On Learning from Collective Data

Liang Xiong

April, 2012

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Jeff Schneider, Chair
Aarti Singh
Eric Xing
Arthur Gretton, University College London

For the degree of Doctor of Philosophy
Thesis research proposal

Copyright © 2012 Liang Xiong
Keywords: Collective data; Grouped data; Low-Rank decomposition; Robust methods; Anomaly/Novelty detection; Collective/Group anomaly; Hierarchical probabilistic models; Topic models; Divergence estimation; Distribution classification; Astronomical data analysis.
Abstract

In many machine learning problems and application domains, the data are naturally organized by groups. For example, an video sequence is a group of images, an image is a group of patches, a document is a group of paragraphs/words, and a community is a group of people. We call them the collective data.

In this thesis, we study how and what we can learn from collective data. Usually, machine learning focuses on individual objects, each of which is given a feature vector and then studied as a point in some metric space. When approaching problems with collective data, researchers often try to reduce the groups into single objects to which traditional methods can be applied. We, on the other hand, will try to develop machine learning methods that respect the collective nature of data.

Several different approaches are taken to address this learning problem. When a group consists of discrete data points, it can naturally be reduced and characterized by its sufficient statistics – the histogram. For this case we develop efficient learning methods based on matrix factorization.

For groups that contain multi-dimensional vectors, such reduction is no longer available. Therefore new methods are needed to learn from those collections directly. To achieve this goal, we develop both generative methods based on topic modeling and discriminative methods using new kernels between groups. With these tools, we can accomplish various tasks such as classification, regression, clustering, anomaly detection, and dimensionality reduction on collective data.

In addition to designing new learning methods, we will use them to help the scientific discovery process. In our collaboration with astronomers and physicists, we see that the new techniques can indeed help scientists make the best of data.
# Contents

1 Introduction .................................................. 1

2 Background and Literature Review ................................. 4
   2.1 Reduced Representations and Matrix Factorization ............... 4
   2.2 Generative Models ............................................ 6
   2.3 Discriminative Methods ...................................... 8
   2.4 Related Fields ................................................ 8
   2.5 Applications on Scientific Data ............................... 9
      2.5.1 Astronomical Surveys .................................. 9
      2.5.2 Particle/Fluid Simulation ............................... 10

3 Learning with Matrix Factorization ............................... 11
   3.1 Direct Robust Matrix Decomposition ........................... 11
      3.1.1 Method Description .................................... 12
      3.1.2 Discussion .............................................. 13
      3.1.3 Experiments ............................................. 14
      3.1.4 Summary ................................................. 17
   3.2 Proposed: Automatic Novelty Discovery for Astronomy ........ 17

4 Learning with Generative Models .................................. 18
   4.1 Genre Models for Collective Data .............................. 18
      4.1.1 Model Specification ..................................... 19
      4.1.2 Applications to Group Anomaly Detection ................ 22
      4.1.3 Experiments ............................................. 22
      4.1.4 Summary ................................................. 25
   4.2 Proposed: More Powerful Topic Models ......................... 25

5 Learning with Discriminative Methods ............................. 26
   5.1 Kernel Machines for Collections .............................. 26
      5.1.1 Problem Definition ...................................... 26
      5.1.2 Nonparametric Kernel Estimation ........................ 27
      5.1.3 Experiments ............................................. 28
      5.1.4 Summary ................................................. 30
Chapter 1

Introduction

The current machine learning paradigm mostly focuses on individual objects. For example in the tasks of classification, regression, and clustering, an object of interest is often described by a “feature vector”, and abstracted as a point in some metric space. Then, the goal is to estimate functions that map these points to target variables such as the class labels or cluster memberships, with the goal of achieving both empirical accuracy on the given data and generalization power over unseen data. This “one point per object” abstraction has led to very concise representations, elegant mathematical theories, and very successful algorithms.

Nevertheless, we also realize that many of the interesting data we see in the real world objects can and should be treated as a collection of constituent objects. For instance, in the field of language modeling and text processing, an article can be considered as a group of paragraph or sections, and further a paragraph is a group of words. In computer vision and image processing, a prevailing assumption is that a visual scene consists of a group of local image patches. In recommendation systems, a user is mainly described by the group of products he/she bought, and in social network studies a community is just a group of users. We call these kinds of data that are organized by groups as the collective data. In the following, we shall use “point” to refer to the basic lower-level entities, and “groups” to refer to the aggregations of points.

The task of learning from collective data arise in many application domains, yet our research is largely motivated by demands from the scientific community. Due to the advancement of sensory systems and the ever increasing computation power, now the scientists are facing data at unprecedented scales. For example in astronomy, modern telescope pipelines like the Sloan Digital Sky Survey[1](SDSS) can produce observations for a vast amount of celestial objects. In physics, large-scale simulation systems such as the JHU Turbulence Database Cluster[2](JTDC) were implemented to study the dynamics of fluid and particles. In these problems, we have huge amount of collective data that are impossible to be examined by experts. Therefore, computational assistance is needed.

In this research, we try to answer the question: how and what can we learn from collective data? We emphasize that it is important to look beyond the point-level behaviors of data when designing

learning algorithms. Consider doing novelty detection on an article that is represented as a bag (group) of words (points). While paragraphs talking about “machine learning” or “gummy bears” will not surprise anyone on their own, an article containing both of them might be interesting. The most tempting approach, and indeed the most used one, is to reduce the groups to single objects and abstracted them as points, so that the traditional point-wise learning techniques can be applied. Yet in many situations this reduction is essentially a feature engineering process that can be domain specific and difficult. Therefore, we aim at general methods that inherently respect the collective nature of data.

We investigate several approaches to learn from different types of collective data. The first approach deals with groups consisting of discrete points that are modeled by categorical random variables. This kind of data are abundant in language modeling and text processing. A popular way to handle them is to assume that the points are exchangeable, and thus we can summarize them by histograms which are the sufficient statistics. This provides a natural reduction of the groups into compact vector representations. We learn from these data under the matrix factorization framework, where the matrices are constructed by stacking the vector representations of the groups. We developed a robust framework that enables reliable factorization/decomposition analysis in the presence of outliers, and use it for anomaly detection purposes.

In more general, and indeed more common, settings, the groups contain points that are described by real-valued multi-dimensional vectors, i.e. each group is a point cloud in some high-dimensional space. In this case, we no longer have a natural way to reduce the groups to some concise representations, and this brings more challenges into our learning tasks. We study both generative and discriminative ways of learning from these groups. From the generative perspective, we can model the generating process of the groups and points, and then use the insights from this model to help us accomplish further learning tasks. Again assuming the exchangeability of points, we devised methods that can capture the multi-level characteristics of the groups based on topic modeling, and then use them for group anomaly detection, clustering, and classification.

We also use discriminative methods to directly learn what we want based on similarity or dissimilarity measures between groups. Assuming that the points are i.i.d., we construct new estimators of kernels between the groups based on a class of new nonparametric divergence estimation methods. These kernel estimators are provably consistent and efficient to compute. Having them, we can take advantage of the existing kernel methods like the support vector machines (SVM) to accomplish various learning tasks including classifications, regression, clustering, dimensionality reduction, and anomaly detection. In our experiments on both synthetic and real-world data sets, these new methods has achieved the state-of-the-art performances.

We believe that our current work is just a beginning and much remains to be done in the future. For example, in presence of temporal data, we can extend the above methods to capture the time-evolving behaviors of the groups to help us better understand the data and make more accurate predictions. Also, considering that real-world data sets almost always contains uncontrollable outliers, we wish to make the methods robust so that the results are more reliable. Our research can also be expanded to situations where the points are even beyond vectors. For example, in astronomy each celestial object (point) has a spectrum that can be considered the noisy observation of a smooth function.
Another promising direction is to consider structured groups where the points and dependent of other points, for example when the points in a group actually forms a sequence as in motion capture data, or when the points forms a graph. It is also important to improve the computational efficiency so that these methods can be used to analyze very large-scale data sets.

Finally, we want to make our research truly useful. We are going to integrate the developed algorithms into scientific data analysis pipelines to accelerate the discovering process.

I summarize the proposed research in Table 1.1. The rest of this proposal is organized as follows. First we introduce some background and review the existing literature In Chapter 2. In Chapter 3 we learn from discrete collective data based on reduced representations using matrix factorization. Chapter 4 and 5 describe the generative and discriminative methods of learning from continuous multi-dimensional data. Finally in Chapter 6 we summarize the current status and plan for the future research.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Methodology</th>
<th>Research Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced groups</td>
<td>Matrix Factorization (Chapter 3)</td>
<td>Efficient robust factorization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Online and pass-efficient algorithms for massive data sets</td>
</tr>
<tr>
<td>Groups of vectors</td>
<td>Generative modeling (Chapter 4)</td>
<td>Genre Models</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Models for flexibility and robustness</td>
</tr>
<tr>
<td></td>
<td>Discriminative learning (Chapter 5)</td>
<td>Nonparametric distribution kernels and support distribution machines</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kernels for structured groups</td>
</tr>
</tbody>
</table>

Table 1.1: Map of the proposed research. Italic font indicates example future work.

A group can naturally be reduced to e.g. a histogram when it contains discrete points. However generally this kind of reduction is not readily available.
Chapter 2

Background and Literature Review

In this chapter, we introduce some background and review the literature in related fields. Section 2.1 describes the learning of collective data that can naturally be reduced to vectorial descriptions. For real-valued vectorial collective data that cannot be reduced, we introduce its generative learning in Section 2.2 and its discriminative learning in Section 2.3. Then we discuss related research from the methodology perspective in Section 2.4. Finally we describe the motivating applications of this research in astronomy and physics in Section 2.5.

First we list some common symbols in Table 2.1. Their meanings will be clear as we introduce more background. For sub-matrices and sub-vectors, we use the Matlab notation, e.g. $X_{1:100,:}$ denotes the first 100 rows of the matrix $X$.

The purpose of our proposed research is to help explore the scientific data sets, therefore we usually motivate the research from the perspective of anomaly/novelty detection. Yet most of the modeling approach can be generalized to address other machine learning problems. To restrict our focus we only consider cases where the groups are pre-defined aggregations of points, i.e. we are not addressing the problem of clustering the points.

