LA-UR- 10-04665

Approved for public release; distribution is unlimited.

Title:	Statistics for characterizing data on the periphery
Author(s):	James Theiler Don Hush
Intended for:	IEEE Int'l Geoscience and Remote Sensing Symposium 25-30 July 2010 Honolulu, HI USA

Los Alamos
NATIONAL LABORATORY
EST. 1943

Form 836 (7/06)

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

ABSTRACT

We introduce a class of statistics for characterizing the periphery of a distribution, and show that these statistics are particularly valuable for problems in target detection. Because so many detection algorithms are rooted in Gaussian statistics, we concentrate on ellipsoidal models of high-dimensional data distributions (that is to say: covariance matrices), but we recommend several alternatives to the sample covariance matrix that more efficiently model the periphery of a distribution, and can more effectively detect anomalous data samples.

Index Terms— anomaly detection, outlier, target detection, probability distribution, robust statistics, Gaussian mixture models, expectation-maximization, leptokurtosis

1. INTRODUCTION

What makes target detection difficult is that the target must be distinguished from the background clutter, and this requires that the background be well characterized. More particularly, when that characterization is a probability distribution, it is the periphery of the background distribution that must be most carefully characterized. Targets in the core of the distribution are impossible to detect; targets far out on the tail of the distribution are easy to detect. It is the targets on the periphery, the targets that are difficult but detectable, that are of most interest to the algorithm developer who wants improved ROC curves.

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 (often implicitly) defined with a distribution that is 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 modeling 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 [2]) are dominated by the high-density core.

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 the desired false alarm rate α. What is of more concern is the $O(p^2)$ parameters, where p is the number of spectral channels, that characterize the shape of the ellipsoid and its centroid.

2. IN DEFIANCE OF ROBUST STATISTICS

The goal of robust statistics is to produce characterizations of data that are insensitive to a few bad data samples. This is typically achieved by discounting (or de-weighting) those samples that, because of their long "lever arm" have undue influence on the estimation. We will consider a contrary approach that puts *extra* weight on points that are far from the centroid.

For estimation of the mean $\boldsymbol{\mu}$ and covariance matrix R, Campbell [3] has recommended equations of the form

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

When the weights are all equal (e.g., $w_i = 1$ for all i), then the standard sample 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

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

distance

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

To make the robust estimator less sensitive to outliers, one discounts the large r samples; for instance [3]:

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

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.

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

$$\text{Anti-robust: } w(r) = \begin{cases} (r/r_o)^\mu & \text{if } r \le r_o \\ (r/r_o)^\nu & \text{if } r > r_o. \end{cases} \quad (4)$$

Here, $\mu = \nu = 0$ corresponds to the standard sample covariance, while $\mu = 0, \nu = -1$ corresponds to the robust estimator suggested by Campbell [3]. An anti-robust estimator takes $\mu > 0$. Note that the choice of a large r_o and a negative ν imbues the estimator with some robustness to very large values of r.

One must also choose a value for the cutoff radius r_o. For a p-dimensional Gaussian, the squared Mahalanobis distance r^2 is chi-squared distributed, with p degrees of freedom. This distribution is approximately Gaussian with mean p and variance $2d$. For our experiments, we take $r_o = \sqrt{p} + b/\sqrt{2}$ with $b = 2$.

In the adaptive version of this scheme, we choose a fraction $\alpha \ll 1$ of the points to emphasize, then (at each iteration) choose r_o so that a fraction α of the data points have Mahalanobis distance larger than r_o.

2.1. Anti-shrinkage estimator

One difficulty with the anti-robust estimators is that the iterations are often unstable. An alternative approach is to estimate a robust covariance matrix and to recognize that the sample covariance is a positive linear combination of the robust and anti-robust estimators. In general, "shrinkage" refers to the statistical approach of modifying an estimator by taking a positive linear combination with a simpler estimator. Since what we want is the anti-robust estimator, we will take a *non*-positive linear combination of the sample covariance and the robust estimator:

$$\hat{R} = \alpha R_{\text{robust}} + (1 - \alpha) R_{\text{sample}} \quad (5)$$

where $\alpha < 0$ is chosen so to optimize an in-sample measure of coverage versus volume, as described in Section 6.

3. EIGENVALUE ADJUSTMENT APPROACH

In the spirit of the anomaly detector suggested by Adler-Golden [4], we will use the sample covariance R to align the covariance matrix, but will adjust the magnitudes within that alignment. Specifically, we will write $R = E\Lambda E$, where E is the matrix of eigenvectors, and Λ is a diagonal matrix whose elements are eigenvalues. In particular, the kth element Λ_{kk} is given by the variance in the $\mathbf{e}_k$ direction, where $\mathbf{e}_k$ is the kth column vector in the matrix E. That is, $\Lambda_{kk} = (1/n)\sum_i (\mathbf{e}_k^T \mathbf{x}_i)^2$. Instead of using variance, we will use inter-percentile difference; we used the square distance between the tth lowest value of $\mathbf{e}_k^T \mathbf{x}_i$ and the tth highest value, thus enclosing a fraction $(n - 2t)/n$ of the samples. In our experiments, we took this fraction to be 0.999. Using these new values $\breve{\Lambda}_{kk}$, we estimate the covariance matrix with $\breve{R} = E\breve{\Lambda}E^T$.

