Subspace Methods for Anomaly Detection in High Dimensional Astronomical Databases

Henrion, Marc
Imperial College London, Department of Mathematics
Huxley Building, South Kensington Campus
London SW7 2AZ, UK
E-mail: marc.henrion03@imperial.ac.uk

Mortlock, Daniel J.
Imperial College London, Department of Physics
Blackett Laboratory, South Kensington Campus
London SW7 2AZ, UK

Hand, David J.
Imperial College London, Department of Mathematics
Huxley Building, South Kensington Campus
London SW7 2AZ, UK

Gandy, Axel
Imperial College London, Department of Mathematics
Huxley Building, South Kensington Campus
London SW7 2AZ, UK

Introduction: Digital Sky Surveys, Virtual Observatories and Anomaly Detection

Survey astronomy consists in observing light emitting sources and recording these measurements in data catalogues. Sources are observed by telescopes which can be ground-based, air-borne or in orbit around the Earth or the Sun. Though some of these telescopes record data over the full light spectrum, for cost and time reasons most telescopes only record certain filter passbands, i.e. measure the total flux received over specific spectrum intervals. Depending on the purpose of a given survey, it can be designed to record flux in Gamma-ray, X-ray, ultraviolet, optical, infrared, microwave or radio passbands. For ground-based telescopes only optical, infrared and radio passbands can be measured as the light emitted by a source in all other parts of the spectrum is absorbed by the atmosphere.

The number of completed or ongoing surveys is very large, and the surveys differ widely in regions of the sky that are mapped, the filter passbands used, the detection limits (survey depth) etc. This is due to different science aims of the different surveys.

But this also means that many surveys overlap, i.e. a given source can be observed in different surveys, depending on which region in the sky it lies, how bright it is and in which parts of the light spectrum it radiates.

This overlap can be exploited by Virtual Observatories (VO), which are simply collections of surveys (with a dedicated web access). The surveys within a VO can be cross-matched (using the objects' coordinates on the sky (typically given in right ascension (ra) and declination (dec))). There are several VOs, a few examples being AstroGrid (also known as the UK Virtual Observatory; http://www.astrogrid.org/), the US National Virtual Observatory (NVO; http://www.us-vo.org/), Euro-VO (http://www.euro-vo.org).

Anomaly Detection is concerned with finding *observations which appear to be inconsistent with the remainder of that set of data* (Barnett and Lewis, 1994). More specifically an anomaly can be defined as *an observation which deviates so much from other observations as to arouse suspicions that*

it was generated by a different mechanism (Hawkins, 1980). Historically the aim was to remove such datapoints (also called outliers) from datasets as they can severely impact the statistical analysis of datasets with outliers. But anomalies can be interesting in themselves as, for example, an anomaly in a credit card dataset can be an indication of credit card fraud. In astronomy an anomaly can be a rare (e.g. quasars, brown dwarfs...) or even an unknown type of object. Finding such objects (and then studying them more closely with follow-up observations) can help to test cosmological models.

Problem Description and Motivation

Our aim is to detect anomalies in data from cross-matched digital sky surveys. There are several challenges that need to be addressed.

Surveys in themselves can be large and high-dimensional (thousands to hundreds of millions of objects; a handful to hundreds of variables). Hence a database compiled by cross-matching surveys from a VO will be large and high dimensional

This is both a curse and a blessing: a curse because of i) computational and methodological issues, ii) the fact that data in high dimensional spaces are sparse (curse of dimensionality); a blessing because the more variables that are recorded for each data point, the more information about source populations we have.

Another property of cross-matched catalogues is that they contain many missing values. Different objects will be observed in different surveys: a given object might have been observed in surveys A, B and C, but not in surveys D and E, while another source is observed in C and E but not A, B and D. Furthermore within each survey there can be missing values as the different bands have different sensitivities and thus not all bands will detect a faint source. However there is a certain structure in the missing detections: if an object is detected in a given survey, it will usually have detections in all bands in that survey (an exception being faint objects near the survey bands' detection limits), whereas it will have all detections missing for a survey in which it has not been observed. This non-totally-at-random missing values structure can be exploited to ease certain methodological and computational issues.

Thus we will have to develop an anomaly detection method which is fast enough to work with large, high-dimensional data, which can handle missing values and which allows a direct comparison of objects with different sets of observed variables.

