Hypergraph-Based Anomaly Detection of High-Dimensional Co-Occurrences

Jorge Silva, *Member*, *IEEE*, and
Rebecca Willett, *Member*, *IEEE*

Abstract—This paper addresses the problem of detecting anomalous multivariate co-occurrences using a limited number of unlabeled training observations. A novel method based on using a hypergraph representation of the data is proposed to deal with this very high-dimensional problem. Hypergraphs constitute an important extension of graphs that allow edges to connect more than two vertices simultaneously. A variational Expectation-Maximization algorithm for detecting anomalies directly on the hypergraph domain without any feature selection or dimensionality reduction is presented. The resulting estimate can be used to calculate a measure of anomalousness based on the false-discovery rate. The algorithm has $O(np)$ computational complexity, where n is the number of training observations and p is the number of potential participants in each co-occurrence event. This efficiency makes the method ideally suited for very high-dimensional settings and requires no tuning, bandwidth, or regularization parameters. The proposed approach is validated on both high-dimensional synthetic data and the Enron e-mail database, where $p > 75,000$, and it is shown that it can outperform other state-of-the-art methods.

Index Terms—Anomaly detection, co-occurrence analysis, unsupervised learning, variational methods, social networks.

1 Introduction

A wide variety of complex systems can be characterized and analyzed using high-dimensional co-occurrences. Co-occurrence data are collected by noting all entities participating in each observed event. For example, each meeting of a group of people in a social network can be considered a co-occurrence [1]. In the context of computer vision, co-occurrences of metadata tags are used to recognize scenes and objects [2], [3]. Co-occurrence data are also widely used in recommender systems and search engine optimization [4], [5], [6] and recently have been used to infer the structure of computer networks [7]. This paper addresses the problem of detecting anomalous co-occurrences based on unlabeled training observations of both "nominal" and anomalous co-occurrences and annotating each observation with a measure of its anomalousness.

Robust statistical analysis of such data is a significant challenge. Consider the following social network example: p people in a social network are being observed, and each time a subset of these people meet, we record the meeting as a co-occurrence event. Using n such observations as training data, we wish to learn a rule for determining whether any future meetings are anomalous. We do not know a priori the distribution of typical meetings or even how many anomalies were present in our n training samples. In addition, it is desirable to have some mechanism for controlling the number of false alarms. Furthermore, there are 2^p conceivable meetings, and n may be very small compared to the size of this domain, resulting in computational and statistical challenges [8], [9]. Parameter tuning via cross validation or related methods can be unstable when n is relatively small. Finally, it is important to have performance bounds and convergence guarantees.

Most existing co-occurrence data analysis methods are limited to pairwise interactions, storing co-occurrences in a matrix, and applying graph-theoretic tools [2], [3], [10]. This is particularly common when analyzing computer, social, and biological networks. However, co-occurrence matrices and graphs are not sufficiently rich to encode potentially critical information about *ensembles* of co-occurring events, in which more that two entities may be participating in each event.

These challenges call for a new paradigm in co-occurrence anomaly detection, and this paper proposes an approach based on *hypergraphs*. Hypergraphs [11] are generalizations of graphs, where the notion of an edge is generalized to that of a *hyperedge*, which may connect more than two vertices. Assume that there exist p entities that can participate in a given event and that we observe n co-occurrence records. Each of these co-occurrences can be represented as a hyperedge in the hypergraph. This paper addresses the problem of detecting anomalous co-occurrences, with special emphasis on the case when p is large relative to n and when p may be in the tens of thousands, *without intermediate feature selection or dimensionality reduction*. The proposed method has several key features:

- The density over the hyperedges can be estimated without evaluating it at all 2^p possible hyperedges.
- The algorithm adapts to an unknown fraction of anomalies contaminating the training data.
- No parameter tuning or bandwidth selection is necessary.
- The computational complexity is $O(np)$.
- The algorithm is implementable and effective even when p is in the tens of thousands.
- The false-discovery rate associated with any candidate anomalous collection of hyperedges can be quickly and accurately calculated.

1.1 Related Work

In anomaly detection, we are interested in identifying sets of hyperedges that have low probability mass, when the number of observations is small and even *evaluating* the probability mass function (pmf) at every possible hyperedge is computationally infeasible. Anomaly detection in general has been addressed using a variety of techniques. In the context of network intrusion detection [12], it is used when the possible types of intrusion are unknown. Ye and Chen [13] assume that all observations are realizations of a Gaussian random variable and then detect anomalies by computing chi-square statistics by inverting the sample covariance matrix, which, as they note, can be difficult. More thorough surveys of anomaly detection in networks, with an emphasis on network security, may be found in [12], [14], [15].

When it is known that the training data are not contaminated with any anomalies, the one-class SVM (OCSVM) can be an effective tool in many contexts [16], [17]. However, it is very sensitive to the choice of kernel and bandwidth parameter, which can be time consuming to learn using cross-validation techniques. Also, its best case computational complexity is $O(n^2p)$. Another common approach is to use kernel density estimation (KDE) to estimate the underlying distribution of the observations and then threshold it. For co-occurrence data, the Aitchison-Aitken (AA) kernel based on the Hamming distance has been proposed in [18]. This approach, however, has several disadvantages. First, as in the OCSVM, the computational complexity is, at best, $O(n^2p)$, and estimating the bandwidth is computationally expensive. Second, it is not clear how to choose the best threshold or how that threshold translates into false-alarm control. Moreover, the error performance of KDE can be poor in the big-p small-n setting [19]. A variant of a KDE-based

- *The authors are with the Department of Electrical and Computer Engineering, Duke University, Box 90291, Durham, NC 27708.*
E-mail: {jg.silva, willett}@duke.edu.

Manuscript received 25 Oct. 2007; revised 29 May 2008; accepted 3 Sept. 2008; published online 16 Sept. 2008.
Recommended for acceptance by M. Pelillo.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2007-10-0726.
Digital Object Identifier no. 10.1109/TPAMI.2008.232.

0162-8828/09/$25.00 © 2009 IEEE Published by the IEEE Computer Society

approach (in a continuous domain) that does account for contamination in the data set is proposed in [20].

Another type of approach, based on graph-theoretic results, can be found in [21], where it is shown that a K-point minimum spanning tree (K-MST) can be used to estimate the nonanomalous training samples. Once such a tree has been extracted from the data using a greedy algorithm (exact computation is computationally intractable), vertices outside of the K-MST can be considered anomalies. A more computationally efficient variant of this approach, called a leave-one-out k-nearest neighbor graph (L1O-kNNG), is proposed in [21]. One of the benefits of the L1O-kNNG is that bounds on the false-alarm level can be derived as a function of K. However, the complexity is $O(pn^2 \log n)$, and it is not clear how to account for contamination in the training data.