In this scheme $\breve{\Lambda}_{kk} > \Lambda_{kk}$ just because the 99.9 inter-percentile distance is larger than the standard deviation, even for Gaussian data. But the overall magnitude of R doesn't matter. We find that the ratio $\breve{\Lambda}_{kk}/\Lambda_{kk}$ tends to be larger for small values of k, consistent with observations made elsewhere that tails are fatter in the high variance directions [4, 5].

We remark that while the eigenvalue adjustment scheme can be applied to the eigenvector matrix E of the original sample covariance, it is also possible to apply this correction to matrices that have been computed by other means, such as the anti-robust covariances in the previous section.

4. GAUSSIAN MIXTURE MODEL APPROACH

Although weighting pixels by Mahalanobis distance makes intuitive sense, we will provide a more formal approach which explicitly models the data with a Gaussian mixture model. If we write

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, R) = (2\pi)^{-d/2} |R|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{x}^T R^{-1} \mathbf{x}\right) \quad (6)$$

as the normal distribution with mean $\boldsymbol{\mu}$ and covariance R, then we will consider a two-component mixture model

$$P(\mathbf{x}) = \underbrace{(1 - \alpha)\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, R_{lo})}_{\text{inner core}} + \underbrace{\alpha\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, R_{hi})}_{\text{periphery}} \quad (7)$$

in which we impose a number of constraints. One, we will take $\boldsymbol{\mu}$ the same for both; that is, they will all be concentric. In fact, for simplicity, we will use the sample mean for $\boldsymbol{\mu}$. Two, we want $\alpha \ll 1$ to be fixed at user-specified values. Three, we want $R_{lo} \ll R_{hi}$, but we will *not* require that the shapes of these covariances be the same. Subject to these constraints, we use the usual expectation-maximization algorithm [6] to estimate R_{lo} and R_{hi}. One minor modification was to used a trimmed estimator that, at each iteration, sets the weights to zero for a tiny fraction ϵ of the points with largest Mahalanobis distance with respect to R_{hi}.

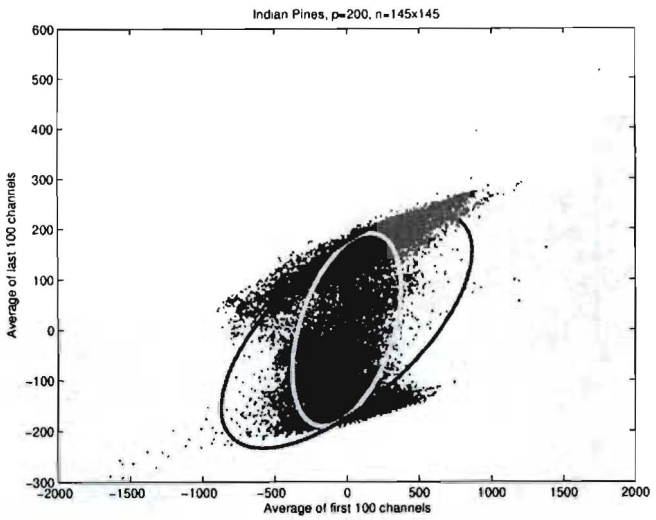

Fig. 1. Illustrating the mixture-of-Gaussians model – the smaller ellipse corresponds to R_{lo} (and is nearly identical to R), and the larger ellipse corresponds to R_{hi}. Although both Gaussians have the same centroid, they have different orientations of their principal axes. The smaller ellipse more efficiently characterizes the core of the distribution, but the larger ellipse better represents the periphery of the data.

5. SUPPORT VECTOR MACHINE APPROACH

As noted in the Introduction, if both the normal and the anomaly distributions were known then their ratio would provide the Bayes optimal anomaly detector. It follows that if we have samples from both distributions then we can design a support vector machine (SVM) to approximate the Bayes optimal detector [7]. In this paper we use a training set that contains both normal samples and synthetically generated anomalies to design a *quadratic* SVM that (approximately) optimizes a *weighted* linear combination of false alarm and missed detection rates. The SVM discriminant function takes the form[1]

$$f(\mathbf{x}) = \mathbf{x}^T Q \mathbf{x} + \mathbf{q}^T \mathbf{x} + q_0 \tag{8}$$