The method we propose below essentially reduces the problem of working in a high-dimensional space to working in many lower-dimensional subspaces. While the reasons for taking this approach are given by the problem above, the specific reasons for working in lower-dimensional data subspaces are four-fold:

- data in high-dimensional spaces are sparse (e.g. Aggarwal et al. (2001)) and hence the local density around every object is low. As a result the very concept of what is an anomaly makes less sense in higher dimensions.

- unless there is a relationship between all the variables in a dataset, anomalies are apparent in subspaces of the data. The more variables there are, the more complex such a relationship will have to be. Also, the more variables are (automatically) collected, the higher the chances of some being independent of each other. For these two reasons, we think, such a complex relationship is increasingly unlikely as the dimensionality increases. Fig. 1 illustrates this point in a two-dimensional setting.

- anomalies might be anomalous in only a subset of variables. In a full-dimensional approach the anomaly score of such anomalies will be less extreme because of the contributions from the

variables the anomalies are not anomalous in, and thus these anomalies can go undetected. A lower-dimensional approach might be able to overcome this.

- a lower-dimensional approach will allow us to deal efficiently and rigorously with missing values: as we can restrict ourselves to the variables in which a particular object has been observed in, there will be no need for imputing missing data, nor will there be information lost due to discarding objects with missing values

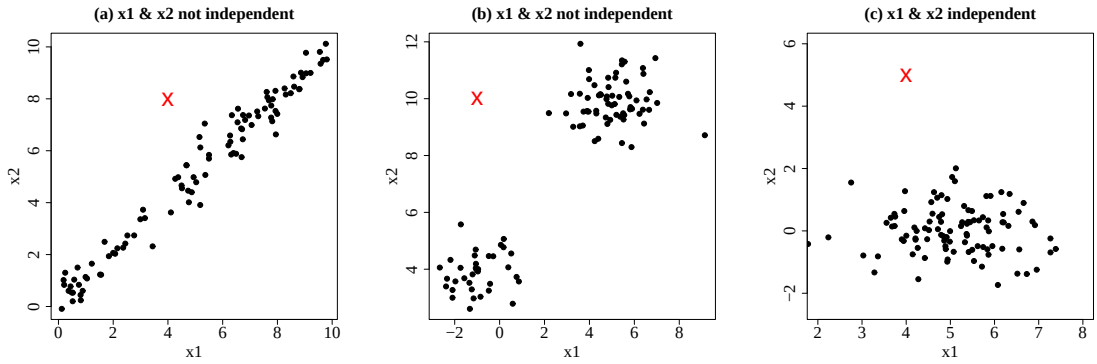

Figure 1: Simple examples of anomalies (red crosses); in figures (a) and (b) there is a relationship between x1 and x2 and we can only spot the anomaly by looking at the two-dimensional space, in figure (c) there is no such relationship and we can spot the anomaly by considering x2 alone.

Proposed Approach: Combining Anomaly Scores from Observed Subspaces

The proposed method to address the anomaly detection problem in datasets obtained by cross-matching astronomical surveys can be summarised in a few easy steps. The main idea consists in looking for anomalies not over the full-dimensional datasets, but in lower-dimensional subspaces of the data. For computational reasons (for this work we want to avoid having to compute some set of 'best' subspaces), we will limit ourselves to the subspaces given by subsets of the data variables.

Our approach is summarised by Algorithm 1, but let us first define some notation:

n - the number of objects (rows) in the dataset
d - the number of variables (columns) of the dataset
D - the maximum dimensionality of the subspaces ($1 \leq D \leq d$)
AS - anomaly score (we assume the more anomalous an object is, the higher its AS)
MV - missing value

Our method is not a novel AS computation algorithm, but attempts to use an AS calculator designed for low-dimensional data on high-dimensional data whilst avoiding the curse of dimensionality. In practice, any AS computation algorithm can be used with our approach. For this work we have used the Local Outlier Factor (LOF; Breunig et al. (2000)).

