Unsupervised Spectral Ranking for Anomaly and Application to Auto Insurance Fraud Detection

Ke Nian[a,1], Haofan Zhang[a,1], Aditya Tayal[a,1], Thomas Coleman[b,2], Yuying Li[a,1,*]

[a] *Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada N2L 3G1*
[b] *Combinatorics and Optimization, University of Waterloo, Waterloo, ON, Canada N2L 3G1*

Abstract

For many data mining problems, obtaining labels is costly and time consuming, if not practically infeasible. In addition, unlabeled data often includes categorical or ordinal features which, compared with numerical features, can present additional challenges. By establishing a connection between unsupervised support vector machine optimization and spectral Laplacian optimization, we propose a new unsupervised spectral ranking method for anomaly (SRA). Using the 1st non-principal eigenvector of the Laplacian matrix directly, the proposed SRA can generate anomaly ranking either with respect to the majority class or with respect to two main patterns. The choice of ranking reference can be made based on whether the cardinality of the smaller class (positive or negative) is sufficiently large. Using an auto insurance claim data set but ignoring labels when generating ranking, we show that our proposed SRA significantly surpasses existing outlier-based fraud detection methods. In addition, based on a couple of additional data sets, for which the rare class can be determined based on the attributes dependence, we illustrate that the performance of SRA for unsupervised ranking of rare classes. Finally we demonstrate that, while our proposed SRA yields good performance for a few similarity measures for the auto insurance claim data, notably ones based on the Hamming distance, choosing appropriate similarity measures for a fraud detection problem remains crucial.

Keywords: unsupervised learning, fraud detection, rare class ranking, similarity measure, kernels, spectral clustering, one-class svm

*Corresponding author.
Email addresses: knian@uwaterloo.ca (Ke Nian), haofan.zhang@uwaterloo.ca (Haofan Zhang), amtayal@uwaterloo.ca (Aditya Tayal), tfcoleman@uwaterloo.ca (Thomas Coleman), yuying@uwaterloo.ca (Yuying Li)

[1]All authors acknowledge funding from the National Sciences and Engineering Research Council of Canada

[2]This author acknowledges funding from the Ophelia Lazaridis University Research Chair. The views expressed herein are solely from the authors.

1. Introduction

According to the 2013 survey from FICO, an American company who provides analytics and decision making services, one in three insurers in North American do not feel adequately protected against fraud, with auto insurance considered as one of the most vulnerable and 58% survey respondents anticipate increase in auto insurance fraud. In addition, many expect to see an increase in auto insurance fraud by organized rings.

Fighting against insurance fraud is a challenging problem both technically and operationally. It is reported in (Tennyson and Salsas-Forn, 2002; Derrig, 2002), approximately $21\% \sim 36\%$ auto-insurance claims contain elements of suspected fraud but only less than 3% of suspected fraud is prosecuted. Traditionally insurance fraud detection relies heavily on auditing and expert inspection. Since manually detecting fraud cases is costly and inefficient, investigative resource is severely constrained. Data mining and machine learning techniques have the potential to detect suspicious cases timely and significantly reduce economic losses, both to the insurers and policy holders. In addition, due to the need to detect fraud prior to the claim payment, data mining analytics is increasingly recognized as a key in fighting against fraud. There is great demand for effectively predicative methods which maximize the true positive detection rate, minimize the false positive rate, and are able to quickly identify new and emerging fraud schemes.

Without available labels, fraud detection can be formulated as an anomaly ranking problem. Anomaly detection constitutes a large collection of data mining problems such as disease detection, credit card fraud detection, and any detection of any new pattern amongst the existing patterns. In addition, comparing to simply providing binary classifications, providing a ranking which represents the degree of relative abnormality is crucial in cost and benefit evaluation analysis, as well as in turning analytic analysis into action.

Many methods for supervised rare class (anomaly) ranking exist in the literature. RankSVM Herbrich et al. (2000) can be applied to a bi-class rare class prediction problem. Unfortunately solving a nonlinear kernel RankSVM problem is computationally prohibitive for large data mining problems. Using SVM ranking loss function, a rare class based nonlinear kernel classification method, RankRC, in proposed recently in Tayal et al. (2013b,a).

Unfortunately it may not be feasible or preferable to use supervised anomaly ranking for fraud detection. Instead, an unsupervised anomaly ranking is more appropriate and beneficial. Due to time and resource cost in obtaining labels for learning, unsupervised learning has the advantage of a faster and more efficient application of knowledge discovery. This is often crucial in time sensitive application domains such as fraud detection. In addition, it may be practically infeasible to obtain accurate labels. In the context of fraud detection, obtaining clearly fraudulent and non-fraudulent samples is very costly, if not impossible. Even if one ignores human investigative costs, it is quite common to find fraud investigators to differ in their claim assessments. This raises issues of data (label) trustworthiness. Moreover, the need to detect fraudulent claims before payments are made and to identify new fraud schemes quickly essentially rule out supervised learning as a candidate solution to effective fraud detection in practice.

In such cases, unsupervised learning is preferable, in order to quickly extract information in available data.

Standard unsupervised anomaly detection methods include clustering analysis and/or outlier detection methods, see, e.g., Hastie et al. (2009). Many outlier detection methods have been proposed in the literature, e.g. k-Nearest Neighbor (k-NN) outlier detection, One-Class Support Vector Machine (OC-SVM) (including kernel-based), and density-based methods such as Local Outlier Factor (LOF). Effectiveness of these methods has been investigated in numerous application domains including network intrusion detection, credit card fraud detection, and abnormal activity detection in electronic commerce. However, many standard outlier detection methods, e.g., one-class SVM, are only suitable for detecting outliers with respect to a single global pattern which we refer to as global outliers in this paper. An implicit assumption in this case is that normal cases are generated from one mechanism and abnormal cases are generated from other mechanisms. These methods work well for some of aforementioned problems, e.g., when only a single majority pattern exists. Density-based methods can be effective in detecting both global outliers and local outliers; but assumption here is that density is the only discriminant for abnormality. Density-based methods fail when small dense clusters also constitute abnormality. In addition, density based methods, e.g., LOF, often require users to define a neighborhood range to compare the density. Tuning these parameters can often be tricky.

Understandably, unsupervised learning is much more difficult than supervised learning since learning targets are not available to guide the learning process. In practice, difficulty in unsupervised learning is further exacerbated by the additional challenge in identifying relevant features for unsupervised learning methods. Based on aforementioned challenges, it then comes to no surprise that the existing literature on auto insurance fraud detection typically formulates the problem as a supervised learning, see, e.g., Phua et al. (2004, 2010) and references therein.

The literature on *unsupervised* auto insurance fraud detection is extremely sparse. To the best of our knowledge, it includes the self-organizing feature map method Brockett et al. (1998) and PRIDIT analysis Brockett et al. (2002); Ai et al. (2012, 2009), which is based on RIDIT Score and Principal Component Analysis. Note that studies in Brockett et al. (1998, 2002); Ai et al. (2012, 2009) have been conducted using the single Personal Injury Protection (PIP) data set, which is provided by Automobile Insurance Bureau (AIB) from Massachusetts Brockett et al. (2002). This data set has been preprocessed by auditors and fraud detection inspectors and the features are the red flag selected by domain experts. Furthermore, values of explanatory variables for fraudulent instances in this data set tend to be smaller than that of the instances from non-fraudulent class, which corresponds to ranking of fraudulent suspiciousness to some degree. Consequently it is reasonable to regard this as a partially supervised learning, which is susceptible to, at least partially, inadequacies in the labeling process for fraud detection. In addition, auto insurance or health insurance claim data often consists of numerical, ordinal, categorical, and text data. Indeed it would be difficult to apply methods in Brockett et al. (1998, 2002); Ai et al. (2012, 2009) without first preprocessing from domain experts. Moreover, when claim data consists of many categorical features, it is reasonable to expect that data can form more than one major pattern. This makes any single class based outlier detection method less likely to be

effective.

The economic value of an efficient unsupervised anomaly ranking, without any help from domain experts, can be substantial. In this paper, we propose a new unsupervised ranking method for anomaly based on spectral analysis. Specifically we consider both rare class ranking, in which anomaly is assessed with respect to a single majority class, as well as anomaly ranking which is assessed with respect to more than one major pattern.

Given a data set of input features, anomaly can be assessed with respect to marginal feature distributions as well as feature dependence relationships, which can be either linear or nonlinear. Following research on use of kernels to measure nonlinear feature dependence, see, e.g., Gretton et al. (2005); Song et al. (2012), we focus on detecting anomaly in feature dependence using similarity kernels, where the similarity captures a dependence relation among input features. We assume that a similarity kernel, which is constructed to capture a feature dependence from given input features, is given. We then use spectral analysis to compute the first bi-modal non-principal eigenvector to generate anomaly ranking. We also compute the 2nd non-principal eigenvector to assist visualization. We also apply several standard anomaly detection methods on auto insurance problem and compare the performance with our proposed ranking algorithms.

The main contribution of this paper is summarized as follows

- We observe a connection between an unsupervised Support Vector Machine (SVM) optimization formulation and the spectral optimization. Specifically we demonstrate that spectral optimization based on the Laplacian matrix can be viewed as a relaxation of unsupervised SVM. Moreover, it can be interpreted as a constant scaling-translation transformation of an approximate optimal bi-class classification function evaluated at given data instances.

- Motivated by the above observation, we propose an unsupervised spectral ranking for anomaly (SRA) algorithm which generates ranking directly from the non-principal bi-modal eigenvectors of a normalized Laplacian matrix.

- Using the real auto insurance claim data in Phua et al. (2004), we evaluate effectiveness of the unsupervised SRA for detecting anomaly with respect to multiple major patterns. We show that the proposed SRA performs significantly better than existing methods. Indeed it comes close to the in-sample training performance of a random forest supervised classification, which can be considered an upper bound for this problem.

- We demonstrate effectiveness of the unsupervised SRA for abnormal detection with respect to a single majority class using a few real data sets from UCI machine learning repository.

