Hierarchical Probabilistic Models for Group Anomaly Detection

Liang Xiong
Machine Learning Department,
Carnegie Mellon University

Barnabás Póczos
Robotics Institute,
Carnegie Mellon University

Jeff Schneider
Robotics Institute,
Carnegie Mellon University

Andrew Connolly
Department of Astronomy,
University of Washington

Jake VanderPlas
Department of Astronomy,
University of Washington

Abstract

Statistical anomaly detection typically focuses on finding individual point anomalies. Often the most interesting or unusual things in a data set are not odd individual points, but rather larger scale phenomena that only become apparent when groups of points are considered. In this paper, we propose generative models for detecting such group anomalies. We evaluate our methods on synthetic data as well as astronomical data from the Sloan Digital Sky Survey. The empirical results show that the proposed models are effective in detecting group anomalies.

1 Introduction

Given a data set, anomaly/novelty detection aims at discovering events that 'surprise' us, since they may have scientific and practical value. We consider the unsupervised detection problem, in which we do not know beforehand which data is normal and which is not. These problems are very common when we have unexplored large-scale data sets, which are more and more frequent thanks to the ever-increasing computing power and ubiquitous data sources.

Most anomaly detection research focuses on finding unusual data points. Nonetheless, in many applications we are more interested in finding *group anomalies*. One type of group anomalies is just a group of individually anomalous points. A more interesting, and

often more difficult case is where the individual data points are normal, but their distribution as a group is unusual. *The contribution* of this paper is to propose methods for detecting both kinds of group anomalies.

Our motivating application is anomaly detection for astronomical data. Contemporary telescopes, such as the *Sloan Digital Sky Survey* (SDSS)[1], produce a vast amount of data. SDSS uses a dedicated telescope to scan the sky and gather astrometric, photometric, and spectroscopic data for celestial objects. The task of finding interesting and scientifically valuable objects in this large pool is of great importance. Moreover, unusual clusters of objects are also valuable for scientific research, since objects in a spatial cluster play important roles in each other's evolution, and the distributions of their features gives insight into how they developed. Similar problems exist in many other domains, such as text and image processing, where aggregated behaviors are of interest.

To solve the group anomaly detection problem, we start from a standard statistical anomaly detection approach of creating a generative model for the data, and then we flag the data that are relatively unlikely to have been generated by that model. We propose two hierarchical probabilistic models for this purpose. We treat each group of instances as a 'bag-of-things', and assume that the points in each group are *exchangeable*. According to the *De Finetti*'s theorem (de Finetti, 1931), the joint distribution of every infinitely exchangeable sequence of random variables can be represented with mixture models, thus we will apply a hierarchical mixture model to represent the data. Having estimated the model, we propose two different scoring functions to detect various anomalies.

The first model is a direct extension of the *Latent*

[1]http://www.sdss.org

Dirichlet Allocation (LDA) model by Blei et al. (2003). We assume that each individual data point falls into one of the several *topics*, and each group is a mixture of topics. The original LDA applies conditional multinomial distributions for generating observations. This is not suitable for us when we have real, vector-valued observations. Hence, we generalize LDA to other parametric distributions such as multivariate Gaussians, which determine the probability of our observations given the corresponding topics. In the astronomical example, each topic can be interpreted as a certain type of galaxy, and each group consists of several types of galaxies. We expect our method to identify groups that contain anomalous points, *and* those whose members are normal, but the topic distribution is unusual.

A drawback of the model above is that it uses a Dirichlet distribution to generate topics distributions. This Dirichlet is uni-modal peaking at a single topic distribution[2], and thus unable to generate multiple normal topic distributions. In other words, there is essentially only one normal topic distribution for the whole data set. This is often too restrictive for real data sets. To address this problem, we propose a second model in which the topic distributions come from a pool of multinomial distributions. This allows multiple types of normal groups that have different topic distributions. Efficient learning algorithms are derived for both models based on variational EM techniques. We demonstrate the performance of the proposed methods on synthetic data sets, and show they are able to identify anomalies that cannot be found by other generative model based detectors. Empirical results are also shown for the SDSS astronomical data.

The paper is structured as follows. In Section 2 we summarize some related work. We formally define the problem set-up in Section 3. The proposed models and how we can learn them are described in Section 4. Experimental results both on simulated problems and on real astronomical data are shown in Section 5. We finish with a short discussion and conclusions (Section 6).

2 Related Work

Typically, the notion of 'anomaly' depends heavily on the specific problem, and various algorithms have been developed for their own purposes. Quite often they are based only on the simple idea that a data point is anomalous if it falls in a low density region of the feature space. For example, Zhao (2009) uses the distances to nearest neighbors as an anomaly score. Breunig et al. (2000) consider the case of non-uniform density of the normal data, and propose a local outlier

---

[2] For Dirichlet parameters greater than 1. In other cases restrictions also exist. See Section 5 for examples.

factor for detecting anomalous instances. We can also explicitly estimate the underlying density function and use statistical tests to find anomalies. To see a more comprehensive summary, readers can refer to the recent survey by Chandola et al. (2009).

Detecting group anomalies is not a new problem, but only a few results have been published on it. One idea is to represent each group as a point, and then apply point anomaly detectors for these groups. To do this, we need to define a set of features for the groups (Chan and Mahoney, 2005; Keogh et al., 2005). A problem with this approach is that it relies heavily on feature engineering, which can be domain specific and difficult. We believe that directly modeling the generative process of the data is more natural, and can help us explore the data sets.