2.1 Reduced Representations and Matrix Factorization

**Reduction of collective data** Many collective data sets contain discrete/categorical points. For example, in text processing we have words that take values from a dictionary. If we assume that these discrete points are infinitely exchangeable, i.e. the order of the points does not affect the nature of the groups, then we can succinctly represent the group with its sufficient statistics: the histogram. Let a group of points be $G = \{x_1, \ldots, x_N\}$ with $x_n \in \{1, \ldots, D\}$. Then $G$ can be represented by a histogram $g = [\sum_{n=1}^{N} I(x_n = 1), \ldots, \sum_{n=1}^{N} I(x_n = D)] \in \mathbb{R}^d$. This approach reduces groups to vectors for which we have mature analysis tools and learning techniques. Note that no information is lost during this reduction process under the exchangeability assumption.

In some problems, even when the nature of the points are not discrete, researchers would still discretize them using techniques like vector quantization [25]. A well-know example is the bag-of-words representation used in image processing and computer vision (e.g. [6] [19] [64]). Inevitably, the information carried by the original data will be compromised during this reduction. However
Table 2.1: Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition and Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>The number of groups.</td>
</tr>
<tr>
<td>$N_m$</td>
<td>The number of points in group $m$.</td>
</tr>
<tr>
<td>$D$</td>
<td>For discrete data, $D$ is the number of possible categorical values. For continuous data, $D$ is the dimensionality of the vector representation.</td>
</tr>
<tr>
<td>$x_{m,n}$</td>
<td>The $n$th point in group $m$. For discrete data, $x_{m,n}$ is a categorical variable taking values from ${1, 2, \ldots, D}$. For continuous data, $x_{m,n} \in \mathbb{R}^D$.</td>
</tr>
<tr>
<td>$G_m$</td>
<td>Group $m$. $G_m = {x_{m,1}, \ldots, x_{m,N_m}}$ contains the set of points in group $m$.</td>
</tr>
<tr>
<td>$g_m$</td>
<td>The vector representation of group $m$. For discrete data, $g_m \in \mathbb{R}^D$ is usually the histogram of points in $G_m$.</td>
</tr>
<tr>
<td>$S^K$</td>
<td>$K$-dimensional probability simplex.</td>
</tr>
<tr>
<td>$\mathcal{N}(\mu, \Sigma)$</td>
<td>Gaussian distribution with mean $\mu \in \mathbb{R}^D$ and covariance $\Sigma \in \mathbb{R}^{D \times D}$.</td>
</tr>
<tr>
<td>$\mathcal{U}(a, b)$</td>
<td>Uniform distribution over the interval $[a, b]$.</td>
</tr>
<tr>
<td>$\mathcal{M}(\theta)$</td>
<td>$K$-dimensional multinomial distribution with parameter $\theta \in S^K$.</td>
</tr>
<tr>
<td>$\mathcal{Dir}(\alpha)$</td>
<td>$K$-dimensional Dirichlet distribution with parameter $\alpha \in \mathbb{R}_+^K$.</td>
</tr>
<tr>
<td>$I(c)$</td>
<td>The indicator function. $I(\text{true}) = 1$ and $I(\text{false}) = 0$.</td>
</tr>
<tr>
<td>$I$</td>
<td>The identity matrix</td>
</tr>
</tbody>
</table>

the resulting vectorial representation is compact and familiar, and researchers can still develop good learning methods based on this reduced representation. In fact, how to effectively reduce the groups makes its own field (e.g. using sparse coding [88]), but it is out of our focus.

Matrix factorization In this research, the learning of the vectorial representation of groups is done under the matrix factorization framework for its simplicity and wide applicability. Matrices are very useful in representing data. Vectors can be stacked to form matrices such as the design matrices in regression and classification, and the document-word matrix in language modeling and text processing. Matrices are also used to describe networks and graphs, as well as preference data in collaborative filtering [42, 66].

We denote a $M \times D$ data matrix as $X \in \mathbb{R}^{M \times D}$. One of the most common analysis for $X$ is factorization/decomposition, such as the principal component analysis (PCA). We assume that $X$ has a low rank $K$ and decomposes as

$$X \approx UV^T, \quad U \in \mathbb{R}^{M \times K}, V \in \mathbb{R}^{D \times K}.$$  \hspace{1cm} (2.1)

For design matrices, PCA reveals the intrinsic linear structure of data. For text data, latent semantic indexing (LSI) [34] and non-negative matrix factorization (NMF) [16] are often applied. The low-rank assumption is also useful in matrix completion [8, 52] and collaborative filtering [66, 68].
More generally, factorization can be written as

$$\min_{L} \|X - L\|_F \quad \text{s.t. rank}(L) \leq K \tag{2.2}$$

where $\|\cdot\|_F$ is the Frobenius norm and $L$ is the low-rank approximation.

The singular value decomposition (SVD) is the most common tool for low-rank analysis. SVD decomposes $X$ as $X = U D_l(s) V^T$, where $l = \min(m, n)$, $s = [s_1, \ldots, s_l]$ contains the singular values in descending order, $D_l(s)$ constructs an $l \times l$ diagonal matrix with $s$, $U$ and $V$ contain the left and right singular vectors. The significance of SVD is that the rank-$K$ truncated SVD $\hat{L}_K = L_{:,1:K} D_l(s_{1:K}) U_{:,1:K}^T$ gives a globally optimal solution to problem (2.2).

Norms Many norms are used in different factorization methods. Firstly the nuclear norm of $X$ is the sum of its singular values $\|X\|_* = \sum_{i=1}^l s_i$, which serves as a convex relaxation of the rank. In (2.2), we used the Frobenius norm a.k.a. the $L_2$-norm, $\|X\|_F = \sqrt{\sum_{i,j} X_{i,j}^2}$. The $L_1$-norm $\|X\|_1 = \sum_{i,j} |X_{i,j}|$ [5, 7] and $L_0$-norm $\|X\|_0 = \sum_{i,j} I(X_{i,j} \neq 0)$, which rigorously is not a norm, is often used for the purposes of robustness and sparsity. We also have structured norms [51, 87]. The $L_{2,1}$-norm $\|X\|_{2,1} = \sum_{m=1}^M \|X_{m,:}\|_2$ is the total length of the matrix’ row vectors, and $L_{2,0}$-norm $\|X\|_{2,0} = \sum_{m=1}^M I(\|X_{m,:}\|_2 > 0)$ is the number of non-zero rows.

Anomalies and robustness Real-world problems almost always involve anomalies or outliers that do not conform to our assumptions. They can severely degrade the models’ quality, or lead to novel discoveries. Thus we want robust methods to find outliers and reduce their impact. The definition of outlier varies depending on specific problems, but in general outliers lie in the low-density regions of data distributions. [11] surveyed outlier detection problems. In matrix factorization, we consider subspace outliers and assume that normal data reside in a low-dimensional linear subspace (the row/column space of the low-rank matrix). For instance in signal processing, a normal signal can be reconstructed by a few bases. If a signal cannot be well reconstructed, it is an outlier.

Factorization methods are often not robust due to the $L_2$-norm used to measure errors [43, 85]. Many robust estimators has been proposed (e.g. [31, 39, 43, 44, 48]). A common approach is to replace the $L_2$-norm with robust norms are insensitive to outliers. For example $L_1$ norm is widely used for robustness [5, 7]. Other measures like the Huber loss [36] and the Geman-McClure function have also been employed [44, 56]. Another strategy is to exclude the outliers: we first guess which data are outliers, and then reduce their influences [43, 85].

2.2 Generative Models

More commonly, points are multi-dimensional vectors. In this case, it is not easy to summarize a group or get sufficient statistics. A straightforward approach is to discretize the points and reduce it to the previous case. However, this discretization step always loses information and might be difficult in high-dimensions. Therefore, we want to attack this problem directly. Our basic assumption is that the points in group $G_m$ are infinitely exchangeable samples from an underlying probability distribution $f_m$. To learn from the groups we should first learn the $f_m$’s.
Group characterization and anomaly detection  To motivate the characterization of groups we consider the problem of finding group anomalies. We consider two types of group anomalies. A point-based group anomaly is a group of individually anomalous points. A distribution-based anomaly is a group where the points are relatively normal, but as a whole they are unusual.

Most existing work on group anomaly detection focuses on point-based anomalies. They first identify anomalous points and then find their aggregations using scan or segmentation methods [13, 14, 32]. Clearly this paradigm will not work for distribution-based anomalies. One solution is to design problem-specific features for groups [10, 41]. However, it relies on feature engineering that is domain specific and can be difficult.

Topic models  We can learn the generating process of the groups using probabilistic models. For this purpose, particular useful are the topic models, among which the probabilistic latent semantic analysis (PLSA) [35] and latent Dirichlet allocation (LDA) [4] are the most well-known. Topic models are originally proposed for text modeling, where we have words as points and document as groups. They are hierarchical mixture models built upon the exchangeability of points i.e. the order of words does not matter. Essentially in LDA, a group $G_m$ is modeled by a mixture density $f_m = \sum_{k=1}^{K} \theta_{mk} \beta_k$. We call the mixture components $\beta_k$'s as the topics and the mixing weights $\theta_m \in S^{K-1}$ as the topic distributions. LDA forces all groups to share the same topics $\{\beta_k\}_{k=1}^{K}$ to share information and enhance statistical power. Topic models are often described by generative schemes. For example a LDA model with topics $\{\beta_k\}_{k=1}^{K}$ and prior topic distribution $Dir(\alpha)$ can be described by Algorithm [1] and the resulting likelihood is $P(G_m, \theta_m, z_m | \alpha, \beta) = Dir(\theta_m | \alpha) \prod_n M(z_{m,n} | \theta_m) \beta_{z_{m,n}}(x_{m,n})$.

**Algorithm 1** The generative process of LDA.

For group $m = 1$ to $M$:

1. Choose the topic distribution $\theta_m \in S^{K-1} \sim Dir(\alpha)$.
2. For points $n = 1$ to $N_m$:
   (a) Choose a (latent) topic $z_{m,n} \in \{1, \ldots, K\} \sim M(\theta_m)$.
   (b) Generate a point $x_{m,n} \sim \beta_{z_{m,n}}$.