Presentation of the paper is organized as follows. In §2, we review similarity measures which are used in our computational investigation. In §3, we summarize the basic idea of spectral analysis. This is followed by presenting spectral optimization as a non-convex relaxation of the unsupervised support vector machine classification in §4. In §5, we propose a new unsupervised spectral ranking for anomaly method (SRA). We

present and discuss, in §6, the results from the proposed ranking method for the auto insurance claim data. Although we focus in this paper in applying our proposed method for auto insurance fraud detection, we note that it can be applied to other anomaly detection problems as well. We recognize however that appropriate choice of similarity measure may depend on specific application context. Brief discussion of applying our methods on other real data sets are included in §7. Concluding remarks are given in §8.

2. Similarity Measures for Categorical Data

In practice, representation of human activities and behavior often leads naturally to categorical (nominal) and ordinal data. Since most existing data mining methods solely focus on numerical values, a common preprocessing technique for categorical data expands categorical features into a set of binary indicators. Unfortunately this can inject unintended order and scale into the representation, which are absent from the original data. This can potentially significantly disturb feature importance in the original data and lead to poor pattern recognition, especially under unsupervised learning settings. However, this crucial understanding is often missing in practice. Indeed we fail to obtain any meaningful results on the auto insurance data set considered in this paper, when we treat categorical data directly by performing binary expansion and apply existing clustering and outlier detection methods.

A more meaningful treatment for categorical data is to use a similarity measure to capture relationship between input categorical features. Indeed similarity measures have been studied for more than a century, see, e.g., Boriah et al. (2008) and references therein. In this paper we consider two types of similarity measures: nominal value definition driven and nominal value distribution driven similarities.

Definition driven similarity is defined directly from the specification of the feature nominal values. For auto insurance fraud detection, match and mismatch in nominal values of categorical features form an intuitive basis for comparing claim patterns. Hence we mainly focus on the overlapping similarity, which is the simplest similarity measure, and its derived kernels. Our subsequent discussion assumes that the data set comes from sampling of random n-dimensional categorical vector $\mathcal{D}$, with the ith feature $\mathcal{D}_i$ having $|\mathcal{D}_i|$ distinct nominal values, $i = 1, 2, \cdots, n$.

Overlapping Similarity

Given two n-dimensional feature vector x and y, overlapping similarity is given by

$$s^O(\mathbf{x}, \tilde{\mathbf{x}}) = 1 - d^H(\mathbf{x}, \tilde{\mathbf{x}})$$

where $d^H(\mathbf{x}, \tilde{\mathbf{x}})$ is the Hamming distance defined as the number of features that $\mathbf{x}$ and $\tilde{\mathbf{x}}$ do not match divided by the total number of features:

$$d^H(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{\sum_{i=1}^{n} \delta(\mathbf{x}_i, \tilde{\mathbf{x}}_i)}{n}$$

where

$$\delta(\mathbf{x}_i, \tilde{\mathbf{x}}_i) = \begin{cases} 1, & \mathbf{x}_i \neq \tilde{\mathbf{x}}_i \\ 0, & \mathbf{x}_i = \tilde{\mathbf{x}}_i \end{cases}$$

Considering the number of nominal values in each feature, a weighted Hamming distance has also been considered in David and Averbuch (2012)

$$d^{WH}(\mathbf{x}, \tilde{\mathbf{x}}) = \sum_{i=1}^{n} \frac{\delta(\mathbf{x}_i, \tilde{\mathbf{x}}_i)}{|\mathcal{D}_i|}.$$

Overlapping similarity is in fact a valid kernel function, see, e.g., Gärtner et al. (2006). Replacing Euclidean distance in a standard Gaussian kernel by the Hamming distance, we immediately obtain a Gaussian kernel derived from the Hamming distance, i.e.,

$$k^{GH}(\mathbf{x}, \tilde{\mathbf{x}}) = e^{-\frac{d^H(\mathbf{x}, \tilde{\mathbf{x}})}{2\sigma^2}} \tag{1}$$

where $\sigma > 0$ is a constant kernel width parameter.

Adaptive Gaussian Kernel

In the Gaussian Hamming Kernel (1), a single bandwidth σ is applied to every data instance. It has been argued in David and Averbuch (2012) that clustering performance can be improved using an adaptive bandwidth. where $|\mathcal{D}_i|$ is the number of distinct values in that dimension. Corresponding to the weighted Hamming distance, one can consider an adaptive kernel

$$k^{WH}(\mathbf{x}, \tilde{\mathbf{x}}) = e^{-\frac{d^{WH}(\mathbf{x}, \tilde{\mathbf{x}})}{\sigma(\mathbf{x}, \tilde{\mathbf{x}})}}$$

where $\sigma(\mathbf{x}, \tilde{\mathbf{x}})$ is a data driven adaptive bandwidth determined by a fixed number of nearest neighbors of data instance $\mathbf{x}$ and $\tilde{\mathbf{x}}$. In the subsequent section, we use β to denote the neighborhood size parameter of this kernel in order to avoid confusion.

Hamming Distance Kernel

Kernels have been shown to be very successful in machine learning and data mining, since they provide a uniform way to handle complicated relationship between features. For categorical data, Hamming distance kernel has been introduced in Couto (2005), which can be viewed as a variant of the string kernel Lodhi et al. (2002).

Let $\mathcal{D}^n$ be the cross product over all n input feature domains. Hamming distance kernel is defined by implicitly considering the $|\mathcal{D}^n|$-dimensional feature space in which each possible feature nominal value combination represents one dimension. For each $\mathbf{u}$ representing one dimension in the $|\mathcal{D}^n|$-dimensional kernel feature space, an explicit mapping is defined to map any instance x into this dimension, i.e., $\theta_{\mathbf{u}}(\mathbf{x})$. For any given original input features $\mathbf{x}$ and $\tilde{\mathbf{x}}$, the kernel function $k(\mathbf{x}, \tilde{\mathbf{x}})$ equals the inner product the mapped features $\{\theta_{\mathbf{u}}(\mathbf{x}), \mathbf{u} \in \mathcal{D}^n\}$ and $\{\theta_{\mathbf{u}}(\tilde{\mathbf{x}}), \mathbf{u} \in \mathcal{D}^n\}$. More specifically, let $\mathcal{D}_i$ be the domain of the ith feature. For each $\mathbf{u} \in \mathcal{D}^n$, given a categorical instance $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, $\mathbf{x}_i \in \mathcal{D}_i$, we define an explicit mapping:

$$\theta_{\mathbf{u}}(\mathbf{x}) = \lambda^{d^H(\mathbf{u}, \mathbf{x})} \tag{2}$$

where $\lambda \in (0,1)$ is a damping parameter. Note that $\theta_u(\mathbf{x})$ is just one dimension of the kernel feature space. Thus the Hamming distance kernel between instances $\mathbf{x}$ and $\mathbf{y}$ is:

$$k^H(\mathbf{x}, \tilde{\mathbf{x}}) = \sum_{u \in \mathcal{D}^n} \theta_u(\mathbf{x}) \theta_u(\tilde{\mathbf{x}}) \tag{3}$$

Directly computing the Hamming distance kernel has an exponential computational complexity. Fortunately a dynamic programming technique can be applied, which allows this kernel to be computed efficiently following the recursive procedure below Couto (2005):

$$k^0(\mathbf{x}, \tilde{\mathbf{x}}) = 1$$
$$k^j(\mathbf{x}, \tilde{\mathbf{x}}) = (\lambda^2(|\mathcal{D}_j| - 1 - \delta(\mathbf{x}_j, \tilde{\mathbf{x}}_j)) + (2\lambda - 1)\delta(\mathbf{x}_j, \tilde{\mathbf{x}}_j) + 1) k^{j-1}(\mathbf{x}, \tilde{\mathbf{x}}), \quad j = 1, \cdots, n$$

and $k^H(\mathbf{x}, \tilde{\mathbf{x}}) = k^n(\mathbf{x}, \tilde{\mathbf{x}})$.

While evaluating the similarity between a pair of instance $\mathbf{x}$ and $\tilde{\mathbf{x}}$, Hamming distance directly compares nominal values of $\mathbf{x}$ and $\tilde{\mathbf{x}}$ to determine the number of different values. In contrast, a Hamming distance kernel assesses the similarity between $\mathbf{x}$ and $\tilde{\mathbf{x}}$ in referencing to all possible nominal value combinations of the categorial attribute. Consequently, a Hamming distance kernel can capture more information than the simple Hamming distance (overlapping similarity).

Distribution driven similarity is determined from distributions of feature values. This type of similarity measures may adjust similarity for rare categorical values, which introduces additional information in representing the relationship of data instances. This can potentially detect different patterns in comparison to simpler feature definition driven similarity measures.

To illustrate, we consider here DISC (Data-Intensive Similarity Measure for Categorical Data) Desai et al. (2011), which is an example of distribution driven similarity measure for categorical data. DISC attempts to represent the underlying semantics of the data by capturing relationships that are inherent in the data. Different from other distribution driven similarities, DISC is based on the cosine similarity of co-occurrence statistics for different feature values. Since the main purpose of using DISC in this paper is to illustrate effect of similarity on the performance of a ranking method, here we simply refer an interested reader to Desai et al. (2011) for a detailed definition and discussion on DISC.

3. Spectral Analysis and Clustering

Spectral clustering is a more recent, popular, and successful clustering technique, which often outperforms traditional clustering techniques such as k-means and hierarchical clustering Hastie et al. (2009). Here we briefly introduce and motivate the method. More detailed discussion can be found in Von Luxburg (2007).

The main objective of clustering is to partition data into groups so that similarity between different groups is minimized. Hence similarity based clustering can be modeled as a graph cut problem. Let each data instance be a vertex and each pair of vertices be connected with an edge with a weight equal to the similarity of the pair. We

can then represent the data and its similarity using an undirected graph $G = (V, E)$, with vertices $V = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n\}$ corresponding to data instances and an adjacency matrix W summarizing similarities, where W_{ij} is the similarity between $\mathbf{v}_i$ and $\mathbf{v}_j$.

Let $\mathbf{d}$ be the degree vector of each vertex with $\mathbf{d}_i = \sum_j W_{ij}$ and D be the diagonal matrix with $\mathbf{d}$ on the diagonal. From the degree matrix D and the weighted adjacency matrix W, a Laplacian matrix, which is fundamental in spectral clustering computation, can be introduced. There are different variations in the definition of Laplacian. The relevant definition to our discussion in this paper is the symmetric normalized Laplacian $L = I - D^{-1/2} W D^{-1/2}$ Ng et al. (2002).