Another approach is to first identify the individual anomaly points, and then try to find aggregations of these points. Scan and segmentation methods are often used for this purpose. On image data, Hazel (2000) applied a point anomaly detector to find anomalous pixels, and then segment the image to find the anomalous group of pixels. Das et al. (2008) first detects interesting points, and then find subsets of the data with a high ratio of anomalous points. Das et al. (2009) proposed a scan statistic-based method to find anomalous subsets of points. In these approaches the anomalousness of a group is determined by the anomalousness of its member points, therefore they cannot find anomalous groups that are unusual only at the group level.

3 Formal Problem Definition

In this section we define formally our problem. For simplicity we will explain the set-up by borrowing terms from astronomy, but our solution to this problem can be used anywhere where the observations can be naturally clustered into groups.

Assume that we have M groups denoted by $\mathbf{G}_1, \ldots, \mathbf{G}_M$. Each group $\mathbf{G}_m$ consists of N_m objects, denoted by $X_{m,n} \in \mathbb{R}^f$, $n = 1, \ldots, N_m$. These are our observations, e.g. $X_{m,n}$ is the $f = 1,000$ dimensional spectrum of the nth galaxy in the mth galaxy group, where these galaxy groups were created based on the spatial positions of the galaxies. Assume further that these $X_{m,n}$ feature vectors are generated by a mixture of K Gaussian distributions, that is, each object (galaxy) $X_{m,n}$ belongs to one of these K types, and if we know its type $Z_{m,n} \in \{1, \ldots, K\}$, then $X_{m,n} \sim \mathcal{N}(\beta^\mu_{Z_{m,n}}, \beta^\Sigma_{Z_{m,n}})$. $\beta = \{\beta^\mu_k, \beta^\Sigma_k\}^K_{k=1}$ is a dictionary of the possible mean values and covariance matrices for the above mentioned Gaussian mixture, where $\beta^\mu_k \in \mathbb{R}^f$, and $\beta^\Sigma_k \in \mathbb{R}^{f \times f}$ is a positive semi definite matrix. For example, when $K = 3$, then we might

think of these objects as 'red', 'blue', and 'emissive' galaxies, and each group $\mathbf{G}_m$ is a set of N_m objects, each object can be one of the K different types. Introduce the $\mathbb{S}^K = \{s \in \mathbb{R}^K | s_k \geq 0, \sum_{k=1}^{K} s_k = 1\}$ notation for the K-dimensional probability simplex, and let $\chi_t \in \mathbb{S}^K$ for all $t = 1, \ldots, T$, and $\chi = \{\chi_1, \ldots, \chi_T\}$ denote the set of T possible non-anomalous distributions (proportions) of the K different objects (red, blue, and emissive galaxies) in the M groups.

Now we can ask the question whether in group $\mathbf{G}_m$ the distribution of these red, blue, and emissive galaxies looks normal, that is, they look similar to a distribution in $\chi = \{\chi_1, \ldots, \chi_T\}$, or we have found a group, where this distribution seems far from the distributions that we can see in the other groups.

In the following sections we will propose two generative probabilistic models that can help us to answer this question and detect anomalous groups.

4 The Hierarchical Models

In this section we introduce our generative models that describe the normal, that is the non-anomalous data, and then we show how we can detect anomalous groups using these models. Our proposed models are inspired by the LDA, however, there are very significant differences that we will explain later.

4.1 The Uni-Modal Model

The LDA model is a generative probabilistic model originally proposed for modeling text corpora. First we briefly review this model, and then explain how we can extend this *discrete* model to be able to find anomalous groups in a data set given by any *real vector-valued* feature representation.

In the original LDA model the data set is a text corpus, that is a collection of M documents. Each document $\mathbf{G}_m$ is a set of N_m words, and each document is represented by a random mixture over latent topics, which is characterized by a distribution over words. Formally, let $\mathcal{D}ir(\pi)$ denote the Dirichlet distribution with parameter π, and let $\mathcal{M}(\theta)$ be the multinomial distribution with parameters $\theta \in \mathbb{S}^K$. In the LDA model given some nonnegative hyperparameters $\pi \in \mathbb{R}_+^K$, we generate first some $\theta_m \in \mathbb{S}^K$ $(m = 1, \ldots, M)$ from the $\mathcal{D}ir(\pi)$ distribution $(\theta_m \sim \mathcal{D}ir(\pi))$. Having these K dimensional θ_m vectors (topic distributions) we generate $Z_{m,n} \sim \mathcal{M}(\theta_m)$ variables $(n = 1, \ldots, N_m)$ indicating which topic is active out of K when we generate the word $X_{m,n} \sim P(\cdot|Z_{m,n}, \beta)$. Here $\beta = \{\beta_1, \ldots, \beta_K\}$ is a dictionary of K f-dimensional probability vectors $(\beta_k \in \mathbb{S}^f)$, and $P(\cdot|Z_{m,n}, \beta) = \mathcal{M}(\beta_{Z_{m,n}})$ is a multinomial distribution with parameters $\beta_{Z_{m,n}}$. While this

model has been shown to be very successful for modeling discrete data, such as text corpora, in its original form it cannot be used for modeling real, vector-valued observations. Thus we modify this model slightly. Instead of using $\mathcal{M}(\beta_{Z_{m,n}})$ for the observations, we assume $\beta_i = \{\beta_i^\mu, \beta_i^\Sigma\}$ to be a mean value $(\beta_i^\mu \in \mathbb{R}^f)$ and a covariance matrix $(\beta_i^\Sigma \in \mathbb{R}^{f \times f})$, and our observations are given by:

$$X_{m,n} \sim P(\cdot|Z_{m,n}, \beta) = \mathcal{N}(\beta_{Z_{m,n}}^\mu, \beta_{Z_{m,n}}^\Sigma).$$

We call this model Gaussian-LDA (GLDA).