1.2 Organization of the Paper

We start, in Section 2, by introducing the hypergraph representation and formulating the problem of anomaly detection on the corresponding discrete space. We then provide in Section 3 a discussion of the key advantages of hypergraphs in this problem and the shortcomings of alternative setups. Next, in Section 4, we propose a variational approximation to the pmf and a resulting $O(np)$ variational Expectation-Maximization (EM) algorithm that automatically learns 1) the parameters of a finite mixture model for the distribution of the observed data and 2) the posterior probabilities of observations being anomalous. The annotation of observations with a measure indicating the degree of anomalousness, based on the positive false-discovery rate (pFDR) [22], is described in Section 5. Section 7 shows experimental results comparing the algorithm to other state-of-the-art anomaly detection algorithms.

2 Problem Formulation

Assume that there are p possible entities that can participate in each co-occurrence event. Thus, each observation is a subset of the p different entities being monitored and can be considered a hyperedge in a hypergraph. More formally, let $\mathcal{H} = \{\mathcal{V}, \mathcal{E}\}$ be a hypergraph [11] with vertex set $\mathcal{V}$ and hyperedge set $\mathcal{E}$. Each hyperedge, denoted $\boldsymbol{x} \in \mathcal{E}$, can be represented as a binary string of length p, where bits set to 1 correspond to vertices that participate in the hyperedge. In this setting, we may approximately[1] equate $\mathcal{E}$ with $\{0,1\}^p$, i.e., the vertices of the binary hypercube of dimension p. The data consists of a multiset $\mathcal{X}_n = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ containing n observed co-occurrences $\boldsymbol{x}_i$, with each $\boldsymbol{x}_i$ being an independent realization of a random variable $\boldsymbol{X} \in \mathcal{E}$.

We define g to be the pmf over $\mathcal{E}$ underlying $\boldsymbol{X}$. We assume the following mixture model:

$$g(\boldsymbol{x}) = (1-\pi)f(\boldsymbol{x}) + \pi\mu(\boldsymbol{x}), \tag{1}$$

where the overall pmf g is a mixture of the nominal distribution f and an anomalous distribution μ with proportion π. To make it possible to learn this mixture, it is necessary to make assumptions on f and μ. First, it is assumed here that μ is known and equal to the uniform distribution on $\mathcal{E}$, which can be shown to be the optimal choice in terms of maximizing the worst case detection rate among all possible anomalous distributions [20], [23]. We also assume that f admits a variational approximation, as we discuss in detail in Section 4. Finally, we assume that π is unknown.

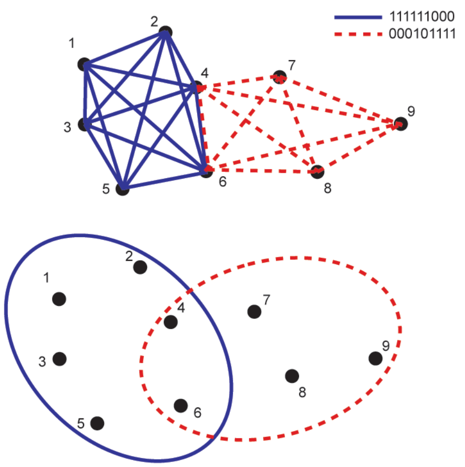


Fig. 1. Modeling two observations, 111111000 and 000101111, with $p = 9$, using (top) a graph and (bottom) a hypergraph. With the graph, representing one observation of a co-occurrence requires multiple edges. With a hypergraph, one hyperedge suffices. The hypergraph is more efficient for storing/representing observations and more informative about the real structure of the data.

A realization $\boldsymbol{x}$ of $\boldsymbol{X}$ is called an anomaly if it is more likely to have been drawn from the distribution μ instead of the nominal distribution f. We can define the anomalous set of hyperedges as

$$\mathcal{A}^* = \{\boldsymbol{x} \in \mathcal{E} : (1-\pi)f(\boldsymbol{x}) < \alpha\pi\mu(\boldsymbol{x})\},$$

where α is a parameter that controls the trade-off between false positives and false negatives. Define $\eta(\boldsymbol{x}) \equiv P(\boldsymbol{X} \sim \mu | \boldsymbol{X} = \boldsymbol{x}, f, \pi)$; we note that it is possible to write $\mathcal{A}^* = \{\boldsymbol{x} \in \mathcal{E} : \eta(\boldsymbol{x}) > \frac{1}{1+\alpha}\}$. Since f and π are unknown, this means that η and $\mathcal{A}^*$ are unknown as well. We use the iid data $\mathcal{X}_n$ to estimate η and, hence, $\mathcal{A}^*$.

3 Graphs versus Hypergraphs

One could consider using a graph to represent co-occurrence data by having each vertex represent one of the p possible co-occurrence entities and using edges to connect vertices associated with observed co-occurrences. This is illustrated in Fig. 1. Note that these two co-occurrence observations require 24 graph edges. Hypergraphs, in contrast to conventional graphs, provide a more natural representation than graphs for multiple co-occurrence data of the type examined in this paper. As Fig. 1 illustrates, two observations can efficiently be represented using only two hyperedges. Furthermore, while a single co-occurrence observation can be represented as a graph, we cannot represent more than one co-occurrence without specialized substructures or significant bookkeeping efforts that basically amount to a nontrivial extension of a typical graph formulation. Even using weighted edges, where the weight is proportional to the number of observed co-occurrences, loses information. In fact, the graph data structure as described above does not encapsulate information about how often more than two vertices may co-occur simultaneously and instead reduces the data to an overly simple pairwise representation. In the context of Fig. 1, this is because the graph representation alone does not encode that vertices 1 and 2 co-occurred at the same time as vertices 3 and 5 but not at the same time as vertices 7 and 8. As a result, the usual data representation for graphs, say, in the form of an adjacency matrix, would simply not work.