The key idea of a spectral clustering algorithm is to determine clustering membership of data instances by applying a simple clustering technique, e.g., k-means, on a subset of eigenvectors (typically the first few non-principal eigenvectors) of a graph Laplacian matrix, see, e.g., Shi and Malik (2000); Ng et al. (2002). Assume that the eigenvalues of L are

$$\lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1}$$

and $\mathbf{g}_k^*$ is an eigenvector associated with λ_k, $k = 0, \cdots, n-1$. Let $\mathbf{e}$ be a n-by-1 vector of ones. Since $LD^{\frac{1}{2}}\mathbf{e} = D^{\frac{1}{2}}\mathbf{e} - D^{-\frac{1}{2}}W\mathbf{e} = 0$ and L is positive semidefinite, $\mathbf{g}_0^* = D^{\frac{1}{2}}\mathbf{e}$ is the principal eigenvector associated with the minimum eigenvalue $\lambda_0 = 0$.

Note that the kth non-principal eigenvector solves

$$\begin{aligned} & \min_{\mathbf{g} \in \Re^n} \mathbf{g}^T L \mathbf{g} \\ \text{subject to} \quad & \mathbf{g} \perp \mathbf{g}_i^*, \quad i = 0, \cdots, k-1 \\ & \mathbf{g}^T \mathbf{g} = \upsilon \end{aligned} \tag{4}$$

where $\upsilon = \sum_{i=1}^n \mathbf{d}_i$.

Since the principal eigenvector corresponding to $\lambda_0 = 0$ is $\mathbf{g}_0^* = D^{\frac{1}{2}}\mathbf{e}$, each non-principal eigenvector $\mathbf{g}_i^*$ satisfies

$$(D^{\frac{1}{2}}\mathbf{e})^T \mathbf{g}_i^* = 0$$

we immediately note that each $\mathbf{g}_i^*$ always contain both positive and negative components, which we can group them into a positive class $\mathcal{C}_+ = \{j : (\mathbf{g}_i^*)_j \geq 0\}$ and a negative class $\mathcal{C}_- = \{j : (\mathbf{g}_i^*)_j < 0\}$.

Indeed, following the typical motivation of spectral clustering, the spectral optimization problem (4) can be regarded as a relaxation of an integer programming problem, which models a normalized graph 2-cut problem. We refer an interested reader to Von Luxburg (2007) for a more detailed discussion. Consequently a spectral clustering based on m eigenvectors can be regarded as a m-step iterative bi-clustering (classification) method, in which each successive iteration looks for a bi-clustering in the space orthogonal to the first $(m-1)$ bi-clustering eigenvector space. To determine clustering membership from the eigenvectors of Laplacian, typically another clustering method, e.g., k-means, is applied. Clustering methods such as k-means typically require the number of clusters k to be specified *a prior*.

The main objective of this paper is to develop a ranking method for anomaly di-

rectly; it is not the determination of cluster membership. In order to achieve this, it is important to have a better understanding on the information which is present in the eigenvector components.

Next we present a new interpretation of the non-principal eigenvectors for Laplacian, which will provide a basis for our proposed spectral ranking method for anomaly.

4. Spectral Laplacian Optimization as Approximation to Unsupervised SVM

Consider the first non-principal eigenvector for the normalized spectral optimization, denoted as $\mathbf{g}^*$ for notational simplicity. Then

$$\begin{aligned} \min_{\mathbf{g}\in\Re^n} \quad & \mathbf{g}^T L\mathbf{g} \\ \text{subject to} \quad & \mathbf{e}^T D^{\frac{1}{2}}\mathbf{g} = 0 \\ & \mathbf{g}^T\mathbf{g} = \upsilon \end{aligned} \tag{5}$$

Recalling that $L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ and ignoring the constant, (5) is equivalent to

$$\begin{aligned} \min_{\mathbf{g}\in\Re^n} \quad & -\mathbf{g}^T D^{-\frac{1}{2}} W D^{-\frac{1}{2}}\mathbf{g} \\ \text{subject to} \quad & \mathbf{e}^T D^{\frac{1}{2}}\mathbf{g} = 0 \\ & \mathbf{g}^T\mathbf{g} = \upsilon \end{aligned} \tag{6}$$

Assume that $\mathbf{z} = D^{\frac{1}{2}}\mathbf{g}$. Then (6) is equivalent to

$$\begin{aligned} \min_{\mathbf{z}\in\Re^n} \quad & -\mathbf{z}^T D^{-1} W D^{-1}\mathbf{z} \\ \text{subject to} \quad & \mathbf{e}^T\mathbf{z} = 0 \\ & \mathbf{z}^T D^{-1}\mathbf{z} = \upsilon \end{aligned} \tag{7}$$

Assuming that W is positive definite in the space $\{\mathbf{z} : (D^{\frac{1}{2}}\mathbf{e})^T\mathbf{z} = 0\}$, the ellipsoidal equality constraint in (7) can be replaced by an inequality constraint

$$\begin{aligned} \min_{\mathbf{z}\in\Re^n} \quad & -\mathbf{z}^T D^{-1} W D^{-1}\mathbf{z} \\ \text{subject to} \quad & \mathbf{e}^T\mathbf{z} = 0 \\ & \mathbf{z}^T D^{-1}\mathbf{z} \leq \upsilon, \end{aligned} \tag{8}$$

since the ellipsoidal constraint in (8) should be active at a solution.

However (8) can be considered as an approximation to the following nonconvex problem

$$\begin{aligned} \min_{\mathbf{z}\in\Re^n} \quad & -\mathbf{z}^T K\mathbf{z} \\ \text{subject to} \quad & \mathbf{e}^T\mathbf{z} = 0 \\ & |\mathbf{z}| \leq C \end{aligned} \tag{9}$$

where $C = \upsilon \cdot \mathbf{d}^{\frac{1}{2}}$ and $K = D^{-1}WD^{-1}$. Here we have approximated the ellipsoidal inequality constraint in (8) by the rectangular constraint in (9).

Interestingly, optimization problem (9) can be regarded as an approximation to the following formulation of the unsupervised support vector machine

$$\min_{\mathbf{y}\in\{\pm 1\}} \max_{\substack{0\le\alpha\le C \\ y^T\alpha=0}} \quad -\frac{1}{2}\alpha^T YKY\alpha + \mathbf{e}^T\alpha \tag{10}$$

where $Y = \text{DIAG}(\mathbf{y})$, and the optimal decision function for the 2-class classification if given by

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i^* \mathbf{y}_i^* K(\mathbf{x}_i, \mathbf{x}) + b^*$$

and the magnitude of the coefficient α_i^* signifies the degree of support of the i-th data instance in the optimal decision function. In particular, the data instance corresponding to $\alpha_i^* > 0$ is termed as a support vector in SVM.

Theorem 1. *Suppose that K is symmetric positive definite. Let* $(\alpha^*, \mathbf{y}^*)$ *be a solution to the unsupervised SVM* (10)*. Let* $\mathbf{z}^*$ *solve the relaxation problem* (9)*. Assume that the solution to*

$$\max_{\substack{0\le\alpha\le C \\ y^T\alpha=0}} -\frac{1}{2}\alpha^T Y_{\mathbf{z}}^* KY_{\mathbf{z}}^*\alpha + \mathbf{e}^T\alpha \tag{11}$$

is achieved at the vertex of $\mathcal{F}_\alpha = \{\alpha : 0 < \alpha \le C\}$*, where* $Y_{\mathbf{z}}^* = \text{DIAG}(\mathbf{y}_{\mathbf{z}}^*)$*. If* $\mathbf{e}^T\alpha^* = \mathbf{e}^T\alpha_{\mathbf{z}}^*$*, then* $\alpha_{\mathbf{z}}^* = |\mathbf{z}^*|$ *and* $\mathbf{y}_{\mathbf{z}}^* = \text{SIGN}(\mathbf{z}^*)$ *solve the unsupervised SVM* (10)*.*

The proof of this theorem is given in Appendix A. We note that in connecting spectral optimization (5) with unsupervised SVM (10), the rectangular constraint in (9) has been approximated by the ellipsoidal constraint in (8).

In addition to being a relaxation of the unsupervised SVM, we also note that the objective function in (9) can be decomposed as

$$\text{sim}(\mathcal{C}_+) + \text{sim}(\mathcal{C}_-) - 2 \times \text{sim}(\mathcal{C}_+, \mathcal{C}_-)$$

where $\mathcal{C}_+ = \{j : \mathbf{z}_j \ge 0\}$, $\mathcal{C}_- = \{j : \mathbf{z}_j < 0\}$, $\text{sim}(\mathcal{C}) = \sum_{i,j\in\mathcal{C}} |\mathbf{z}_i||\mathbf{z}_j|K_{ij}$ measures similarity of instance in $\mathcal{C}$ ($\mathcal{C}$ can be either $\mathcal{C}_+$ or $\mathcal{C}_-$, and $\text{sim}(\mathcal{C}_+, \mathcal{C}_-) = \sum_{i\in\mathcal{C}_+, j\in\mathcal{C}_-} |\mathbf{z}_i||\mathbf{z}_j|K_{ij}$ measures similarity between $\mathcal{C}_+$ and $\mathcal{C}_-$. Consequently, (9) corresponds to maximizing a measure of cluster quality.

Theorem 1 suggests that the normalized spectral optimization problem (5) can be regarded as an approximation to the unsupervised SVM problem (10) with the kernel $K = D^{-1}WD^{-1}$ and $C = \upsilon \cdot \mathbf{d}^{\frac{1}{2}}$.

Since the optimal separating hypothesis from the unsupervised SVM has the form

$$f(\mathbf{x}) = \left(\sum_{j=1}^{n} y_j^* \alpha_j^* \cdot K(\mathbf{x}, \mathbf{x}_j) + b^* \right)$$

and a non-principal eigenvector of the normalized spectral clustering $\mathbf{z}^*$ yields an approximation $|\mathbf{z}^*| \approx \alpha^*$ and $\text{SIGN}(\mathbf{z}^*) \approx y^*$, which are the coefficients of the bi-class separating optimal decision function, $|\mathbf{z}_j^*|$ provides a measurement of the strength of support from the jth data point on the two class separation decision. We note however that, due to the use of the ellipsoidal constraint rather than rectangular constraints and other approximations, $\mathbf{z}^*$ is different from the exact SVM decision function coefficients. In particular, the eigenvector is likely to have mostly nonzero components, i.e., every data point offers support to some degree in this two clusters separation.

