ANNOTATED MINIMUM VOLUME SETS FOR NONPARAMETRIC ANOMALY DISCOVERY

Clayton D. Scott[*]

University of Michigan
Dept. of Elec. Eng. and Comp. Sci.
Ann Arbor, MI 48105

Eric D. Kolaczyk[†]

Boston University
Dept. of Mathematics and Statistics
Boston, MA 02215

ABSTRACT

We consider an anomaly detection problem, wherein a combination of typical and anomalous data are observed and it is necessary to identify the anomalies in this particular dataset without recourse to labeled exemplars. We take as our goal to produce an annotated ranking of the observations, indicating the relative priority for each to be examined further as a possible anomaly, while making no assumptions on the distribution of typical data. We propose a framework in which each observation is linked to a corresponding minimum volume set and, implicitly adopting a hypothesis testing perspective, each set is associated with a test. An inherent ordering of these sets yields a natural ranking, while the association of each test with a false discovery rate yields an appropriate annotation. The combination of minimum volume set methods with false discovery rate principles, in the context of data contaminated by anomalies, is new and estimation of the key underlying quantities requires that a number of issues be addressed. We offer some solutions to the relevant estimation problems, and illustrate the proposed methodology on synthetic and computer network traffic data.

Index Terms— minimum volume sets, false discovery rate, nonparametric outlier detection, multiple level set estimation, monotone density estimation

1. INTRODUCTION

Anomaly detection problems can be characterized as coming in two flavors. In anomaly *prediction*, one observes training data that represent typical or normal measurements of a system. The goal is to use the training data to construct a prediction rule that will distinguish typical data from anomalous data in the future. In anomaly *discovery*, data are observed that contain a combination of normal and abnormal data, and the goal is to identify the anomalies in that particular dataset. The former learning problem is supervised in the sense that all data are known to be typical. Our concern is with the latter problem, which is unsupervised and more difficult.

By way of motivation, consider the data in Fig. 1, representing measurements gathered on Internet traffic flowing over links in the Abilene network, discussed in Section 4. Each point corresponds to the total traffic volumes (measured in bytes) for a given ten minute interval over a pair of links to a given node in the network. In this example, the node corresponds to Atlanta, and the links correspond to routes in the Abilene network to Atlanta from Houston and Washington. The goal is to design a system that takes this collection of roughly 1000 measurements and identifies the extent to which each point may represent potentially anomalous behavior. The output of this automated system would be transmitted, for example, to a network operator who might then conduct follow-up experiments on the nature of the most suspect data. An essential feature of this system is that it make no assumptions about the distribution of the typical data. Additionally, the system is required to apply equally well to data at other nodes in the network, whose distributional characteristics will be markedly different from those of the data for the Atlanta node.

In this paper we propose a framework that meets these quite general requirements. Formally, we suppose we observe independent and identically distributed measurements $X_i \in \mathbb{R}^d, i = 1, \ldots, n$ from a mixture distribution i.e.,

$$X_i \sim Q = \pi\mu + (1-\pi)P, \tag{1}$$

where μ is the distribution on anomalies, P is the distribution on typical data, and π is the *a priori* probability of an anomaly. The challenge here is that we assume P is unknown, as is π as well. However, we will allow that the user be willing or able to specify μ, which we will therefore consider known. For example, in the absence of any detailed information on the nature of anomalies, it is natural to assume that μ is simply a uniform distribution. All numerical work herein was done with this assumption, but our overall framework and all stated analytical results hold for arbitrary μ. We also assume the supports of μ and P are bounded, with the former

[*]Supported in part by NSF grant no. 0240058

[†]Supported in part by NSF grant CCR-0325701 and ONR award N000140610096. The authors thank Mark Crovella and Parminder Chhabra for helpful discussions. They thank P. Chhabra for supplying the network traffic data and known anomalies.

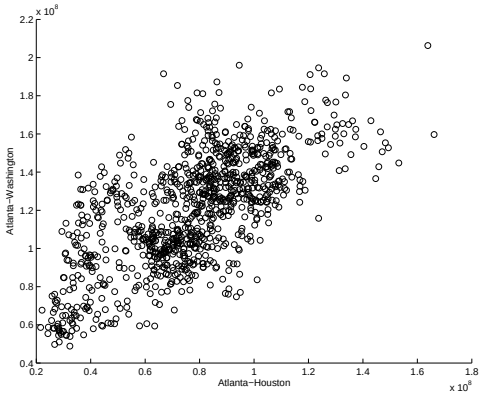


Fig. 1. Scatterplot of volume levels for traffic passing through select pairs of links at Atlanta, in the Abilene network, over consecutive 10 minute intervals.

containing the latter.

