with the exponential family, graphical model and information geometry. Each of these fields has extraordinary rich contents and they are all united under the

framework of maximum entropy learning. This unique property of KL-divergence makes it preferable to other measures as a robust surrogate function in machine learning and statistics. For our work, maximum entropy learning plays a key role, related to the role of maximum likelihood estimation in machine learning and statistics. As seen in Chapter 2, maximum entropy learning and maximum likelihood estimation are conjugate dual to each other from the perspective of convex analysis and information geometry. However, the role of data in both problems are different: in maximum likelihood, the learning objective to choose a parametric model that matches the empirical distribution of data; in maximum entropy, data are used to generate constraints and an non-parametric model is learned from prior distribution and data constraints. In the former case, data are trusted and dominate the learning process, but in the latter case, data constraints are allowed to fluctuate, allowing more flexibility in modeling. Maximum entropy learning explicitly separates out the task-related structural information and the task-independent model information. It then allows us to extend and combine different learning tasks into a single unified framework. This is the basis for our development in Chapter 3 and Chapter 4.

One of the principal contributions of this thesis is the development of multi-view interpretation of graph signal analysis. This provides a fresh perspective at the intersection of the fields of network analysis and graphical models. The former focuses on the representation, inference and characterization of network topology. The latter focuses on the representation and inference of the structure of high dimensional data. The definition of smooth graph signals in graph signal processing (GSP) brings together the correlation between the view of graph and the view of data on graph: the behavior of a datum is closely associated with the position of its corresponding vertex within the network. Unlike multi-view learning algorithms discussed in Chapter 1, the multi-view graph signal learning is formulated as a hierarchical model, with a graph layer on the top of a data layer. Before GSP, the idea of coupling graph and data views was pursued independently in social network analysis through definitions of various centrality measures [Barthelemy, 2004, Newman, 2005, Brandes, 2008, Jackson, 2010], and in graphical models through inverse covariance estimation, correlation screening [Hero and Rajaratnam, 2011, Firouzi et al., 2013] and model selection methods [Ravikumar et al., 2008, Anandkumar et al., 2011, Ravikumar et al., 2010]. For GSP, such correlation naturally resides in the eigenvalues and eigenvectors of graph Laplaican matrix. The eigenvalues of graph Laplacian reflect the invariant and geometric property of the graph and the eigenvectors of graph Laplacian form an orthonormal basis in a function space of graph signals. The development of alternative hierarchical models in analyzing the coupling effect between data and network topology is an important extension of this thesis. This serves as a future research direction.

To some extent, the effort of graph signal processing is still unidirectional: the graph signal analysis is based on prior knowledge on the underling network topology. The alternative direction which seeks to infer graph topology given data is also important in knowledge discovery and data representation. In graphical models, this is the problem of the graphical model selection [Ravikumar et al., 2008, Anandkumar et al., 2011, Ravikumar et al., 2010]. However, for general graphical models, learning graphical model topology is still challenging for large-scale data. Even for Gaussian graphical models, where efficient learning algorithms exist, it is difficult to impose additional topological constraint on learning task. In terms of this, our work in Chapter 5 generalizes the standard Gaussian graphical model by imposing a sparsity constraint on the cross edges between two clusters. We show that by borrowing the strength of multi-view learning graphical models with specific topological structure is possible.

We summarize the main contributions of the thesis as follows.

- Chapter 3 was dedicated to robust maximum entropy learning. In particular, we solved a classification problem when the underling feature distribution is a mixture of anomalous and nominal distributions. The proposed GEM-MED generalizes the standard maximum entropy discrimination (MED) by minimizing the generalization error of the classifier with respect to a nominal data distribution. To circumvent the difficulty in learning the support of nominal distribution for high dimensional data, we exploited the versatility of the kernel method in combination with the power of minimal-entropy-sets. This allows one to perform classification and anomaly detection simultaneously under a unified maximum entropy learning framework. We demonstrated its performance advantages in terms of both classification accuracy and detection rate on a simulated data set and on a real footstep data set, compared to the state-of-the-art robust learning algorithms.

- Chapter 4 addressed the problem of label uncertainty and multi-view inconsistency in multi-view classification problem. In many applications such as video surveillance and multi-media retrieval, labels are collected via crowd sourcing or from less controlled environments such as the Internet. To avoid using corrupted labels directly, it is assumed that we are provided a set of label distributions associated with the dataset. The label distribution measures the uncertainty of assigning label to each given sample. In Chapter 4, we proposed a multi-view maximum entropy learning model on statistical manifolds via stochastic consensus constraints. In particular, the Kullback-Liebler divergence was used to measure the dissimilarity of information contents in different views. The resulting consensus-view distribution is the *Karcher*

mean [Nielsen and Bhatia, 2013, Nielsen et al., 2013] of multiple view-specific distributions on the statistical manifold. An efficient algorithm based on constrained EM was proposed to learn the consensus-view distribution and multiple view-specific distributions iteratively. Experiments showed that the proposed COM-MED method is robust in the presence of corruption and outliers and it achieved superior classification performance over other multi-view learning methods.

- Chapter 5 extended the problem of multi-view learning to network topology inference problem. Relational database naturally has a two-view representation. It consists of a set of measurements taken at nodes of a graph whose edges specify pairwise node dependencies. In Chapter 5, the joint distribution of the measurements is assumed to be Gaussian distributed with sparse inverse covariance matrix whose zero entries are specified by the topology of the graph. The objective is to estimate the (non-marginal) sub-graph associated with the set of directly measured nodes. With the help of an external source, which provides a noisy summary of dependency structure among latent data, we proposed the DiLat-GGM, which generalizes the existing LV-GGM by taking into account the local effect of the latent variables explicitly. The proposed DiLat-GGM includes a latent feature selection procedure by introducing additional row sparsity structure on the conditional cross-covariance matrix. From a multi-view learning perspective, DiLat-GGM is seen as learning the sub-network by combining both the reliable proprietary information from an internal source and the unreliable information from an external source. Experiments on synthesis dataset showed that DiLat-GGM improve over LV-GGM and GLasso by a margin in terms of the edge selection accuracy. The proposed model is well-suited for applications such as competitive pricing models where two companies operate in a market where they can only directly measure the behaviors of their own customers.

Despite our contributions in this thesis that advanced the state-of-the-art, the understanding of robust multi-view learning remains incomplete. In the next section, we will discuss some interesting research topics that might be fruitful pursuits.

6.2 Directions for Future Research

In this section, several interesting topics and projects are provided.

6.2.1 Multi-view Gaussian Graphical Model Selection

In Chapter 4, we proposed a multi-view maximum entropy discrimination model for categorical data analysis. It is natural to extend this work to handle high-dimensional real-valued data. In specific, assume that the joint distribution of response and covariates are Gaussian with sparse inverse covariance matrix whose support set defines the topology of a graph. Also assume that data are collected from different information sources, and they shared the common response variables. Our task is to fuse Gaussian graphical models from different source in order to learn a consensus graphical model whose structure reflects the shared information among different views. This work can be used to learn shared information for multi-layer network, which is useful in biological analysis, social network analysis and sensor network analysis.