In addition, assume that $\mathbf{g}_1^*$ is the first non-principal eigenvector of the unnormalized Laplacian L, i.e., $L = I - W$, with the eigenvalue λ_1. Then $K = W$ and $\mathbf{z}_1^* = \mathbf{g}_1^*$. Under this assumption, it can be easily verified that

$$K\mathbf{z}_1^* = (1+\lambda_1)\mathbf{z}_1^*.$$

Consequently

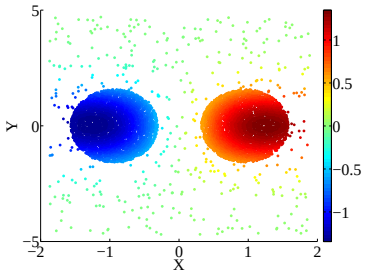
$$(1+\lambda_1)\mathbf{z}_1^* = K\mathbf{z}_1^* \approx \begin{pmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_n) \end{pmatrix} - b^*$$

Consequently $\mathbf{z}_1^*$ can be interpreted as a constant scaling-translation mapping of the approximate optimal bi-class separation function $f(\mathbf{x})$ evaluated data instances. Hence it is reasonable to use spectral optimization solution $\mathbf{z}^*$ as the ranking for the bi-cluster separation.

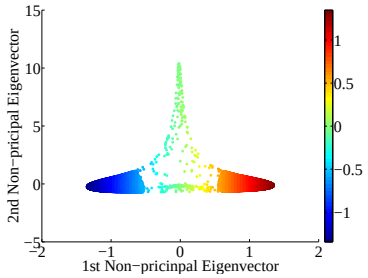
5. A New Spectral Ranking for Anomaly

In standard spectral analysis, a k-means clustering method is typically applied to non-principal eigenvectors to determine clustering memberships. Our discussion in §4 suggests that the components of a non-principal eigenvector $\mathbf{z}^*$ has meaning beyond indicating a cluster membership. In fact $|\mathbf{z}^*|$ provides a bi-class clustering strength measure in an optimal bi-class clustering in the high dimensional feature space according to the assumed similarity. We now graphically illustrate in details the information in a non-principal eigenvector $\mathbf{z}^*$ and motivate how the information can be used to rank anomaly. Furthermore, we allow a choice of the reference in the assessment of anomaly ranking. If the minority cluster class does not have a sufficient mass, one can choose to assess anomaly likelihood with respect to a single majority class and ranking is generated suitably with this view. Otherwise, anomaly is assessed with two main clusters.

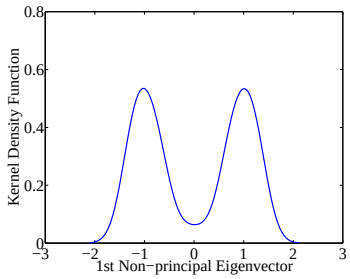
Figure 1 presents a visual illustration of the clustering strength information in the first non-principal eigenvector for a balanced two-cluster data set. We consider the 1st and the 2nd non-principal eigenvector of the Laplacian corresponding to the Gaussian kernel with bandwidth $\sigma = 1$ as the similarity. Subplot (a) graphs the original 2-D data while Subplot (b) graphs points in the space of the 1st and 2nd non-principal eigenvectors. To see how the original data points correspond to points in the eigenvector space, we assign each data point in the dimension of the 1st non-principal eigenvector,



(a) Synthetic Data



(b) Visualization of $\mathbf{z}_1^* = D^{1/2}\mathbf{g}_1^*$ and $\mathbf{z}_2^* = D^{1/2}\mathbf{g}_2^*$



(c) Kernel Density Estimation

Figure 1: Visualizing Information in the 1st Non-principal Eigenvector: two balanced clusters with noise

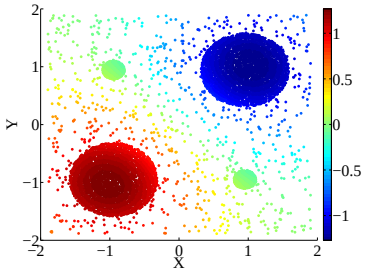
$\mathbf{z}^* = D^{1/2}\mathbf{g}_1^*$, a unique color with the darker color corresponds to a larger magnitude in the 1st non-principal eigenvector component. The color of the original data point in Subplot (a) is the same as the color of the corresponding point in the eigenvectors in Subplot (b). The colormap is shown at the right side of Subplot (a) and Subplot (b) in Figure 1. We also graph the density of the 1st non-principal eigenvector in Subplot (c), which clearly indicating presence of two clusters in this case.

From Figure 1, it can be observed that the bi-class clustering strength $|\mathbf{z}^*|$ of global outliers, typically corresponding to colors in yellow and green, is the smallest. In addition, data points which are closer to the other cluster have the light blue or light red colors; this suggest that they offer less information in defining the clusters and consequently the corresponding clustering strength $|\mathbf{z}^*|$ is smaller than that of the cluster cores, colored as dark blue and dark red.

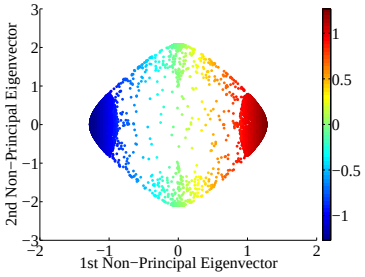
In Figure 2 we consider instead a synthetic example which includes major cluster patterns in combination with small cluster patterns. Subplots (a), (b), and (c) are similarly generated as in Figure 1. From Figure 2, data points in two smaller clusters lie closer to the origin and are depicted in lighter colors. In addition, similar to Figure 1, we observe that global outliers mostly have smallest clustering strength support and lie near the origin (in yellow and green). In addition, data points which are closer to the other cluster offer less information in defining the clusters and consequently the

corresponding clustering strength $|\mathbf{z}^*|$ is smaller. We can also see from the darkness of red and blue color that the edge of the major clusters has relatively smaller clustering strength($|\mathbf{z}|$) than the core of the major clusters. Note that the density plot of the 1st non-principal eigenvector in Subplot (c) now has three modes and indicates presence of additional small clusters.

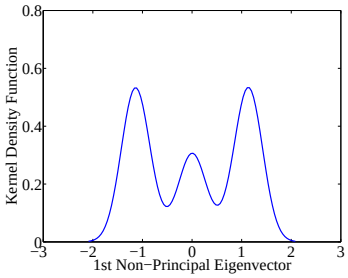
Figure 1 and 2 demonstrate that the 1st non-principal eigenvector indeed contains major bi-class clustering support strength, which can be used to produce ranking for anomaly, either global outliers or small anonymous patterns relative to major normal clusters. In this paper, we propose to define the anomaly score as $\mathbf{f}^* = \|\mathbf{z}^*\|_\infty - |\mathbf{z}^*|$, which yields larger positive values for points closer to the origin, e.g., global outliers or anonymous small clusters. The edges of the major clusters will be ranked as more abnormal next and the core of the major clusters will be ranked as least abnormal. This indeed is reasonable and represents our intention.



(a) Synthetic Data



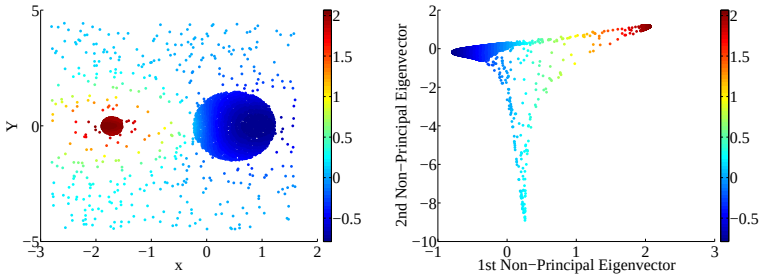
(b) Visualization of $\mathbf{z}_1^* = D^{1/2}\mathbf{g}_1^*$ and $\mathbf{z}_2^* = D^{1/2}\mathbf{g}_2^*$



(c) Kernel Density Estimation

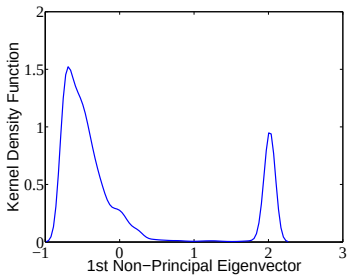
Figure 2: Visualizing Information in the 1st Non-principal Eigenvector: two major patterns and small clusters

In practice there often can be cases when there is only one major pattern for the normal class with small clusters representing a rare class (possibly with global outliers). Figure 3 illustrates such an example. Under this scenario, it is more reasonable to regard one of the two clusters represented in the 1st non-principal eigenvector as anonymous with respect to the major pattern. In this case, the appropriate ranking score can be generated from either $\mathbf{f}^* = \mathbf{z}^*$ or $\mathbf{f}^* = -\mathbf{z}^*$, depending on whether the smaller cluster



(a) Synthetic Data

(b) Visualization of $\mathbf{z}_1^* = D^{1/2}\mathbf{g}_1^*$ and $\mathbf{z}_2^* = D^{1/2}\mathbf{g}_2^*$



(c) Kernel Density Estimation

Figure 3: Visualizing Information in the 1st Non-principal Eigenvector: one major pattern and anonymous small clusters

corresponds to positive or negative values. Note that global outliers, if present, will be ranked as next abnormal cases. The major cluster will have the smallest eigenvector components and will be ranked as the least abnormal.