We then take as our goal to produce an annotated ranking of the observations X_i, indicating the relative priority for each to be examined further as a possible anomaly. We have in mind that the ranking serve to impose priorities through a basic ordering, while the annotations should provide some indication of the likelihood that observations are actually anomalous. Our approach in this paper is to link each observation to a corresponding minimum volume (MV) set and, implicitly adopting a hypothesis testing perspective, to associate each set with a test. An inherent ordering of these sets yields a natural ranking, while the association of each test with a false discovery rate (FDR) yields an appropriate annotation.

The assumption of known μ can be justified on a number of grounds. In many situations, the assumption of uniform μ is intuitive and natural, and has recently been shown in a certain setting to optimize the worst-case detection rate among all choices for the unknown anomaly distribution [1]. Furthermore, previous works that apply MV sets to anomaly prediction implicitly assume μ is a known distribution on anomalies [2, 3, 4, 5, 6]. That is, these predictors are only optimal when in fact the anomalies truly follow the volume-defining measure μ. Finally, ranking with respect to MV sets and uniform μ coincides with the ranking determined by the so-called likelihood data-depth [7, 8]. The connection to data depth is discussed further in the concluding section.

Connections to previous work on MV sets and FDR, proofs of results, additional experimental results, and all omitted details may be found in [9].

2. METHODOLOGY

Recall the model in Eqn. (1). Define $G_{P,\beta}$ to be the set with minimal μ-measure containing at least $\beta \in [0,1]$ probability mass under P i.e.,

$$G_{P,\beta} = \arg\min\{\mu(G) : P(G) \geq \beta\} \ . \tag{2}$$

This is the MV set under P, where volume[1] is assessed with respect to μ. It is easily seen that MV sets coincide with density level sets of P, provided the density exists with respect to μ. Each mass β corresponds to a certain level of the density of P, and as β ranges from 1 to 0, the density level ranges from 0 to the maximum value of the density. Although we refer to MV and density level sets interchangeably, we favor the former here because our method features the masses and volumes that MV sets make explicit.

As stated previously, our goal is to produce an annotated ranking of the observations. We consider the task of ranking first. For $i = 1, \ldots, n$, let

$$\beta_i \equiv \inf_{0 \leq \beta \leq 1} \{\beta : X_i \in G_{P,\beta}\}. \tag{3}$$

That is, letting β vary freely between 0 and 1, we assign to X_i a set G_{P,β_i} that, among all MV sets, just barely includes X_i. Ordering the β_i as $\{\beta_{(n)}, \ldots, \beta_{(1)}\}$, from largest to smallest, naturally induces a ranking $\{X_{(1*)}, \ldots, X_{(n*)}\}$ of the observations, where $(i*)$ denotes the index of i-th most potentially anomalous observation.

Our choice of approach here may be motivated by considering the problem of formally testing the null hypothesis $H_0 : X_i \sim P$ versus the alternative hypothesis $H_1 : X_i \sim \mu$, for each $i = 1, \ldots, n$. If we choose to use a test of size α i.e., with a Type I error rate $\Pr\{\text{Reject } H_0 | H_0 \text{ True}\} = \alpha$, then the set $G^c_{P,1-\alpha}$ is in fact the rejection region for the most powerful test of this size. Then, if instead of making a hard decision of H_0 versus H_1, we report the corresponding statistical p-value under this class of tests, that p-value is simply $1 - \beta_i$. Therefore, our proposed ranking follows the reverse ordering of the observed p-values.

Now consider the issue of annotation of our ranked observations. The values β_i are themselves an obvious, and indeed not unreasonable, candidate for such an annotation. However, there is the need to interpret these values and, although the values β_i are well-defined probabilities in the context of the individual hypothesis tests for their corresponding observations X_i, they are not designed to be meaningfully interpreted *en masse* when simultaneously conducting multiple hypothesis tests. This observation is a variation on the issue at the heart of the so-called 'multiple testing problem' in statistics. Stated simply, the problem is that, whereas standard testing theory dictates that one should choose the size α of a single test to control the chance of an incorrectly rejected null hypothesis i.e., a 'false discovery', in contexts where a large number of such tests are to be conducted, one will end up with a correspondingly large number of false discoveries purely by chance. Such an outcome is often unsatisfactory, particularly when nontrivial amounts of energy are expected to be used to follow up on discoveries, as is often the case in anomaly detection problems.

