STATISTICS FOR CHARACTERIZING DATA ON THE PERIPHERY

James Theiler and Don Hush

Space and Remote Sensing Sciences
Los Alamos National Laboratory
Los Alamos, NM 87545

1. INTRODUCTION

The detection of anomalies (and of anomalous changes) requires that the samples that are anomalous be distinguished from the samples that are normal [1]. One way this can be achieved is by identifying two probability distributions: one for normal data and one for anomalies. The normal data distribution is generally fit to the data, while the anomalies are defined (often implicitly) with a distribution that is usually much broader and flatter than the normal data distribution. If both distributions were precisely known, then their ratio would provide the Bayes optimal detector of those anomalies.

While the choice of distribution for modelling the anomalies does require some care, the main technical challenge in anomaly detection is the characterization of the normal data distribution. The more "tightly" fit the distribution is to the normal data, the more accurately one can detect those data that do not fit the normal model.

For anomaly detection problems, very low false alarm rates are desired. Thus the challenge is even greater because we need to characterize the density in regions where the data are sparse; that is, on the periphery (or the ”tail”) of the distribution. Yet, traditional density estimation methods for anomaly detection (e.g. the simplest and most common approach is to fit a single Gaussian to the data) are dominated by the high-density core - where most of the data samples are located.

In this work, we will investigate two approaches for characterizing the periphery of a data distribution, and evaluate their performance for a set of anomaly detection and anomalous change detection problems.

In all of the examples here, our model for characterizing the periphery of a multivariate distribution will be an ellipsoid; our aim then, is to estimate a covariance matrix that characterizes that ellipsoid. We remark that the overall scale of the covariance is not of particular concern to us; for the single scalar measure of overall size, we can adjust the parameter to achieve whatever desired false alarm rate α that is desired. What is of more concern to us is the $O(d^2)$ parameters, where d is the number of spectral channels, that characterize the centroid and shape of the ellipsoid.

This work was supported by the Laboratory Directed Research and Development (LDRD) program at Los Alamos National Laboratory.

2. IN DEFIANCE OF ROBUST STATISTICS

For robust statistical estimation of the mean $\boldsymbol{\mu}$ and covariance matrix R, one employs equations of the form [2]:

$$\begin{aligned} \boldsymbol{\mu} &= \sum_{i=1}^{m} w_i \mathbf{x}_i \Big/ \sum_{i=1}^{m} w_i, \\ R &= \sum_{i=1}^{m} w_i^2 (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \Big/ \sum_{i=1}^{m} w_i^2. \end{aligned} \tag{1}$$

When the weights are all equal (e.g., $w_i = 1$ for all i), then the standard estimators for mean and covariance are obtained. For a robust estimator, one can alter these weights depending on how far the samples are from the mean. Distance to the mean is measured in terms of the Mahalanobis distance

$$r_i = \left[(\mathbf{x}_i - \boldsymbol{\mu})^T R^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\right]^T . \tag{2}$$

To make the robust estimator less sensitive to outliers, one discounts the large r samples; for instance:

$$\text{Robust: } w(r) = \begin{cases} 1 & \text{if } r \le r_o \\ r_o/r & \text{if } r > r_o. \end{cases} \tag{3}$$

To use this in practice requires an iterative approach, since the weights depend on Mahalanobis distance; Mahalanobis distance depends on $\boldsymbol{\mu}$ and R; and $\boldsymbol{\mu}$ and R depend on the weights. One must also choose a value for the cutoff radius r_o. For a d-dimensional Gaussian, the squared Mahalanobis distance r^2 is chi-squared distributed, with d degrees of freedom. The mean of r^2 is d and the variance is $2d$; it follows that r will have mean $d^{1/2}$ and standard deviation $1/\sqrt{2}$. This suggests that a good choice for r_o is of the form $r_o = \sqrt{d} + n/\sqrt{2}$ where n is of the order of a few. We used $n = 2$ in our numerical experiments.