With GLDA we can model real, vector-valued observations, but it has a serious problem when we want to apply it for group anomaly detection. GLDA learns that each group is a certain mixture of K Gaussian components, but it also assumes that there is only one "best" mixture (topic distribution) for all groups, because $\mathcal{D}ir(\pi)$, the distribution of topic distributions $\theta \in \mathbb{S}^K$, is uni-modal *i.e.* it peaks at a single point. While this is acceptable when used as the prior in LDA, it is too restrictive when used to model multimodal distributions of topic distributions. To address this issue we extend the GLDA model with the previously mentioned χ term, the set of the typical topic distributions (proportions of the Gaussian components).

4.2 The Multi-Modal Model

In this section we introduce the Mixture of Gaussian Mixture Model (MGMM) model that extends GLDA with a set of typical topic mixtures/distributions, and hence can resolve the previously mentioned unimodality problem. The graphical representation of this new model can be seen in Figure 1.
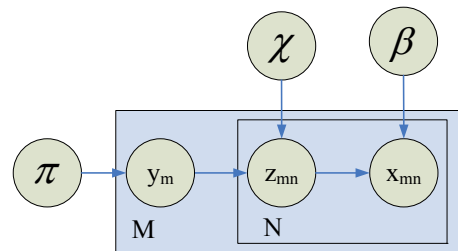

Figure 1: The MGMM Model

Let again $\chi_t \in \mathbb{S}^K$ for all $t = 1, \ldots, T$, and $\chi = \{\chi_1, \ldots, \chi_T\}$ denote the set of possible non-anomalous probability distributions of the K different topics (red, blue, and emissive galaxies) in the M groups. Let $\pi \in \mathbb{S}^T$ denote a distribution vector on the set χ, and let $\beta = \{\beta_k^\mu, \beta_k^\Sigma\}_{k=1}^K$ be a dictionary of the possible mean values and covariance matrices.

The generative process of the MGMM model is described in Algorithm 1. Note that this model is differ-

---

Algorithm 1 Generative process for MGMM

  for $m = 1$ to M **do**
- Choose a group type $\{1, \ldots, T\} \ni Y_m \sim \mathcal{M}(\pi)$
- Let the topic distribution $\theta_m \doteq \chi_{Y_m} \in \mathbb{S}^K$.
- Choose N_m, the number of points in the group $\mathbf{G}_m$. (N_m can be random, e.g. sampled from a Poisson distribution).

    for $n = 1$ to N_m **do**
- Choose a galaxy type $Z_{m,n} \in \{1, \ldots, K\}$, $Z_{m,n} \sim \mathcal{M}(\theta_m)$.
- Generate a galaxy feature $X_{m,n} \in \mathbb{R}^f$, $X_{m,n} \sim P(X_{m,n}|\beta, Z_{mn}) = \mathcal{N}(\beta^\mu_{Z_{m,n}}, \beta^\Sigma_{Z_{m,n}})$.

    end for
  end for

---

ent from the other *mixture of Gaussian mixture models* introduced by Li (2001), since we require that the points in the same group should come from a single Gaussian mixture model.

Our strategy for group anomaly detection is as follows. Using the training set $\{X_{m,n}\}$, we first learn the hyperparameters $\{\pi, \chi, \beta\}$ of the model. If a group $\mathbf{G}$ is not compatible with our model, then it will lead to a small likelihood $P(\mathbf{G}|\pi, \chi, \beta)$ compared to that of the other groups, and we can detect it as an anomalous group. Unfortunately, direct maximization of the likelihood function, as in many hierarchical models, is intractable, thus we resort to variational EM methods (Jordan, 1999) for inference and learning.

4.3 Inference and Learning

For the sake of brevity, introduce the shorthands $\mathbf{G}_m = \{X_{m,n}\}_{n=1}^{N_m}$, and $Z_m = \{Z_{m,n}\}_{n=1}^{N_m}$. Given the observations and latent variables, the complete likelihood of a group $\mathbf{G}_m$ is as follows.

$$P(Y_m, Z_m, \mathbf{G}_m | \pi, \chi, \beta) \qquad (1)$$
$$= P(Y_m|\pi) \prod_{n=1}^{N_m} P(Z_{m,n}|Y_m, \chi) P(X_{m,n}|Z_{m,n}, \beta)$$
$$= \mathcal{M}(Y_m|\pi) \prod_{n=1}^{N_m} \mathcal{M}(Z_{m,n}|Y_m, \chi) P(X_{m,n}|Z_{m,n}, \beta)$$
$$= \pi_{Y_m} \prod_{n=1}^{N_m} \chi_{(Y_m, Z_{m,n})} \mathcal{N}\left(X_{m,n}|\beta^\mu_{Z_{m,n}}, \beta^\Sigma_{Z_{m,n}}\right).$$