[1] Here "volume" connotes a probability measure, in contrast to traffic volumes in the network example.

This problem has received a great deal of attention in the statistical literature over the past decade, since the seminal paper of Benjamini and Hochberg [10]. Their proposal for this problem effectively boils down to focusing attention not on the size α of individual tests, but rather the *rate* of false discoveries across tests. Since their paper, an entire subliterature has evolved on the topic of FDR's, including a number of extensions in which analogues of the model in Eqn. (1) are assumed (see references in [9]). From among these various contributions, we choose to adopt the so-called positive FDR[2] statistic of Storey [11] as a natural one for our problem. In our context, this statistic is written as a probability

$$\text{pFDR}(G) = \Pr\{X \sim P \,|\, X \notin G\} \quad , \tag{4}$$

where G denotes an arbitrary set and $X \sim Q$. This is the probability that, given a 'discovery' is made (i.e., $X \notin G$), that in fact this discovery is false.

Storey [11] also proposes a corresponding analogue of the p-value, which he calls a q-value. This statistic, in our context, takes the form $\text{pFDR}(G_{P,\beta_i})$. We therefore propose, as a more meaningful alternative to the values β_i, to annotate our ranked observations by the values

$$\gamma_i = 1 - \text{pFDR}(G_{P,\beta_i}).$$

There immediately arises the question of whether the values of the γ_i's are consistent with the ranking arising from the β_i's. The following result addresses this concern, in the affirmative.

Proposition 1 *Let $\beta(t) = \sup\{\beta : \mu(G_{P,\beta}) \leq t\}$, for $t \in [0,1]$. Assume $\beta(t)$ is concave. Then the ordered sequences $\{\beta_{(n)}, \ldots, \beta_{(1)}\}$ and $\{\gamma_{(n)}, \ldots, \gamma_{(1)}\}$ produce the same rank ordering $\{X_{(1*)}, \ldots, X_{(n*)}\}$ of the observations $X_1, \ldots, X_n$.*

Note that $\beta(t)$ may be thought of as the optimal receiver operating characteristic (ROC) curve of a testing problem have μ as the distribution under the null and P the distribution under the alternative. (This is the reverse of what we consider throughout the paper, and is only used here as an analytical device.) As a result, we know that $\beta(t)$ is nondecreasing. The assumption that $\beta(t)$ is concave is satisfied, for example, when P is a continuous distribution and μ uniform.

3. ESTIMATION

There remains the issue of computing the annotations γ_i, or even the rankings through β_i, since both rely on P, which we assume unknown. Instead, all we have at our disposal are the observations $X_1, \ldots, X_n$, which are from the mixture distribution Q defined in (1), and our assumed knowledge of the contaminating distribution μ. In analogy to Eqn. (2), for $0 \leq \tilde{\beta} \leq 1$ define the MV set under Q at level $0 \leq \tilde{\beta} \leq 1$ as

$$G_{Q,\tilde{\beta}} = \arg\min\{\mu(G) : Q(G) \geq \tilde{\beta}\} \ .$$

The following result is fundamental to the practical implementation of our proposed methodology, in relating the MV sets under P to those under Q.

Proposition 2 *If $0 \leq \beta \leq 1$ and*

$$\tilde{\beta} \equiv \tilde{\beta}_{P,\beta} := \pi\mu(G_{P,\beta}) + (1-\pi)\beta \ , \tag{5}$$

then $G_{Q,\tilde{\beta}} = G_{P,\beta}$. Conversely, if $0 \leq \tilde{\beta} \leq 1$ and

$$\beta \equiv \beta_{Q,\tilde{\beta}} := \frac{\tilde{\beta} - (1-\pi)\mu(G_{Q,\tilde{\beta}})}{\pi} \ , \tag{6}$$

then $G_{P,\beta} = G_{Q,\tilde{\beta}}$.