LOF generalises distance-based outliers (DB-outliers), introduced by Knorr et al. (2000). The DB-outliers technique computes the number of neighbours within a certain radius of a given object. If that number is less than a threshold, the object is flagged as anomalous. Alternatively the inverse of the number of neighbours within a chosen radius of an object can be used as AS. LOF looks at the local density around an object. The LOF score is essentially the average of the ratios of the average distance to the k nearest neighbours of the k nearest neighbours of a given object and the average distance to the k nearest neighbours of this object (though there is some smoothing for small distances involved as well). Connectivity-based Outlier Factor (COF; Tang et al. (2001)) improves on LOF by

Algorithm 1 Proposed approach

1. for i in $1 : D$
2. for j in $1 : \binom{d}{i}$
3. compute AS for objects with no MVs in j^{th} i-dimensional subspace
4. store the AS vector for this subspace
5. end for j
6. combine the AS vectors for all i-dimensional subspaces
7. end for i
8. output D AS vectors or D lists of anomaly candidates

computing the neighbourhood set of a given object in an incremental fashion. While this improves LOF, it is also much more computation-intensive and seems to outperform LOF only on contrived datasets (such as straight lines). Since LOF is much more widely used and studied than COF, we have used it with our method for this work. Local Density Factor (LDF; Latecki et al. (2007)) is very similar to LOF, but uses kernel density functions to compute density estimates, rather than just distances to nearest neighbours. LDF can outperform LOF, but does so at the cost of an extra parameter: in addition to the number of nearest neighbours (also used by LOF and COF), the bandwidth used with the kernel functions needs to be set as well. In practice, it can prove difficult to set this parameter and since our method will involve computing anomaly scores across many subspaces for which there might not be one best parameter, we have preferred to use LOF with our algorithm. LOF, COF and LDF are *density-based* anomaly detection methods and, as such, compute distances between objects. As distances between objects with missing values are not well-defined, these methods cannot be used (at least without modification) on data with missing values.

The key step in our approach is step 6 in algorithm 1 above. It is by – sensibly – combining, for each object, the AS of the subspaces the object has been observed in, that we can directly compare the anomalousness of objects with different sets of observed variables. If one were to, say, sum all the AS from the observed subspaces then objects with many observed variables are more likely to have high AS than objects with few observed variables and objects are not directly comparable. Hence we need to impose restrictions on what constitutes a valid combination function.

We will use the following notation:

$d_D = \binom{d}{D}$, the number of subspaces of dimension D in a d-dimensional dataset

$X = (AS_{i,j})_{\substack{1 \le i \le n \\ 1 \le j \le d_D}}$, a matrix of AS, with $AS_{ij} \in \mathbb{R} \cup \{NA\}$, X_i the i^{th} row of X

$\mathcal{G} = \{X \,|\, X \text{ an } n \times d_D \text{ AS matrix}\}$, the set of all $n \times d$ AS matrices

We define a *combination function* to be a function $\rho : \mathcal{G} \to (\mathbb{R} \cup \{\text{NA}\})^n$ which satisfies properties 1 and 2 below. If, in addition, a combination function satisfies properties 3-6 below, it is termed *well-behaved*.

Let $\mathcal{F}$ be the set of all combination functions and let $\rho \in \mathcal{F}$.

Property 1 (Putting objects with different numbers of missing values on the same scale)

Let $x_0 \in \mathbb{R}$ be a constant. Let $\mathcal{G}_0 = \{X \in \mathcal{G} \,|\, \forall i \in \{1, \ldots, n\},\ \forall j \in \{1, \ldots, d_D\},\ X_{i,j} \in \{x_0, \text{NA}\}\}$.

Then $\exists\, c \in \mathbb{R}$ so that $\forall X \in \mathcal{G}_0$, $\forall i \in \{1, \ldots, n\}$, $\rho_i(X) = \begin{cases} c & \text{if } \exists\, j \in \{1, \ldots, d_D\} \text{ so that } X_{i,j} = x_0 \\ \text{NA} & \text{otherwise} \end{cases}$.

This property guarantees that objects with many missing values have combined AS on the same scale as objects with few missing values and thus that objects with different sets of observed variables are directly comparable through their combined AS. For example, if we were to combine AS by summing all the non-missing AS for each object, then an object with, say, 5 non-missing AS will automatically have a much larger AS than an object with only 1 non-missing AS. This needs to be avoided and therefore property 1 is needed.

Property 2 (No non-missing combined AS for objects with at least one non-missing AS)

$\forall X, Y \in \mathcal{G}$ so that $\forall i, j \in \{1, \ldots, n\} \quad X_{i,j} = \text{NA} \Leftrightarrow Y_{i,j} = \text{NA}$
then $\forall i \in \{1, \ldots, n\} \rho_i(X) = \text{NA} \Leftrightarrow \rho_i(Y) = \text{NA}$

This property means an object has a missing combined AS if and only if all of its subspace-specific AS are missing and thus guarantees that each object which has at least one non-missing AS, also has a non-missing combined AS.