In what follows, instead of using $\mathcal{N}\left(X_{m,n}|\beta^\mu_{Z_{m,n}}, \beta^\Sigma_{Z_{m,n}}\right)$ we will use the more general $P(X_{m,n}|Z_{m,n}, \beta)$ term. The marginal likelihood of the observations $\mathbf{G}_m = \{X_{m,n}\}_{n=1}^{N_m}$ is

$$P(\mathbf{G}_m|\pi, \chi, \beta) = \sum_{t=1}^{T} \pi_t \prod_{n=1}^{N_m} \sum_{k=1}^{K} \chi_{tk} P(x_{mn}|z_{mn}, \beta).$$

To learn the hyperparameters $\{\pi, \chi, \beta\}$ using maximum likelihood estimation, we want

$$\arg\max_{\pi, \chi, \beta} \prod_{m=1}^{M} P(\mathbf{G}_m|\pi, \chi, \beta).$$

The traditional EM method is intractable here, thus we make use of the variational approach. That is, instead of maximizing the exact likelihood, we will only maximize a lower bound of it.

Denote the hyperparameters by $\Theta = \{\pi, \chi, \beta\}$. According to the Jensen inequality, for any $\{q_m(Y, Z)\}_{m=1}^{M}$ set of distributions we have that

$$\sum_{m=1}^{M} \log P(\mathbf{G}_m|\Theta)$$
$$\geq \sum_{m=1}^{M} \int d(Y, Z) q_m(Y, Z) \log \frac{P(Y, Z, \mathbf{G}_m|\Theta)}{q_m(Y, Z)}$$
$$= \sum_{m=1}^{M} \mathbb{E}_{q_m}[\log P(Y, Z, \mathbf{G}_m|\Theta)] - \mathbb{E}_{q_m}[\log q_m(Y, Z)],$$

[3]with equality iff $q_m(Y, Z) = P(Y, Z|\mathbf{G}_m, \Theta)$. This posterior distribution has difficult, intractable form, thus instead of the direct maximization of $\sum_{m=1}^{M} \log P(\mathbf{G}_m|\Theta)$, we will solve only the

$$\arg\max_{\Theta, \{q_m\}} \sum_{m=1}^{M} \mathbb{E}_{q_m}[\log P(Y, Z, \mathbf{G}_m|\Theta)] - \mathbb{E}_{q_m}[\log q_m] \quad (2)$$

problem, where we look for the surrogate distribution q_m in a special parametric form:

$$q(Y_m, Z_m | \gamma_m, \phi_m) = q(Y_m|\gamma_m) \prod_{n=1}^{N_m} q(Z_{m,n}|\phi_{m,n}).$$

Here $\gamma_m \in \mathbb{S}^T$ and $\phi_{m,n} \in \mathbb{S}^K$ are the variational parameters, and $q(Y_m|\gamma_m) = \mathcal{M}(\gamma_m)$, $q(Z_{m,n}|\phi_{m,n}) = \mathcal{M}(\phi_{m,n})$ are multinomial distributions. Using Eq. (1) and Eq. (2), we have that the variational learning problem we need to solve is

$$\arg\max_{\{\gamma_m\}, \{\phi_m\}, \Theta} \sum_{m=1}^{M} L_m(\gamma_m, \phi_m, \Theta),$$

where $\Theta = \{\pi, \chi, \beta\}$, and L_m has the following form:

$$L_m(\gamma_m, \phi_m; \pi, \chi, \beta) =$$
$$= \mathbb{E}_q[\log P(Y_m, Z_m, \mathbf{G}_m|\pi, \chi, \beta)] - \mathbb{E}_q[\log q(Y_m, Z_m)]$$
$$= \mathbb{E}_q[\log P(Y_m|\pi)] + \sum_{n=1}^{N_m} \mathbb{E}_q[\log P(Z_{m,n}|Y_m, \chi)]$$
$$+ \sum_{n=1}^{N_m} \mathbb{E}_q[\log P(X_{m,n}|Z_{m,n}, \beta)] - \mathbb{E}_q[\log q(Y_m|\gamma_m)]$$
$$- \sum_{n=1}^{N_m} \mathbb{E}_q[\log q(Z_{m,n}|\phi_{m,n})].$$

---

[3]$\mathbb{E}_q$ denotes the expected value w.r.t. distribution q.

We need to maximize this L_m function. Here we just show the end results, the details of the calculations can be found in the Appendix.

$$\phi^*_{m,n,k} = \frac{\exp\left(\sum_{t=1}^{T} \gamma_{m,t} \log \chi_{t,k} + \log P\left(X_{m,n}|\beta_k\right)\right)}{\sum_{j=1}^{K} \exp\left(\sum_{t=1}^{T} \gamma_{m,t} \log \chi_{t,j} + \log P\left(X_{m,n}|\beta_j\right)\right)},$$

$$\gamma^*_{m,t} = \frac{\exp\left(\log \pi_t + \sum_{n=1}^{N} \sum_{k=1}^{K} \phi_{m,n,k} \log \chi_{t,k}\right)}{\sum_{\tau=1}^{T} \exp\left(\log \pi_\tau + \sum_{n=1}^{N} \sum_{k=1}^{K} \phi_{m,n,k} \log \chi_{\tau,k}\right)},$$

$$\pi^*_t = \left(\sum_{\tau=1}^{T} \sum_{m=1}^{M} \gamma_{m,\tau}\right)^{-1} \sum_{m=1}^{M} \gamma_{m,t},$$

$$\chi^*_{t,k} = (\sum_{j=1}^{K} \sum_{m=1}^{M} \gamma_{m,t} \sum_{n=1}^{N_m} \phi_{m,n,j})^{-1} \sum_{m=1}^{M} \gamma_{m,t} \sum_{n=1}^{N_m} \phi_{m,n,k}.$$