In addition, the edge structure of a graph is usually represented as a $p \times p$ symmetric adjacency matrix with $\frac{p}{2}(p-1)$ distinct elements, so that even converting observations into a collection of edge weights could be enormously challenging computationally. Modifications to the standard graph, such as

[1] We say "approximately" due to the existence of prohibited hyperedges, namely, the origin, $\boldsymbol{x} = 0$, and all $\boldsymbol{x}$ within Hamming distance 1 from the origin, which correspond to "co-occurrences" with zero or one entities. The impact of this precluded set becomes negligible for a large p and is omitted from this paper for simplicity of presentation.

edge-labeled multigraphs, might be devised; however, the hypergraph is a natural representation that admits effective computational methods, as described below.

4 Variational Expectation-Maximization

We choose an estimate of f from the class $\mathcal{F}$ of distributions with two key properties. First, each $\widetilde{f} \in \mathcal{F}$ can be expressed as the product of its marginals. Hence,

$$\widetilde{f}(\boldsymbol{x}) = \prod_{j=1}^{p} \widetilde{f}_j(x_j) = \prod_{j=1}^{p} \theta_j^{x_j}(1-\theta_j)^{(1-x_j)}, \tag{2}$$

where each $\widetilde{f}_j : \{0,1\} \longrightarrow \mathbb{R}_0^+$ is a bona fide univariate probability mass function that sums to one of the form $\widetilde{f}_j(x_j) = \theta_j^{x_j}(1-\theta_j)^{(1-x_j)}$ for $\theta_j \in [0,1]$, with x_j being a realization of a Bernoulli random variable X_j, that corresponds to the participation of the jth entity observed in a co-occurrence. Second, members of $\mathcal{F}$ have no uniform marginals (i.e., $\theta_j \neq 1/2$; this is a technical condition that ensures identifiability). All members of $\mathcal{F}$ are log-concave unimodal distributions.

Approximating the true f by members of $\mathcal{F}$ is an example of a *variational approximation*, which has been used in machine learning in several contexts. For example, in Bayesian networks [24], [25], such a factorization of the class-conditional densities leads to the well-known "naïve" Bayes classifier. For a thorough introduction to statistical inference in large-scale problems with variational methods, see [26]. In very recent work [27], it has been shown that linear combinations of N separable functions, of which product measures are a particular case, constitute a surprisingly rich and useful function class, even for a very small N. Setting $N = 1$ is a reasonable first approximation that has some history in machine learning, even though it restricts the estimated distributions to be unimodal.

An EM algorithm can be used for learning the finite mixture parameters f and π in (1) [28], [29]. A particular implementation has been previously derived for Bernoulli product measures [30] and can be applied in the present setting. Let $\eta_i \equiv \eta(\boldsymbol{x}_i)$; then, we have:

E-step:

$$\widehat{\eta}^{(t+1)}(\boldsymbol{x}_i) = \frac{\widehat{\pi}^{(t)}\mu(\boldsymbol{x}_i)}{(1-\widehat{\pi}^{(t)})\widehat{f}^{(t)}(\boldsymbol{x}_i) + \widehat{\pi}^{(t)}\mu(\boldsymbol{x}_i)}, \tag{3}$$

M-step:

$$\begin{aligned} \widehat{\pi}^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n} \widehat{\eta}_i^{(t+1)}, \\ \widehat{\theta}_j^{(t+1)} &= \frac{\sum_{i=1}^{n}\left(1-\widehat{\eta}_i^{(t+1)}\right)x_{i,j}}{\sum_{i=1}^{n}\left(1-\widehat{\eta}_i^{(t+1)}\right)}, \\ \widehat{f}^{(t+1)}(\boldsymbol{x}_i) &= \prod_{j=1}^{p}\left(\widehat{\theta}_j^{(t+1)}\right)^{x_{i,j}}\left(1-\widehat{\theta}_j^{(t+1)}\right)^{(1-x_{i,j})}. \end{aligned} \tag{4}$$

In the M-step, $x_{i,j}$ denotes the value of the jth bit of pattern $\mathbf{x}_i$. Estimating each of the p marginals of f in the M-step requires only the sum of n terms, leading to $O(np)$ complexity, and protection from overfitting comes as a natural consequence of the restricted class of estimates. If a hard decision is necessary, it is natural to threshold the η_i at $\frac{1}{1+\alpha}$, where α controls the trade-off between false positives and detection failures, as seen in Section 2.

5 Annotations

One of the key facets of the approach proposed in this paper is the annotation of observations. The annotations, which are scalars in the [0,1] interval, allow the observations to be ranked and provide some measure of how anomalous they appear to be under the model. Most methods in the existing literature cannot readily accomplish this task. In our case, the posterior probabilities η_i constitute one possible type of annotations. In this section, we describe an alternative.

Starting from premises similar to our own, [20] propose first learning π separately and then assigning annotations γ_i to each $\boldsymbol{x}_i$, equal to

$$\gamma_i = 1 - \mathrm{pFDR}(\mathcal{A}_i), \tag{5}$$

where pFDR is the positive False Discovery Rate [22] associated with the set $\mathcal{A}_i$, which in turn is defined as

$$\mathcal{A}_i = \{\boldsymbol{x} \in \mathcal{E} : f(\boldsymbol{x}) < f(\boldsymbol{x}_i)\}. \tag{6}$$

Note that $\mathcal{A}_i$ can be thought of as a level set (with respect to f) of low-probability hyperedges, which barely excludes $\boldsymbol{x}_i$, so if $\mathcal{A}_i$ is larger (i.e., has a larger μ-measure), that suggests that $\boldsymbol{x}_i$ is less anomalous. Further note that the $\mathcal{A}_i$s constitute a collection of nested level sets of f. To see the relationship between the $\mathcal{A}_i$s and $\mathcal{A}^*$, let $\mathcal{A}_{(k)}$ denote the kth largest $\mathcal{A}_i$ according to the μ-measure (i.e., $\mathcal{A}_{(k)}$ corresponds to the $\boldsymbol{x}_i$ with the kth largest $f(\boldsymbol{x}_i)$). Then, there exists some k^* such that

$$\mathcal{A}_{(1)} \supseteq \cdots \supseteq \cdots \supseteq \mathcal{A}_{(k^*-1)} \supseteq \mathcal{A}^* \supseteq \mathcal{A}_{(k^*)} \cdots \supseteq \mathcal{A}_{(n)}.$$

In other words, the level set $\mathcal{A}^*$ contains a nested collection of $\mathcal{A}_i$s. The value of k^* depends on α, the parameter that controls for the compromise between false alarms and detection failures, and on the mixture parameters f and π. The pFDR, for a set $\mathcal{A}$, is defined as follows:

$$\mathrm{pFDR}(\mathcal{A}) = P\{\boldsymbol{X} \sim f | \boldsymbol{X} \in \mathcal{A}\}. \tag{7}$$