Property 3 (AS inequality for comparable objects)

$\forall\, X \in \mathcal{G},\ \forall\, i_1, i_2 \in \{1, \ldots, n\}$ so that

$$\begin{cases} X_{i_1,j} \leq X_{i_2,j} & \forall j \in \{k \,|\, k \in \{1, \ldots, d_D\} \text{ and } X_{i_1,k} \neq \text{NA}, X_{i_2,k} \neq \text{NA}\} \\ X_{i_1,j} = X_{i_2,j} = \text{NA} & \forall j \in \{k \,|\, k \in \{1, \ldots, d_D\} \text{ and } X_{i_1,k} = X_{i_2,k} = \text{NA}\} \end{cases}$$

we have that

$$\rho_{i_1}(X) \leq \rho_{i_2}(X).$$

This property simply means that if an object's AS are each less than or equal to those of another object, then its combined AS should be less than or equal to that other object's combined AS.

Property 4 (Effect on the AS of other objects)

$\forall X, \tilde{X} \in \mathcal{G}$ so that $\exists (i_0, j_0) \in \{1, \ldots, n\} \times \{1, \ldots, d_D\}$

$$\begin{cases} X_{i,j} = \tilde{X}_{i,j} & \forall (i,j) \in \{1, \ldots, n\} \times \{1, \ldots, d_D\} \setminus \{(i_0, j_0)\} \\ X_{i_0,j_0} \leq \tilde{X}_{i_0,j_0} \end{cases}$$

we have that

$$\begin{cases} \rho_i(X) \geq \rho_i(\tilde{X}) & \forall i \in \{1, \ldots, n\} \setminus \{i_0\} \\ \rho_{i_0}(X) \leq \rho_{i_0}(\tilde{X}) \end{cases}.$$

This property means that if we change an AS matrix so that we only change one object's AS, in particular by increasing one of its AS (i.e. by making that object more anomalous in one subspace), then the combined AS for all other objects should remain unchanged or decrease (i.e. stay equally anomalous or become less anomalous) whereas, obviously, the combined AS for the object in question increases.

Property 5 (Preservation of order)

Let $f : \mathbb{R} \to \mathbb{R}$ be a monotonically increasing function, and let $f(X) = \left(f\left(X_{i,j}\right)\right)_{\substack{1 \leq i \leq n \\ 1 \leq j \leq d_D}} \quad \forall\, X \in \mathcal{G}$.
Then
$\rho_i(X) \leq \rho_j(X) \Rightarrow \rho_i(f(X)) \leq \rho_j(f(X))$.

This property simply guarantees preservation of ranks of anomalousness.

Property 6 (Continuity)

For every choice of elements of an AS matrix that are missing, ρ is continuous on the non-missing components.

Combination functions which satisfy property 6 will be called *continuous*.

Examples of combination functions

Let $S_i = \{X_{i,j} | j \in \{1, \ldots, d_D\} \text{ and } X_{i,j} \neq \text{NA}\}, i = 1, \ldots, n$.

- Selecting the highest AS:

$$\rho^{(ext)}{}_i(X) = \max\{X_{i,j} | X_{i,j} \in S_i\} \quad \forall i \in \{1, \ldots, n\}.$$

- Averaging the AS:

$$\rho^{(avg)}{}_i(X) = \sum_{X_{i,j} \in S_i} X_{i,j}/|S_i| \quad \forall i \in \{1, \ldots, n\}.$$

- Averaging the top N AS:

$$\rho^{(topN)}{}_i(X) = \sum_{j=0}^{N-1} X_{i,(|S_i|-j)}/N \quad \forall i \in \{1, \ldots, n\}$$

where $X_{i,(j)}$ is the j^{th} order statistic of the AS scores of object i. (N.B. if an object has less than N AS, the combined AS is the average of all the available AS.)

- Sum of the excess above a certain quantile:
For each $j \in 1, \ldots, d_D$ let $q_j^{(1-\alpha)}$ be the $(1-\alpha)$ quantile of the AS recorded for subspace j. For all j, we subtract $q_j^{(1-\alpha)}$ from the AS for that subspace. Finally, for each object, we sum the non-negative values.

$$\rho^{(topquant)}{}_i(X) = \sum_{j \in S_i} \left(X_{i,j} - q_j^{(1-\alpha)}\right) I\left(X_{i,j} \geq q_j^{(1-\alpha)}\right) \quad \forall i \in \{1, \ldots, n\}$$

where $I(.)$ is the indicator function.