Finally, to calculate β, we need to solve

$$\arg\max_{\beta_k} \sum_{m=1}^{M} \sum_{n=1}^{N_m} \sum_{k=1}^{K} \phi_{m,n,k} \log P\left(X_{m,n}|\beta_k\right).$$

Specially, when $P\left(X_{m,n}|\beta_k\right) = \mathcal{N}\left(X_{m,n}|\beta_k^\mu, \beta_k^\Sigma\right)$, then learning $\left(\beta_k^\mu, \beta_k^\Sigma\right)$ is the same as fitting Gaussians in a mixture of Gaussians model with $\phi_{m,n,k}$ being the mixture proportions (Mclachlan and Krishnan, 1996).

4.4 Detection Criterions

In this section we discuss how to define scoring functions that can detect group anomalies. Having learned the parameters Θ, a natural choice is to score a group by its likelihood under the model. We define the *likelihood score* of a group $\mathbf{G}$ simply as $-\ln P(\mathbf{G}|\Theta)$. This likelihood score is able to find anomalous groups that either contain anomalous points or have strange group-level behaviors *i.e.* topic distributions.

Despite its generality, the likelihood score focuses more on the effects of individual points, instead of the groups' topic distributions. For example, one single extreme outlier can inflate the anomaly score of the whole group to infinity, and hence we find that the effect of anomalous topic distributions are often overshadowed by anomalous points. Moreover, the likelihood score might misclassify some cases. For example, suppose that the model learned two topics $\{T_1, T_2\}$ that both appear with probability $1/2$. Then any group that consists of m_1 topics T_1 and m_2 topics T_2 has the same likelihood: $1/2^{(m_1+m_2)}$. However, if we observe a group that only contains topic T_1, it is clearly more anomalous than those that have both topics.

To overcome this difficulty, we propose to score only the topic distribution in each group: we first infer the posterior distributions of the topics given the data, and then compute the expected likelihood of the topic distributions. Formally, for the MGMM model the *topic score* is defined as

$$\mathbb{E}_{\mathbf{Z}_m}[-\ln P(\mathbf{Z}_m|\Theta)] = -\sum_{\mathbf{Z}_m} P(\mathbf{Z}_m|\Theta, \mathbf{G}_m) \ln P(\mathbf{Z}_m|\Theta), \tag{3}$$

where $\ln P(\mathbf{Z}_m|\Theta) = \ln \sum_t \pi_t \mathcal{M}(\mathbf{Z}_m|\chi_t)$ is a mixture of multinomials. This score finds groups whose topic variables $\mathbf{Z}_m$ are not compatible with any of the stereotypical topic distributions in χ learned by MGMM. For GLDA, we can similarly define the *topic score* as

$$\mathbb{E}_{\theta_m}[-\ln P(\theta_m|\Theta)] = -\int_{\theta_m} P(\theta_m|\Theta, \mathbf{G}_m) \ln P(\theta_m|\Theta)\, d\theta. \tag{4}$$

In practice, we use the topic score to find anomalous group-level behaviors, and the likelihood score to find aggregations of anomalous points. We can also use a weighted combination of the likelihood score and the topic score depending on the types of anomalies we are looking for.

To simplify computation, we use the variational distributions $q_m(\cdot)$ to replace the corresponding posteriors $P(\mathbf{Z}_m|\Theta, \mathbf{G}_m)$ in (3) and $P(\theta_m|\Theta, \mathbf{G}_m)$ in (4). The integrations then can be done by *Monte Carlo* method using samples drawn from the approximate posteriors.

4.5 Model Selection

One limitation of the MGMM model is that T and K need to be assigned by the user. To automatically determine their values, we can use either model scoring methods such as BIC (Schwarz, 1974), or AIC (Akaike, 1974), or we can resort to nonparametric Bayesian modeling. In this paper we investigate the first way for model selection. The definition of BIC score is given by $BIC(X, \Theta) = \ln L(X, \Theta) - \frac{1}{2}\ln(|X|)|\Theta|$, where $|\cdot|$ stands for the number of free parameters. Similarly, the AIC score is given by $AIC(X, \Theta) = \ln L(X, \Theta) - |\Theta|$. We can then use these two scoring functions to perform a two dimensional search for the best T and K values.

5 Numerical Experiments

We show some experimental results to demonstrate the effectiveness of the proposed GLDA and MGMM models. We compared them with two other *point-wise* detectors: a simple *Gaussian mixture model* (GMM) based density estimator, which scores points by their negative log-density, and the *KNN* algorithm proposed

by Zhao (2009), which scores points by their distance to their nearest neighbors. The anomaly score of a group from GMM and KNN is the mean anomaly scores of its member points. For GLDA and MGMM, we combine the likelihood score and the topic score to detect both point and group anomalies by first scaling both scores to fit the range $[0, 1]$ and then add them.