Thus, if we declare observations that lie in $\mathcal{A}_i$ to be "discovered" anomalies, then $\mathrm{pFDR}(\mathcal{A}_i)$ is the probability that those observations arise from the nominal distribution f. It can be shown that $\gamma_i = \pi\mathbb{U}(\mathcal{A}_i)/\mathbb{G}(\mathcal{A}_i)$, where $\mathbb{U}(.)$ and $\mathbb{G}(.)$ refer to the probability measures associated with μ and g. Denoting the f-measure by $\mathbb{F}(.)$, we can estimate γ_i by

$$\widehat{\gamma}_i = \frac{\widehat{\pi}\widehat{\mathbb{U}}(\mathcal{A}_i)}{\widehat{\mathbb{G}}(\mathcal{A}_i)} = \frac{\widehat{\pi}\widehat{\mathbb{U}}(\mathcal{A}_i)}{(1-\widehat{\pi})\widehat{\mathbb{F}}(\mathcal{A}_i) + \widehat{\pi}\widehat{\mathbb{U}}(\mathcal{A}_i)}, \tag{8}$$

where $\widehat{\mathbb{F}}$, $\widehat{\mathbb{G}}$, and $\widehat{\mathbb{U}}$ are Monte Carlo estimates.

Most density estimation methods do not provide an "easy-to-sample" form of the pmf. Drawing samples from distributions estimated using nonparametric methods such as KDE require involved MCMC techniques whose convergence is hard to assess. In our case, however, at the conclusion of the EM algorithm, we may use $\widehat{f}^{(t+1)}$ to estimate the γ_i for $i = 1, \ldots, n$ using a very computationally efficient Monte Carlo procedure. Specifically, we sample from the fully factorized f and μ distributions—this amounts to sampling from p independent Bernoulli distributions using a standard random number generator, as suggested, e.g., in [31]—and then compute the empirical $\mathbb{F}$ and $\mathbb{U}$ measures of $\mathcal{A}_i$, for each $\boldsymbol{x}_i$. Afterward, we plug the estimates into (8), thus obtaining $\widehat{\gamma}_i$.

6 Key Advantages of Proposed Method

The proposed variational approximation has a number of key advantages:

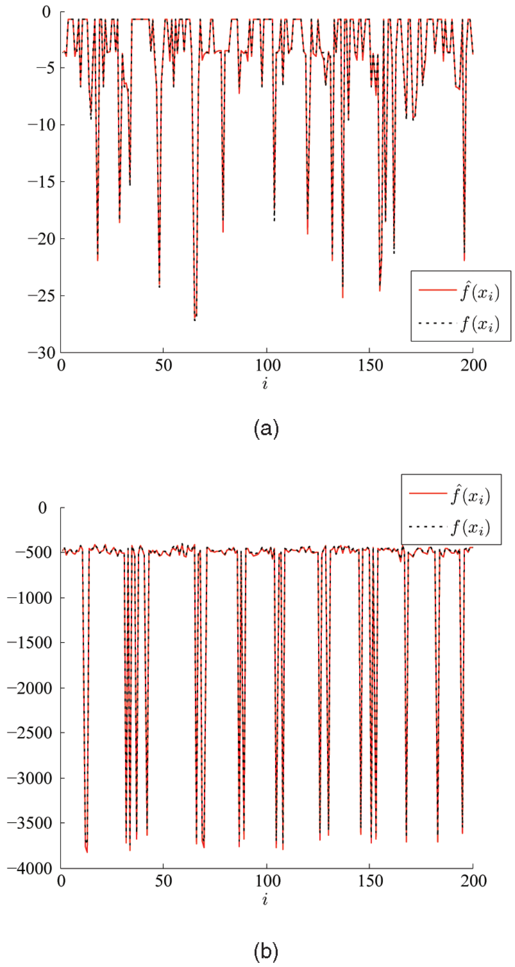


Fig. 2. Log of true pmfs and variational EM pmf estimates, with $n = 200$ and $p_H = 0.95$. (a) $p = 10$. (b) $p = 2{,}000$. The plots show $f(\boldsymbol{x}_i)$ and $\widehat{f}_n(\boldsymbol{x}_i)$, in log scale, for $\boldsymbol{x}_i$ in the test set.

- The variational approximation leads to a very computationally efficient M-step.
- The γ_is and the η_is can be computed very easily and rapidly.
- The pmf only has to be computed at the n $\boldsymbol{x}_i$ locations, rather than at all 2^p hyperedges for anomaly detection.
- Annotations and rankings of observations can be computed via Monte Carlo, which is particularly efficient since we simply draw from p independent Bernoulli distributions.
- Unlike OCSVM, the proposed method returns posterior probabilities rather than simple hard decisions.
- Unlike most KDE-based methods, a principled criterion for making a decision about each observation, based on the pFDR, is available.
- Unlike the general case of mixtures of distributions from the exponential family [32], our model is *identifiable*, i.e., there is a unique maximizer of the likelihood. Furthermore, the EM algorithm enjoys *local consistency* [33], [34] in our setting, which means that EM will reach the unique maximizer if initialized close enough. This is hard to verify for general mixtures. To the best of our knowledge, these results have not been derived before for mixtures of multivariate Bernoulli distributions. Due to space considerations, we do not include proofs; see [35] for details.

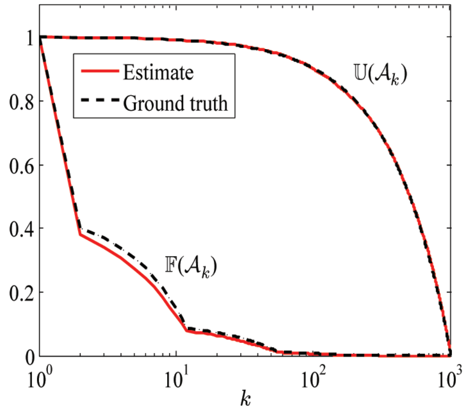


Fig. 3. True and estimated nominal $(\mathbb{F}(\mathcal{A}_k))$ and anomalous $(\mathbb{U}(\mathcal{A}_k))$ measures of nested anomalous level sets $\mathcal{A}_k$. Estimates are computed using 10,000 Monte Carlo samples of $\widehat{f}_n$ and μ for $p = 10$. Ground-truth measures are computed by evaluating f and μ on all 2^p hyperedges.