The above synthetic examples illustrate the information in the first non-principal eigenvector described in §4. Using the cluster support strength information in the non-principal eigenvector, we propose a new Spectral Ranking for Abnormality (SRA) method, which is summarized in Algorithm 1. If one of $\mathcal{C}^+$ and $\mathcal{C}^-$ is sufficiently small in size relative to the other, the proposed SRA provides anomaly ranking score $\mathbf{f}^*$ with respect to one majority class (rare class ranking). Otherwise SRA yields an anomaly ranking with respect to the two ($\pm$) major classes. Specifically, let $\mathcal{C}_+ = \{i, (\mathbf{g}_1)_i^* \geq 0\}$ and $\mathcal{C}_- = \{i, (\mathbf{g}_1^*)_i < 0\}$ denote data instance index sets corresponding to non-negative and negative value in $\mathbf{g}_1^*$ respectively. In addition, assume that an a priori upper bound for the anomaly ratio χ is given. If $\min\{\frac{|\mathcal{C}_+|}{n}, \frac{|\mathcal{C}_-|}{n}\} \geq \chi$, then neither the set $\mathcal{C}_+$ nor $\mathcal{C}_-$ is considered as an anomaly class and SRA outputs ranking with respect to multiple patterns. Otherwise, SRA outputs anomaly ranking with respect to a single majority class. If the 1st non-principal eigenvector is relatively balanced, $|\mathcal{C}_+| \approx |\mathcal{C}_-|$, SRA outputs anomaly ranking with respect to multiple patterns.

One of the main advantages of the proposed SRA over existing anomaly ranking methods is that our proposed SRA can distinguish simultaneously small clusters and

global outliers from majority patterns. The existing methods typically require a user to target one instead of both cases. In addition, our proposed SRA distinguishes edges of the main clusters from the core of the main clusters. Finally, SRA can be easily applied to both cases when there are multiple patterns(clusters) for the normal class and cases when only one major pattern exists.

Algorithm 1: Spectral Ranking for Abnormality (SRA)

Input: W: An n-by-n similarity matrix W.
χ: Upper bound of the ratio of anomaly
Output: $\mathbf{f}^* \in \Re^n$: A ranking vector with a larger value representing more abnormal
mFLAG : A flag indicating ranking with respect to multiple major patterns or a single major pattern
begin
Form Laplacian $L = I - D^{-1/2} W D^{-1/2}$;
Compute $\mathbf{z}^* = D^{\frac{1}{2}} \mathbf{g}_1^*$ where $\mathbf{g}_1^*$ is the 1st non-principal eigenvector for L ;
Let $\mathcal{C}_+ = \{i : \mathbf{z}_i^* \geq 0\}$ and $\mathcal{C}_- = \{i : \mathbf{z}_i^* < 0\}$;
if $min\{\frac{|\mathcal{C}_+|}{n}, \frac{|\mathcal{C}_-|}{n}\} \geq \chi$ **then**
mFLAG = 1, $\mathbf{f}^* = \max(|\mathbf{z}^*|) - |\mathbf{z}^*|$;
else if $|\mathcal{C}_+| > |\mathcal{C}_-|$ **then**
mFLAG = 0, $\mathbf{f}^* = -\mathbf{z}^*$;
else
mFLAG = 0, $\mathbf{f}^* = \mathbf{z}^*$;
end
end

6. SRA for an Auto Insurance Fraud Detection

To illustrate the effectiveness of the proposed SRA, we apply it to a fraud detection problem from auto insurance claim data which has been used for the supervised detection in Phua et al. (2004). This is the only publicly available auto insurance fraud data that we can find from the academic literature. This data set, which is provided by Angoss KnowledgeSeeker software, consists of 15420 claim instances from January 1994 to December 1996. The data set consists of 6% fraudulent labels and 94% legitimate labels, with an average of 430 claims per month. In addition, the data set has 6 ordinal features and 25 categorical features.

This insurance claim data, used in our subsequent study, consists of 31 features, all of which can be considered categorical or ordinal. Feature examples include base policy, fault, vehicle category, vehicle price (6 nominal values), month of accidents, make of the car, accidental area, holiday, sex. Intuitively, anomaly in this case should be assessed with respect to nominal value combinations. Consequently the Hamming distance and Hamming distance based kernels are reasonable similarities to use.

In Phua et al. (2004), this data set is used to assess achievable prediction quality from the supervised learning. Since labels for auto insurance claims are generally not available at the detection time, here we apply our proposed *unsupervised* SRA to the claim data set. In other words, the labels are used only for performance evaluation.

The Receiver Operating Characteristic (ROC) can be used to obtain a class skew independent measure Provost et al. (1997); Bradley (1997); Metz (1978). The ROC graph is obtained by plotting the true positive rate (number of true positives divided by m_+) against the false positive rate (number of false positives divided by m_-) as the threshold level is varied. The true positive rate (also known as sensitivity) is one evaluation criterion and the false positive rate (or one minus specificity) is the second evaluation criterion.

The ROC curves (or true-positive and false-positive frontiers) depict the trade-off between the two criteria, benefits (true positive) and costs (false positives), for different choices of the threshold. Thus it does not depend on a priori knowledge to combine the two objectives into one. A ROC curve that dominates another provides a better solution at any cost point and it corresponds to a higher area under the curve (AUC). In addition, AUC yields the probability that the generated ranking places a positive class sample above a negative class sample when the positive sample is randomly drawn from the positive class and the negative class sample is drawn randomly from the random class respectively DeLong et al. (1988). Thus, the ROC curves and AUC can be used as a criteria to measure how well an algorithm performs on certain data sets. In the subsequent section, we will use AUC and ROC curves to compare different methods. The target class is the rare class.

When performing unsupervised fraud detection on this data, we recall two major challenges which have been briefly mentioned in previous sections. Firstly, most of the features in this dataset are categorical or ordinal. Secondly, unlike common anomaly detection problems, the claim data forms multiple patterns. Consequently a single cluster based global outlier detection method generally produces unsatisfactory results.

For comparison, we include performance of two existing popular unsupervised outlier detection methods, one-class SVM (OC-SVM) and Local Outlier Factor (LOF). In addition, we train supervised Random Forest (RF) on the full dataset. The training accuracy of RF can then be used as an upper bound for evaluations of unsupervised learning methods. For the Random Forest (RF) computational results, the number of trees built is 500 and the number of features used for building each tree is 6. The resultant votes are used as the decision value. Since $|\mathcal{C}_+| \approx |\mathcal{C}_-|$ for all the similarities we have experimented with, we always report ranking with regard to multiple patterns (mFLAG=1) for this auto insurance fraud detection problem. Table 1 summarizes the AUCs from different unsupervised methods using different similarities.

Next we present more detailed discussions on the comparison in terms of ROCs for different kernels and different methods. Unless otherwise noted, we always include performance of the supervised RF to show the upper bound.

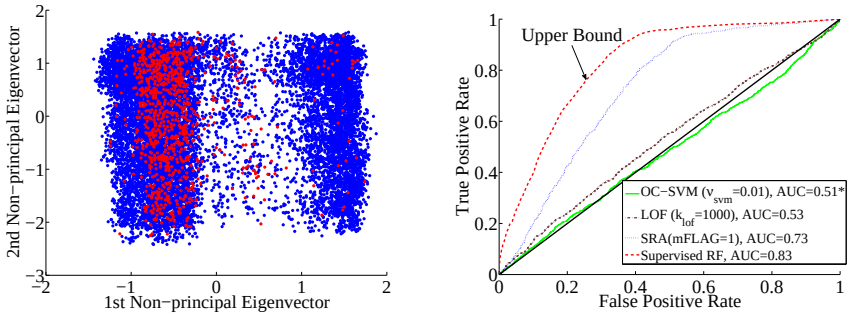
Comparisons with LOF and OC-SVM for Overlapping Similarity

We first compare performance of SRA with that of LOF, OC-SVM on overlapping similarity. We note that, in contrast to SRA which does not require any parameter itself,

Table 1: Summary of the AUCs for the Auto Insurance Fraud Detection Data Set

Automobile Fraud Detection Data Set										
Method			OS	AGK				HDK		DISC
				β				λ		
				10	100	1000	3000	0.5	0.8	
LOF	k_{svm}	10	0.53	0.5	0.52	0.58	0.64	0.52	0.53	0.55
		100	0.51	0.51*	0.54	0.58	0.67	0.51	0.52	0.57
		500	0.53	0.52*	0.55	0.59	0.68	0.51	0.51	0.57
		1000	0.53	0.52*	0.53	0.59	0.69	0.5	0.5	0.56
		3000	0.5	0.58*	0.55	0.58	0.69	0.54*	0.55*	0.53
OC-SVM	ν_{svm}	0.01	0.51*	0.53*	0.51*	0.54	0.59	0.51*	0.52*	0.53*
		0.05	0.51*	0.53*	0.51*	0.55	0.59	0.52*	0.53*	0.52*
		0.1	0.51*	0.54*	0.51*	0.55	0.59	0.53*	0.54*	0.56*
SRA	mFLAG	1	**0.73**	**0.74**	**0.74**	**0.66**	**0.74**	**0.74**	**0.74**	**0.66**

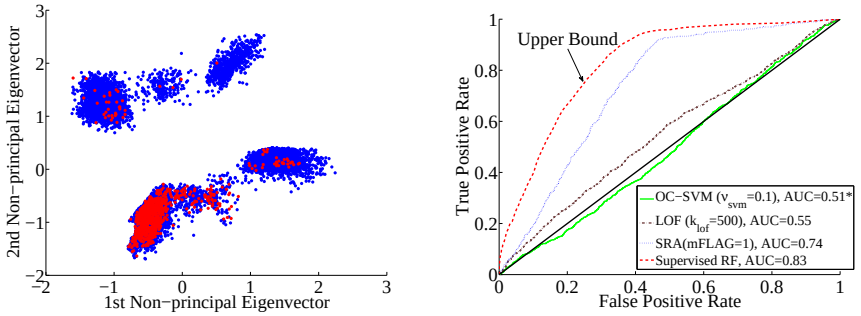
For entries marked by *, AUC reported is one minus the actual AUC (which is below 0.5).
OS:Overlapping Similarity, AGK: Adaptive Gaussian Kernel
HDK:Hamming Distance Kernel, DISC: DISC Similarity



(a) Visualization of $\mathbf{z}_1^* = D^{1/2}\mathbf{g}_1^*$ and $\mathbf{z}_2 = D^{1/2}\mathbf{g}_2^*$

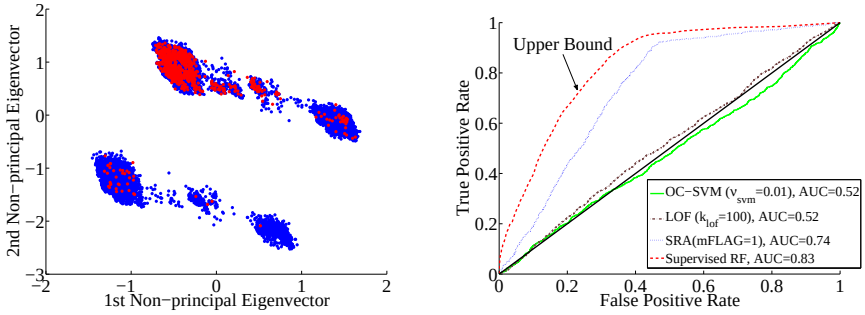
(b) ROC Curves

Figure 4: Comparing SRA, LOF, and OC-SVM Based on the Overlapping Similarity. SRA clearly dominates LOF and OC-SVM.