This result links Q-MV sets to P-MV sets in an explicit fashion. In particular, it implies that the ordering given by the P-MV sets coincides with that of the Q-MV sets. Intuitively, the smallest P-MV set containing X_i is also the smallest Q-MV set containing X_i. More formally, define

$$\tilde{\beta}_i \equiv \inf_{0 \leq \beta \leq 1} \{\tilde{\beta} : X_i \in G_{Q,\tilde{\beta}}\}. \tag{7}$$

in analogy to the β_i defined for P in Eqn. (3). Then $\beta_{(n)}, \ldots, \beta_{(1)}$ and $\tilde{\beta}_{(n)}, \ldots, \tilde{\beta}_{(1)}$ define the same rank ordering of the X_is.

Furthermore, using Bayes' rule in conjunction with the expression in (4) and the definiton of γ_i, our proposed annotations may be expressed in the form

$$\begin{aligned} \gamma_i &= \pi\mu(G^c_{P,\beta_i})/Q(G^c_{P,\beta_i}) \\ &= \pi\mu(G^c_{Q,\tilde{\beta}_i})/Q(G^c_{Q,\tilde{\beta}_i}). \end{aligned} \tag{8}$$

Hence, we have the key insight that, since μ is known, to estimate γ_i we need only estimate $G_{Q,\tilde{\beta}_i}$, $Q(G_{Q,\tilde{\beta}_i})$, and π.

3.1. Estimation of $G_{Q,\tilde{\beta}_i}$

For convenience, write $G_i = G_{Q,\tilde{\beta}_i}$, the smallest Q-MV set containing X_i. Suppose $\{\hat{G}_\lambda\}_{\lambda\in\Lambda}$ is a family of set estimates such that (a) each $\hat{G}_\lambda$ estimates some Q-MV set, and (b) Λ is such that the range of MV sets estimated is sufficiently rich to reasonably approximate $G_{Q,\tilde{\beta}}$ for any $0 \leq \tilde{\beta} \leq 1$. Then a natural estimator for G_i is $\hat{G}_i := \hat{G}_{\hat{\lambda}_i}$, where $\hat{\lambda}_i := \arg\min\{\mu(\hat{G}_\lambda) : \lambda \in \Lambda, X_i \in \hat{G}_\lambda\}$.

We now briefly discuss two examples of such families $\{\hat{G}_\lambda\}_{\lambda\in\Lambda}$. The first requires solving the intermediate task of

[2]The pFDR is so named because it happens to be equal to the expected fraction of false discoveries, conditional on a positive number of discoveries having been made.

density estimation, while the second operates on the principle of direct set estimation.[3]

First, suppose a nonparametric estimate $\hat{f}(x)$ of the density f of Q is computed, such as a kernel density estimate. Then the sets $\hat{G}_\lambda = \{x : \hat{f}(x) \geq \lambda\}$ estimate the level sets of f, which coincide with the MV sets of Q. This approach has the advantage that the estimated sets are guaranteed to be nested. Therefore, the smallest such set containing a given X_i can be computed rapidly via a bisection search on λ.

The second example is based on the one-class support vector machine (OCSVM) with Gaussian kernel [3]. Here $\hat{G}_\lambda$ is the OCSVM with regularization parameter λ. It has been shown [12] that for each λ, $\hat{G}_\lambda$ is a consistent estimator of the λ level set of Q. As λ varies through its range, all MV sets of Q are accounted for. For more on this approach, see [13], where the algorithm of Hastie et al. [14] is used to efficiently compute the entire family $\{\hat{G}_\lambda\}_{\lambda \in \Lambda}$.

Other methods for direct set estimation readily follow from classification algorithms having the ability to control the tradeoff between false positives and false negatives [15, 16]. If λ is a parameter that controls such a tradeoff, then $\hat{G}_\lambda$ may be identified with the classifier that discriminates $X_1, \ldots, X_n$ from an artificially generated sample from μ [2, 5].

3.2. Estimation of $Q(G_{Q,\tilde{\beta}_i})$

Given estimates $\hat{G}_i$ of the sets $G_i = G_{Q,\tilde{\beta}_i}$, we may then estimate $\tilde{\beta}_i = Q(G_i)$ and related quantities through $\hat{Q}(\hat{G}_i)$, where

$$\hat{Q}(G) = \frac{|\{i : X_i \in G\}|}{n}.$$

Since X_i is on the boundary of $G_{Q,\tilde{\beta}_i}$, we average the two possible empirical estimates so that $\hat{Q}(G_{Q,\tilde{\beta}_{(i)}}) = (i - 1/2)/n$.