5.1 Synthetic Problems

First, we test the effectiveness of the algorithms on synthetic data sets. These experiments are designed particularly to demonstrate the differences between the models and scoring functions.

We generate the data sets according to the process described in Algorithm 1. The points are sampled from three 2-dimensional Gaussian components (*i.e.* $K = 3$), whose means are $[-1.7, -1]$, $[1.7, -1]$, $[0, 2]$, and the covariances are all $\Sigma = 0.2 \times \mathbf{I}_2$, where $\mathbf{I}_2$ denotes the identity matrix. These components are the *'topics'*, *i.e.* the types of the galaxies. Then we design two normal *group types* ($T = 2$), which are specified by two different sets of mixing weights ($\chi_1, \chi_2 \in \mathbb{S}^3$). We generated $M = 50$ groups, and $N_m \sim Poisson(100)$ points in each. The resulting points individually are all normal, *w.r.t.* other points.

To test the detection performance, we inject two types of anomalies. The first kind is a group of point anomalies, which is a group of points sampled from $\mathcal{N}([0, 0], \mathbf{I}_2)$ (the anomalous topic). We corrupted one group with this anomaly. The second kind is the group anomaly, where the points are individually normal, but together as a group look anomalous. We construct these anomalies by using points from the normal topics, but their topic distributions are different from the normal ones (χ_1, χ_2).

First, we test the performances on a data set with a uni-modal distribution of topic distributions, which has only one normal topic distribution $\chi = (0.33, 0.33, 0.33)$, *i.e.* there are about the same amount of points from each topic in a normal group. We corrupt two more groups with injected group anomalies, whose topic distributions are $(0.85, 0.08, 0.07)$ and $(0.04, 0.48, 0.48)$, respectively. Thus overall we corrupt 3 groups (one point anomaly, and two group anomalies) out of the $M = 50$ groups.

The detection results are shown in Figure 2. Each box contains a group, and we show 12 out of the 50 groups. We draw black boxes for normal groups, green boxes for groups of point anomalies, and yellow/magenta boxes for group anomalies. The points of the groups are plotted and colored according to the anomaly scores (darker color indicates higher anomaly score). The anomaly detection is successful, if the
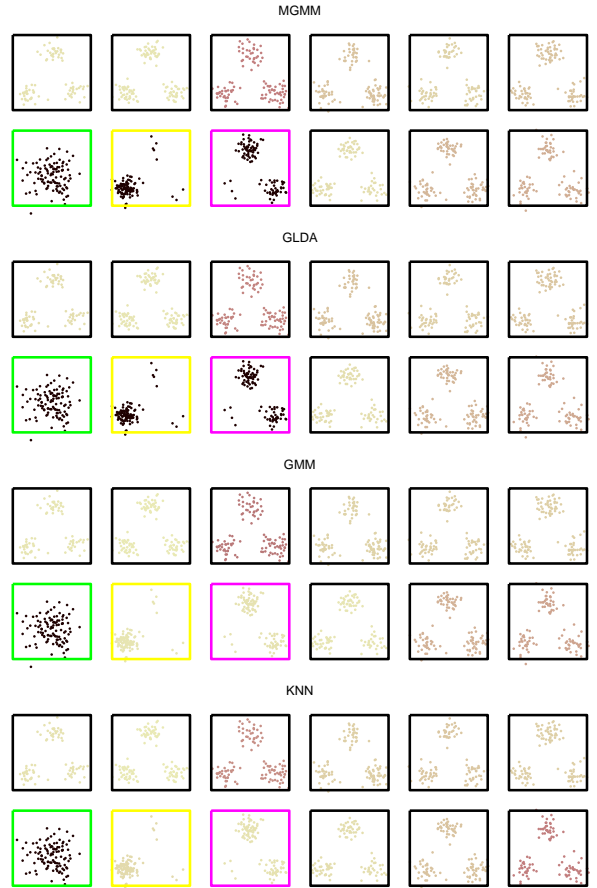

Figure 2: Detection results of MGMM, GLDA, GMM, and KNN methods on a data set with a uni-modal distribution of topic distributions. Inject anomalies are in the lower-left corner of each plot.

green, yellow, magenta boxes contain dark points, and the black boxes contain light gray points.

We can see that the group of point anomalies is easily identified by all methods, but the point-wise detectors (GMM, KNN) failed to detect the group anomalies, since these groups contain points that are individually normal. On the other hand, the proposed MGMM and GLDA models both examine the topic distributions of each group, and are able to discover the eccentric behaviors at the group level.

Next, we show that the uni-modal GLDA is not effective in more general cases. We create a data set with a multi-modal distribution of topic distributions. The two normal group types have topic distributions $\chi_1 = (0.33, 0.64, 0.03)$ and $\chi_2 = (0.33, 0.03, 0.64)$, and the group type distribution is $\pi = (0.48, 0.52)$. According to these parameters, a normal group should either consist mainly of topics 1&2, or mainly of topics 1&3. We corrupt three groups again in the same
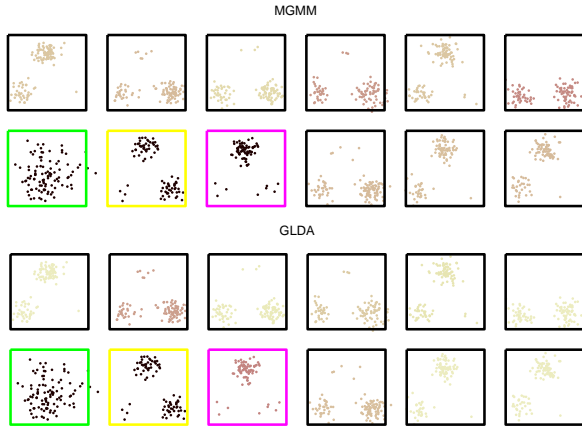
MGMM