But for problems which depend primarily on the periphery of the distribution, this scheme seems to be getting it backwards: it discounts just the data that we most need to pay attention to. Instead, we considered a weighting scheme that discounts the *small* Mahalanobis points:

$$\text{Fragile: } w(r) = \begin{cases} (r/r_o)^2 & \text{if } r \le r_o \\ 1 & \text{if } r > r_o. \end{cases} \tag{4}$$

Now, although we do want to characterize the periphery, we don't want to be unduly influenced by the actual outliers (*a.k.a.* anomalies) in the data, so we actually considered the following scheme:

$$\text{Periphery-weighted: } w(r) = \begin{cases} (r/r_o)^2 & \text{if } r \le r_o \\ r_o/r & \text{if } r > r_o. \end{cases} \tag{5}$$

The points for which $r \approx r_o$ will be the most heavily weighted in this scheme, and this suggests a strategy for choosing r_o. If we desire a false alarm rate in the regime of $\alpha \ll 1$, then choose r_o so that a fraction α of the data points have Mahalanobis distance larger than r_o.

3. SUPPORT VECTOR MACHINE

The idea that the boundary of the distribution should depend on points near that boundary, and not be influenced by the details of how the points in the core are distributed, motivates the use of a support vector machine for estimating mean and

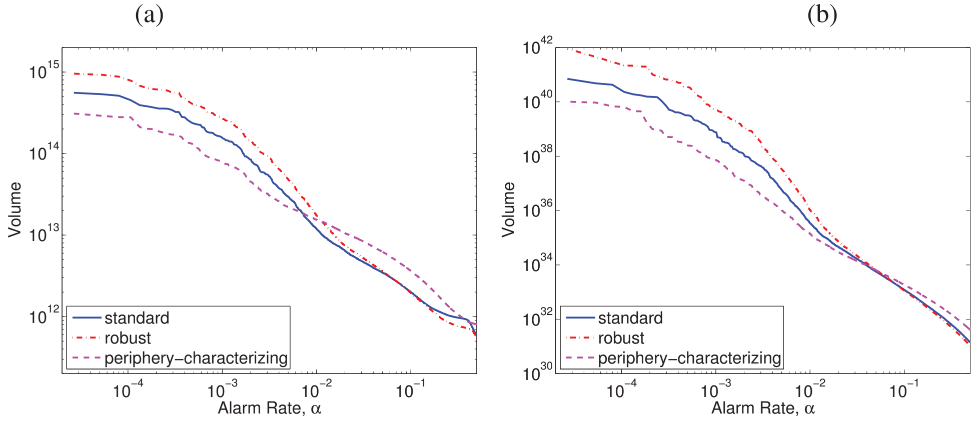

Fig. 1. Coverage plots show how the volume V of the ellipsoid increases as the fraction of uncovered data (the alarm rate) α decreases, using three different methods to estimate the ellipsoid shape. The two panels are for (a) the first three and (b) the first ten $d = 10$ principal components of the AVIRIS (Airborne Visual/InfraRed Imaging Spectrometer [4]) hyperspectral image of the Florida coastline, from data set f960323t01p02_r04_sc01. Half of the points are used to estimate covariance, and the other half are used to estimate performance, so these are out-of-sample results. The standard estimator uses Eq. (1) with all weights equal to one. The robust estimator uses weights given by Eq. (3), with $r_o = \sqrt{d} + 2/\sqrt{2}$. The "periphery-characterizing" estimator uses Eq. (5) with $r_o = \sqrt{d} + 2/\sqrt{2}$.

covariance. Instead of computing moments (or Mahalanobis distance weighted moments), the support vector machine estimates only the boundary between the two distributions. Employing the scheme detailed in Ref. [3], and constraining the solution to be quadratic[1], we are able to produce a mean and covariance estimator that depends, quite formally[2], only on the points on the periphery of the distribution.