7 Experiments

In order to validate our algorithm, we have conducted experiments using both a synthetic and a real data set. The real data comes from the Enron e-mail database [36], which is publicly available at http://www.cs.cmu.edu/~enron.

7.1 Synthetic Data

For the first experiment, we created a synthetic data set consisting of a mixture of nominal and anomalous co-occurrences, distributed according to (1). The data set was split into training and test sets, each of size n. The nominal samples were generated by dividing the vertex set $\{1, \dots, p\}$ of the hypergraph into high-probability $(\{1, \dots, \tilde{p}\})$ and low-probability vertices $(\{\tilde{p}+1, \dots, p\})$, which are active with probabilities p_H and p_L, respectively. We used $p = 10$ and 2,000, with $n = 200$, $\tilde{p} = \frac{p}{2}$, $p_H = 0.95$, $p_L = 0.05$, and $\pi = 0.1$. The variational EM algorithm was initialized with $\hat{\pi}^{(0)} = \frac{1}{2}$ and $\hat{\theta}_j^{(0)} = \frac{1}{2}, \forall_j$. The pmf estimate obtained after convergence is shown in Fig. 2, which depicts the estimated and true densities, $\widehat{f}$ and f, evaluated at the test data locations. Fig. 2a corresponds to $p = 10$, and Fig. 2b corresponds to $p = 2{,}000$.

Also shown, in Fig. 3, are the measures $\mathbb{F}(\mathcal{A}_k)$ and $\mathbb{U}(\mathcal{A}_k)$ and their empirical estimates, obtained by Monte Carlo with 10,000 samples of $\widehat{f}_n$. Ground-truth probability masses were exhaustively computed for all hyperedges. This was done for $p = 10$ only since $p = 2{,}000$ is impractical. It is clear that the proposed method is highly successful in estimating both f and the measure of its level sets.

Additionally, we have evaluated the performance of our algorithm in estimating π, as a function of the sample size, for 100 Monte Carlo runs of the synthetic model just described. We have used $p = 2{,}000$ for every run. As Fig. 4 shows, the mean squared error of $\widehat{\pi}$ steadily decreases as a function of n.

For comparison, the OCSVM [16], with both the Gaussian and the AA kernels, and L1O-kNNG [21] algorithms were applied to the same data. The parameters for OCSVM are ν, which controls regularization, and γ, which is the kernel bandwidth. We used cross validation with a two-dimensional grid search to select the best ν and γ. The results for all methods, using the test set, are displayed in Fig. 5, where the first image corresponds to $p = 10$ and the second image corresponds to $p = 2{,}000$. The top plot in each image corresponds to the "ground truth," defined in this example as $y_i = I_{\{X \sim \mu\}}$. As illustrated in Fig. 5, OCSVM succeeds in detecting most of the anomalies but at the cost of a high number of false positives—less so with the AA kernel—whereas the L1O-kNNG algorithm performs better than OCSVM, achieving essentially the

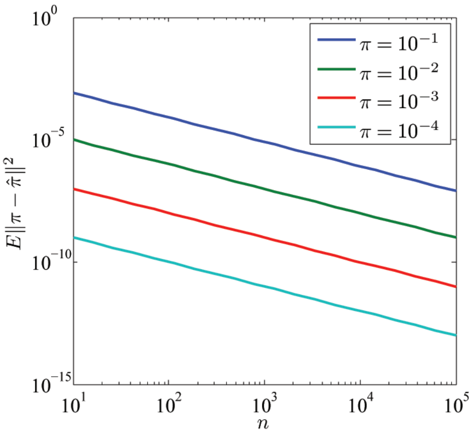


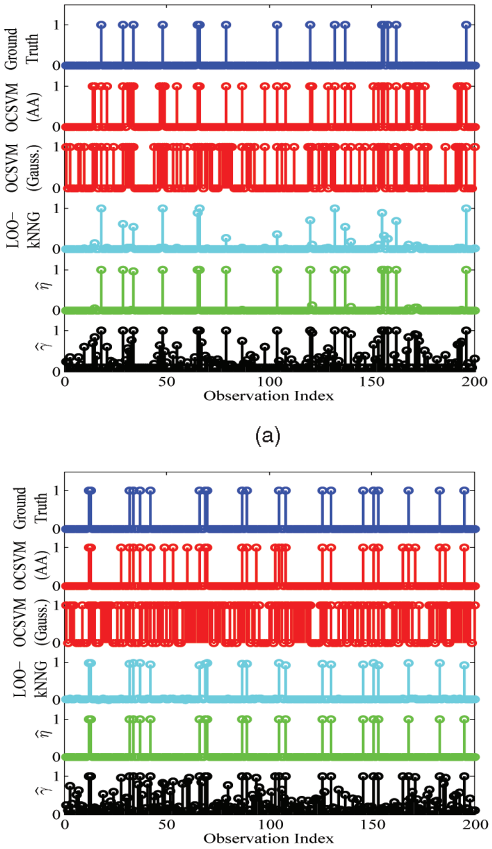
Fig. 4. Mean squared error of $\widehat{\pi}$ as a function of n, for different values of π, showing an increase in accuracy as n increases. All curves are obtained using $p = 2,000$ and averaging over 100 Monte Carlo runs.

same performance as our variational EM for $p = 2,000$ and performing slightly worse for $p = 10$. It should be taken into account that the computational complexity of L1O-kNNG is, at best, $O(pn^2 \log n)$, compared to $O(np)$ for variational EM. Also, unlike variational EM, while L1O-kNNG returns scalar "scores" in the [0, 1] interval for each test observation, they do not have a clear interpretation, particularly when the training data are contaminated as in our example.

7.2 Enron E-Mail Database

For the second experiment, we have utilized the Enron corpus, which consists of approximately 500,000 e-mails involving 151 known employees and more than 75,000 distinct addresses[2] between the years 1998 and 2002. We have organized the data in two different ways for two different analyses: by message subject and by day. For the first analysis, we examined co-occurrences of e-mail addresses that sent or received messages having the same subject line (we have trimmed the prefixes "Re:" and "Fw:", and removed pathological cases such as blank subject lines). For the second analysis, we examined co-occurrences of e-mail addresses that sent or received messages in the same day. Thus, we have records for two types of co-occurrences between users, and we have attempted to discover anomalies among these co-occurrences.