(a) Visualization of $\mathbf{z}_1^* = D^{1/2}\mathbf{g}_1^*$ and $\mathbf{z}_2^* = D^{1/2}\mathbf{g}_2^*$ (b) ROC Curves

Figure 5: Comparisons Based on an Adaptive Gaussian Kernel with $\beta = 100$



(a) Visualization of $\mathbf{z}_1^* = D^{1/2}\mathbf{g}_1^*$ and $\mathbf{z}_2^* = D^{1/2}\mathbf{g}_2^*$ (b) ROC Curves

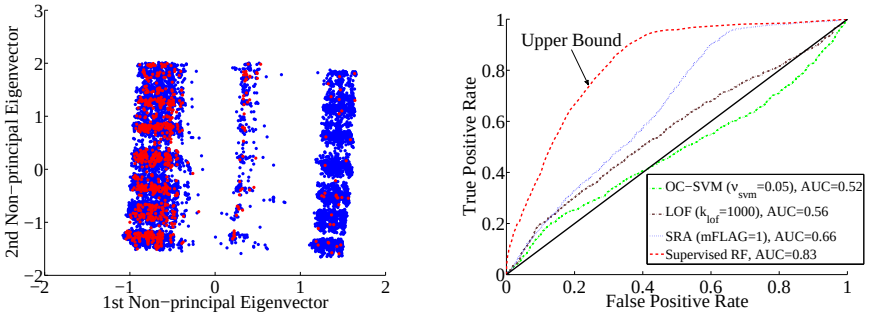
Figure 6: Comparisons Based on a Hamming Distance Kernel with $\lambda = 0.8$

LOF requires an addition parameter for the number of neighborhood n_{LOF} and OC-SVM requires an additional width parameter μ_{SVM}. This can present more challenges for unsupervised learning since there is no mechanism to decide how to choose their values.

Figure 4 presents ROC curves for (supervised) RF, SRA, LOF, and OC-SVM, where latter three used overlapping similarity. We emphasize that the only parameter required by SRA is the upper bound on the anonymous rate, which we believe that it is reasonable to expect a crude approximation at least in practice. For the auto fraud insurance data set used here, since $|\mathcal{C}^+|$ and $|\mathcal{C}^-|$ are always roughly balanced, we always choose to report rankings with multiple patterns and the choice of χ is practically irrelevant. For LOF and OC-SVM however, parameter choices are important and we report the optimal AUCs among different parameters we have experimented with.

Comparisons Using Adaptive Gaussian Kernel with Weighted Hamming Distance.

Figure 5 shows ROC curves for SRA, LOF, OC-SVM achieved with the Adaptive Gaussian Kernel with the weighted Hamming distance and neighborhood size $\beta = 100$.



(a) Visualization of $\mathbf{z}_1^* = D^{1/2}\mathbf{g}_1^*$ and $\mathbf{z}_2^* = D^{1/2}\mathbf{g}_2^*$

(b) ROC curves

Figure 7: Comparisons Based on DISC Similarity

We observe that AUC for each of SRA, LOF and OC-SVM is improved. We conjecture that the improvement comes from the facts that the weighted hamming distance is a better distance measures than hamming distance since information on the number of distinct value in each feature is also included. From Figure 5, we also observe that, using the adaptive Gaussian kernel, clusters are also more distinct comparing to the overlapping kernel.

Comparisons Using the Hamming Distance Kernel

Hamming distance kernel is defined based on the overlapping similarity measure. However, as discussed previously, a Hamming distance kernel can capture more information than the simple Hamming distance (overlapping similarity). This is because that, while Hamming distance directly compare nominal values of x and y to determine the number of different values, In contrast, a Hamming distance kernel assesses the similarity between x and y in referencing to all possible nominal value combinations of categorial attributes. Indeed, from Figure 6 which compares performance of SRA, LOF and OC-SVM using a hamming distance kernel, we observe a complex cluster structure, even though SRA achieves a similar 0.74 AUC.

Comparisons Using DISC Similarity

We have also experimented with a few distribution driven similarity measures, e.g., Lin similarity Boriah et al. (2008), and the results are consistently worse (less than 0.6 AUC) in comparison to overlapping, adaptive gaussian and hamming distance kernel. The only exception is that of the DISC similarity measure. Figure 7 compares AUCs of SRA, LOF, OC-SVM using DISC similarity. We observe here that a different cluster structure is revealed using DISC. Once again, we observe that SRA significantly dominates LOF and OC-SVM.

Understanding Cluster Structure: Further Validation of SRA Ranking

To further justify the SRA ranking generated, we seek to understand the cluster structure revealed. Specifically we investigate significant features under which highly

ranked claims are different from the majority. If deviation from the majority is based on features which are expected to lead to suspiciousness, this provides supports for the generated ranking. As an example, we consider clusters identified using Hamming kernel with $\lambda = 0.8$.

Figure 8 is identical to Subplot (a) in Figure 6, except that clusters formed in the space of the 1st and 2nd non-principal eigenvectors are depicted with different colors. We report the fraud ratio of each cluster in Table 2. It can be seen from the plot and the table that 92% of the fraudulent cases actually reside in the regions colored in black, brown and purple. These are also the regions that have relatively high anomaly score from SRA. This intrigues us to further understand each cluster, especially the one with the highest fraud ratio.

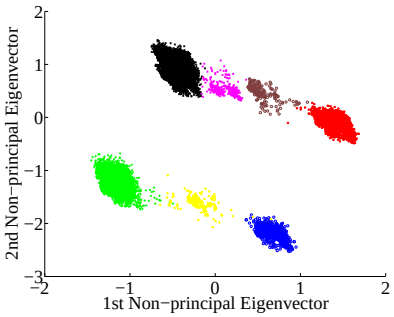


Figure 8: Clusters With Color

n_I: # instance n_L: # legal cases
n_F: # fraud χ_f: fraud ratio

Cluster	n_I	n_L	n_F	χ_f
1 (blue)	723	722	1	0.0014
2 (green)	3261	3228	33	0.0101
3 (red)	4266	4231	35	0.0082
4 (black)	6423	5622	761	0.1185
5 (yellow)	206	203	3	0.0146
6 (purple)	340	294	46	0.1353
7 (brown)	201	157	44	0.2189

Table 2: Cluster Summary Information

For this purpose, we build an unpruned standard CART (classification and regression tree) to determine the decision rule for each cluster. The training labels for CART are the cluster labels we manually identify from the 1st and 2nd non-principal eigenvectors and the training data is the whole data set. Figure 9 describes the computed CART tree.

We can discover several rules from Figure 9 for the clusters (colored in black, brown and purple) for which SRA has assigned high ranks:

- If the insurance policy is for collision or all perils and it is the policy holder who causes the accident(policy holder at fault), the corresponding claim would belong to the clusters with high fraud ratio (colored in black, brown and purple).
- If the insurance policy is for liability and/or and it is **not** the policy holder who causes the accident(third party at fault), the corresponding claim would belong to the other clusters (colored in blue, red, yellow and green).
- Following the first rule, If the policy holder drives sports car, the corresponding claim would belong to cluster 7(colored in brown). If the policy holder drive utility car, the corresponding claim would belong to cluster 6(colored in purple). Otherwise, the corresponding claim would belong to cluster 4 (colored in black)

Indeed these rules identify cases with reasonable doubt. In addition, from training *supervised* random forest in the previous section, we have learnt that the three most

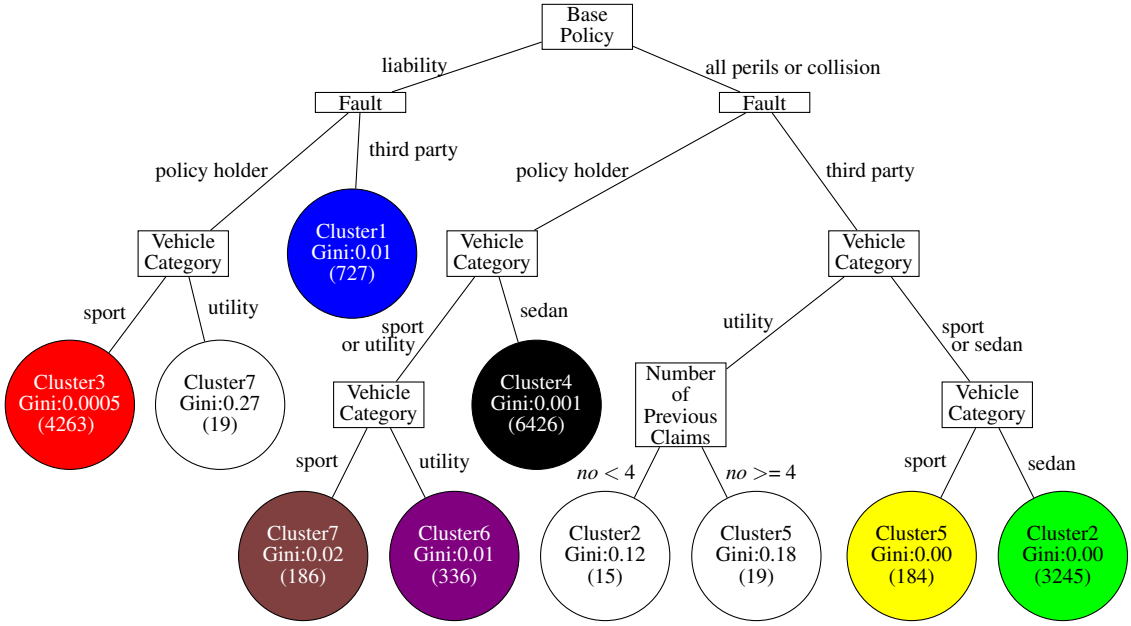


Figure 9: Decision Tree: rules leading to the colored clusters

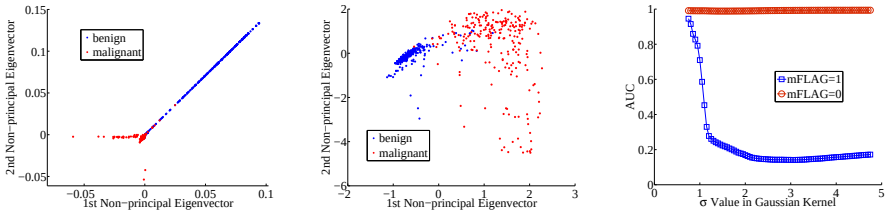
important features in classify fraudulent cases against legal cases are **base policy**, **car types** and **fault**. These three features are actually the features used in defining the clusters. This analysis supports that the ranking from SRA is meaningful and reasonable.