3.3. Estimation of π

The estimation of π is facilitated by a transformation of variables. Specifically, define $Y_i = \mu(G_i)$, where recall that $G_i = G_{Q,\tilde{\beta}_i}$. Writing $G_i = G(X_i)$ now to emphasize the dependence on X_i, we consider $Y = \mu(G(X))$ as a univariate random variable on the interval $[0, 1]$ resulting from transformation of the generic random variable $X \sim Q$. The following result shows π to be related to the density of Y in a simple manner.

Proposition 3 *Let*

$$D(t) := \inf\{\beta : \mu(G_{P,\beta}) \leq t\}$$

and

$$\tilde{D}(t) := \inf\{\tilde{\beta} : \mu(G_{Q,\tilde{\beta}}) \leq t\} \ .$$

[3]Note that estimating every level set of a density is equivalent to estimating the density itself, so there is no clear advantage of one approach over another.

Assume $D(t)$ to be differentiable in t. Additionally, assume $D'(t) \to 0$ as $t \to 1$. Then the density of Y is $\tilde{D}'(t) = \pi + (1 - \pi)D'(t)$, and therefore $\pi = \tilde{D}'(1-)$.

Thus $\tilde{D}(t)$ is the cumulative distribution function of Y. The assumption $D'(t) \to 0$ as $t \to 1$ holds provided it is not possible to write $P = (1-\theta)P_0 + \theta\mu$ for some distribution P_0 and for $\theta > 0$. Otherwise, P has a uniform component and it is impossible to resolve π accurately. Proof of the proposition employs arguments that, similar to those of Proposition 3, rely on ROC curves of optimal tests, only in this case in a dual sense, with P and μ switched in their roles as null and alternative. The reader is referred to [9] for details.

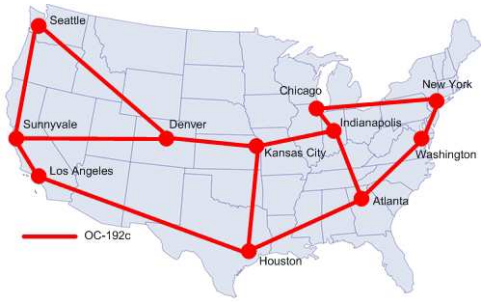
The obvious strategy now is to estimate π by estimating $\tilde{D}'(1-)$ based on the values $Y_1, \ldots, Y_n$. Note, however, that we do not in fact have access to the Y_i, given a lack of knowledge of Q. We propose therefore to estimate each Y_i by the value $\hat{Y}_i := \mu(\hat{G}_i)$ once the estimates $\hat{G}_i$ are computed and to proceed accordingly.

In the event that $D(t)$ (and hence $\tilde{D}(t)$) is concave in addition to being differentiable, estimating π amounts to estimating the value of a monotone decreasing density at the right boundary of its support. A consistent estimator for this problem has been studied in [17]. Practical estimators have also been developed in recent work on multiple testing [18, 11] where they are used to estimate the proportion of true null hypothesis. There the p-values of a test play a role similar to our Y_i; under a null hypothesis, p-values are uniform, just as our Y_i's are uniform under $X \sim \mu$.

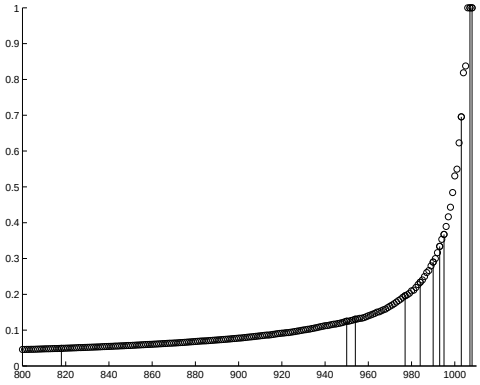
4. NETWORK ANOMALY DETECTION

Now return to the problem of detecting anomalous Internet traffic on a given network, described at the start of this paper. Fig. 2 (a) shows a map of the Abilene network, the 'backbone' network serving most universities and research labs in the United States. Developed as part of the Internet2 project [19], a project devoted to development of the 'next-generation' Internet, Abilene and Abilene data frequently serve as a testbed for development methodologies. Typically measurements on a network like Abilene are most easily available locally at network nodes (e.g., routers, regional aggregation points, etc.). So a natural way to approach the problem of anomaly detection is to seek to determine, at a given point in time, whether the traffic through a given network node is anomalous in nature or not. This problem is made challenging by many issues, particularly the facts that (i) traffic at a network node is a combination of the traffic from a number of incoming and outgoing links, and (ii) traffic on fixed links has been found to have subtle combinations of various characteristics, and hence is not highly amenable to simple parametric modeling (e.g., [20, 21]).