- Sum of the excess above a certain quantile and below another one:
We choose $0 \leq \alpha_2 < \alpha_1 \leq 1$ and compute, for each j, $q_j^{(1-\alpha_1)}$ and $q_j^{(1-\alpha_2)}$. For all j, we set all those AS exceeding $q_j^{(1-\alpha_2)}$ equal to $q_j^{(1-\alpha_2)}$ and then subtract the amount by which they exceed $q_j^{(1-\alpha_2)}$, i.e. for all i so that $X_{i,j} > q_j^{(1-\alpha_2)}$ we set $X_{i,j} = q_j^{(1-\alpha_2)} - (X_{i,j} - q_j^{(1-\alpha_2)})$. Then, for all j, we subtract $q_j^{(1-\alpha)}$ from all the AS for that subspace. Finally, for each object, we sum the non-negative values.

$$\rho^{(midquant)}{}_i(X) = \sum_{j \in S_i} \Big[\left(X_{i,j} - q_j^{(1-\alpha_1)}\right) I\left(q_j^{(1-\alpha_1)} \leq X_{i,j} \leq q_j^{(1-\alpha_2)}\right)$$
$$+ \left(2q_j^{(1-\alpha_2)} - X_{i,j} - q_j^{(1-\alpha_1)}\right) I\left(X_{i,j} > q_j^{(1-\alpha_2)} \text{ and } X_{i,j} \leq q_j^{(1-\alpha_2)} - q_j^{(1-\alpha_1)}\right)\Big]$$
$\forall i \in \{1, \ldots, n\}$.