GLDA

Figure 3: Detection results of MGMM and GLDA on a data set with a multi-modal distribution of topic distributions. The uni-modal GLDA breaks down on this data set.

way as in the previous experiment. The detection results are shown in Figure 3. Results from GMM and KNN are not shown because they failed again on this task and produced similar results as in Figure 2. The GLDA model can no longer effectively detect all the group anomalies because the uni-modal Dirichlet cannot accommodate multiple normal group types. Lacking of this flexibility, GLDA learned a model (Figure 4b) that misclassified one group anomaly as normal. On the other hand, MGMM is able to learn the true model (Figure 4c) and detect all anomalies, since its multi-modality admits multiple normal group types.
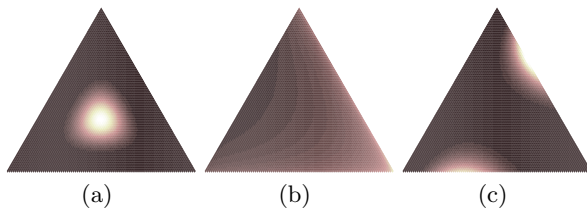

Figure 4: (a): the Dirichlet distribution learned from the uni-modal data. (b): the Dirichlet learned from the multi-modal data. Observe that this distribution is flat and assigns large probability to anomalous topic distributions in the corner. (c): the shape of the multi-modal distribution learned by MGMM.

Finally, we demonstrate the effects of the likelihood score and the topic score in details. Figure 5a shows the MGMM result on the multi-modal data using only the likelihood score. The magenta anomaly (third box) was misclassified because of the effect described in Section 4.4. Figure 5b shows the MGMM result on the uni-modal data using the topic score only: the green

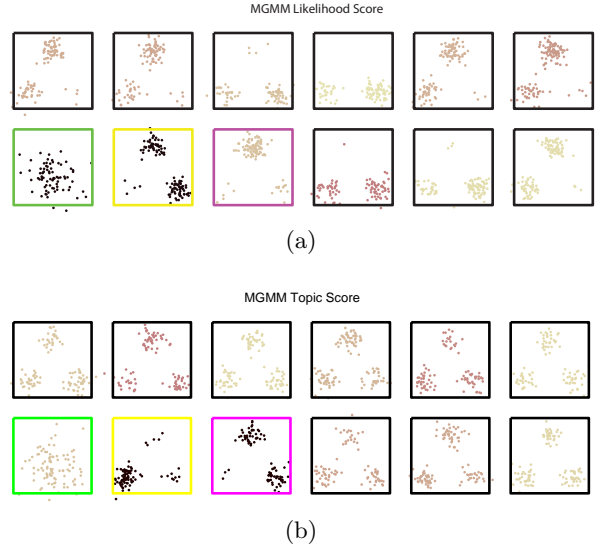MGMM Likelihood Score

(a)

MGMM Topic Score

(b)

Figure 5: Detection results of MGMM using different scoring functions. (a): result using the likelihood score only. (b): result using the topic score only.

anomaly (point anomalies) was missed. The reason behind this is that the topic score only examines the topic distribution without point-level details. In this contrived example, the point anomalies happened to be in the middle of the normal topics, so MGMM infers that this group consists of equal amount of points from each topic, which is exactly the normal behavior. From this, we can see that the topic score only focuses on the group-level behaviors. Combining it with the likelihood score, we can detect both types of anomalies.

5.2 Anomaly Detection in Astronomical Data

In this experiment, we use the algorithms on the *Sloan Digital Sky Survey* (SDSS) data set to find group anomalies. SDSS produces a large amount of data for celestial objects and gives them high-dimensional feature descriptions. Figure 6 shows one sample object from SDSS. Here we are interested in the galaxies in the SDSS. This subset contains about 7×10^5 objects that were identified by the SDSS pipeline as galaxies, and each object has a 4000-dimensional spectrum, which we down-sampled to get a 1000-dimensional feature vector for each galaxy.

To find the spatial clusters of galaxies, we first construct a neighborhood graph by adding edges between nearby galaxies (closer than 1 megaparsecs), and then treat the connected components in the graph as spatial clusters. This step produces 505 spatial clusters (7530 galaxies), each cluster contains about 10–50 galaxies. Then we reduced the 1000-dimensional features to 22-dimensional vectors by PCA to preserve 95% of the
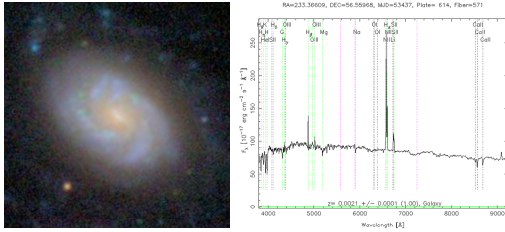
Figure 6: One object from the SDSS data set. The first image is the photometric observation, and the second image is the spectroscopic feature.

variance. This step helps the models get more reliable estimates of the Gaussians and accelerates the computation. For MGMM and GLDA, the topic score is used since we only want to find group anomalies. For all methods, we use BIC to select their parameters K and T.