Assume that the response $\boldsymbol{y} \in \mathbb{R}^d$ is high-dimensional. For each view i, the covariates are denoted as $\boldsymbol{x}^i \in \mathbb{R}^s, i = 1, \ldots, V$. Assume that $p_i(\boldsymbol{y}, \boldsymbol{x}^i) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Theta}^i)$ where $\boldsymbol{\Theta}^i := \begin{bmatrix} \boldsymbol{\Theta}^i_y & \boldsymbol{\Theta}^i_{y,x} \\ \boldsymbol{\Theta}^i_{x,y} & \boldsymbol{\Theta}^i_x \end{bmatrix}$ is the precision matrix for view i. Thus the predictive distribution $p_i(\boldsymbol{y}|\boldsymbol{x}^i) = \mathcal{N}(\boldsymbol{T}_i^T \boldsymbol{x}^i, \boldsymbol{\Theta}^i_y), i = 1, \ldots, V$ for some $\boldsymbol{T}_i = \boldsymbol{\Theta}^i_{x,y}(\boldsymbol{\Theta}^i_y)^{-1}$. Denote $\boldsymbol{\Theta}^i_{x,y} := \boldsymbol{B}_i$ so that $\boldsymbol{B}_i^T \boldsymbol{x}^i = \boldsymbol{\Theta}^i_y \boldsymbol{T}_i^T \boldsymbol{x}^i := \boldsymbol{\Theta}^i_y \boldsymbol{\mu}_{y|x^i}$. We have no knowledge regarding the conditional cross covariance $\boldsymbol{\Theta}^{i,j}$ between two different views $i \neq j$ and it is not easy to estimate due to the high dimensionality of remaining data. Using the KL-divergence, it is known that the Karcher mean of multiple Gaussian graphical models is also Gaussian. Denote the mean of consensus Gaussian model as $\boldsymbol{\mu}_c$ and precision matrix $\boldsymbol{\Theta}_c$. Our task is to infer the predictive Gaussian graphical model for each view $p_i(\boldsymbol{y}|\boldsymbol{x}^i; \boldsymbol{B}_i, \boldsymbol{\Theta}^i_y)$ and a consensus distribution $q(\boldsymbol{y}; \boldsymbol{\mu}_c, \boldsymbol{\Theta}_c)$.

Following the formulation of COM-MED in Chapter 4, we can formulate a multi-view predictive Gaussian graphical model learning as

$$
\begin{aligned}
\min_{\substack{\boldsymbol{B}_i, \boldsymbol{\Theta}^i_y, i=1,\ldots,V \\ \boldsymbol{\mu}_c, \boldsymbol{\Theta}_c, q \in \Delta}} \quad & \sum_{i=1}^{V} \left\{ \mathcal{L}_i(\boldsymbol{y}, \boldsymbol{x}^i; \boldsymbol{B}_i, \boldsymbol{\Theta}^i_y) + \alpha_i \left\| \boldsymbol{\Theta}^i_y \right\|_1 \right\} + \beta \left\| \boldsymbol{\Theta}_c \right\|_1 \\
\text{s.t.} \quad & \boldsymbol{\Theta}^i_y \succeq \boldsymbol{0}, \quad i = 1, \ldots, V \\
& \boldsymbol{\Theta}_c \succeq \boldsymbol{0} \\
& \sum_{i=1}^{V} \mathbb{KL}\left(q(\boldsymbol{y}; \boldsymbol{\mu}_c, \boldsymbol{\Theta}_c) \parallel p_i(\boldsymbol{y}|\boldsymbol{x}^i; \boldsymbol{B}_i, \boldsymbol{\Theta}^i_y) \right) \leq \rho
\end{aligned}
$$

where $\mathcal{L}_i(\boldsymbol{y}, \boldsymbol{x}^i; \boldsymbol{B}_i, \boldsymbol{\Theta}^i_y) = -\log p_i(\boldsymbol{y}|\boldsymbol{x}^i; \boldsymbol{B}_i, \boldsymbol{\Theta}^i_y)$ is the conditional negative log-likelihood loss function in view i. $\alpha_i, \beta, \rho > 0$ are all fixed non-negative parameters, They can be set

by cross-validations.

This problem can be seen as a hierarchical graphical model selection problem with a linear mapping $\boldsymbol{B}_i$ connecting the layer of covariates $\boldsymbol{x}^i$ and the layer of response $\boldsymbol{y}$. Moreover, the conditional negative log-likelihood function $\mathcal{L}(\boldsymbol{y}, \boldsymbol{x}^i; \boldsymbol{B}_i, \boldsymbol{\Theta}_y^i)$ is non-convex in $(\boldsymbol{B}_i, \boldsymbol{\Theta}_y^i)$ due to the log-partition function. Note that the Karcher mean constraint is convex-concave, since it is convex in q and concave in each p_i.

Note that the Karcher mean of zero mean Gaussian graphical model on statistical manifold is equivalent to the Karcher mean of a set of positive definite matrices on matrix manifold. Matrix information geometry [Ando, 1979, Bhatia, 2003, Bhatia and Holbrook, 2006, Nielsen and Bhatia, 2013, Nielsen et al., 2013, Cherian et al., 2013] has provided a solid mathematical foundation for research in this field. It can also be seen as a two-layer multi-view generative neural network with linear neurons. Therefore, this work can be viewed as learning multi-layer network with hierarchical models.

6.2.2 Multi-view Generative Adversarial Network

Neural network are essentially hierarchical graphical models with non-linear connections between successive layers. Learning a multi-view neural network to combine data of different type is a challenging task, especially when data contain images. In the past five years, the most significant advances in the field of deep learning is the development of Generative Adversarial Network (GAN) in [Goodfellow et al., 2014]. GAN trains two neural networks with competitive goals: a generative model that captures the data distribution, and a discriminative model that estimates the probability that a sample came from the training data rather than the generative model. During the training step, the generative model maximizes the error probability of the discriminative model. This framework corresponds to a minimax two-player game [Nisan et al., 2007]. GAN has been the state-of-the-art algorithm in domain adaptation [Ajakan et al., 2014, Ganin et al., 2016], text-to-image synthesis [Reed et al., 2016], semi-supervised learning [Springenberg, 2015] and multi-modal learning [Liu and Tuzel, 2016]. For instance, in [Springenberg, 2015], a discriminative classifier is learned using GAN under the semi-supervised setting. Their method is based on an objective function that trades-off mutual information between observed examples and their predicted label distribution, against robustness of the classifier to an adversarial generative model. In [Reed et al., 2016], the GAN is used to learn a conditional multi-modal distribution to generate image from text data. According to all these recent developments, GAN is a promising method for robust multi-view learning model and it is capable to learn high dimensional multi-modal distribution.

The idea of GAN resembles the use of KL-divergence as a consensus measure in Chapter 4, so we can use our multi-view learning framework to learn *multi-view GAN*. Note that the KL-divergence is also a min-max objective with respect to the consensus view and each individual view. As opposed to our work in Chapter 4, we can learn a multi-view model using competitive objectives. In particular, we train a set of discriminative view-specific models and a generative consensus-view model. The goal for each individual model to maximize their disagreement with the consensus model. And the role of generative consensus-view model is to fool these adversarially-trained view-specific classifiers into predicting that the synthesis data are real. The primary goal is to learn a multi-modal consensus-view model from multi-view data that captures the feature characteristic of each view. This will be one of interesting projects to pursue in future.