Our methodology, which makes no assumptions on the distribution P of normal network traffic, is natural for this



(a)



(b)

Fig. 2. (a) Abilene backbone (b) Ranked annotations produced by our method for the Atlanta data, with vertical stems corresponding to known anomalies [21].

task[4]. Fig. 2 (b) shows the largest 208 (out of 1008 total data points) annotations $\hat{\gamma}_i$ for the data at the Atlanta router shown in Fig. 1. The vertical stems are anomalies that were detected using a global method (having access to all data on all links in the network) and serve as ground truth anomalies for our purposes [20, 21]. There are 11 anomalies total. We see that 8 out of 11 occur past the 'knee' of the curve at roughly 0.2, and three are in the top six.

Additional results, applied to synthetic and flow cytometry data, as well as details of our implementation may be found in [9].

5. REFERENCES

[1] Ran El-Yaniv and Mordechai Nisenson, "Optimal single-class classification strategies," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. MIT Press, Cambridge, MA, 2007.

[2] J. Theiler and D. M. Cai, "Resampling approach for anomaly detection in multispectral images," in *Proc. SPIE*, 2003, vol. 5093, pp. 230–240.

[3] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1472, 2001.

[4] Gert R. G. Lanckriet, Laurent El Ghaoui, and Michael I. Jordan, "Robust novelty detection with single-class mpm," in *Advances in Neural Information Processing Systems 15*, S. Thrun S. Becker and K. Obermayer, Eds., pp. 905–912. MIT Press, Cambridge, MA, 2003.

[5] I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection," *J. Machine Learning Research*, vol. 6, pp. 211–232, 2005.

[6] C. Scott and R. Nowak, "Learning minimum volume sets," *J. Machine Learning Res.*, vol. 7, pp. 665–704, 2006.

[7] R. Fraiman and J. Meloche, "Multivariate L-estimation," *Test*, vol. 8, no. 2, pp. 255–317, 1999.

[8] R. Liu, J. Parelius, and K. Singh, "Multivariate analysis by data depth: Descriptive statistics, graphics, and inference (with discussion)," *Ann. Stat.*, vol. 27, pp. 783–858, 1999.

[9] C. Scott and E. Kolaczyk, "Nonparametric assessment of contamination in multivariate data using minimum-volume sets and FDR," Tech. Rep., Univ. Michigan, 2007, Available at http://www.eecs.umich.edu/∼cscott/.

[10] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Statist. Soc B*, vol. 57, no. 1, pp. 289–300, 1995.

[11] J.D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society, Series B*, vol. 64, pp. 479–498, 2002.

[12] R. Vert and J.-P. Vert, "Consistency and convergence rates of one-class SVM and related algorithms," *J. Machine Learning Research*, pp. 817–854, 2006.

[13] G. Lee and C. Scott, "The one class support vector machine solution path," *Proc. Int. Conf. Acoust. Speech. Sig. Proc.*, 2007.

[14] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, "The entire regularization path for the support vector machine," *J. Machine Learning Research*, pp. 1391–1415, 2004.

[15] C. Elkan, "The foundations of cost-sensitive learning," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, Seattle, Washington, USA, 2001, pp. 973–978.

[16] F. R. Bach, D. Heckerman, and E. Horvitz, "Considering cost asymmetry in learning classifiers," *J. Machine Learning Research*, pp. 1713–1741, 2006.

[17] V. Kulikov and H. Lopuhaä, "The behavior of the NPMLE of a decreasing density near the boundaries of the support," *Ann. Stat.*, 2006, in press.

[18] M. Langaas, B. H. Lindqvist, and E. Ferkingstad, "Estimating the proportion of true null hypotheses, with application to DNA microarray data," *Journal of the Royal Statistical Society, Series B*, vol. 67, pp. 555–572, 2005.

[19] Internet2, "http://www.internet2.org," .

[20] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. Kolaczyk, and N. Taft, "Structural analysis of network traffic flows," in *Proc. ACM SIGMETRICS/Performance*, 2004.

[21] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *Proc. ACM SIGCOMM*, 2004.

[4] Technically the datapoints in this example are correlated, due to temporal correlations in the underlying traffic flows. However, we ignore these correlations here for the purpose of illustration.