Since LDA, various improvements have been proposed. Many of them enhanced the flexibility of generating mechanism of topic distributions [19, 40] or capture correlation between topics [3, 50]. On the other hand, [18] allow the topics to vary for different groups in order to account for the burstiness of words. These ideas are helpful ingredients in creating a model that can thoroughly capture how groups are generated.

Topic models for continuous data  Although topic models are proposed for discrete data like text, it is straightforward to use them for continuous multi-dimensional points by replacing the topics $\{\beta_k\}$ in Algorithm [1] with multivariate distributions such as Gaussians. The idea stays the same: we approximate the underlying distribution $f_m$ with a mixture model $\sum_k \theta_{mk} \beta_k$, and try to figure out how they are it is generated.
2.3 Discriminative Methods

Discriminative methods can also be used, in which we circumvent the generating process of data and aim directly at what we want to learn, e.g., the class label of a group. Here we focus on discriminative methods that are based on similarity measures, such as the SVM.

Set similarities Many methods have been proposed to measure the similarity between sets of vectors. [83] compared several traditional sets distances such as the Hausdorff distance for this purpose. [28, 29] proposed the pyramid matching kernels between vector sets based on hierarchical approximate matching of points. [82] measures group similarities based on the angles between the subspaces spanned by the points from different groups. [72] proposed algebraic kernels between matrices that represents sets of vectors.

Distribution similarities Alternatively, we can measure the similarity between the underlying distribution \(f_m's\). In statistics, this question is answered by two-sample tests such as the Student t-test, the Kolmogorov-Smirnov test, and the permutation test. However, these methods either rely on parametric assumptions, only use limited statistics, or have difficulties in high-dimensions.

Another way is to estimate densities \(f_m\) first and then measure similarities. [67] compute divergences by discretizing the continuous densities. [37] defines Fisher kernels between parametric densities. [55] fits Gaussian mixture models (GMM) to compute the Kullback-Leiber (KL) divergences. [38] fits exponential family densities, and then compute product kernels between these densities. [15] defined a kernel on the level-sets of fitted densities. The problem with these methods is that density estimation is itself notoriously difficult and parametric methods often introduce uncontrollable bias. [61, 62] proposed nonparametric methods to estimate divergences between distributions based on samples while avoiding density estimation.

Kernel methods [73, 74, 75] tries to embed probability distributions into reproducing kernel Hilbert spaces (RKHS). In these methods, a density \(f\) is mapped to a mean function \(\mu_f\) in a RKHS \(\mathcal{H}_k\) induced by a kernel \(k(\cdot, \cdot)\) as \(\mu_f(\cdot) = \mathbb{E}_{x \sim f}(k(x, \cdot))\). Then the density discrepancy is defined as the distance between the mean functions, and the inner-product of densities is just the inner-product of the mean functions, which is equivalent to the average kernel values between each pair of inter-group point pairs.

2.4 Related Fields

Many other research also closely related to learning from collective data. Statistical relational learning (SRL) [26] enhances point-centered machine learning by consider a group of point and their relationships altogether. For example, SRL studies the collective classification problem, where the goal is to simultaneously classify several objects based on their attributes and relations. This problem is also studied under the name of structural prediction typically using Markov networks [76] or large-margin approaches [27]. Even though collective behaviors are considered, SRL still assigns labels to points, whereas our research shall only focus on the groups.
Multiple instance learning (MIL) \cite{90} also tries to classify groups/bags of points. In MIL, a group is positive if at least one of its point is positive; otherwise it is negative. Consequently, in MIL the nature of a group is determined by a few of its points. By comparison, we consider that it is the holistic behavior of all the points that characterizes a group. Yet sometimes the methods for learning from collective data indeed overlap with methods for MIL \cite{2,30}.

Another related field is on graph kernels. Graph kernels studies the similarities between graphs, which are defined by a set of nodes and edges. Graphs can also be considered as collective data. But unlike before, graph data are structured and the elements in a graph can no longer be considered i.i.d. or exchangeable. Graph kernels often count the intersection of sub-structures between graphs. For example, the random walk kernel \cite{21,79} measures the path similarity of random walks in different graphs. \cite{1} designed kernels between groups based on graph kernels.

### 2.5 Applications on Scientific Data

The proposed research is intended to help us learn from scientific data sets. Here we describe two ongoing projects that involves astronomical surveys and particle/fluid simulation systems. In these projects, we first want to conduct explorative study, to obtain the overall data profile and pick out potentially interesting things using novelty detection.

#### 2.5.1 Astronomical Surveys

Astronomical surveys provide a holistic view of the universe by imaging a large portion of the sky at a relatively low resolution. The Sloan Digital Sky Survey\cite{1}(SDSS) project has imaged more than 35% of the sky and gives millions of observations for stars, galaxies, and quasars. Astronomers are also planning for even more powerful survey telescopes such as the Large Synoptic Survey Telescope\cite{2}(LSST) that can scan the sky faster and deeper. The massive amount of data produced by these surveys calls for the assistance of computational methods.

We focus on the spectroscopic observations in the SDSS data set. SDSS provides for each object a 3700-dimensional spectrum as shown in Figure 2.1b, and optionally its spatial location for galaxies and quasars. We directly take these spectra as the feature vectors for these objects.

Two novelty detection tasks are currently considered on this data set. The first one is to find individually anomalous objects such as planetary nebulae. This task is traditional in that we want to find unusual vector points, yet it still poses unique challenges. For example, the spectra have high-dimensionality which makes it difficult to use density-based methods, and they usually contain emission lines (see the spikes in Figure 2.1b) that could easily distort models. We focus on subspace outliers assuming that normal spectra can be reconstructed by a few bases, and develop robust methods to address the emission lines.

The second task is to detect special clusters of objects. Based on the 3-D spatial locations, we can find nearby objects and put them into clusters/groups. These clusters could shed light on the development of the universe \cite{80}, and it will be valuable to find interesting ones for the

\begin{itemize}
  \item \url{http://www.sdss.org}
  \item \url{http://www.lsst.org}
\end{itemize}
astronomers. In this case, each cluster contains a set of spectrum vectors, and we shall address this problem using group anomaly detection methods.

### 2.5.2 Particle/Fluid Simulation

In physics, researchers often simulate particle or fluid systems at a very large scale. For example, the [JHU Turbulence Database Cluster](http://turbulence.pha.jhu.edu) (JTDC) provides open access to $1024^4$ space-time points in a fluid simulation. At each point, the 3-D velocity as well as other information including pressure and temperature are recorded. See Figure 2.2 for examples. Our task is again to detect interesting phenomena.

In these systems, a single particle is seldom interesting, but a group of particles can form interesting phenomena like the vortices as in Figure 2.2c. This can again be framed as a group anomaly detection problem. We treat points in a local region as a group, and aims at finding interesting collective motion patterns characterized by the distribution of locations and velocities.
Chapter 3

Learning with Matrix Factorization

In this chapter we describe methods for learning from vectorial group representations such as histograms using matrix factorization. As described in Chapter 2, groups with discrete points can be naturally and safely reduced to histograms, and we can further stack these histogram into matrices. The purpose of matrix factorization is to learn the low-dimensional linear structure of these vectors, and its result can facilitate further analysis. The results obtained here can be apply to general matrix factorization problems.

The proposed method will be used to detect novelties in astronomical data sets. At the end of this chapter we describe our plan to build a automatic system for real-time novelty discovery from astronomical survey data.

3.1 Direct Robust Matrix Decomposition

Matrix factorization has been extensively studied. Here we consider the robustness problem motivated by our novel spectra detection problem. We assume that the data such as the been organized into a data matrix $X$, and it is approximately low-rank. Further, we assume a small portion of this matrix has been corrupted by some arbitrary outliers e.g. the large emission line spikes in Figure 2.16. The goal of the proposed algorithm is to get a reliable estimation of the true low-rank structure of this matrix and to identify the outliers. To achieve this, our basic idea is to exclude the outliers from the model estimation.

We formulate this problem as a constrained optimization problem. This formulation aims at minimizing the $L_2$ error of the low-rank approximation subject to that a small number of outliers can be ignored. No further assumptions were made. This formulation reflects our direct understanding of outliers and robust estimation. Thus we call it direct robust matrix factorization (DRMF). DRMF is based on efficient routines like SVD and thresholding, therefore it is simple to implement, efficient, and easy to use. DRMF is also very flexible: we can impose additional constraints on both the factorization (e.g. non-negative factors [46]) and the outliers (e.g. outlier columns instead of entries [86]) to incorporate knowledge for better performance.
3.1.1 Method Description

We define the robust low-rank approximation as the answer to the question: if you are allowed to ignore some data as outliers, what is the best low-rank approximation? A directly formulation of this question is the DRMF problem:

$$\min_{L,S} \|X - S - L\|_F \quad \text{s.t. } \text{rank}(L) \leq K, \|S\|_0 \leq e,$$

(3.1)

where $L$ is the low-rank approximation, $K$ is the maximal rank, $S$ is the outliers matrix, and $e$ is the maximal number of non-zeros in $S$ i.e. the maximal number of outliers allowed. Comparing DRMF to the regular matrix factorization (2.2), the only difference is that outliers $S$ can be excluded from the approximation, as long as the number of outliers is not too large. Note that we do not need the actual number of outliers. Instead, we only use $e$ to put an upper limit on it.

By excluding the outliers, we can ensure the reliability of the estimated low-rank structure. On the other hand, the number of outliers is constrained so that the estimation is still faithful to the data. DRMF is advantageous over existing methods in its simplicity and directness: no special robust error measurement is introduced, nor do we make assumptions about the outliers beyond necessity.

Algorithm 2 Direct Robust Matrix Factorization (DRMF)

1. **Input:** $X$ the data matrix; $K$ the maximal rank of the factorization; $e$ the maximal number of outliers; $S$ the initial outliers.

2. While not converged, iterate:

   $$L = \arg\min_L \|C - L\|_F, \quad C = X - S, \quad \text{s.t. } \text{rank}(L) \leq K$$

   (3.2)

   $$S = \arg\min_S \|E - S\|_F, \quad E = X - L, \quad \text{s.t. } \|S\|_0 \leq e$$

   (3.3)

3. **Output:** $L$ the robust low-rank approximation; $S$ the outliers.