6.2.3 Dimensionality Reduction of Graph Signal with Gaussian Graphical Models

A topic not directly addressed in this thesis is the problem of *dimensionality reduction on high dimensional graph signals*. That is, we are given a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a high-dimensional graph signal data $\boldsymbol{x} \in \mathbb{R}^p$ whose dependency is depicted by $\mathcal{G}$. The data are indexed by $\mathcal{V}$. Our goal is to learn a proximity matrix that preserve both the similarity of data and the topological structure of the network. We can use some of the results in this thesis to achieve this using Gaussian graphical model and kernel approximation.

Dimensionality reduction [Peason, 1901, Kruskal, 1964, Schölkopf et al., 1997, Tipping and Bishop, 1999, Tenenbaum et al., 2000, Roweis and Saul, 2000, Jolliffe, 2002, Donoho and Grimes, 2003, Belkin and Niyogi, 2003, Lawrence, 2004, 2005, Lafon and Lee, 2006, van der Maaten et al., 2009] is an indispensable technique for data analyst to understand the structure of high dimensional data. To achieve dimensionality reduction, it is commonly required to pre-compute a proximity matrix to measure all pair-wise distances between sample objects. Then the low-dimensional representations of data are learned under the condition that the pairwise proximity is preserved. If the high-dimensional data lie in a low-dimensional smooth manifold, measures such as the k-nearest-neighbor distance can be used to compute proximity matrix that resembles the geodesic distance on the manifold [Tenenbaum et al., 2000, Roweis and Saul, 2000]. Other way to find such proximity is by metric learning [Xing et al., 2002, Weinberger et al., 2006, Davis et al., 2007], which learns a positive definite matrix under the constraints that the cluster information is preserved under the projected space. According to spectral graph theory [Chung, 1997, Coifman and Lafon, 2006], a graph is a discrete approximation of a smooth manifold, where the shortest

path between vertices resembles the geodesic curve between two points on manifold. In terms of this, Gaussian graphical models, which describe a random smooth mapping from graph vertex domain to the Euclidean space, can be used to approximate the manifold coordinate map as well. In [Lawrence, 2004], Lawrence proposed the Gaussian process latent variable model (GPLVM). GPLVM learns the proximity matrix directly from the sample covariance matrix of high dimensional data. In addition to this, it uses a kernel function to project the high dimensional data in $\mathbb{R}^p$ to a low dimensional Euclidean space $\mathbb{R}^d$ induced by kernel function. The learned projection preserves the proximity measure. The problem of GPLVM is that the sample covariance matrix cannot characterize the topological structure of the underling graph, thus the learned proximity cannot capture the existing topological information in the network.

Denote the proximity matrix as a symmetric matrix $\boldsymbol{H} \in \mathbb{R}^{n\times n}$, where $H_{i,j} := \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2$, where $\boldsymbol{z} \in \mathbb{R}^d$ is some low-dimensional representation of data $\boldsymbol{x} \in \mathbb{R}^p$. Let $\boldsymbol{x}$ be a graph signal on graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Denote $\boldsymbol{M}_{\mathcal{E}} = [\mathbb{1}\{(i,j) \in \mathcal{E}\}]_{i,j}$ as a mask of edge set $\mathcal{E}$. Assume that $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is a set of n i.i.d data. Denote the sample Gram matrix $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}^T/p$. Define a kernel map $K_\theta : \mathbb{R} \to \mathbb{R}$, which is operated on each entry of $\boldsymbol{H}_{i,j} = \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2$ independently. For instance, for heat kernel, $K_{i,j} = \exp(-\theta\, \boldsymbol{H}_{i,j})$. Assume that $K \in \mathcal{K}_\theta$, where $\mathcal{K}_\theta$ is the class of kernel functions parameterized by θ. The problem of learning a proximity matrix can be solved under the framework of Gaussian graphical model inference, i.e.

$$
\begin{aligned}
\min_{\theta,\boldsymbol{H}} \quad & -\log\det K_\theta(\boldsymbol{H}) + \operatorname{tr}\left(\boldsymbol{S}K_\theta(\boldsymbol{H})\right) \\
\text{s.t.} \quad & K_\theta(\boldsymbol{H}) \succeq \boldsymbol{0} \\
& [K_\theta(\boldsymbol{H})]_{i,j} \le \epsilon, \quad (i,j) \notin \mathcal{E} \\
& \operatorname{rank}(\boldsymbol{H}) \le r \\
& 0 \le \boldsymbol{H}_{i,j} \le \rho(\epsilon),
\end{aligned}
$$

where $r, \rho, \epsilon > 0$ are fixed parameters. The first two constraints ensures that the learned kernel matrix is a valid precision matrix and fit the given graph topology. The last two constraints finds a valid proximity matrix. Note that for the distance-induced matrix $\boldsymbol{H}$, it should be low-rank. The choice of ρ and ϵ are not independent, since the kernel function is monotonic decreasing with respect to the distance proximity.

The work of *Monotonic Single-Index model* [Kakade et al., 2011, Foster et al., 2013, Ganti et al., 2015] provides similar formulation as above, which may be served as a starting point of our model. Note that compared with Monotonic Single-Index model, our model

use a LogDet Divergence (Chapter 2). It is thus expected to achieve more robustness from the above formulation. Future work remains to find an efficient solution for above problem.

BIBLIOGRAPHY

Marcel R Ackermann and Johannes Blömer. Bregman clustering for separable instances. In *Scandinavian Workshop on Algorithm Theory*, pages 212–223. Springer, 2010.

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

Pankaj K Agarwal, Siu-Wing Cheng, Yufei Tao, and Ke Yi. Indexing uncertain data. *Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 137–146, 2009.

Ameya Agaskar and Yue M Lu. A spectral graph uncertainty principle. *IEEE Transactions on Information Theory*, 59(7):4338–4356, 2013.

Charu C Aggarwal, Joel L Wolf, Kun-Lung Wu, and Philip S Yu. Horting hatches an egg: A new graph-theoretic approach to collaborative filtering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 201–212. ACM, 1999.

Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 131–142, 1966.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

Animashree Anandkumar, Vincent Tan, and Alan S Willsky. High-dimensional graphical model selection: tractable graph families and necessary conditions. *Advances in Neural Information Processing Systems*, pages 1863–1871, 2011.

Rie Kubota Ando and Tong Zhang. Learning on graph with laplacian regularization. *Advances in neural information processing systems*, 19:25, 2007.

Tsuyoshi Ando. Concavity of certain maps on positive definite matrices and applications to hadamard products. *Linear Algebra and its Applications*, 26:203–241, 1979.

Aamir Anis, Akshay Gadde, and Antonio Ortega. Towards a sampling theorem for signals on arbitrary graphs. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3864–3868. IEEE, 2014.