For the subject analysis, if all 75,000 users are considered, it is possible to identify approximately 130,000 distinct subject lines. Memory considerations have forced us to restrict ourselves to a subset of $p = 1,151$ users, comprised of the 151 known employees and an additional 1,000 other e-mail addresses, randomly selected. This leads to 30,841 distinct subject lines. Of these, $n = 20,000$ were used for training, and the remaining 10,841 were used for testing. The estimated nominal and anomalous pmf values for the test set are depicted in the top plot of Fig. 6, with the Hamming norm $\|\boldsymbol{x}_i\|$ shown below for comparison. Clearly, the estimate $\widehat{f}(\boldsymbol{x}_i)$ is more informative for anomaly detection than $\|\boldsymbol{x}_i\|$; in particular, the most anomalous observation according to $\widehat{f}(\boldsymbol{x}_i)$ is has a relatively small Hamming norm. The most anomalous co-occurrence we detect, at $i = 5,090$, has a subject line of "Demand Ken Lay Donate Proceeds from Enron Stock Sales." The reason why this subject is considered the most anomalous is *not* that it involves a large number of users but rather that a significant number of previously inactive addresses participated in the e-mail thread.



(a)

(b)

Fig. 5. Simulation results on the test set for (a) $p = 10$ and (b) $p = 2,000$. The horizontal axis is observation index i. Top plot: ground truth, $y_i = I_{\boldsymbol{x}_i \sim \mu}$. Second and third plots: $\widehat{y}_i$ estimated by OCSVM using the AA and Gaussian kernels, respectively. There are several false alarms (with both $p = 10$ and $p = 2,000$) and zero missed detections. Fourth plot: Anomalousness scores estimated by L1O-kNNG, which, when thresholded at 0.5, contains zero false alarms and zero missed detections with $p = 2,000$ and a number of missed detections with $p = 10$. Fifth plot: $\widehat{\eta}_i$ computed by the proposed variational EM algorithm; setting $\widehat{y}_i = I_{\{\widehat{\eta}_i > 1/2\}}$ results in zero false alarms and zero missed detections with both $p = 10$ and $p = 2,000$. Bottom plot: $\widehat{\gamma}_i$.

For the date analysis, we have used e-mail time stamps in order to record users that were active in each day, either sending or receiving e-mails. This was done for 1,177 days, starting from January 1, 1999. Of these, $n = 800$ days were randomly selected for training, and the remaining 377 were used for the test set. In this situation, we were able to utilize the full set of users, leading to a dimensionality of $p = 75,511$. The corresponding pmf estimates are illustrated in Fig. 7. The density estimate is more informative than simply $\|\boldsymbol{x}_i\|$, allowing the identification of a highly anomalous date at $i = 324$, corresponding to 1,125 days past 1 January 1999, i.e., 30 January 2002. On this date, most messages again contain the subject line "Demand Ken Lay Donate Proceeds from Enron Stock Sales." It is interesting to note that this happens to be the week of the Enron CEO's resignation and that, in the period of 27-30 January 2002, a television interview and a number of articles focusing on his personal finances had appeared in the media [37].

Also shown in Figs. 6 and 7 are the estimated η_i values. It is apparent that based on the $\widehat{\eta}_i$, a larger number of observations would be declared anomalous relative to the OCSVM, although the latter also detects the anomalies at $i = 5,090$ and $i = 324$ for the subjectwise and datewise analyses, respectively. Also, in Fig. 7, it can be seen that the frequency of high $\widehat{\eta}_i$ values increases near periods of high activity.

[2] Note that, due to the existence of aliases and multiple e-mail addresses for the same user, there are significantly fewer than 75,000 actual users. We have *not* attempted to consolidate the data set to take this into account.

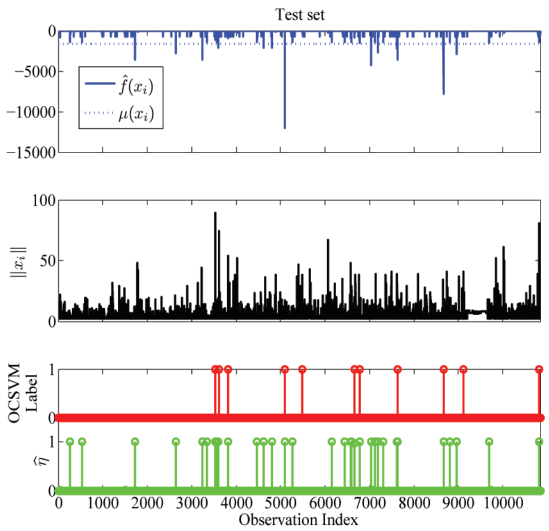


Fig. 6. Subject line analysis for the Enron data set, with $p = 1,151$. The results are shown for the test set. First plot from the top: $\widehat{f}(x_i)$, in logarithmic scale. Second plot: Hamming distance to the origin, $\|\boldsymbol{x}_i\|$, shown for comparison. Note that the most anomalous message subject, at $i = 5,090$, does not coincide with the maximum of $\|x_i\|$. The corresponding subject text is "Demand Ken Lay Donate Proceeds from Enron Stock Sales." (Bottom plot) Top: $\widehat{y}_i$ obtained using OCSVM. Bottom: Estimated $\widehat{\eta}_i$ values. Both algorithms detect the most anomalous message subject at $i = 5,090$.

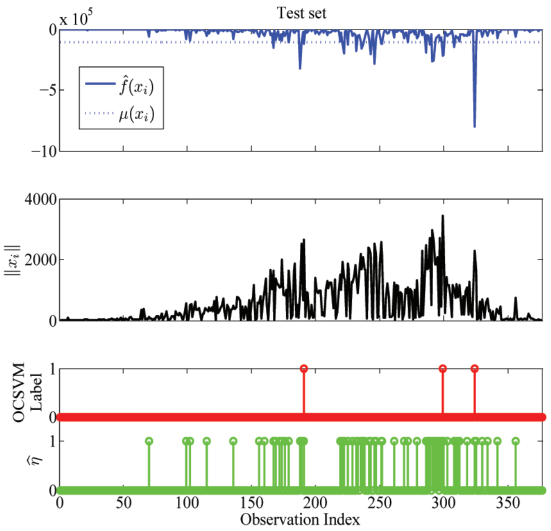


Fig. 7. Date analysis for the Enron data set, with $p = 75,511$. The results are shown for the test set. First plot from the top: $\widehat{f}(x_i)$, in logarithmic scale. Second plot: Hamming distance to the origin, $\|\boldsymbol{x}_i\|$, shown for comparison. Note that the most anomalous observation, at $i = 324$, again does not coincide with the maximum of $\|x_i\|$. The corresponding date is 1,125 days after 1 January 1999, which coincides with that of most messages bearing the subject line "Demand Ken Lay Donate Proceeds from Enron Stock Sales." Third plot: top: $\widehat{y}_i$ obtained using OCSVM. Bottom: Estimated $\widehat{\eta}_i$ values. Both algorithms detect the most anomalous day, at $i = 324$.