The solution to the low-rank approximation problem (3.2) is directly given by SVD. Since only the first $K$ singular vectors are required, we can accelerate the computation using partial SVD algorithms such as PROPACK [45].

The outlier detection problem (3.3) can also be solved efficiently using the thresholding below

$$S_{i,j} = \begin{cases} E_{i,j} & |E_{i,j}| \geq |E|_0 \varepsilon \\ 0 & \text{otherwise,} \end{cases}$$

(3.4)
where $|E|_{(e)}$ is the $e$-th largest absolute value in $E$. This result is very intuitive: in each round, large errors are considered outliers and put into $S$ to be excluded from the low-rank fitting. The above results give the global optima to each step in Algorithm 2. They are guaranteed to improve the objective within the feasible region, thus the algorithm is going to converge. Using partial SVD, the complexity of each iteration is $O(MD(K + \log(e)))$, which is quite efficient.

To choose the parameters $e$, a rule of thumb is to set it to e.g. $5\%$ of the data. Note that this number doesn’t have to be accurate. Like most factorization methods, $K$ has to be selected by prior knowledge, heuristics, cross-validation, and computational resource. For example, we can observe the singular values and use $K$ to preserve a certain percentage of data variability.

### 3.1.2 Discussion

An limitation of traditional methods is that performance cannot be guaranteed in high dimensions \[17, 86\]. Recently, constraining the matrix nuclear norm \[8, 52\] instead of its rank became a popular strategy to overcome this problem \[7, 86, 91\], and has been shown to outperform traditional methods. These methods can be summarized as the nuclear norm minimization (NNM). To compare, we rewrite DRMF in one of its Lagrangian form, and show them in Table 3.1. We can immediately see the relationship: DRMF minimizes the rank, while NNM minimizes the nuclear norm; DRMF measures outliers by the $L_0$-norm, while NNM uses the $L_1$-norm. In fact, the nuclear norm and the $L_1$-norm are proposed as convex relaxations of the rank and the $L_0$-norm in the first place. In this sense, DRMF is the original problem that NNM is trying to solve.

By using the relaxations, NNM is convex and the globally optimal solutions can be found. In addition, theories have been provided for choosing $\lambda$ to guarantee correct recovery under certain conditions \[7, 86\]. Yet, it is unknown how tight these relaxations are in general. On the other hand, the DRMF problem is non-convex and the initial $S$ affects the final solution of DRMF. For moderate situations we found $S = 0$ works well. But in extreme cases where SVD is completely disrupted, this simple guess would lead DRMF to bad local minima. As a remedy, we can leverage the convexity of NNM and initialize DRMF by the NNM results to obtain results that are better than using either NNM or DRMF alone. The theoretical properties of DRMF are difficult to analyze due to the non-convex and non-continuous nature of the rank and the $L_0$-norm. Yet we show that DRMF can achieve better empirical performance than the relaxed NNM.

Recently we noticed a parallel work GoDec \[89\] that shares the same idea with DRMF. By comparison, DRMF extends to structured outliers, and addresses the non-convexity problem.

<table>
<thead>
<tr>
<th>NNM</th>
<th>DRMF</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\min_{L,S} |L|_* + \lambda |S|_1$</td>
<td>$\min_{L,S} \text{rank}(L) + \lambda |S|_0$</td>
</tr>
<tr>
<td>s.t. $|X - L - S|_F \leq \sigma$</td>
<td>s.t. $|X - L - S|_F \leq \sigma$</td>
</tr>
</tbody>
</table>

Table 3.1: Comparing the nuclear norm minimization (NNM) problem and DRMF. $L$ is low-rank; $S$ is the sparse outlier. $\sigma$ is the allowed approximation error.
Extensions to Incorporating Prior Knowledge

In many situations, additional knowledge is available for us to find outliers. For example, our collective data analysis, if one point has been corrupted, then it is likely that the entire group is unreliable. In this case, we should look for outlier rows so that evidences of outliers can aggregated. DRMF can easily be extended to handle this. We consider the outlier patterns to be groups of entries. Instead of counting the number of outlier entries, we count the number of outlier patterns using structured norms such as the $L_{2,0}$-norm. Concretely, the following DRMF-Row (DRMF-R) problem handles row outliers:

$$\min_{L,S} \|X - S - L\|_F \quad \text{s.t.} \quad \text{rank}(L) \leq K, \|S\|_{2,0} \leq e,$$

where $e$ is the maximal number of outlier rows allowed. DRMF-R can easily be solved by replacing (3.2) in Algorithm 2 with:

$$S = \arg \min_S \|X - L - S\|_F \quad \text{s.t.} \quad \|S\|_{2,0} \leq e.$$

which again can be solved using an efficient thresholding procedure. Row-wise outliers has also been considered in outlier pursuit (OP) [86]. OP extends NNM by using the $L_{2,1}$-norm to capture outlier rows. Not surprisingly, OP is the convex relaxation of the DRMF-R problem (3.5).

The low-rank component of DRMF can also be extended. For example, we can require the factors in (2.1) to be non-negativity as in non-negative matrix factorization (NMF) [16]. DRMF can also be extended to handle missing values in collaborative filtering. Fast and pass-efficient algorithms like [57] can be integrated into DRMF to do robust analysis on massive data sets.

3.1.3 Experiments

We show the empirical performance of DRMF on both simulation and real-world data sets. We compare DRMF to some state-of-the-art competitors:

- **Robust PCA (RPCA) [7]** The efficient “inexact augmented Lagrange multiplier” code from http://perception.csl.uiuc.edu/matrix-rank is used.

- **Stable principal component pursuit (SPCP) [91]** We implemented SPCA in Matlab using the proximal gradient method according to [20].

- **Outlier Pursuit (OP) [86]** Using the proximal gradient method according to [86].

In terms of Table 3.1, RPCA and SPCP are NNM problems with $\sigma = 0$ and $\sigma > 0$ respectively; OP is NNM with $\|S\|_{2,1}$ measurement for outlier and $\sigma = 0$. Truncated SVD are also used.

DRMF and DRMF-R are implemented in Matlab. Partial SVD is done using PROPACK [45]. We terminate the iteration when the change of the objective value is diminishing. DRMF, SPCP, and OP are initialized by 10 iterations of RPCA. For DRMF, we always set the maximal number of outliers to be $e = 5\%$ of the data unless indicated otherwise.

Simulation Data

We first study the performances on simulated data sets. We follow the set up in [7] to create the matrix. We generate a rank-$K$ matrix as $L = UV^T \in \mathbb{R}^{M \times M}$, where entries of $U, V$ are
i.i.d. samples from $\mathcal{N}(0, 1/K)$. To generate $S$, we select $\gamma M^2$ entries from $S$ and then draw their values from $\mathcal{U}(-\sigma_o, +\sigma_o)$, where $\sigma_o$ is the magnitude of outliers. Finally, we add i.i.d. noise to the entries to get $X = L + S + \mathcal{N}(0, \sigma_n^2)$, where $\sigma_n$ is the noise level.

We first test how well these methods can recover $L$. The quality of recovery is measured by the root mean squared error (RMS) between the estimated $\hat{L}$ and $L$ as $\sqrt{\frac{\sum_{ij} (\hat{L}_{ij} - L_{ij})^2}{m^2}}$. The simulation parameters are $K = 0.05M, \gamma = 0.05, \sigma_o = 1$. For RPCA, we use parameter $\lambda = 1/\sqrt{M}$ as in [7]. The ranks of SVD and DRMF are set to the true $K$. Square matrices of sizes $M$ between $[100, 2000]$ are used. We report the mean performances of 8 random runs.

First we test the entry outliers, noiseless case where entries in $S$ are randomly selected and $\sigma_n = 0$. Here the assumption of RPCA is satisfied. We compare SVD, RPCA, and DRMF. From Figure 3.1a, we see that RPCA and DRMF achieved perfect performances while SVD failed, showing the necessity and effectiveness of robust factorization.

Next we test the entry outliers, noisy case with $\sigma_n = 0.1$ and keeping other settings. This violates the condition of RPCA. We compare SVD, RPCA, SPCP, and DRMF. For SPCP, the parameters are set according to [91]. Figure 3.1b shows that DRMF performs the best again. RPCA performs poorly due to the noise, which inflates the estimated rank dramatically. SPCP shows better results but is a bit better than SVD and much worse than DRMF.

Further we test the row outliers, noisy case. We randomly select $0.05M$ rows in $S$ and fill in $\mathcal{U}(-1, 1)$ outliers. This violates the assumption of RPCA and SPCP, and invalidates the condition of using RPCA and OP. We compare SVD, RPCA, SPCP, OP, DRMF, and DRMF-R here. For OP, we use parameter $\lambda = 2/\sqrt{\gamma n}$ from [86] and tuning. Results are in Figure 3.1c. Facing row outliers, SVD failed on large matrices. OP shows an worse performance than DRMF and unstable. RPCA, SPCP, DRMF, and DRMF-R show stable performances and DRMF-R beats the others by a large margin. This verifies that utilizing knowledge about outlier patterns helps robust modeling.

We also examine how the magnitude of outliers affects the recovery quality. We simulate noiseless matrices with entry outliers, using $n = 400, K = 20$, and $\gamma = 0.05$. Then we change $\sigma_o$ from 1 to $10^5$, and compute the RMS between $\hat{L}$ and $L$. Results by RPCA and DRMF are shown in Figure 3.1d. We see that DRMF is not affected by the magnitude of outliers at all. On the other hand, the $L_1$-norm in RPCA is robust but still influenced linearly by outliers magnitude.
Video Background Modeling and Activity Detection

Estimating video background is important in computer vision but also difficult due to its variability and the foreground activities such as moving people. We use matrix factorization to approach this problem assuming that the background is of low-rank (i.e. it can be approximated by the combinations of a few “bases” images), and the activities are sparse outliers. Then the low-rank component can capture the background, while the foreground activities will be recognized as outliers.

Video sequences “Hall” (size $128 \times 160$, frames 2100-2400), “Lobby” (size $144 \times 176$, frames 1300-1700), “Restaurant” (size $120 \times 160$, frames 2500-3000), and “Shopping Mall” (size $128 \times 160$, frames 1500-2000) from [49] are used. We flatten and stack the frames as rows in a matrix, and then apply RPCA and DRMF to obtain the low-rank component as the background. For RPCA we use the suggested settings again, and for DRMF we select the rank $k$ by preserving $99.9\%$ variance of the background found by RPCA.