Aamir Anis, Aly El Gamal, Salman Avestimehr, and Antonio Ortega. Asymptotic justification of bandlimited interpolation of graph signals for semi-supervised learning. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5461–5465. IEEE, 2015.

Andreas Argyriou, Mark Herbster, and Massimiliano Pontil. Combining graph laplacians for semi–supervised learning. In *Advances in Neural Information Processing Systems*, pages 67–74, 2005.

Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1): 1–106, 2012.

Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh, Srujana Merugu, and Dharmendra S Modha. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. *Journal of Machine Learning Research*, 8(Aug):1919–1986, 2007.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre dAspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

Marc Barthelemy. Betweenness centrality in large complex networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):163–168, 2004.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

Michèle Basseville. Divergence measures for statistical data processingan annotated bibliography. *Signal Processing*, 93(4):621–633, 2013.

Homayoon Beigi. Metrics and divergences. *Fundamentals of Speaker Recognition*, pages 301–311, 2011.

Jan Beirlant, Edward J Dudewicz, László Györfi, and Edward C Van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. *Journal of Computer and System Sciences*, 74(8):1289–1308, 2008.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.

Rudolf Beran. Minimum hellinger distance estimates for parametric models. *The Annals of Statistics*, pages 445–463, 1977.

Alessandro Bergamo and Lorenzo Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. *Advances in Neural Information Processing Systems*, pages 181–189, 2010.

Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

Dimitri P Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

Dimitri P Bertsekas. *Convex optimization algorithms*. Athena Scientific Belmont, 2015.

Rajendra Bhatia. On the exponential metric increasing property. *Linear Algebra and its applications*, 375:211–220, 2003.

Rajendra Bhatia and John Holbrook. Riemannian geometry and matrix geometric means. *Linear algebra and its applications*, 413(2-3):594–618, 2006.

Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, pages 401–406, 1946.

A Bhattachayya. On a measure of divergence between two statistical population defined by their population distributions. *Bulletin Calcutta Mathematical Society*, 35:99–109, 1943.

Türker Biyikoğu, Peter F Stadler, and Josef Leydold. *Laplacian eigenvectors of graphs: Perron-Frobenius and Faber-Krahn Type Theorems. Lecture Notes in Mathematics*. Springer, 2007.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

David M Blei, Michael I Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128, 2006.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

Joseph Bockhorst and Mark Craven. Exploiting relations among concepts to acquire weakly labeled training data. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 43–50. Morgan Kaufmann Publishers Inc., 2002.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Ulrik Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145, 2008.

Samuel L Braunstein and Carlton M Caves. Statistical distance and the geometry of quantum states. *Physical Review Letters*, 72(22):3439, 1994.

Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. LOF: identifying density-based local outliers. *ACM Sigmod Record*, 29(2):93–104, 2000.

S Fi Burch, SF Gull, and John Skilling. Image restoration by a powerful maximum entropy method. *Computer Vision, Graphics, and Image Processing*, 23(2):113–128, 1983.

Kevin M Carter, Raviv Raich, and Alfred O Hero. Spherical laplacian information maps (slim) for dimensionality reduction. *Statistical Signal Processing, 2009. SSP'09. IEEE/SP 15th Workshop on*, pages 405–408, 2009.

Kevin M Carter, Raviv Raich, William G Finn, and Alfred O HeroIII. Information-geometric dimensionality reduction. *IEEE Signal Processing Magazine*, 28(2):89–99, 2011.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2): 572–596, 2011.

Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, 2012.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

Eugene Charniak. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 132–139. Association for Computational Linguistics, 2000.

Sotirios Chatzis. Infinite markov-switching maximum entropy discrimination machines. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 729–737, 2013.

CL Philip Chen and Chun-Yang Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275:314–347, 2014.

Siheng Chen, Rohan Varma, Aliaksei Sandryhaila, and Jelena Kovačević. Discrete signal processing on graphs: Sampling theory. *IEEE Transactions on Signal Processing*, 63 (24):6510–6523, 2015.

Yilun Chen, Ami Wiesel, and Alfred O Hero. Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Transactions on Signal Processing*, 59(9):4097–4107, 2011.

Sundeep Prabhakar Chepuri, Sijia Liu, Geert Leus, and Alfred O Hero III. Learning sparse graphs under smoothness prior. *arXiv preprint arXiv:1609.03448*, 2016.

Anoop Cherian, Suvrit Sra, Arindam Banerjee, and Nikolaos Papanikolopoulos. Jensen-bregman logdet divergence with application to efficient similarity search for covariance matrices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2161–2174, 2013.

Myung Jin Choi, Venkat Chandrasekaran, and Alan S Willsky. Gaussian multiresolution models: Exploiting sparse markov and covariance structure. *IEEE Transactions on Signal Processing*, 58(3):1012–1024, 2010a.

Seung-Seok Choi, Sung-Hyuk Cha, and Charles C Tappert. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48, 2010b.

C Christoudias, Raquel Urtasun, and Trevor Darrell. Multi-view learning in the presence of view disagreement. *Proceedings of the Converence on Uncertainty in Artificial Intelligence (UAI)*, 2008.

Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

Andrzej Cichocki, Rafal Zdunek, Seungjin Choi, Robert Plemmons, and Shun-Ichi Amari. Non-negative tensor factorization using alpha and beta divergences. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 3, pages III–1393. IEEE, 2007.

David A Cieslak, T Ryan Hoens, Nitesh V Chawla, and W Philip Kegelmeyer. Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, 24(1):136–158, 2012.

Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

Michael Collins and Yoram Singer. Unsupervised models for named entity classification. *Proceedings of the joint SIGDAT conference on empirical methods in natural language processing and very large corpora*, pages 100–110, 1999.

Thomas H Cormen. *Introduction to algorithms*. MIT press, 2009.

Jose A Costa and Alfred O Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing*, 52(8):2210–2221, 2004.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew McCallum, Tom Mitchell, Kamal Nigam, and Seán Slattery. Learning to construct knowledge bases from the world wide web. *Artificial intelligence*, 118(1):69–113, 2000.

I Csisz et al. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Publ. Math. Inst. Hungar. Acad.*, 8: 95–108, 1963.

Adele Cutler and Olga I Cordero-Brana. Minimum hellinger distance estimation for finite mixture models. *Journal of the American Statistical Association*, 91(436):1716–1723, 1996.

Rui Dai and Ian F Akyildiz. A spatial correlation model for visual information in wireless multimedia sensor networks. *IEEE Transactions on Multimedia*, 11(6):1148–1159, 2009.

Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, pages 193–200. ACM, 2007.

Thyagaraju Damarla. Seismic and ultrasonic data analysis for characterizing people and animals. *SPIE Defense, Security, and Sensing*, 2012.

Thyagaraju Damarla, Asif Mehmood, and James Sabatier. Detection of people and animals using non-imaging sensors. In *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pages 1–8. IEEE, 2011.

Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

Sanjoy Dasgupta, Michael L Littman, and David McAllester. PAC generalization bounds for co-training. *Advances in neural information processing systems*, 1:375–382, 2002.

Alexandre d'Aspremont, Onureena Banerjee, and Laurent El Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30 (1):56–66, 2008.

Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, pages 101–126, 2006.

Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007.

Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Inderjit S Dhillon and Suvrit Sra. Generalized nonnegative matrix approximations with bregman divergences. In *Advances in neural information processing systems*, volume 18, 2005.

Steven Diamond and Stephen Boyd. CVXPY: A python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Guoru Ding, Qihui Wu, Yu-Dong Yao, Jinlong Wang, and Yingying Chen. Kernel-based learning for statistical signal processing in cognitive radio networks: Theoretical foundations, example applications, and future directions. *IEEE Signal Processing Magazine*, 30(4):126–136, 2013.

Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst. Learning laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 64(23):6160–6173, 2016.

David L Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100 (10):5591–5596, 2003.

John Duchi, Stephen Gould, and Daphne Koller. Projected subgradient methods for learning sparse gaussians. *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 153–160, 2008.

Shinto Eguchi and Shogo Kato. Entropy and divergence associated with power function and the statistical application. *Entropy*, 12(2):262–274, 2010.

Jason Farquhar, David Hardoon, Hongying Meng, John S Shawe-taylor, and Sandor Szedmak. Two view learning: SVM-2K, theory and practice. In *Advances in neural information processing systems*, pages 355–362, 2005.

Jacques Ferber. *Multi-agent systems: an introduction to distributed artificial intelligence*, volume 1. Addison-Wesley Reading, 1999.

Raihana Ferdous et al. An efficient k-means algorithm integrated with jaccard distance measure for document clustering. In *Internet, 2009. AH-ICI 2009. First Asian Himalayas International Conference on*, pages 1–6. IEEE, 2009.

Hamed Firouzi, Bala Rajaratnam, and Alfred Hero III. Predictive correlation screening: Application to two-stage predictor design in high dimension. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 274–288, 2013.

Pedro A Forero, Vassilis Kekatos, and Georgios B Giannakis. Robust clustering using outlier-sparsity regularization. *IEEE Transactions on Signal Processing*, 60(8):4163–4177, 2012.

Jared C Foster, Jeremy MG Taylor, and Bin Nan. Variable selection in monotone single-index models via the adaptive lasso. *Statistics in medicine*, 32(22):3944–3954, 2013.

Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

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

Robert G Gallager. *Information theory and reliable communication*, volume 2. Springer, 1968.

Kuzman Ganchev, João V Graça, John Blitzer, and Ben Taskar. Multi-view learning over structured and non-identical outputs. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2008.

Kuzman Ganchev, Joao Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 11:2001–2049, 2010.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Ravi Sastry Ganti, Laura Balzano, and Rebecca Willett. Matrix completion under monotonic single index models. In *Advances in Neural Information Processing Systems*, pages 1873–1881, 2015.

Izrail Moiseevitch Gelfand, Richard A Silverman, et al. *Calculus of variations*. Courier Corporation, 2000.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734. IEEE, 2005.

M Grant, S Boyd, V Blondel, S Boyd, and H Kimura. CVX: Matlab software for disciplined convex programming, version 2.0. *Recent Advances in Learning and Control*, pages 95–110, 2012.

Peter D Grünwald and A Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *Annals of Statistics*, pages 1367–1433, 2004.

Stephen F Gull and John Skilling. Maximum entropy method in image processing. *Communications, Radar and Signal Processing, IEE Proceedings F*, 131(6):646–659, 1984.

David L Hall and James Llinas. An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1):6–23, 1997.

David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.

David Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

Ernst Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271, 1909.

Alfre O Hero, Bing Ma, Olivier JJ Michel, and John Gorman. Applications of entropic spanning graphs. *IEEE signal processing magazine*, 19(5):85–95, 2002.

Alfred Hero and Bala Rajaratnam. Large-scale correlation screening. *Journal of the American Statistical Association*, 106(496):1540–1552, 2011.

Alfred O Hero. Geometric entropy minimization (GEM) for anomaly detection and localization. *Advances in Neural Information Processing Systems*, pages 585–592, 2006.

Alfred O Hero and Olivier JJ Michel. Asymptotic theory of greedy approximations to minimal k-point random graphs. *IEEE Transactions on Information Theory*, 45(6):1921–1938, 1999.

Alfred O Hero, Bing Ma, Olivier Michel, and John Gorman. Alpha-divergence for classification, indexing and retrieval. *Communication and Signal Processing Laboratory, Technical Report CSPL-328, U. Mich*, 2001.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Cho-Jui Hsieh, Inderjit S Dhillon, Pradeep K Ravikumar, and Mátyás A Sustik. Sparse inverse covariance matrix estimation using quadratic approximation. *Advances in neural information processing systems*, pages 2330–2338, 2011.

Cho-Jui Hsieh, Mátyás A Sustik, Inderjit S Dhillon, Pradeep K Ravikumar, and Russell Poldrack. Big & quic: Sparse inverse covariance estimation for a million variables. *Advances in Neural Information Processing Systems*, pages 3165–3173, 2013.

Cho-Jui Hsieh, Mátyás A Sustik, Inderjit S Dhillon, and Pradeep Ravikumar. Quic: quadratic approximation for sparse inverse covariance estimation. *Journal of Machine Learning Research*, 15(1):2911–2947, 2014.

Po-Sen Huang, Thyagaraju Damarla, and Mark Hasegawa-Johnson. Multi-sensory features for personnel detection at border crossings. In *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pages 1–8. IEEE, 2011.

David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, 1925.

Yuri Ivanov, Bruce Blumberg, and Alex Pentland. Expectation maximization for weakly labeled data. *ICML*, pages 218–225, 2001.

Tommi Jaakkola, Marina Meila, and Tony Jebara. Maximum entropy discrimination. *Advances in Neural Information Processing Systems*, 1999.

Paul Jaccard. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz, 1901.

Matthew O Jackson. *Social and economic networks*. Princeton university press, 2010.

Nathalie Japkowicz and Mohak Shah. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4): 620, 1957a.

Edwin T Jaynes. Information theory and statistical mechanics. ii. *Physical review*, 108(2): 171, 1957b.

Tony Jebara. Multitask sparsity via maximum entropy discrimination. *The Journal of Machine Learning Research*, 12:75–110, 2011.

Jiwoon Jeon and R Manmatha. Using maximum entropy for automatic image annotation. In *International Conference on Image and Video Retrieval*, pages 24–32. Springer, 2004.

Apoorva Jindal and Konstantinos Psounis. Modeling spatially-correlated sensor network data. In *Sensor and Ad Hoc Communications and Networks, 2004. IEEE SECON 2004. 2004 First Annual IEEE Communications Society Conference on*, pages 162–171. IEEE, 2004.

Apoorva Jindal and Konstantinos Psounis. Modeling spatially correlated data in sensor networks. *ACM Transactions on Sensor Networks (TOSN)*, 2(4):466–499, 2006.

Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

Dan Jurafsky and James H Martin. *Speech and language processing*, volume 3. Pearson, 2014.

Thomas Kailath. The divergence and Bhattacharyya distance measures in signal selection. *Communication Technology, IEEE Transactions on*, 15(1):52–60, 1967.

Sham M Kakade and Dean P Foster. Multi-view regression via canonical correlation analysis. *Learning Theory*, pages 82–96, 2007.

Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pages 927–935, 2011.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

Won Hwa Kim, Seong Jae Hwang, Nagesh Adluru, Sterling C Johnson, and Vikas Singh. Adaptive signal recovery on graphs via harmonic analysis for experimental design in neuroimaging. In *European Conference on Computer Vision*, pages 188–205. Springer, 2016.

Lawrence A Klein. *Sensor and data fusion: a tool for information assessment and decision making*, volume 324. SPIE press Bellingham WA, 2004.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Nir Krause and Yoram Singer. Leveraging the margin more carefully. *Proceedings of the twenty-first international conference on Machine learning*, page 63, 2004.

Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

Brian Kulis, Mátyás A Sustik, and Inderjit S Dhillon. Low-rank kernel learning with bregman matrix divergences. *Journal of Machine Learning Research*, 10(Feb):341–376, 2009.

Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Sanjiv Kumar and Martial Hebert. Discriminative fields for modeling spatial dependencies in natural images. In *NIPS*, volume 16, pages 1531–1538, 2003.

Morton Kupperman. Probabilities of hypotheses and information-statistics in sampling from exponential-class populations. *The Annals of Mathematical Statistics*, pages 571–575, 1958.

Nicholas Kushmerick. Learning to remove internet advertisements. *Proceedings of the third annual conference on Autonomous Agents*, pages 175–181, 1999.

John D Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers Inc., 2001.

Stephane Lafon and Ann B Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE transactions on pattern analysis and machine intelligence*, 28(9):1393–1403, 2006.

Gert R Lanckriet and Bharath K Sriperumbudur. On the convergence of the concave-convex procedure. In *Advances in neural information processing systems*, pages 1759–1767, 2009.

Kenneth Lange, David R Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics*, 9(1):1–20, 2000.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

Neil Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6(Nov):1783–1816, 2005.

Neil D Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, 16(3):329–336, 2004.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. A maximum entropy framework for part-based texture and object recognition. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 832–838. IEEE, 2005.

Wenke Lee and Dong Xiang. Information-theoretic measures for anomaly detection. In *Security and Privacy, 2001. S&P 2001. Proceedings. 2001 IEEE Symposium on*, pages 130–143. IEEE, 2001.

M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Bruce G Lindsay. Efficiency versus robustness: the case for minimum hellinger distance and related methods. *The annals of statistics*, pages 1081–1114, 1994.

Thomas Lipp and Stephen Boyd. Variations and extension of the convex–concave procedure. *Optimization and Engineering*, 17(2):263–287, 2016.

Meizhu Liu and Baba C Vemuri. Robust and efficient regularized boosting using total bregman divergence. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2897–2902. IEEE, 2011.

Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 469–477, 2016.

Philip M Long and Rocco A Servedio. Random classification noise defeats all convex potential boosters. *Machine Learning*, 78(3):287–304, 2010.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

Clifford Lynch. Big data: How do your data grow? *Nature*, 455(7209):28–29, 2008.

Shiqian Ma, Lingzhou Xue, and Hui Zou. Alternating direction methods for latent variable gaussian graphical model selection. *Neural computation*, 25(8):2172–2198, 2013.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997.

SZ Mahmoodabadi, A Ahmadian, and MD Abolhasani. Ecg feature extraction using daubechies wavelets. In *Proceedings of the fifth IASTED International conference on Visualization, Imaging and Image Processing*, pages 343–348, 2005.

P Martin Mai and Gregory C Beroza. A spatial random field model to characterize complexity in earthquake slip. *Journal of Geophysical Research: Solid Earth*, 107(B11), 2002.

Stéphane Mallat. *A wavelet tour of signal processing*. Academic press, 1999.

Robert Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–7. Association for Computational Linguistics, 2002.

Gideon S Mann and Andrew McCallum. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *The Journal of Machine Learning Research*, 11:955–984, 2010.

Christopher D Manning, Hinrich Schütze, et al. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

Goran Marjanovic and Alfred O Hero. ℓ_0 sparse inverse covariance estimation. *IEEE Transactions on Signal Processing*, 63(12):3218–3231, 2015.

Antonio G Marques, Santiago Segarra, Geert Leus, and Alejandro Ribeiro. Stationary graph processes and spectral estimation. *arXiv preprint arXiv:1603.04667*, 2016.

Hamed Masnadi-Shirazi and Nuno Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. *Advances in neural information processing systems*, pages 1049–1056, 2009.

Viktor Mayer-Schönberger and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

Rahul Mazumder and Trevor Hastie. The graphical lasso: New insights and alternatives. *Electronic journal of statistics*, 6:2125, 2012.

Andrew McAfee, Erik Brynjolfsson, Thomas H Davenport, DJ Patil, and Dominic Barton. Big data. *The management revolution. Harvard Bus Rev*, 90(10):61–67, 2012.

Jonathan Mei and José MF Moura. Signal processing on graphs: Causal modeling of unstructured data. *IEEE Transactions on Signal Processing*, 2016.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

Zhaoshi Meng, Brian Eriksson, and Al Hero. Learning latent variable gaussian graphical models. *Proceedings of The 31st International Conference on Machine Learning*, pages 1269–1277, 2014.

Paul Mermelstein. Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, 116:374–388, 1976.

Gerald Minerbo. Ment: A maximum entropy algorithm for reconstructing a source from projection data. *Computer Graphics and Image Processing*, 10(1):48–68, 1979.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

Kevin Moon and Alfred Hero. Multivariate f-divergence estimation with confidence. *Advances in Neural Information Processing Systems*, pages 2420–2428, 2014a.

Kevin R Moon and Alfred O Hero. Ensemble estimation of multivariate f-divergence. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 356–360. IEEE, 2014b.

Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. *Advances in neural information processing systems*, pages 10–18, 2012.

Peter Müller and Fernando A Quintana. Nonparametric bayesian data analysis. *Statistical science*, pages 95–110, 2004.

Noboru Murata, Takashi Takenouchi, Takafumi Kanamori, and Shinto Eguchi. Information geometry of u-boost and bregman divergence. *Neural Computation*, 16(7):1437–1481, 2004.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Ion Muslea, Steven Minton, and Craig A Knoblock. Active+ semi-supervised learning= robust multi-view learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 435–442. ACM, 2002.

Mohammad Mahdi Naghsh, Mahmoud Modarres-Hashemi, Shahram ShahbazPanahi, Mojtaba Soltanalian, and Petre Stoica. Unified optimization framework for multi-static radar code design using information-theoretic criteria. *IEEE Transactions on Signal Processing*, 61(21):5401–5416, 2013.

Sunil K Narang, Akshay Gadde, and Antonio Ortega. Signal processing techniques for interpolation in graph structured data. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5445–5449, 2013a.

Sunil K Narang, Akshay Gadde, Eduard Sanou, and Antonio Ortega. Localized iterative methods for interpolation in graph structured data. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 491–494. IEEE, 2013b.

AS Nemirovskii. Interior point polynomial time methods in convex programming, 2004. *Lecture Notes*, 2004.

Mark EJ Newman. A measure of betweenness centrality based on random walks. *Social networks*, 27(1):39–54, 2005.

Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 689–696, 2011.

Nam H Nguyen, Nasser M Nasrabadi, and Trac D Tran. Robust multi-sensor classification via joint sparse representation. In *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pages 1–8. IEEE, 2011.

Frank Nielsen and Rajendra Bhatia. *Matrix information geometry*. Springer, 2013.

Frank Nielsen and Sylvain Boltz. The burbea-rao and bhattacharyya centroids. *Information Theory, IEEE Transactions on*, 57(8):5455–5466, 2011.

Frank Nielsen, Meizhu Liu, and Baba C Vemuri. Jensen divergence-based means of spd matrices. In *Matrix Information Geometry*, pages 111–122. Springer, 2013.

Kamal Nigam and Rayid Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 86–93. ACM, 2000.

Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. *Algorithmic game theory*, volume 1. Cambridge University Press Cambridge, 2007.

Caleb C Noble and Diane J Cook. Graph-based anomaly detection. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636. ACM, 2003.

Richard Nock and Frank Nielsen. Bregman divergences and surrogates for learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2048–2059, 2009.

Robert Nowak, Urbashi Mitra, and Rebecca Willett. Estimating inhomogeneous fields using wireless sensor networks. *IEEE Journal on Selected Areas in Communications*, 22 (6):999–1006, 2004.

Sebastian Nowozin, Christoph H Lampert, et al. Structured learning and prediction in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 6(3–4): 185–365, 2011.

James M Ortega and Werner C Rheinboldt. *Iterative solution of nonlinear equations in several variables*. SIAM, 2000.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

Eduardo Pavez and Antonio Ortega. Generalized laplacian precision matrix estimation for graph signal processing. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6350–6354. IEEE, 2016.

K Peason. On lines and planes of closest fit to systems of point in space. *Philosophical Magazine*, 2(11):559–572, 1901.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Kaare Brandt Petersen and Michael Syskind Pedersen. The matrix cookbook. version: November 15, 2012, 2012.

Barnabás Póczos and Jeff G Schneider. On the estimation of alpha-divergences. In *AISTATS*, pages 609–617, 2011.

Lawrence Rabiner and Biing-Hwang Juang. Fundamentals of speech recognition. 1993.

CE Rasmussen and CKI Williams. Gaussian processes for machine learning. *Adaptive computation and machine learning*, 2006.

Adwait Ratnaparkhi et al. A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing*, volume 1, pages 133–142. Philadelphia, USA, 1996.

Gunnar Rätsch, Manfred K Warmuth, and Karen A Glocer. Boosting algorithms for maximizing the soft margin. *Advances in neural information processing systems*, pages 1585–1592, 2007.

Pradeep Ravikumar, Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Model selection in gaussian graphical models: High-dimensional consistency of ℓ_1-regularized mle. *Advances in Neural Information Processing Systems*, pages 1329–1336, 2008.

Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3): 1287–1319, 2010.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 3, 2016.

Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Christian P Robert and George Casella. *Monte Carlo statistical methods*. Springer, 1999.

Jonathan Root, Jing Qian, and Venkatesh Saligrama. Learning efficient anomaly detectors from k-nn graphs. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 790–799, 2015.

Adam J Rothman, Peter J Bickel, Elizaveta Levina, Ji Zhu, et al. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. CRC Press, 2005.

Aliaksei Sandryhaila and José MF Moura. Discrete signal processing on graphs. *IEEE transactions on signal processing*, 61(7):1644–1656, 2013.

Aliaksei Sandryhaila and Jose MF Moura. Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure. *IEEE Signal Processing Magazine*, 31(5):80–90, 2014a.

Aliaksei Sandryhaila and Jose MF Moura. Discrete signal processing on graphs: Frequency analysis. *IEEE Transactions on Signal Processing*, 62(12):3042–3054, 2014b.

Raúl Santos-Rodríguez, Alicia Guerrero-Curieses, Rocío Alaiz-Rodríguez, and Jesús Cid-Sueiro. Cost-sensitive learning based on bregman divergences. *Machine Learning*, 76 (2-3):271–285, 2009.

Mahadev Satyanarayanan. Mobile computing: the next decade. *ACM SIGMOBILE Mobile Computing and Communications Review*, 15(2):2–10, 2011.

Robert E Schapire and Yoav Freund. *Boosting: Foundations and algorithms*. MIT press, 2012.

Louis L Scharf. *Statistical signal processing*, volume 98. Addison-Wesley Reading, MA, 1991.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels*. The MIT Press, 2002.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *Artificial Neural Networks, ICANN'97*, pages 583–588. Springer, 1997.

Bernhard Schölkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John C Platt. Support vector method for novelty detection. *Advances In Neural Information Processing Systems*, 12:582–588, 1999.

Clayton D Scott and Robert D Nowak. Learning minimum volume sets. *The Journal of Machine Learning Research*, 7:665–704, 2006.

Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. *Proceedings of the 15th international conference on Multimedia*, pages 357–360, 2007.

Ervin Sejdić, Igor Djurović, and Jin Jiang. Time–frequency feature representation using energy concentration: An overview of recent advances. *Digital Signal Processing*, 19 (1):153–183, 2009.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

Pannagadatta K Shivaswamy and Tony Jebara. Laplacian spectrum learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 261–276. Springer, 2010.

David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

Vikas Sindhwani and David S Rosenberg. An RKHS for multi-view learning and manifold co-regularization. In *Proceedings of the 25th international conference on Machine learning*, pages 976–983. ACM, 2008.

Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML Workshop on Learning with Multiple Views*, pages 74–79. ACM, 2005.

Satish Sinha, Partha S Routh, Phil D Anno, and John P Castagna. Spectral decomposition of seismic data with continuous-wavelet transform. *Geophysics*, 70(6):P19–P25, 2005.

John Skilling and RK Bryan. Maximum entropy image reconstruction: general algorithm. *Monthly notices of the royal astronomical society*, 211(1):111–124, 1984.

Cees GM Snoek and Marcel Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia tools and applications*, 25(1):5–35, 2005.

Cees GM Snoek, Bauke Freiburg, Johan Oomen, and Roeland Ordelman. Crowdsourcing rock n'roll multimedia retrieval. *Proceedings of the international conference on Multimedia*, pages 1535–1538, 2010.

Qing Song, Wenjie Hu, and Wenfang Xie. Robust support vector machine with bullet hole image classification. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 32(4):440–448, 2002.

Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*, 2015.

Kumar Sricharan and Alfred Hero. Efficient anomaly detection using bipartite k-NN graphs. *Advances in Neural Information Processing Systems*, pages 478–486, 2011.

Kumar Sricharan and Alfred O Hero. Ensemble weighted kernel estimators for multivariate entropy estimation. *Advances in Neural Information Processing Systems*, pages 566–574, 2012.

Kumar Sricharan, Raviv Raich, and Alfred O Hero III. Empirical estimation of entropy functionals with confidence. *arXiv preprint arXiv:1012.4188*, 2010.

Kumar Sricharan, Raviv Raich, and Alfred O Hero. Estimation of nonlinear functionals of densities with confidence. *IEEE Transactions on Information Theory*, 58(7):4135–4159, 2012.