We do not report $\widehat{\gamma}_i$ for either situation; unlike the synthetic example above, the Enron data set exhibits very high sparsity, i.e., the vast majority of individual users have very low probability of participating in a co-occurrence. Thus, f is extremely small for all but very few hyperedges, which in turn causes the estimated measure of the anomalous set to quickly tend to zero. This drives the $\widehat{\gamma}_i$ to virtually one for many of the observations. Thus, in this case, it is more fruitful to use $\widehat{f}(\boldsymbol{x}_i)$ for visualizing the results. We note that the rank ordering given by $f(\boldsymbol{x}_i)$ necessarily coincides with that given by the true γ_i as $\mathcal{A}_i$ are nested.

7.3 Running Times

We present, in Table 1, runtimes of all tested algorithms, for both the synthetic data and the Enron corpus. Note that the entries corresponding to the L1O-kNNG for the Enron data set were left blank due to the extremely long runtimes—we aborted execution after 1 hour, when only 5 percent of the test observations had been processed. For OCSVM, we have used the well-known "libsvm" toolbox [38]. The reported times include cross validation, which was carried out using a coarser grid search than that in Section 7.1, leading to more favorable times for OCSVM. For L10-kNNG, we used the code provided by Prof. Al Hero. Table 1 shows that, for a very small sample size, OCSVM and L1O-kNNG outperform variational EM, but that quickly changes with a slightly larger n. This is in line with the known quadratic dependence on n for those two methods. In the larger Enron data set, the $O(np)$ complexity of variational EM clearly asserts itself.

8 Conclusions

We have proposed a scalable algorithm that detects anomalies on hypergraphs, with only $O(np)$ computational complexity. We model the data as a two-component mixture and learn all the parameters using EM with a variational approximation. Unlike the general mixture model case, our model is identifiable under very mild assumptions, and the proposed EM algorithm enjoys local consistency, as detailed in [35]. The algorithm allows annotations (γ_is) related to the pFDR to be computed more efficiently than alternative procedures. Furthermore, the algorithm improves upon classification-based approaches like OCSVM by providing more information than simply an all-or-nothing decision and on KDE by providing principled criteria for choosing a decision threshold. The proposed procedure has been validated on *very* high-dimensional examples and compared favorably with other state-of-the-art methods. As the results show, our method can outperform alternatives in terms of estimation error, for a useful class of distributions, while computationally scaling considerably better.

It is noteworthy that variational EM can, in some of our experiments, achieve better discriminative ability than OCSVM, even though it is well known that pmf estimation is, in many senses, a harder task than discrimination. Furthermore, OCSVM is directly attempting to minimize the probability of error, while variational EM, in contrast, takes a more indirect maximum-likelihood approach. The strong performance of variational EM compared to OCSVM is due to the combination of the following three factors: 1) OCSVM provides an estimate from a richer class than variational EM. Thus, the OCSVM suffers from a higher estimation error, even though asymptotically, with a sufficiently high n, its generalization error could eventually become smaller than that of variational EM. 2) It has been shown that for a small n, density-based classifiers can actually reach their own (less favorable in general) asymptotic error *faster*. An interesting

TABLE 1
Runtimes, in Seconds, for All Tested Algorithms

	Variational EM	OCSVM	L1O-kNNG†
Synthetic ($n = 100$, $p = 2,000$)	31.67	6.29	3.2
Synthetic ($n = 300$, $p = 2,000$)	32.70	57.21	41.17
Enron (by subject, $n = 20,000$, $p = 1,151$)	129.36	2303.22	$\gg$ 3600
Enron (by day, $n = 800$, $p = 75,511$)	503.96	1182.5	$\gg$ 3600

(†) *LIO-kNNG was too slow to run to completion with the Enron data set.*

treatment of the subject is given in [39]. 3) For a finite sample size, SVM minimizes the empirical classification error plus a regularization term. In the one-class version, the regularization term encodes a particular false-alarm probability but does not take into account the proportion of anomalies, since the training sample is assumed to only contain data from the nominal distribution.

Another interesting observation is that the pmf estimates are *better* for a higher p, as can be seen in Fig. 2. This is because the nominal and anomalous distributions are more separable in high dimensions, i.e., the measure of the set where f and μ have similar values vanishes when p increases, which means that a given observation will, with high probability, be unambiguously either anomalous or nominal.

In all of the above work, we assume binary observations, where no information is available about each entity's participation (such as a measure of correlation or co-occurrence significance). In many practical applications (*cf.* [20] and [21]), however, the measured data is continuous or multinomial. For example, we may know how active a participant was in a meeting or how many times a particular feature appears in each image in a database. Extending our methods to this setting is possible but introduces new challenges. One could maintain the variational approximation with continuous data and, in each coordinate direction, perform univariate density estimation (e.g., using KDE) to estimate f. The difficulty, however, is twofold: 1) We would lose local consistency guarantees, and 2) the complexity would be $O(n^2p)$ instead of $O(np)$ if kernels were used, making the approach less scalable. In contrast, quantizing continuous-domain data to be multinomial and then using variational EM is computationally very feasible; if we use a K-bin histogram in each coordinate direction, the complexity becomes $O(npK)$. An upper bound on the mean squared error scales like $O(K/n + 1/K^2)$, which in turn will reduce the accuracy of $\widehat{f}$ [19]. Further investigations into extending the applicability of the techniques proposed in this paper to broader settings remain an important component of future work.

ACKNOWLEDGMENTS

The authors wish to thank Professor Al Hero for providing the code for the L1O-kNNG algorithm. This work was supported by US National Science Foundation CAREER Award CCF-06-04397 and US Defense Advanced Research Projects Agency Grant HT0011-07-1-003.

REFERENCES

[1] A. Ozgur, B. Cetin, and H. Bingol, "Co-Occurrence Network of Reuters News," http://arxiv.org/abs/0712.2491, Dec. 2007.

[2] A. Globerson, G. Chechik, F. Pereira, and N. Tishby, "Euclidean Embedding of Co-Occurrence Data," *J. Machine Learning Research,* vol. 8, pp. 2265-2295, 2007.