Results are shown in Figure [3.2]. Since the true background is unknown, we shall resort to visual inspection instead of quantitative evaluation. Both methods achieved impressive results. However, it is also clear that RPCA results contains more artifacts than DRMF except for the
“Lobby” sequence, which is simple enough that both methods are nearly perfect.

### 3.1.4 Summary

We developed the direct robust matrix factorization (DRMF) algorithm as a simple and effective way for robust low-rank factorizations and outlier detection on matrices. DRMF is conceptually simple (excluding the outliers), easy to implement (less than 10 lines of Matlab code), efficient (linear complexity \( w.r.t. \) number of entries), and flexible (easy to incorporate prior knowledge about the outliers and the low-rank structure). Empirically DRMF achieved superior results than the recently proposed nuclear norm minimization (NNM) family methods.

### 3.2 Proposed: Automatic Novelty Discovery for Astronomy

In the future, we will develop an automatic system for real-time novelty discovery from astronomical survey data. From the ongoing SDSS III project\(^1\) we can get daily updates of new objects observed by the telescope. The goal is to develop a system that can examine these new objects in real-time, and automatically pick out the potentially interesting ones to present them to the astronomers for further examination. Then we collect feedback from the astronomers to support further studies.

We aim at detecting *subspace outliers*. The assumption behind this choice is that normal spectra lie in a low-dimensional linear subspace. In other words, we can find a small number of bases whose linear combinations can approximate normal spectra. On the contrary, anomalous spectra contain unusual spectral patterns that cannot be reconstructed by these bases.

The idea above can be realized by the DRMF algorithm. DRMF is able to alleviate the impact of emission lines in the spectra and obtain reliable subspace models. It is also efficient enough to process the large amount of data. We can use the existing data to learn a reliable subspace that contains most of the normal spectra. Once we have this subspace model, anomalous spectra can be found outside of this subspace. Ideally, we want the subspace model to be updated in real-time after we have seen new spectra. This requires a online learning algorithm for robust matrix factorization that is in our plan.

Anomaly detection is just the first step in analyzing astronomical data. After the astronomers have examined and labeled these anomalies, we shall have the “seed” label information to support further learning tasks such as *active learning* and *classification*. To facilitate this process, We are developing a website to present the detection results and to collect feedbacks. A prototype of this website has been developed. This website enables easy communication and collaboration. We can inform the astronomers of the latest results and they can give us feedback on how we are doing. In the future, this website will also be used for active learning and other tasks.

\(^1\)http://www.sdss3.org
Chapter 4

Learning with Generative Models

From now on, we consider groups of real-valued, multi-dimensional points. For these groups, there is no easy way to reduce them into vectorial representations. In this chapter, we describe parametric generative models to directly capture the generating process of the groups. Having these models, we can do classification, clustering, anomaly detection, and so on for groups.

4.1 Genre Models for Collective Data

Again we assume that points are infinitely exchangeable, and call a group under this assumption a “bag-of-vectors” (BoV). Then we can use mixture models, more specifically the topic models, to capture the generative process of the data groups. We motivate our approaches with the purpose of group anomaly detection. A straightforward solution to the collective/group anomaly detection is that, if we have a model to generate normal data, then we can mark the groups that are unlikely under this model as anomalies.

As mentioned before two types of group anomalies are considered. A point-based group anomaly is a group of individually anomalous points. A distribution-based anomaly is a group where the points are relatively normal, but as a whole they are unusual. Although topic models has been extensively studied, existing methods are incapable of detecting group anomalies effectively and comprehensively. To detect anomalies, the model should be flexible enough to capture complex behaviors. It should be able to model complex and multi-modal distributions of topic distributions. LDA, however, only uses a single uni-modal Dirichlet distribution to generate topic distributions, thus cannot effectively define normality. LDA also uses the same topics for all groups, therefore these topics are not adapted to each group and the groups are indistinguishable at the topics level.

Based on the Latent Dirichlet Allocation (LDA) [4], we progressively propose two probabilistic hierarchical models designed specifically for the purpose of detecting group anomalies. The first model focuses on distribution-based anomalies and enriches LDA by allowing a flexible way of generating topic distributions. The second model further takes both distribution-based and point-based anomalies into account, forming a very elastic topic model and a comprehensive anomaly detector. We demonstrate the effectiveness of the these models on synthetic and on real-world turbulence data.
4.1.1 Model Specification

We extend LDA to address several of its weaknesses in modeling continuous, complex collective data. To address the problem of uni-modality of topic distributions in LDA, we introduce the concept “genres” to characterize the topic distributions so that complex normal behaviors can be recognized. A genre, intuitively, is a typical/normal topic distribution, and we allow a group to derive its topic distribution from one of the several genres. Moreover, the assumption that topics are shared globally in topic models is also relaxed to further enhance the flexibility.

Multinomial Genre Models

To start, we let the genres be the typical topic distributions themselves. In other words, we make a pool of multinomial distributions, and each group can select one of these multinomials as its topic distribution. We call this model the Multinomial Genre Model (MGM). We assume that the points are generated from Gaussian distributions (topics) and $\beta = \{\mu_k, \Sigma_k\}^K_{k=1}$ are the mean and covariance of the Gaussians. Let $\alpha_t \in S^K$ (genre $t$) denote a typical distribution of topics, and $\alpha = \{\alpha_1, \ldots, \alpha_T\}^T_{t=1}$ denote the set of $T$ genres. Let $\pi \in S^T$ denote a distribution over the genres. The generative process of MGM is described in Algorithm 3.

Algorithm 3 The generative process of MGM.

```plaintext
for groups $m = 1$ to $M$ do
  • Choose a genre $\{1, \ldots, T\} \ni y_m \sim \mathcal{M}(\pi)$. Let the topic distribution $\theta_m$ be $\alpha_{y_m} \in S^K$.
  for $n = 1$ to $N_m$ do
    • Choose a topic $z_{m,n} \in \{1, \ldots, K\}$, $z_{m,n} \sim \mathcal{M}(\theta_m)$.
    • Generate a vector $x_{m,n} \in \mathbb{R}^D$, $x_{m,n} \sim P(x_{m,n}|\beta, z_{m,n})$.
```

Under MGM, the complete and marginal likelihood of group $G_m$ are

$$P(y_m, z_m, G_m|\pi, \alpha, \beta) = \mathcal{M}(y_m|\pi) \prod_{n=1}^{N_m} \mathcal{M}(z_{m,n}|y_m, \alpha)P(x_{m,n}|z_{m,n}, \beta), \quad (4.1)$$

$$P(G_m|\pi, \alpha, \beta) = \sum_{t=1}^{T} \pi_t \prod_{n=1}^{N_m} \sum_{k=1}^{K} \alpha_{tk}P(x_{m,n}|z_{m,n}, \beta). \quad (4.2)$$

To learn the parameters $\{\pi, \alpha, \beta\}$, we can use the Expectation-Maximization (EM) method. Details are omitted here.

Flexible Genre Models

Although the MGM model addressed some flexibility issues of LDA, it is still inadequate for group anomaly detection. By modeling the genres by a pool of multinomial distributions, MGM does not take the uncertainty of topic distributions into account. MGM also inherits the assumption that the topics are shared globally, therefore it cannot capture the point level behaviors. Other topic models also try to enhance LDA. [40] also used a pool of multinomials like MGM, but it does not
Figure 4.1: The Flexible Genre Model (FGM).

handle continuous data and group anomaly detection problems. The Theme Model (ThM) \[19\] lets a mixture of Dirichlets generate the topic distributions and then uses the mixture memberships to cluster groups. This idea is useful for modeling group-level behaviors but still cannot capture point-level behaviors as the topics are shared globally. On the other hand, \[18\] proposed to use different topics for different groups to account for the burstiness of the words (points). These adaptive topics are useful in recognizing point-level anomalies, but cannot detect anomalous behavior at the group level. In this section we further take these ingredients and improve our model to form a comprehensive model for detecting both distribution-based and point-based anomalies.

At the group level, “genres” are still used to model the topic distributions. Instead of multinomials, we use one Dirichlet distributions for each genre to model a typical distribution of topic distributions. At the point level, each group has its own topics to accommodate the variations of its points, while global topic information is still shared via the “topic generator” components. We call this model the Flexible Genre Model (FGM). Given a group of points, we can examine whether or not it conforms to the normal behavior defined by the learned genres and topics; A point-based anomaly contains points from unusual topics other than \(\{\beta_k\}\), while a distribution-based anomaly contains normal points but has an unusual topic distribution \(\theta_m\).

Algorithm 4 The generative process of FGM.

```latex
\begin{algorithm}
\begin{algorithmic}
\For {groups \(m = 1\) to \(M\)}
\State Choose a genre \(\{1, \ldots, T\} \ni y_m \sim \mathcal{M}(\pi)\).
\State Choose a topic distribution according to the genre \(y_m\): \(\mathcal{S}^K \ni \theta_m \sim \text{Dir}(\alpha_{y_m})\).
\State Choose \(K\) topics \(\{\beta_{m,k} \sim P(\beta_{m,k}|\eta_k)\}_{k=1}^K\).
\For {points \(n = 1\) to \(N_m\)}
\State Choose a topic \(\{1, \ldots, K\} \ni z_{m,n} \sim \mathcal{M}(\theta_m)\).
\State Generate a vector \(x_{m,n} \sim P(x_{m,n}|\beta_{m,z_{m,n}})\).
\EndFor
\EndFor
\end{algorithmic}
\end{algorithm}
```

The generative process of FGM is presented in Algorithm 4. A graphical representation of FGM is given in Figure 4.1. We let \(\mathcal{M}(\pi)\) be distribution of genres. Each genre is a Dirichlet distribution for generating the topic distributions \(\theta_m\), and \(\alpha = \{\alpha_t\}_{t=1}^T\) is the set of genre parameters. Each group has \(K\) topics \(\beta_m = \{\beta_{m,k}\}_{k=1}^K\). The “topic generators”, \(\eta = \{\eta_k\}, \{P(\cdot|\eta_k)\}_{k=1,\ldots,K}\), are the global distributions for generating the corresponding topics. Having the topic distribution \(\theta_m\) and the topics \(\{\beta_{m,k}\}\), points are generated as in LDA.