We presented the detection results by MGMM on this data set to the astronomers and received positive feedbacks. Using the settings as above, the top anomalies found by MGMM are largely dense clusters of star-forming galaxies and irregular galaxies. Their existence is rare and indicates ongoing large scale events. We are still actively studying the meaning of them and other anomalies we found.

To be able to get a statistically meaningful comparison of the algorithms, we again use artificial anomaly injections due to the lack of labels. To evaluate the ability to detect group anomalies, injections are constructed using randomly selected galaxies, so that they look the same as the real data at the point-level, but their topic distributions were different than those of in the real groups. We compared the MGMM, GLDA, and GMM models in this experiment. The performances are measured by the *average precision* (AP) and *area under the ROC curve* (AUC) of retrieving the injected anomalies. In each run we inject 10 such random anomalies, so that the whole data set contains 515 groups. The results from 30 random runs are shown in Figure 7.

We can see that MGMM and GLDA both significantly outperforms the GMM model, whose performance is close to a baseline detector returning uniformly random results. AUC performances indicate that GLDA and MGMM tends to give the anomalies high scores. Further, the AP of MGMM is much higher than GLDA, showing that MGMM is able to detect the top anomalies much earlier. Note that the performances have large variances because each time the injections are random and we only injected 2% anomaly groups *w.r.t.* the whole data set. However, the improvement is significant. For the AP performances,
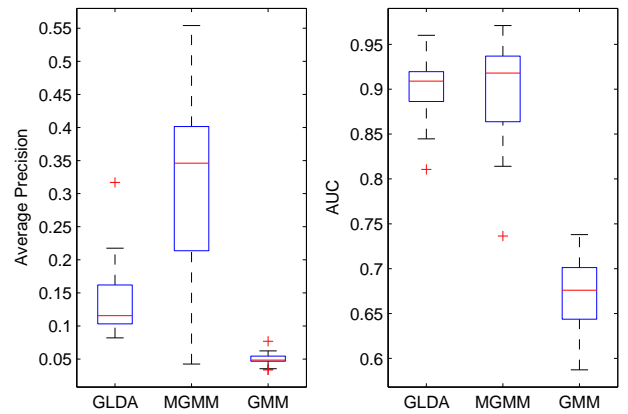

Figure 7: Anomaly detection performance on the SDSS galaxy cluster data.

paired t-tests gives significance values 4.9×10^{-11} for GLDA vs. GMM and 1.6×10^{-8} for MGMM vs. GLDA.

6 Discussion and Conclusions

In this paper we investigated how to use hierarchical probabilistic models for the group anomaly detection problem. Following the paradigm of topic modeling, two models are proposed to capture the generative process of both the individual points and the groups. The first model, called Gaussian LDA (GLDA), is effective for uni-modal group behaviors. Its extended version, the MGMM model, can also handle multi-modal group behaviors. The use of likelihood in group anomaly detection has also been investigated. The proposed scoring functions are able to detect both the point-level and group-level anomalous behaviors. Our experiments on both synthetic and real data sets show that the proposed models are effective in characterizing the data, and detecting anomalies.

Our future plan is to apply full Bayesian treatment for the current models, so that we can account for the uncertainty of the parameters, and get better results in the high-dimensional, small-sample scenarios. We can also use non-parametric Bayesian techniques, such as the *Hierarchical Dirichlet Process* (HDP) by Teh et al. (2006) to implement automatic complexity control.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, (19-6):716–723.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *JMLR*, 3:993–1022.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: Identifying density-based local outliers. In *SIGMOD*.

Chan, P. K. and Mahoney, M. V. (2005). Modeling multiple time series for anomaly detection. In *IEEE International Conference on Data Mining*.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41-3.

Das, K., Schneider, J., and Neill, D. (2008). Anomaly pattern detection in categorical datasets. In *Knowledge Discovery and Data Mining (KDD)*.

Das, K., Schneider, J., and Neill, D. (2009). Detecting anomalous groups in categorical datasets. Technical Report 09-104, CMU-ML.

de Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Academia Nazionale dei Lincei, Serie 6. Memorie, Classe di Scienze Fisiche, Mathematice e Naturale*, 4:251–299.

Hazel, G. G. (2000). Multivariate gaussian MRF for multispectral scene segmentation and anomaly detection. *IEEE Trans. Geoscience and Remote Sensing*, 38-3:1199 – 1211.

Jordan, M. I., editor (1999). *Learning in Graphical Models*. MIT Press, Cambridge, MA.

Keogh, E., Lin, J., and Fu, A. (2005). Hot sax: Efficiently finding the most unusual time series subsequence. In *IEEE International Conference on Data Mining*.

Li, J. (2001). Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics*, 14-3:547 – 568.

Mclachlan, G. J. and Krishnan, T. (1996). *The EM Algorithm and Extensions*. John Wiley and Sons.

Schwarz, G. E. (1974). Estimating the dimension of a model. *Annals of Statistics*, (6-2):461–464.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet process. *Journal of the American Statistical Association*, 101:1566 – 1581.

Zhao, M. (2009). Anomaly detection with score functions based on nearest neighbor graphs. In *NIPS*.