[3] N. Jhanwar, S. Chaudhuri, G. Seetharaman, and B. Zavidovique, "Content Based Image Retrieval Using Motif Cooccurrence Matrix," *Proc. Fourth Indian Conf. Computer Vision, Graphics and Image Processing,* vol. 22, no. 14, pp. 1211-1220, 2004.

[4] E. Garcia, "Targeting Documents and Terms: Using Co-Occurrence Data, Answer Sets and Probability Theory," http://www.miislita.com/semantics/c-index-3.html, May 2008.

[5] M. Li, B. Dias, W. El-Deredy, and P.J.G. Lisboa, "A Probabilistic Model for Item-Based Recommender Systems," *Proc. ACM Int'l Conf. Recommender Systems),* 2007.

[6] H. Li and N. Abe, "Word Clustering and Disambiguation Based on Co-Occurrence Data," *Proc. 19th Int'l Conf. Computational Linguistics,* 2002.

[7] M. Rabbat, M. Figueiredo, and R. Nowak, "Network Inference from Co-Occurrences," *IEEE Trans. Information Theory,* vol. 54, no. 9, pp. 4053-4068, 2008.

[8] P.D. Hoff, A.E. Raftery, and M.S. Handcock, "Latent Space Approaches to Social Network Analysis," *J. Am. Statistical Assoc.,* vol. 97, no. 460, pp. 1090-1099, 2002.

[9] M.E.J. Newman, "The Structure and Function of Complex Networks," *SIAM Rev.,* vol. 45, pp. 167-256, 2003.

[10] T. Hofmann and J. Puzicha, "Statistical Models for Co-Occurrence Data," Technical Report AIM-1625, Massachusetts Inst. of Technology, citeseer.ist.psu.edu/article/hofmann98statistical.html, 1998.

[11] C. Berge, *Hypergraphs: Combinatorics of Finite Sets.* North Holland, 1989.

[12] W. Lee and S. Stolfo, "Data Mining Approaches for Intrusion Detection," *Proc. Seventh Usenix Security Symp.,* 1998.

[13] N. Ye and Q. Chen, "An Anomaly Detection Technique Based on a Chi-Square Statistic for Detecting Intrusions into Information Systems," *Quality and Reliability Eng. Int'l,* vol. 17, pp. 105-112, 2001.

[14] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava, "A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection," *Proc. Third SIAM Int'l Conf. Data Mining,* May 2003.

[15] T. Ahmed, B. Oreshkin, and M. Coates, "Machine Learning Approaches to Network Anomaly Detection," *Proc. Second Workshop Tackling Computer Systems Problems with Machine Learning,* Apr. 2007.

[16] B. Schölkopf, J.C. Platt, J. Shawne-Taylor, A.J. Smola, and R.C. Williamson, "Estimating the Support of a High-Dimensional Distribution," *Neural Computation,* vol. 13, pp. 1443-1471, 2001.

[17] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, "A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data," *Applications of Data Mining in Computer Security,* D. Barbara and S. Jajodia, eds., chapter 4, Kluwer Academic, 2002.

[18] J. Aitchison and C.G.G. Aitken, "Multivariate Binary Discrimination by the Kernel Method," *Biometrika,* vol. 63, pp. 413-420, 1976.

[19] D.W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley & Sons, 1992.

[20] C. Scott and E. Kolaczyk, "Nonparametric Assessment of Contamination in Multivariate Data Using Minimum Volume Sets and FDR," technical report, Univ. of Michigan, 2007.

[21] A.O. Hero, "Geometric Entropy Minimization (GEM) for Anomaly Detection and Localization," *Advances in Neural Information Processing Systems,* 2007.

[22] J. Storey, "The Positive False Discovery Rate: A Bayesian Interpretation of the q-Value," *Annals of Statistics,* vol. 31, no. 6, pp. 2013-2035, 2003.

[23] R. El-Yaniv and M. Nisenson, "Optimal Single-Class Classification Strategies," *Advances in Neural Information Processing Systems,* 2007.

[24] A. McCallum and K. Nigam, "A Comparison of Event Models for Naïve Bayes Text Classification," *Proc. AAAI Workshop Learning for Text Categorization,* Technical Report WS-98-05, 1998.

[25] K. Humphreys and D.M. Titterington, "Improving the Mean-Field Approximation in Belief Networks Using Bahadur's Reparameterisation of the Multivariate Binary Representation," *Neural Processing Letters,* vol. 12, pp. 183-197, 2000.

[26] M.J. Wainwright and M.I. Jordan, "Graphical Models, Exponential Families, and Variational Inference," technical report, Dept. of Statistics, Univ. of California, Berkeley, 2003.

[27] G. Beylkin, J. Garcke, and M.J. Mohlenkamp, "Multivariate Regression and Machine Learning with Sums of Separable Functions," submitted, 2007.

[28] G. McLachlan and D. Peel, *Finite Mixture Models.* John Wiley & Sons, 2000.

[29] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions.* Wiley-Interscience, 1996.

[30] J.H. Wolfe, "Pattern Clustering by Multivariate Mixture Analysis," *Multivariate Behavioral Research,* vol. 5, pp. 329-350, 1970.

[31] W.K. Hastings, "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika,* vol. 57, no. 1, pp. 97-109, 1970.

[32] N. Atienza, J. García-Heras, J.M. Muñoz-Pichardo, and R. Villa, "On the Consistency of MLE in Finite Mixture Models of Exponential Families," *J. Statistical Planning and Inference,* vol. 137, pp. 496-505, 2007.

[33] D.M. Titterington, A.F.M. Smith, and U.E. Makov, *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, 1985.

[34] R.A. Redner and H.F. Walker, "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Rev.,* vol. 26, pp. 195-239, 1984.

[35] J. Silva and R. Willett, "Hypergraph-Based Anomaly Detection in Very Large Networks," Technical Report ECE-2008-01, Duke Univ., 2008.

[36] B. Klimt and Y. Yang, "The Enron Corpus: A New Dataset for E-Mail Classification Research," *Proc. 15th European Conf. Machine Learning,* 2004.

[37] R. Abelson, "Enron's Many Strands: Ex-Chief's Holdings; Putting 'Lost Everything' in Perspective," *New York Times,* Jan. 2002.

[38] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," http://www.csie.ntu.edu.tw/cjlin/libsvm, 2001.

[39] A. Ng and M.I. Jordan, "On Discriminative versus Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes," *Advances in Neural Information Processing Systems,* 2002.