By comparing FGM to LDA: (i) in FGM, each group has a latent genre \(y_m\), which determines how its topic distribution should look like \((\text{Dir}(\alpha_{y_m}))\), and (ii) each group has its own topics
\{\beta_{m,k}\}_{k=1}^K$, but they are still tied through the global generator $P(\cdot|\eta)$. Thus, the topics can adapted to local group data, but information is still shared globally. Moreover, the topic generators $P(\cdot|\eta)$ determine how the topics $\{\beta_{m,k}\}$ should look like. If a group uses unusual topics to generate its points, it can be identified.

For computational convenience, the topic generators are Gaussian-Inverse-Wishart (GIW) distributions parameterized by $\eta_k = \{\mu_{0k}, \kappa_{0k}, \Psi_{0k}, \nu_{0k}\}$ \textsuperscript{23}, which are conjugate to the Gaussian topics. Let $\Theta = \{\pi, \alpha, \eta\}$ denote the model parameters. The complete likelihood of data and latent variables in group $G_m$ under FGM is:

$$
P(G_m, y_m, \theta_m, z_m, \beta_m|\Theta) = \mathcal{M}(y_m|\pi) \text{Dir}(\theta_m|\alpha_{y_m}) \prod_k \text{GIW}(\beta_{m,k}|\eta_k) \prod_n \mathcal{M}(z_{mn}|\theta_m) \mathcal{N}(x_{mn}|\beta_{m,z_{mn}}).$$

By integrating out $\theta_m, \beta_m$ and summing out $y_m, z$, we get the marginal likelihood of $G_m$:

$$
P(G_m|\Theta) = \sum_t \sum_k \int \text{Dir}(\theta_m|\alpha_t) \prod_k \text{GIW}(\beta_{m,k}|\eta_k) \prod_n \sum_k \theta_{mk} \mathcal{N}(x_{mn}|\beta_{m,k}) d\beta_m d\theta_m.$$  \textbf{(4.4)}

FGM can be used to accomplish various learning tasks. For example, anomalies can be detected by examine the likelihood of topics and topic distributions. We can also train different models for different classes of groups and then do classification in a Naive Bayes way.

\textbf{Inference and Learning}

The parameters of FGM can be learned via MLE. The inferred latent states can be used for detecting anomalies and exploring the data. Nonetheless, the inference and learning under FGM is intractable, so we train FGM using an approximate method described below.

Approximate inference in FGM can be done using Gibbs sampling \textsuperscript{24}. In Gibbs sampling, we update one variable at a time by drawing samples from its distribution conditioned on all the other variables. Thanks to the use of conjugate distributions, Gibbs sampling in FGM is straightforward. The sampling distributions of the latent variables in group $m$ are given below. We use $P(\cdot|\sim)$ to denote the distribution of one variable conditioned on all the others.

$$
P(z_{mn} = k|\sim) \propto P(x_{mn}|z_{mn} = k, \beta_m) P(z_{mn} = k|\theta_m) = \theta_{m,k} \mathcal{N}(x_{mn}|\beta_{m,k}),$$

$$
P(y_m = t|\sim) \propto P(\theta_m|\alpha_t) P(y_m = t|\pi) = \pi_t \text{Dir}(\theta_m|\alpha_t),$$

$$
P(\theta_m|\sim) \propto P(z_m|\theta_m) P(\theta_m|\alpha, y_m) = \mathcal{M}(z_m|\theta_m) \text{Dir}(\theta_m|\alpha_{y_m}) = \text{Dir}(\alpha_{y_m} + h_m),$$

$$
P(\beta_m|\sim) \propto P(x_m^{(k)}|\beta_{m,k}) P(\beta_{m,k}|\eta_k) = \mathcal{N}(x_m^{(k)}|\beta_{m,k}) \text{GIW}(\beta_{m,k}|\eta_k) = \text{GIW}(\beta_{m,k}|\eta_k^t),$$

where $h_m \in \mathbb{R}^K$ denotes the histogram of the values in $z_m$, $x_m^{(k)}$ are points in group $G_m$ from topic $k$. $\eta_k^t$ is the parameter of the posterior GIW distribution given $x_m^{(k)}$; its form can be found in standard statistics textbooks \textit{e.g.} \textsuperscript{23}.

The parameters of FGM characterize the data’s normal behaviors. Each of genres $\alpha = \{\alpha_t\}_{t=1}^T$ captures one typical distribution of topic distributions. The topic generators $\eta = \{\eta_k\}_{k=1}^K$ determine
how the normal topics \( \{ \beta_{m,k} \} \) should look like. To learn them, we can use single-sample Monte Carlo EM [9], which updates the parameters to their MLE based on the Gibbs samples. \( \pi \) can be estimated from the histogram of \( y \)'s. \( \alpha_t \) is learned by the MLE of a Dirichlet distribution given the multinomials \( \{ \theta_m | y_m = t, m = 1, \ldots, M \} \), which can be solved using gradient-based methods [54]. The \( k \)th topic-generator’s parameter \( \eta_k = \{ \mu_{0k}, \kappa_{0k}, \Psi_{0k}, \nu_{0k} \} \) is the MLE of a GIW distribution given \( \{ \beta_{m,k} = (\mu_{m,k}, \Sigma_{m,k}) \}_{m=1,\ldots,M} \) (the \( k \)th topics of all groups), and we have derived an efficient solution for this problem. 

To select appropriate values for the parameters \( T \) and \( K \) (the number of genres and topics), we can apply the Bayesian information criterion (BIC) [71], or use the values that maximize the likelihood of a held-out validation set. In addition to Gibbs sampling, we also developed variational methods to learn FGM. Details are quite lengthy and omitted here.

### 4.1.2 Applications to Group Anomaly Detection

FGM can easily be used for anomaly detection. We first infer a group’s latent states including the topics \( \beta \) and the topic distribution \( \theta \), and then examine if they are compatible with the model.

Point-based anomalies can be found by examining the topics. If a group contains anomalous points, then the topics that generated these points will deviate from the topic generators \( \eta \). Let \( P(\beta_m | \Theta) = \prod_{k=1}^K GIW(\beta_{m,k} | \eta_k) \). The point-based anomaly score (PB score) is

\[
\mathbb{E}_{\beta_m} [ - \ln P(\beta_m | \Theta) ] = - \int_{\beta_m} P(\beta_m | \Theta, G_m) \ln P(\beta_m | \Theta) d\beta_m. \tag{4.9}
\]

The posterior \( P(\beta_m | \Theta, G_m) \) can again be approximated using Gibbs sampling, and the expectation can be done by Monte Carlo integration.

Distribution-based anomalies can be detected by examining the topic distributions. The genres \( \{ \alpha_t \}_{t=1,\ldots,T} \) capture the typical distribution of topic distributions. If a group’s topic distribution \( \theta_m \) is unlikely under these genres, we call it anomalous. Let \( P(\theta_m | \Theta) = \sum_{t=1}^T \pi_t Dir(\theta_m | \alpha_t) \). The distribution-based anomaly score (DB score) is

\[
\mathbb{E}_{\theta_m} [ - \ln P(\theta_m | \Theta) ] = - \int_{\theta_m} P(\theta_m | \Theta, G_m) \ln P(\theta_m | \Theta) d\theta_m. \tag{4.10}
\]

Again, this expectation can be approximated using Gibbs sampling and Monte Carlo integration. Similar scoring functions can be designed for MGM.

### 4.1.3 Experiments

#### Synthetic Data

First we compare MGM, FGM, and an adaptation of the Theme Model (ThM) [19] on synthetic data sets. We use the data likelihood of ThM as a scoring function to detect point-based anomalies.

Using the synthetic data sets described below, we demonstrate the behavior of the different models and scoring functions. We generated the data using 2-dimensional GMMs as in [84].
Each group is generated by a GMM. All GMMs share three Gaussian components with covariance $0.2 \times I_2$ and centered at $(-1.7, -1)$, $(1.7, -1)$, and $(0, 2)$, respectively. The groups’ mixing weights are randomly chosen from $w_1 = [0.33, 0.33, 0.33]$ or $w_2 = [0.84, 0.08, 0.08]$. Thus, a group is normal if its points are from the three Gaussians, and their mixing weights are close to $w_1$ or $w_2$.

To test the detectors, we injected both point-based and distribution-based anomalies. The point-based anomaly contains points from $\mathcal{N}((0, 0), I_2)$. Distribution-based anomalies were generated by GMMs with normal Gaussian components but mixing weights $[0.33, 0.64, 0.03]$ and $[0.08, 0.84, 0.08]$, which were different from $w_1$ and $w_2$. We generated $M = 50$ groups with sizes $N_m \sim \text{Poisson}(100)$. One point-based and two distribution-based anomalies were injected.

The detection results of MGM, ThM, and FGM are shown in Figure 4.2. Normal groups have black solid boxes, point-based anomalies have green dashed boxes, and distribution-based anomalies have red/magenta dashed boxes. Points are colored by the anomaly scores of the groups (darker color means more anomalous). An ideal detector would make dashed boxes’ points dark and solid boxes’ points light gray.

![Detection results on synthetic data.](image)

Figure 4.2: Detection results on synthetic data.

We see that all models can find the distribution-based anomalies since they can learn the normal topic distributions. However, MGM and ThM miss the point-based anomaly because the anomalous points are in the middle of the topics, thus the inferred topic distribution is around $[0.33, 0.33, 0.33]$, which is exactly $w_1$. This example shows one problem of scoring based on topic distributions only. On the contrary, using the sum of PB and DB scores, FGM found all of the anomalies thanks to its characterization of groups at both the point-level and the group-level. We also show the result of scoring the groups by the data likelihood under ThM. Only point anomalies are found. This is because the data likelihood under ThM is dominated by the anomalousness of points, thus a few eccentric points will overshadow group-level behaviors.

Figure 4.3 shows the density estimated by MGM, ThM, and FGM, respectively, for the point-based anomalous group. We can see that FGM gives a better estimation due to its adaptive topics, while MGM and ThM are limited to use their global topics.