7. Illustrating SRA: Beyond Insurance Fraud Detection

We have so far focused on applying SRA to the challenging auto insurance fraud detection problem. While a future thorough study is necessary, using two real rare class classification data sets from UCI machine learning repository, we demonstrate here that the proposed SRA can potentially be applied for rare class classification problems when rare class can be distinguished from the major class based on attributes dependence. We also illustrate that SRA can be applied for problems with numerical features, how the choice of kernel parameters affect the results of the SRA ranking, and potential enhancement by using multiple non-principal eigenvectors.

SRA With Respect to One Major Pattern

We first consider the breast cancer Wisconsin data. The breast cancer Wisconsin data set has 699 instances, with 458 cases are benign and 241 cases are malignant. The malignant cases can be regarded as the rare class and anonymous with respect to a pattern of normal cases. All the features are numerical. We apply a Gaussian kernel on the standardized data set. The visualizations of eigenvectors ($\mathbf{z}_1^* = D^{1/2}\mathbf{g}_1^*$) are shown in Figure 10 for different σ values. In addition, Subplot (c) in Figure 10 illustrates



(a) Eigenvectors for $\sigma = 0.75$ (b) Eigenvectors for $\sigma = 6$ (c) AUC as σ varies

Figure 10: Breast Cancer: Visualization of $\mathbf{z}_1^* = D^{1/2}\mathbf{g}_1^*$ and $\mathbf{z}_2^* = D^{1/2}\mathbf{g}_2^*$

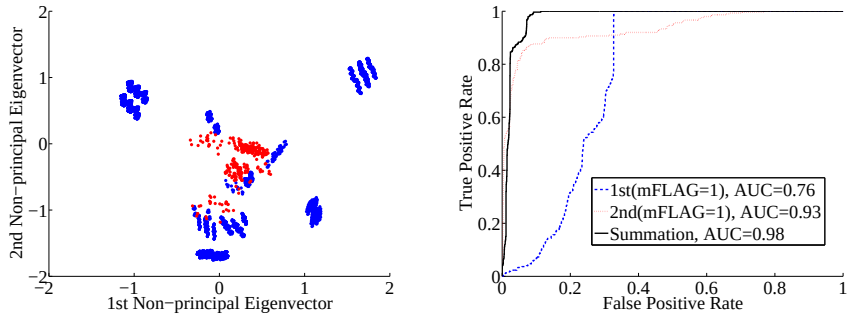
how AUC changes with σ for Gaussian kernel, with reference both to multiple main clusters (mFLAG=1) and single main cluster (mFLAG=0) cases are reported. As can be seen, for this data set, since the distribution of the 1st non-principal eigenvector is always skewed, output rank against two major patterns (mFLAG=1) will not provide satisfactory results. However, we can choose the perspective of rare class ranking easily by setting the upper bound anonymous ratio χ, e.g., setting $\chi > \frac{241}{699} \approx 34\%$ will set set mFLAG=0 and allow adoption of this perspective.

Effect of the Gaussian Kernel Parameter

For numerical data, the most often used similarity is a distance based Gaussian kernel, which has a width parameter σ. This parameter may change clustering patterns: when $\sigma \to 0$, data points become further apart and there appears to be more clusters. When $\sigma \to +\infty$, data points all become similar and data appears to form a single cluster. It can be expected, for certain data, when the parameter σ is small multiple patterns for normal class emerge while with large σ value usually only one pattern exists.

Using Multiple Non-principal Eigenvectors

Up until now, we have only utilized the 1st non-principal eigenvector when generating the ranking. In fact subsequent bi-modal eigenvectors can potentially be used to gain further performance improvement. To illustrate we consider the mushroom data set as an illustration of the potential improvement that can be achieved by considering more than one non-principal eigenvectors in generating the ranking vector. The original data set consists of 8124 instances. We create normal cases by including all 4208 edible samples and randomly selecting 300 poisonous samples. All features in this data set are categorical and we use Hamming distance kernel($\lambda = 0.5$) as similarity for SRA. Figure 11 presents visualization of eigenvectors. Using the 1st non-principal eigenvector only, the SRA ranking yields 0.76 AUC. Using the 2nd non-principal eigenvector, the corresponding ranking yields 0.93 AUC. Using the 1-norm of the both eigenvectors, i,e., summation of the absolute values of the two score vectors, AUC is increased to 0.98. We leave how to best utilize multiple eigenvectors in generating ranking to future studies.



(a) Visualization of $\mathbf{z}_1^* = D^{1/2}\mathbf{g}_1^*$ and $\mathbf{z}_2^* = D^{1/2}\mathbf{g}_2^*$ (b) ROC curves

Figure 11: Mushroom (Hamming Kernel Similarity)

8. Concluding Remarks

In this paper, we propose a spectral ranking method for anomaly (SRA). By establish a relationship between unsupervised SVM optimization and spectral optimization formulations, we observe that the spectral optimization can be considered as an approximation (relaxation) of the unsupervised SVM. For example, the absolute value of a non-principal eigenvector is a measure of strength of support in a bi-clustering and a non-principal eigenvector can be regarded as a ranking vector, which is an approximation to the optimal hypothesis for a bi-class unsupervised SVM evaluated at the given data instances. Based on this information in the eigenvector, a data instance is more likely to be an anomaly if its magnitude is smaller. Furthermore, we allow a choice of the reference in the assessment of anomaly ranking. If the minority cluster class does not have a sufficient mass, one can choose to assess anomaly likelihood with respect to a single majority class and ranking is generated suitably with this view. Otherwise, anomaly is assessed with two main clusters.

As an illustration the proposed SRA, we consider the challenging auto fraud insurance detection problem based on a real claim data set. Since obtaining such labels are time consuming, costly, and error prone in real applications, we model the problem as unsupervised learning and ignore labels when generating ranking using SRA, even though fraud labels are given for this particular data set. Since data attributes are categorical, we assess anomaly in nominal value combinations which lead to suspiciousness of the claim. We choose Hamming distance and Hamming distance based kernels in generating spectral ranking for this data set. SRA yields an impressive 0.74 AUC, which is close to 0.83 AUC generated by the supervised RF. We also illustrate unsupervised ranking performance on a couple real data problems which demonstrate the potential and possible further enhancement of the proposed SRA.

References

Ai, J., P. L. Brockett, and L. L. Golden (2009). Assessing consumer fraud risk in insurance claims: An unsupervised learning technique using discrete and continuous

predictor variables. *North American Actuarial Journal* 13(4), 438–458.

Ai, J., P. L. Brockett, L. L. Golden, and M. Guillén (2012). A robust unsupervised method for fraud rate estimation. *Journal of Risk and Insurance* .

Boriah, S., V. Chandola, and V. Kumar (2008). Similarity measures for categorical data: A comparative evaluation. In *SDM*, pp. 243–254. SIAM. ISBN 978-0-89871-654-2, 978-1-61197-278-8.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30, 1145–1159.

Brockett, P. L., R. A. Derrig, L. L. Golden, A. Levine, and M. Alpert (2002). Fraud classification using principal component analysis of ridits. *Journal of Risk and Insurance* 69(3), 341–371.

Brockett, P. L., X. Xia, and R. A. Derrig (1998). Using kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. *Journal of Risk and Insurance* pp. 245–274.

Couto, J. (2005). Kernel k-means for categorical data. In *Advances in Intelligent Data Analysis VI*, pp. 46–56. Springer.

David, G. and A. Averbuch (2012). Spectralcat: Categorical spectral clustering of numerical and nominal data. *Pattern Recognition* 45(1), 416–433.

DeLong, E. R., D. M. DeLong, and D. L. Clarke-Pearson (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* 44(3), 837–845.

Derrig, R. A. (2002). Insurance fraud. *Journal of Risk and Insurance* 69(3), 271–287.

Desai, A., H. Singh, and V. Pudi (2011). Disc: data-intensive similarity measure for categorical data. In *Advances in Knowledge Discovery and Data Mining*, pp. 469–481. Springer.

Gärtner, T., Q. V. Le, and A. J. Smola (2006). A short tour of kernel methods for graphs. Technical report, NICTA, Australia, Canberra.

Gretton, A., O. Bousquet, A. J. Smola, and B. Schölkopf (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In S. Jain, H. U. Simon, and E. Tomita, eds., *Proceedings of the International Conference on Algorithmic Learning Theory*, pp. 63–77. Springer-Verlag.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning*. Springer.

Herbrich, R., T. Graepel, and K. Obermayer (2000). *Large Margin Rank Boundaries for Ordinal Regression*. MIT Press.

Lodhi, H., C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins (2002). Text classification using string kernels. *The Journal of Machine Learning Research* 2, 419–444.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in nuclear medicine* 8(4), 283–298.

Ng, A. Y., M. I. Jordan, Y. Weiss, et al. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* 2, 849–856.