and can be converted to a Mahalanobis distance classifier using

$$R = Q^{-1}, \qquad \mu = -\frac{1}{2} Q^{-1} \mathbf{q} \,. \tag{9}$$

Instead of computing moments (or Mahalanobis distance weighted moments), the support vector machine more directly estimates the decision boundary between the two distributions. Increasing the weight on false alarms moves the decision boundary toward the periphery of the data so that the solution has fewer false alarms, though at the expense of more missed detections. Furthermore the SVM solution for Q takes the form

$$Q = \sum_{\mathbf{x}_i \in \text{data}} a_i \mathbf{x}_i \mathbf{x}_i^T - \sum_{\mathbf{x}_i \in \text{anomalies}} a_i \mathbf{x}_i \mathbf{x}_i^T \tag{10}$$

where all $a_i \geq 0$. The *support vector property* of SVM solutions implies that the nonzero coefficients in the first sum correspond to normal samples that lie near or beyond the decision boundary. Thus the solution is defined explicitly in terms of the peripheral normal samples.

The SVM approach requires us to generate samples from the anomaly distribution. The results in this paper we obtained using random samples from a uniform distribution over a hyper-rectangle that encompasses the normal data. Although increasing the number of samples promises more accurate solutions, it also increases the computational demand, and so the number of samples must be chosen to balance these two concerns. The results in this paper were obtained using approximately fives times as many anomalous samples as normal samples.

6. A MEASURE OF PERFORMANCE FOR ANOMALY DETECTION

Because anomalies are rare, measuring the performance of an anomaly detection algorithm can be 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 will 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, since the anomalies that are inside the volume V are the ones that will *not* be detected.

Fig. 2 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 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.

7. 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

[1] This form can be realized by using a quadratic kernel, or by quadratically extending the original training vectors and using a linear kernel.

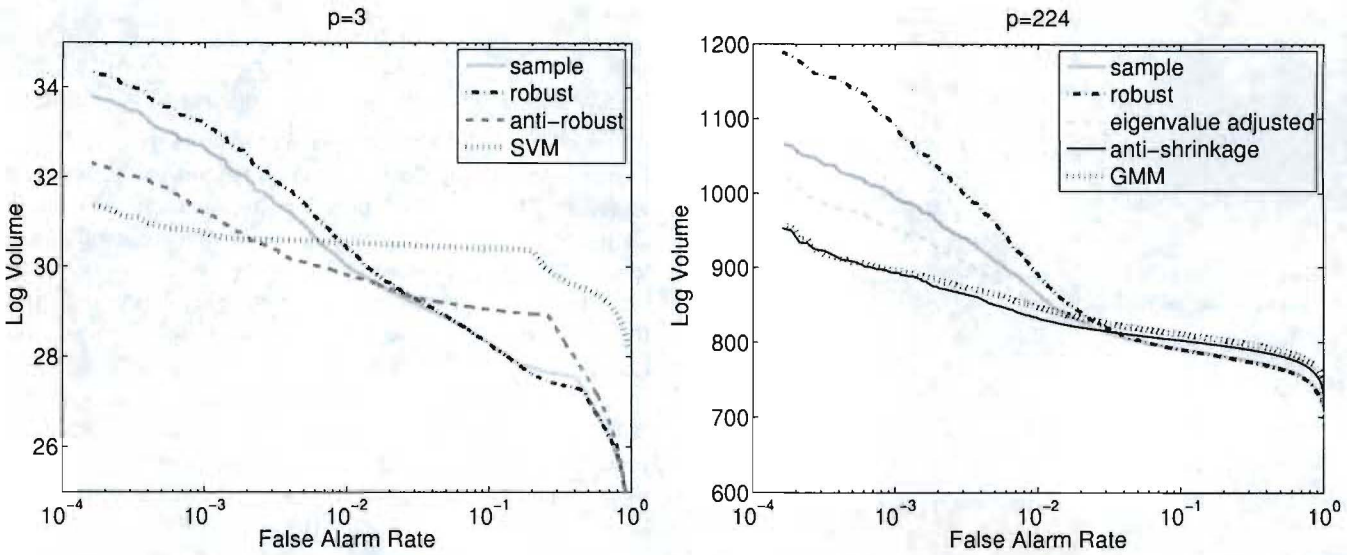

Fig. 2. Coverage plots show how the volume V of the ellipsoid increases as the fraction of uncovered data (the alarm rate) α decreases, using various algorithms to to estimate the covariance matrix. The top panel is for the first $p = 3$ principal components, and the bottom panel is all $p = 224$ spectral channels of the AVIRIS (Airborne Visual/InfraRed Imaging Spectrometer [8]) 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 sample estimator uses Eq. (1) with all weights equal to one.

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.

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 [4, 5, 9]. 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 [10] – rather that attempt to characterize the full distribution, we seek instead to characterize only the contour on the periphery.

8. REFERENCES

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

[2] I. S. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 38, pp. 1760–1770, 1990.

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

[4] 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.

[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] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm (with discussion)," *Journal of the Royal Statistical Society B*, vol. 39, pp. 138, 1977.

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

[8] 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.

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

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