**Turbulence Data**

We present an explorative study of detecting group anomalies on turbulence data from the JHU Turbulence Database Cluster (JTDC) [59]. JTDC simulates fluid motion through time on a 3-dimensional grid, and here we perform our experiment on a continuous $128^3$ sub-grid. At each
time step and each vertex, JTDC records the velocity of the fluid. We consider the vertices in a local cubic region as a group, and the goal is to find groups of vertices whose velocity distributions (i.e. moving patterns) are unusual. The following steps were used to extract the groups: (1) We chose the $\{(8i, 8j, 8k)\}_{i,j,k}$ grid points as centers of our groups. Around these centers, the points in $7^3$ sized cubes formed our groups. (2) The feature of a point in the cube was its velocity relative to the velocity at its cube’s center point. After these pre-processing steps, we had $M = 4096$ groups, each of which had 342 3-dimensional feature vectors.

We applied MGM, ThM, and FGM to find anomalies in this group data. $T = 4$ genres and $K = 6$ topics were used for all methods. We do not have a ground-truth for anomalies in this data set. However, we can compute the “vorticity score” [53] for each vertex that indicates the tendency of the fluid to “swirl”. Vortices and especially their interactions are uncommon and of great interest in fluid dynamics. This vorticity can be considered as a hand crafted anomaly score based on expert knowledge. We do not want an anomaly detector to match this score perfectly because there are other “non-vortex” anomalous events it should find as well. However, we do think higher correlation with this score indicates better anomaly detection performance.

Figure 4.4 visualizes the DB scores of FGM and the vorticity. We can see that the two pictures are highly correlated, implying that FGM was able to find interesting turbulence activities based on velocity only and without using the definition of vorticity or any other expert knowledge. Correlation values between vorticity and the MGM, ThM, and FGM scores from 20 random runs are displayed in Fig. 4.4c, showing that FGM is better at finding regions with high vorticity.

(a) FGM-DB Score  (b) Vorticity  

Figure 4.4: (a) & (b) FGM-DB anomaly score and vorticity visualized on one slice of the cube. (c) Correlations between the anomaly scores with the vorticity.
4.1.4 Summary

We presented an parametric, generative approach to model collective data, and use it for the group anomaly detection problem. Compared to traditional topic models, the proposed Multinomial Genre Model (MGM) Flexible Genre Model (FGM) are able to capture complex group behaviors at multiple levels. Empirical results show that FGM achieves good performances.

4.2 Proposed: More Powerful Topic Models

Topic modeling is a very active field recently. Thanks to the extreme flexibility of probabilistic models, there are many opportunities to develop further enhancements that can model collective data better.

More flexibility For example, we can derive even more expressive models by taking the correlation between topics into account like in the correlated topic models (CTM) [3]. Since the generative mechanism in CTM still very limited (e.g. uni-modal), we will use more flexible priors to generate topic distributions. One possible way of doing this is to use Gaussian distributions as genres, and then let them generate topic distributions through the multi-logistic transformation as in CTM. This genre model will be able to capture the multi-modality of topic distributions and also the topic correlation.

Robust genre models Robustness is another desired property for generative models. In anomaly detection, the data we used to train the normal model are very like to have been contaminated by anomalies already. Hence, we hope that the estimated model would not be affected. To achieve this goal, we can use robust distributions such as the t-distribution as genres in genre models.

Functional observations The modeling capability of topic models can be extended beyond vectorial points. In astronomy, the spectra are actually noisy observations of some potentially smooth functions. Due to the limitation of the sensors and the fact that objects have different red-shift, the observations will contain much noise and many missing parts. We are investigating the possibility to combine topic models with Gaussian process [65] to learn better from these functional data.

Relational learning on groups We also plan to study relational learning methods at the group level. Like in statistical relational learning, we can learn from relations among the groups. For example, relations can be derived from the temporal evolution of groups. We can also form relations based on geographic locations, where it is reasonable to believe that groups from nearby locations were correlated.
Chapter 5

Learning with Discriminative Methods

Unlike generative models, discriminative methods tries to learn target concepts directly regardless of the generating mechanism of data. In this chapter we describe discriminative ways of learning from collective data based on similarity or dissimilarity measures between groups. The advantage of this approach is its great flexibility, since we can take advantage of the existing tools that rely on similarities to accomplish a vast variety of tasks on groups.

5.1 Kernel Machines for Collections

In this section, we describe how to generalize kernel machines from finite-dimensional vector spaces to the domain of finite sample sets. Our assumptions is that group $G_m$ is characterized by its underlying distribution $f_m$, from which its points are i.i.d. samples.

We develop methods for the classification of distributions. In the classification problem our goal is to find a map from the space of distributions to the space of class, while in the anomaly detection problem we want to find distributions that are unlike others. Note that only finite i.i.d. samples are observed from these distributions. For this purpose we extend the support vector machines (SVM) to the space of distributions. In our framework, some of the distributions in the training data will play the role of support vectors.

5.1.1 Problem Definition

We denote underlying distribution of group $G_m$ as $f_m$, and the space of sample sets as $\mathcal{G}$. In the supervised learning problem we have $\{(G_n, Y_n)\}_{n=1}^N$ (input, output) pairs. The output domain is discrete, i.e. $Y_n \in \mathcal{Y} = \{y_1, \ldots, y_K\}$. We are looking for a function $h : \mathcal{G} \rightarrow \mathcal{Y}$, such that for a new input and output pair $(G, Y) \in \mathcal{G} \times \mathcal{Y}$ when the classification is perfect we have that $h(G) = Y$.

We will use ideas from SVM to perform this classification. For simplicity, we only discuss binary classification problems, which can be easily extended to multiclass situations.

Let $\mathcal{P}$ denote the set of density functions, $\mathcal{H}$ be a Hilbert-space with inner product $\langle \cdot, \cdot \rangle_\mathcal{H}$, and $\phi : \mathcal{P} \rightarrow \mathcal{H}$ denote an map from the densities to the feature space $\mathcal{H}$. The dual form of the “soft
margin SVM” for classification is [69]:

$$\hat{\alpha} = \arg \max_{\alpha \in \mathbb{R}^M} \sum_{m=1}^{M} \alpha_m - \frac{1}{2} \sum_{i,j}^{M} \alpha_i \alpha_j y_i y_j K_{ij},$$

subject to $$\sum_{m} \alpha_m y_m = 0, 0 \leq \alpha_m \leq C$$, where $$C > 0$$ is a parameter, $$y_m \in \{-1, 1\}$$ are the class labels, and $$K \in \mathbb{R}^{M \times M}$$ is the Gram/Kernel matrix: $$K_{ij} = \langle \phi(f_i), \phi(f_j) \rangle_H = k(f_i, f_j)$$. Now, the predicted class label of a test density $$f$$ is simply $$h(f) = \text{sign}(\sum_{i=m}^{M} \alpha_m y_m K(f_i, f) + b)$$, where the bias term $$b$$ can be obtained by averaging $$b = y_j - \sum_i y_i \alpha_i G_{ij}$$ over all points with $$\alpha_j > 0$$.

There are many tools available to solve the quadratic programming task in (5.1). All that remains is to compute $$\{k(f_i, f_j)\}_{i,j}$$ and $$\{K(f_i, f_j)\}_{i,j}$$ based on the $$i.i.d.$$ samples. We propose estimators for these kernels in Section 5.1.2.

The SVM formulation can also be used to find anomalous distributions using e.g. one-class SVM [70] in unsupervised settings. Like before, the only thing we have to do is to find a good kernel matrix for the groups.

### 5.1.2 Nonparametric Kernel Estimation

To use the methods above, we must estimate $$k(p, q)$$, the kernel value between distributions $$p$$ and $$q$$ using two finite $$i.i.d.$$ sample sets from them. Many kernels, i.e. positive semi-definite (PSD) functionals of $$p$$ and $$q$$ can be constructed from

$$D_{\alpha, \beta}(p\|q) = \int p^\alpha(x) q^\beta(x) \, dx,$$  \hspace{1cm} (5.2)

where $$\alpha, \beta \in \mathbb{R}$$. Linear ($$K(p, q) = \int pq$$), polynomial ($$K(p, q) = (\int pq + c)^s$$), and Gaussian kernels ($$K(p, q) = \exp(-\frac{1}{2} \mu^2(p, q)/\sigma^2)$$, where $$\mu(p, q) = \int p^2 + q^2 - 2pq$$) are examples of these. For the Gaussian kernel, one can also use other “distances”, e.g. the Hellinger distance where $$\mu(p, q) = 1 - \int p^{1/2} q^{1/2}$$, the Rényi-$$\alpha$$ divergence, where $$\mu(p, q) = \frac{1}{\alpha-1} \log \int p^\alpha q^{1-\alpha}$$, or the KL-divergence, which is its $$\alpha \to 1$$ limit case.

To estimate $$D_{\alpha, \beta}(p\|q)$$ for some $$\alpha, \beta$$ values, we use the tools that have been applied for Rényi entropy [47], Shannon entropy [27], KL divergence [81], and Rényi divergence estimation [60]. We show how to estimate $$D_{\alpha, \beta}(p\|q)$$ in an efficient, nonparametric, and consistent way. Based on [61], we can estimate $$D_{\alpha, \beta}(p\|q)$$ using

$$\hat{D}_{\alpha, \beta} = \frac{1}{n} \sum_{i=1}^{n} (n - 1)^{-\alpha} m^{-\beta} \rho_k^{-d\alpha}(i) \rho_k^{-d\beta}(i) B_{k, \alpha, \beta},$$  \hspace{1cm} (5.3)

where $$B_{k, \alpha, \beta} \doteq \sum_{i=1}^{n} (n - 1)^{-\alpha} m^{-\beta} \rho_k^{-d\alpha}(i) \rho_k^{-d\beta}(i)$$.

Under certain conditions $$\hat{D}_{\alpha, \beta}$$ is a consistent estimator of $$D_{\alpha, \beta}$$, and thus by plugging these estimators into kernels we get consistent estimators for those kernels. It means that the more sample points we have the better the quality of the kernel estimation is, and eventually it is converging to the correct value.
However, consistency of our estimators does not guarantee that the estimated Gram matrix is positive semi-definite or even symmetric because of the nature of the divergences and the estimation errors. Nevertheless, we can project the estimated Gram matrix to the cone of PSD matrices so that it will be a valid input to SVM. This can be done by truncating the negative eigenvalues of the Gram matrix. Details about this can be found in [33].