Shiliang Sun and Guoqing Chao. Multi-view maximum entropy discrimination. *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1706–1712, 2013.

Shiliang Sun and Feng Jin. Robust co-training. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(07):1113–1126, 2011.

Ashwin Swaminathan, Yinian Mao, and Min Wu. Robust and secure image hashing. *IEEE Transactions on Information Forensics and Security*, 1(2):215–230, 2006.

Alexander Szalay and Jim Gray. 2020 computing: Science in an exponential world. *Nature*, 440(7083):413–414, 2006.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Roberto Toldo and Andrea Fusiello. Robust multiple structures estimation with j-linkage. In *European conference on computer vision*, pages 537–547. Springer, 2008.

Eran Treister and Javier S Turek. A block-coordinate descent approach for large-scale sparse inverse covariance estimation. *Advances in neural information processing systems*, pages 927–935, 2014.

Mikhail Tsitsvero, Sergio Barbarossa, and Paolo Di Lorenzo. Signals on graphs: Uncertainty principle and sampling. *IEEE Transactions on Signal Processing*, 64(18):4845–4860, 2016.

Koji Tsuda, Gunnar Rätsch, and Manfred K Warmuth. Matrix exponentiated gradient updates for on-line learning and bregman projection. *Journal of Machine Learning Research*, 6(Jun):995–1018, 2005.

David E Tyler. Robust statistics: Theory and methods. *Journal of the American Statistical Association*, 103(482):888–889, 2008.

Laurens JP van der Maaten, Eric O Postma, and H Jaap van den Herik. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(1-41): 66–71, 2009.

Vladimir Naumovich Vapnik and Vlamimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4): 395–416, 2007.

Mehmet C Vuran, Özgür B Akan, and Ian F Akyildiz. Spatio-temporal correlation: theory and applications for wireless sensor networks. *Computer Networks*, 45(3):245–259, 2004.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Chengjing Wang, Defeng Sun, and Kim-Chuan Toh. Solving log-determinant optimization problems by a newton-CG primal proximal point algorithm. *SIAM Journal on Optimization*, 20(6):2994–3013, 2010.

Lei Wang, Huading Jia, and Jie Li. Training robust support vector machine with smooth ramp loss in the primal space. *Neurocomputing*, 71(13):3020–3025, 2008.

Xiaohan Wang, Pengfei Liu, and Yuantao Gu. Local-set-based graph signal reconstruction. *IEEE Transactions on Signal Processing*, 63(9):2432–2444, 2015.

Larry Wasserman. *All of Nonparametric Statistics*. Springer, 2010.

Kilian Q Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. *Advances in neural information processing systems*, 18:1473, 2006.

Paul Whitla. Crowdsourcing and its application in marketing activities. *Contemporary Management Research*, 5(1), 2009.

Ami Wiesel and Alfred O Hero. Distributed covariance estimation in gaussian graphical models. *IEEE Transactions on Signal Processing*, 60(1):211–220, 2012.

Ami Wiesel, Yonina C Eldar, and Alfred O Hero III. Covariance estimation in decomposable gaussian graphical models. *IEEE Transactions on Signal Processing*, 58(3): 1482–1492, 2010.

Yichao Wu and Yufeng Liu. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479), 2007.

Tianpei Xie, Nasser M Nasrabadi, and Alfred O Hero. Learning to classify with possible sensor failures. *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2395–2399, 2014.

Tianpei Xie, Nasser M Nasrabadi, and Alfred O Hero. Semi-supervised multi-sensor classification via consensus-based multi-view maximum entropy discrimination. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1936–1940. IEEE, 2015.

Tianpei Xie, Nasser M Nasrabadi, and Alfred O Hero. Learning to classify with possible sensor failures. *IEEE Transactions on Signal Processing*, 65(4):836–849, 2017.

Eric P Xing, Michael I Jordan, Stuart Russell, and Andrew Y Ng. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, pages 505–512, 2002.

Ning Xiong and Per Svensson. Multi-sensor management for information fusion: issues and approaches. *Information fusion*, 3(2):163–186, 2002.

Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

Kevin S Xu and Alfred O Hero. Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing*, 8(4):552–562, 2014.

Linli Xu, Koby Crammer, and Dale Schuurmans. Robust support vector machine training via convex outlier ablation. In *Proceedings of the 21st national conference on Artificial intelligence*, volume 1, pages 536–542. AAAI Press, 2006.

Pan Xu, Jian Ma, and Quanquan Gu. Speeding up latent variable gaussian graphical model estimation via nonconvex optimizations. *arXiv preprint arXiv:1702.08651*, 2017.

Ronald R Yager. On the dempster-shafer framework and new combination rules. *Information sciences*, 41(2):93–137, 1987.

Min Yang, Linli Xu, Martha White, Dale Schuurmans, and Yao-liang Yu. Relaxed clipping: A global training method for robust regression and classification. *Advances in neural information processing systems*, pages 2532–2540, 2010.

Yee Hwa Yang and Terry Speed. Design issues for cdna microarray experiments. *Nature Reviews Genetics*, 3(8):579–588, 2002.

Soo Sung Yoon, Hosik Sohn, Yong Ju Jung, and Yong Man Ro. Inter-view consistent hole filling in view extrapolation for multi-view image generation. *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 2883–2887, 2014.

Shipeng Yu, Balaji Krishnapuram, Harald Steck, RB Rao, and Rómer Rosales. Bayesian co-training. *Advances in Neural Information Processing Systems*, pages 1665–1672, 2007.

Shipeng Yu, Balaji Krishnapuram, Rómer Rosales, and R Bharat Rao. Bayesian co-training. *Journal of Machine Learning Research*, 12(Sep):2649–2680, 2011.

Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(Aug):2261–2286, 2010.

Ming Yuan. Discussion: Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1968–1972, 2012.

Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, pages 19–35, 2007.

Alan L Yuille and Anand Rangarajan. The concave-convex procedure. *Neural computation*, 15(4):915–936, 2003.

Alan L Yuille, Anand Rangarajan, and AL Yuille. The concave-convex procedure (cccp). *Advances in neural information processing systems*, 2:1033–1040, 2002.

Zhao Zhang, Mingbo Zhao, and Tommy WS Chow. Graph based constrained semi-supervised learning framework via label propagation over adaptive neighborhood. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2362–2376, 2015.

Manqi Zhao and Venkatesh Saligrama. Anomaly detection with score functions based on nearest neighbor graphs. *Advances in Neural Information Processing Systems*, pages 2250–2258, 2009.

Zhi-Hua Zhou and Ming Li. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 24(3):415–439, 2010.

Jun Zhu, Ning Chen, and Eric P Xing. Infinite svm: a dirichlet process mixture of large-margin kernel machines. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 617–624, 2011.

Jun Zhu, Ning Chen, and Eric P Xing. Bayesian inference with posterior regularization and applications to infinite latent SVMs. *Journal of Machine Learning Research*, 15: 1799–1847, 2014.

Xiaofan Zhu and Michael Rabbat. Approximating signals supported on graphs. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3921–3924. Citeseer, 2012.