4. A MEASURE OF PERFORMANCE FOR ANOMALY DETECTION

Because anomalies are rare, measuring the performance of an anomaly detection algorithm is somewhat problematic. Rather than concentrate on the anomalies, however, we will emphasize how well the model fits the normal data. In particular, given an alarm rate α (the rate at which normal samples are predicted to be anomalous), we will compute the volume $V(\alpha)$ of the ellipsoid which contains a fraction $1 - \alpha$ of the data. We will plot V versus α and our best algorithms with give the smallest values of V at low α. As we adjust the overall radius of the ellipsoid whose shape is specified by a given covariance matrix, we will trace out a curve in the V-versus-α space that has the flavor of a ROC curve. In fact, the α directly corresponds to false alarm rate. The V corresponds to a kind of missed detection rate, for the anomalies that are inside the volume V are the ones that will *not* be detected.

Fig. 1 shows such curves. As the alarm rate decreases, the volume necessary for achieving that alarm rate increases. For the low alarm rates, we see that the periphery-characterizing estimates outperform the standard and robust estimates. The

[1] This is *effectively* done by using quadratic kernels, though the implementation is slightly different from that.

[2] Points that are inside the boundary by a distance larger than the "margin" are not support vectors, and do not influence the fitting of the boundary to the data.

robust estimate is worse than the standard estimate at low α, but for larger $\alpha \approx 0.5$, the robust is slightly better. That is: the robust estimator better characterizes the core of the distribution while the periphery-characterizing estimates are better at, well, characterizing the periphery.

5. DISCUSSION AND CONCLUSIONS

In the ideal case of a multivariate Gaussian distribution, the contours are concentric ellipsoids, fully characterized by a mean vector and covariance matrix. Furthermore, the optimal estimator of these parameters are just the sample mean and sample covariance from classical statistics. These statistics give equal weight to all data samples, whether they are from the core or the periphery of the distribution.

But for deviations from this ideal, it may be preferable to preferentially use the data in the periphery of the distribution to estimate the shape of the contour in the periphery. This is done explicitly in the weighting function shown in Eq. (5), and implicitly when a support vector machine is used to learn that contour.

It is widely recognized that hyperspectral data is generally more fat-tailed than a Gaussian distribution, but it has recently become apparent that the "fatness" of those tails is different in different directions [5, 6, 7]. A consequence of this observation is that the best covariance matrix for characterizing the core of the data may differ from the best covariance matrix for characterizing the periphery. The approach we suggest here follows Vapnik's dictum [8] – rather that attempt to characterize the full distribution, we seek instead to characterize only the contour on the periphery.

6. REFERENCES

[1] A. Schaum, "Hyperspectral anomaly detection: Beyond RX," *Proc. SPIE*, vol. 6565, 2007.

[2] N. A. Campbell, "Robust procedures in multivariate analysis I: Robust covariance estimation," *Applied Statistics*, vol. 29, pp. 231–237, 1980.

[3] Ingo Steinwart, Don Hush, and Clint Scovel, "A classification framework for anomaly detection," *J. Machine Learning Research*, vol. 6, pp. 211–232, 2005.

[4] G. Vane, R. O. Green, T. G. Chrien, H. T. Enmark, E. G. Hansen, and W. M. Porter, "The Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)," *Remote Sensing of the Environment*, vol. 44, pp. 127–143, 1993.

[5] J. Theiler, B. R. Foy, and A. M. Fraser, "Characterizing non-Gaussian clutter and detecting weak gaseous plumes in hyperspectral imagery," *Proc. SPIE*, vol. 5806, pp. 182–193, 2005.

[6] P. Bajorski, "Maximum Gaussianity models for hyperspectral images," *Proc. SPIE*, vol. 6966, pp. 69661M, 2008.

[7] S. M. Adler-Golden, "Improved hyperspectral anomaly detection in heavy-tailed backgrounds," *Proc. First IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 2009, Digital Object Identifier 10.1109/WHISPERS.2009.5289019.

[8] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 2nd edition, 1999.