As we reviewed in Section 2.3 most divergence estimators rely on density estimation which is notoriously difficult. By contrast, our method avoids this intermediate and nuisance step and is completely nonparametric and provides provably consistent kernel estimations.

5.1.3 Experiments

Scene classification using “bag-of-patches” representations is a well-studied problem and many methods have been proposed (e.g. [6,19]). Here we test the performance of our non-parametric kernels in this task.

The “bag-of-vectors” (BoV) representation of images is based on the dense SIFT descriptors. We put a regular 2D grid with step size 10 on each image, and compute 128-dimensional SIFT descriptors on the grid nodes. In an attempt for scale invariance, we compute three SIFT descriptors with sizes of \{6, 9, 12\} pixels at each point. Following [6], we can also include color information in the SIFT features. Finally, we use PCA to reduce the feature vectors’ dimensionality. Our implementation uses the PHOW function of the VLF EAT package [78] for feature extraction.

To convert the BoV data to the “bag-of-words” (BOW) representation, these SIFT vectors are quantized by K-means into visual words, for which the vocabulary size (number of K-means clusters) is 1000 for color images and 500 for gray-scale images. For PLSA the number of topics is 25, as suggested by [6]. This discretization step inevitably loses information, and we will show that our non-parametric kernels outperform BOW in most cases.

The compared algorithms are listed as below:

**Nonparametric divergence kernels** These kernels are based on the proposed nonparametric Rényi-\(\alpha\) divergence estimators (NPR-\(\alpha\)) and Hellinger distance estimators (NPH). We use the \(k = 5\)th nearest neighbors in these estimators. For NPR, we test the performance with \(\alpha \in \{0.5, 0.8, 0.9, 0.99\}\). Note that when \(\alpha = 0.99\) the Rényi-divergence approximates the KL divergence, and when \(\alpha = 0.5\) it is twice the Bhattacharyya distance.

**Parametric kernels** These kernels are based on Gaussian or GMM assumptions. We first fit a Gaussian/GMM density to each group, and then compute the KL-divergence (G-KL, GMM-KL) [55] and product probability kernels (G-PPK, GMM-PPK) [38] with \(\alpha = 0.5\) between the groups (therefore they are actually the Bhattacharyya Coefficients between Gaussians). Tuning the number of GMM components for each group is not feasible, so we always use 3 components. Note that GMM-KL has no analytic form, so we use a Monte Carlo method with 500 samples.

**BOW kernels** We can convert continuous points to discrete ones by quantizing the points, then we can compute the histogram of words for each group. By doing this we get the “bag-of-words” representation. The chi-square distance between these BOW histograms is used to construct the Gaussian kernel. The histograms can be further processed by PLSA [35] and then used to build Gaussian kernels based on Euclidean distance.
Pyramid matching kernel We also use the vocabulary-guided pyramid matching kernel \textbf{PMK} [29] to construct the kernel matrices of the groups. This version is used because it performs better for high-dimensional data. The authors’ implementation \textit{libpmk} and the suggested values of parameters are used.

We use \textit{LibSVM}’s [12] multi-class SVM for classification. The parameter \(C\) (penalty to points within the margin) is chosen from \(\{2^{-9}, 2^{-6}, \cdots, 2^{20}\}\) by cross-validation. For \textit{PPK} and \textit{PMK}, we use their kernel values directly. For the other kernels based on divergences, we use the Gaussian kernels \(\exp\left(-\frac{1}{2} \mu^2 / \sigma^2\right)\), where \(\mu\) is the divergence/distance, estimated with Equation (5.5). The kernel width \(\sigma\) is also chosen by cross-validation from \(\sigma_0 \times \{2^{-4}, 2^{-2}, \cdots, 2^{10}\}\), where \(\sigma_0\) is the mean of the divergences between each pair of groups.

The \textit{OT} data set from [58] is used. This dataset contains 8 outdoor scene categories: \textit{coast, mountain, forest, open country, street, inside city, tall buildings, and highways}. There are 2688 images in total, each about 256 \times 256 pixels. Sample images from the \textit{OT} data set are shown in Figure 5.1. The goal here is to classify test images into one of the 8 categories.

![Figure 5.1: Images from the 8 OT scene categories: coast, forest, highway, inside city, mountain, open country, street, tall building.](image)

First, we use the gray-scale images. The features are reduced to 19 dimensions using PCA preserving 70\% of variance, so that a typical image is represented by 1542 18-dimensional points. The accuracies of 16 random 2-fold cross-validations are shown in Figure 5.2a. Our Rényi-divergence kernels achieve the best overall accuracy, reaching 88.8\% when \(\alpha = 0.8\). It outperformed the \textit{BOW} kernel, which is also very accurate (88.6\%); a paired t-test shows a \(p\)-value of 0.06.

![Figure 5.2: Accuracies on the OT data set.](image)

We can use color information to enhance the features. It is also possible to incorporate spatial information. Originally, an image is characterized by the distribution of its local features. By

\[ \text{http://people.csail.mit.edu/jjl/libpmk} \]
concatenating the $x$ and $y$ coordinates of patches with the local feature vectors, we can cope with the joint distribution of local appearances and their locations. We can also include larger regions into the feature set, so that co-occurrences of local objects can be captured at larger scales.

We combine the above approaches to increase the classification accuracy on the OT data set. We extract color SIFT features with the bin sizes of $\{6, 12, 18, 24, 30\}$ so that a typical image contains 1815 points. By including color, the SIFT features are now 384-dimensional, which are reduced by PCA to 53 dimensions, preserving 70% of the variance. Then the $y$ coordinates of patches are appended to these feature vectors. Finally, each dimension of the feature vectors is normalized to have zero mean and unit variance.

The accuracies of 16 random runs are shown in Figure 5.2b. Here results of 10-fold cross-validations are used so that we can directly compared to other published results. We can see adding the extra information greatly increased classification accuracies. NPR-0.99 achieved the best average accuracy of 92.2%, which is much better than BOW’s 90.1%. Notably, this 92.2% accuracy surpasses the best previous result of which we are aware, 91.57% [63]. For comparison, in 2-fold cross-validations the average accuracies of NPR-0.99 and BOW are 90.6% and 88.5% respectively. GMM-PPK is not shown because it is too low.

5.1.4 Summary

We proposed support distribution machines to learning from distributions based on $i.i.d.$ sample sets and provided non-parametric methods for estimating the necessary kernel matrices. We showed that our methods achieved the state of the art performances in image classification tasks.

5.2 Proposed: Learning from Structured Collections

So far we have been focusing on groups with exchangeable or $i.i.d.$ points. It will open a whole new world if we drop this assumption and consider structure groups in which there are dependencies between the points.

5.2.1 Time Series Data Analysis

We can approach time series analysis using the methods for collective data. Essentially, a length $l$ $D$-dimensional temporal sequence can be considered as $l$ non-$i.i.d.$ samples from a $D + 1$-dimensional joint distribution of the features and the time. The challenge, however, will be how to make use the temporal nature of the data to make extrapolations and predict into the future.

5.2.2 Learning from Graphs

As we mentioned before, graphs can be considered as a special form of collective data where points are connected by edges. Traditional graph kernels focus on the topological structure of the graphs [22], yet ignore the nodes’ attributes. It would be interesting to study how these two streams of research could converge.
Chapter 6

Summary and Research Timeline

6.1 Summary

We propose the research of machine learning for collective data i.e. data organized by groups. Unlike traditional learning methods that focus on points with vectorial descriptions, we focus on groups of points and study their collective behaviors. The goal is to develop novel learning algorithms that respect and exploit the collective nature of data to achieve superior empirical performances. Several promising results have been obtained in this direction, and we plan to do more interesting and useful work in the future.

The resulting methods is widely applicable in many real-world problems. Yet our primary intention is to use them to help the scientific discovery process. We are planning to develop automatic discovery and learning systems for data sets from astronomy and physics based on the proposed research.

6.2 Research Timeline

The future research in this thesis is planned in Table 6.1:

| Spring 2012 | • Enhance the generative models for more flexibility and robustness.  
|            | • Enhance the efficiency of the matrix factorization algorithm and implement the automatic discovery system for astronomy. |
| Fall 2012  | • Extend the generative models for functional data.  
|            | • Study how to do relational learning on groups.  
|            | • Explore the possibility of learning from structured groups. |
| Spring 2013| • Continue to study group-level relational learning and structured groups.  
|            | • In the last three months wrap up the work and complete the thesis. |

Table 6.1: Planned research time line.
Bibliography

[1] Francis R. Bach. Graph kernels between point clouds. In ICML, 2008. 2.4
[7] E. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? preprint, 2009. 2.1 2.1 3.1.2 3.1.2 3.1.3 3.1.3
[13] Kaustav Das, Jeff Schneider, and Daniel Neill. Anomaly pattern detection in categorical datasets. In Knowledge Discovery and Data Mining (KDD), 2008. 2.2


[34] Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, 1999. 2.1


[41] Eamonn Keogh, Jessica Lin, and Ada Fu. Hot sax: Efficiently finding the most unusual time series subsequence. In *IEEE International Conference on Data Mining*, 2005. 2.2


[50] Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic
correlations. In *ICML*, 2006. 2.2


[56] Minh Hoai Nguyen and Fernando De la Torre. Robust kernel principal components analysis. In *NIPS*, 2009. 2.1


[61] B. Póczos, L. Xiong, and J. Schneider. Nonparametric divergence estimation with applications to machine learning on distributions. In *27th Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain*, 2011. 2.3 5.1.2


[67] Yasushi Sakurai and Rosalynn Chong. Efficient distribution mining and classification. In


[74] Le Song, Jonathan Huang, and Alex Smola Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In ICML, 2009.


[84] Liang Xiong, Barnabás Póczos, and Jeff Schneider. Hierarchical probabilistic models for
group anomaly detection. In *International conference on Artificial Intelligence and Statistics (AISTATS)*, 2011. 4.1.3

[85] Huan Xu, Constantine Caramanis, and Shie Mannor. Principal component analysis with contaminated data: The high dimensional case. In *Annual Conference on Learning Theory (CoLT)*, 2010. 2.1


37