Phua, C., D. Alahakoon, and V. Lee (2004). Minority report in fraud detection: classification of skewed data. *ACM SIGKDD Explorations Newsletter* 6(1), 50–59.

Phua, C., V. Lee, K. Smith, and R. Gayler (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119* .

Provost, F., T. Fawcett, and R. Kohavi (1997). The case against accuracy estimation for comparing induction algorithms. In *In Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 445–453.

Shi, J. and J. Malik (2000). Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22(8), 888–905.

Song, L., A. Smola, A. Gretton, J. Bedo, and K. Borgwardt (2012). Feature selection via dependence maximization. *Journal of Machine Learning Research* 13, 1393–1434.

Tayal, A. D., T. F. Coleman, and Y. Li (2013a). Bounding the difference between rankrc and ranksvm and an application to multi-level rare class kernel ranking. *Journal on Machine Learning Research* submitted.

Tayal, A. D., T. F. Coleman, and Y. Li (2013b). Rankrc: Large-scale nonlinear rare class ranking. *IEEE Transactions on Knowkedge and Data Engineering* submitted.

Tennyson, S. and P. Salsas-Forn (2002). Claims auditing in automobile insurance: fraud detection and deterrence objectives. *Journal of Risk and Insurance* 69(3), 289–308.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing* 17(4), 395–416.

Appendix A. A Nonconvex Relaxation of the Unsupervised SVM

Here we present a nonconvex relaxation for the unsupervised SVM.

We start with the supervised SVM formulation. Assume that $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)\}$ is a set of n training examples, where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$, $y_i \in \{+1, -1\}$. Let $\phi : \mathcal{D} \to \mathcal{F}$ be a non-linear map from the input space $\mathcal{D}$ to a (potentially infinite dimensional) feature space $\mathcal{F}$ derived from feature inputs. Let $K(\mathbf{x}, \mathbf{x}') \equiv \phi(\mathbf{x})^T \phi(\mathbf{x}')$ be the corresponding kernel, where $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$.

Given labels y, SVM solves

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} c_i \xi_i, \\ \text{subjecto to} \quad & y_i \left(\mathbf{w}^T \phi(\mathbf{x}_i) + b\right) \geq 1 - \xi_i, \quad i = 1, ..., n, \\ & \xi_i \geq 0, \quad i = 1, ..., n, \end{aligned} \tag{A.1}$$

where ξ_i's are slack variables when training instances are not linearly separable. The regularization weight $C = (c_1, \cdots, c_n) \geq 0$ is a penalty, associated with margin violations, which determines the trade-off between model accuracy and complexity.

The optimal decision function $f(\mathbf{x})$ then has the form

$$f(\mathbf{x}) = \left(\sum_{j=1}^{n} y_j \alpha_j K(\mathbf{x}, \mathbf{x}_j) + b \right) \tag{A.2}$$

where α is computed by solving the SVM below

$$\begin{aligned} \min_{\mathbf{u},b,\xi} \quad & \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{n} c_i \xi_i, \\ \text{subject to} \quad & y_i \left(\sum_{j=1}^{n} y_j \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) + b \right) \geq 1 - \xi_i, \quad i = 1, ..., n, \\ & \xi_i \geq 0, \quad i = 1, ..., n, \end{aligned} \tag{A.3}$$

Note that $\{\mathbf{x}_i : \alpha_i \neq 0\}$ are the support vectors for the classification problem.

For unsupervised SVM learning, the objective becomes determining the optimal labels based on margin maximization SVM. The basic idea is to optimally choose the labels $\mathbf{y}$ so that the corresponding margin maximization optimal hypothesis yields the minimum SVM objective. In other words, unsupervised SVM solves the following nested minimization problem

$$\min_{y_i \in \{\pm 1\}} \left\{ \min_{w,\xi,b, y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \xi_i \geq 0} \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} c_i \xi_i \right\} \tag{A.4}$$

Replacing the inner optimization problem by its dual, unsupervised SVM (A.4) is

equivalent to the following minmax problem:

$$\min_{y_i\in\{\pm 1\}} \max_{\substack{0\le\alpha\le C\\ y^T\alpha=0}} \quad -\frac{1}{2}\alpha^T YKY\alpha+\mathbf{e}^T\alpha \tag{A.5}$$

where $Y = \text{DIAG}(y)$.
For any given y, we can introduce the following transformation:

$$\mathbf{z}_i = \alpha_i \cdot y_i, \qquad\qquad i = 1,\cdots,n$$

For any $\alpha_i \neq 0$, we have

$$y_i = \text{sign}(z_i), \quad i = 1,\cdots,n \tag{A.6}$$

In addition,

$$\alpha^T YKY\alpha = \mathbf{z}^T K\mathbf{z}, \quad \mathbf{e}^T\alpha = \mathbf{e}^T|\mathbf{z}|, \quad \text{and} \quad y^T\alpha = \mathbf{e}^T\mathbf{z}$$

Therefore, for any given y, the inner optimization problem in (A.5) is equivalent to

$$\begin{aligned} \max_{\mathbf{z}} \quad & \mathbf{e}^T|\mathbf{z}| - \frac{1}{2}\mathbf{z}^T K\mathbf{z} \\ \text{subject to} \quad & \mathbf{e}^T\mathbf{z} = 0, \\ & |\mathbf{z}| \le C \end{aligned} \tag{A.7}$$

We note that $\mathbf{e}^T|\mathbf{z}|$ is convex and (A.7) has many local maximizers.

Now assume that K is positive definite in the space $\{\mathbf{z} : \mathbf{e}^T\mathbf{z} = 0\}$. Then the local minimizers of

$$\begin{aligned} \min_{\mathbf{z}} \quad & -\frac{1}{2}\mathbf{z}^T K\mathbf{z} \\ \text{subject to} \quad & \mathbf{e}^T\mathbf{z} = 0, \\ & |\mathbf{z}| \le C \end{aligned} \tag{A.8}$$

are at the boundary of $|\mathbf{z}| \le C$. If all local maximizers of (A.8) have the same value for the term $\mathbf{e}^T|\mathbf{z}|$, then unsupervised SVM (A.5) is equivalent to (A.8). Note that (A.8) remains an NP-hard problem.

Theorem 2. *Suppose that K is symmetric positive definite. Let (α^*, y^*) be a solution to the unsupervised SVM* (A.5)*. Let $\mathbf{z}^*$ solves the relaxation problem* (A.8)*. Assume that the solution to*

$$\max_{\substack{0\le\alpha\le C\\ y^T\alpha=0}} -\frac{1}{2}\alpha^T Y_{\mathbf{z}}^* KY_{\mathbf{z}}^*\alpha+\mathbf{e}^T\alpha \tag{A.9}$$

is achieved at the vertex of $\mathcal{F}_\alpha = \{\alpha : 0 < \alpha \le C\}$, where $Y_{\mathbf{z}}^* = \text{DIAG}(y_{\mathbf{z}}^*)$. If $\mathbf{e}^T\alpha^* = \mathbf{e}^T\alpha_{\mathbf{z}}^*$, then $\alpha_{\mathbf{z}}^* = |\mathbf{z}^*|$ and $y_{\mathbf{z}}^* = \text{SIGN}(\mathbf{z}^*)$ solves the unsupervised SVM (A.5)*.*

PROOF. Since K is a symmetric positive definite (SPD), both Y^*KY^* and $Y_{\mathbf{z}}^*KY_{\mathbf{z}}^*$ are SPD. Hence (A.9) has a unique solution which is the unique vertex of $\mathcal{F}_\alpha$.

Since K is positive definite, a solution $\mathbf{z}^*$ to (A.8) must be at a vertex of the region $\mathcal{F}_{\mathbf{z}} = \{\mathbf{z} : \mathbf{e}^T\mathbf{z} = 0 \ |\mathbf{z}| \leq C\}$. Hence $\alpha^*_{\mathbf{z}}$ satisfies ${y^*_{\mathbf{z}}}^T\alpha = 0$ and is at the unique vertex of $\mathcal{F}_\alpha = \{\alpha : \ 0 < \alpha \leq C\}$.

Following the assumption, the solution to (A.9) satisfies $\alpha^T y^*_{\mathbf{z}} = 0$ and is at the unique vertex of $\mathcal{F}_\alpha = \{\alpha : \ 0 < \alpha \leq C\}$. This implies that $\alpha^*_{\mathbf{z}}$ solves (A.9).

Since (α^*, y^*) solves the unsupervised SVM (A.4), we have

$$\left(-\frac{1}{2}{\alpha^*}^T Y^* K Y^* \alpha^* + \mathbf{e}^T\alpha^*\right) \leq \left(-\frac{1}{2}{\alpha^*_{\mathbf{z}}}^T Y^*_{\mathbf{z}} K Y^*_{\mathbf{z}} \alpha^*_{\mathbf{z}} + \mathbf{e}^T\alpha^*_{\mathbf{z}}\right) \tag{A.10}$$

where $Y^* = \text{DIAG}(y^*)$. Following the assumption $e^T\alpha^* = e^T\alpha^*_{\mathbf{z}}$, we have

$$-\frac{1}{2}{\alpha^*}^T Y^* K Y^* \alpha^* \leq -\frac{1}{2}{\alpha^*_{\mathbf{z}}}^T Y^*_{\mathbf{z}} K Y^*_{\mathbf{z}} \alpha^*_{\mathbf{z}}. \tag{A.11}$$

It is clear that $Y^*\alpha^*$ is a feasible point for the relaxation problem (A.8). Since z^* solves (A.8),

$$-\frac{1}{2}{\alpha^*_{\mathbf{z}}}^T Y^*_{\mathbf{z}} K Y^*_{\mathbf{z}} \alpha^*_{\mathbf{z}} \leq -\frac{1}{2}{\alpha^*}^T Y^* K Y^* \alpha^*. \tag{A.12}$$

From (A.11), (A.12), and $\mathbf{e}^T\alpha^* = \mathbf{e}^T\alpha^*_{\mathbf{z}}$, we conclude that $\alpha^*_{\mathbf{z}} = |\mathbf{z}^*|$ and $y^*_{\mathbf{z}} = \text{SIGN}(\mathbf{z}^*)$ solves the unsupervised SVM (A.5). This completes the proof.