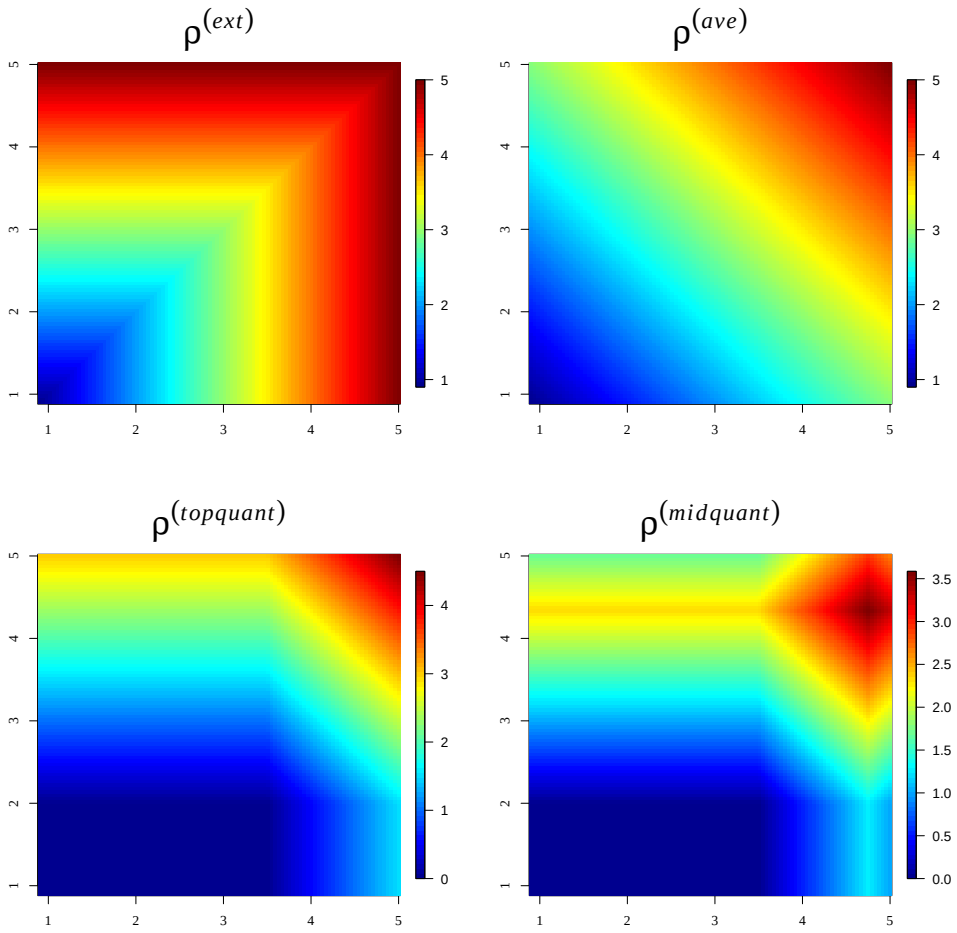

Figure 2: Examples of combination functions; the axes represent two different AS and the colour scale indicates the magnitude of the combined AS.

All of the above are valid combination functions ($\rho^{(ext)}$, $\rho^{(avg)}$, $\rho^{(topN)}$ satisfy property 1 with $c = x_0$ and $\rho^{(topquant)}$ and $\rho^{(midquant)}$ with $c = 0$ and property 2 is obviously met). The first four are also well-behaved. $\rho^{(midquant)}$ is not well-behaved as it does not satisfy properties 3, 4 and 5. It is, however, continuous. If we had not subtracted the quantiles $q_j^{(1-\alpha)}, q_j^{(1-\alpha_1)}$ for the last two combination functions respectively, they would not have been continuous; idem if we had simply set the AS above the second quantile equal to zero for $\rho^{(midquant)}$.

As we have already explained, property 1 needs to be met to guarantee comparability of the combined AS and property 2 guarantees non-missing combined AS for objects with at least one non-missing AS. Properties 3-5 intuitively appear desirable. And indeed they would be if there would only be ordinary objects and anomalies in a dataset. However, in practice, it is often the case that there are spurious objects (e.g. cosmic rays in astronomical datasets) or objects badly affected by observational noise (e.g. sources near large stars which get affected by diffraction spikes). Such noise objects have often very extreme measurements and result in very high anomaly scores. Although they do not satisfy properties 3-5, combination functions such as $\rho^{(midquant)}$ above allow one to effectively discard AS which are too extreme and focus on sources with consistently high but not extreme AS. Property 5 is usually desirable. For example, choosing the quantiles for $\rho^{(topquant)}$ and $\rho^{(midquant)}$ is an arbitrary process. Having *soft* thresholds (i.e. continuous combination functions) moderates the arbitrariness of such choices. But if there is a specific reason why a hard threshold might be appropriate for a combination function for a particular dataset, then property 6 would not be needed.

Results

This is still work in progress and, as a result, we will limit ourselves to giving a preview of how our method works in practice. The presentation at the conference will contain a more detailed performance assessment of our method and feature results from astronomy data.

Performance on simulated data

We compare our method to LOF and LDF on different sets of simulated data with no MVs. LDF with Mahalanobis distance proved to be computationally prohibitive for running large numbers of simulations, and so we have used LDF with Euclidean distance only. As our method involves computing $\binom{d}{D}$ AS vectors (whereas LOF and LDF, when applied directly to the d-dimensional data have to compute just one such vector), it is slower than both LOF and LDF (with Euclidean distance), even though the distances it has to compute are in lower-dimensional spaces. However, LOF and LDF do not work on data with MVs, whereas our method does.

For assessing the performance of the different anomaly detection techniques, we will use the receiver operating characteristic (ROC) curve and the true positive rate (or sensitivity). For the latter we will rank the AS and flag the objects with the highest scores as anomalous.

Dataset 1 consists of $15'188$ objects with 188 anomalies. The data are sampled from a multivariate normal distribution and the variables are pair-wise linearly correlated. The anomalies are anomalous in only 3 out of 35 variables. For the anomalies, these variables are not correlated with any of the other variables. Figure 3 shows that all three anomaly detection methods struggle with this type of data, but using $\rho^{(ext)}$ as combination function our method is able to outperform both LOF and LDF.

Dataset 2 consists of $10'100$, 60-dimensional data points with 100 anomalies. The normal data points are sampled from two distinct multivariate normal distributions and the anomalies do not lie in any of these two clusters of data points. For this data, our method is outperformed by both LOF and LDF as shown on Figure 4.

Dataset 3 consists of $25'250$, 20-dimensional data with 250 anomalies. The data points form again two distinct clusters, but this time the anomalies are anomalous in only half of the variables and there are three types of anomalies. The first type of anomalies (83 data points) lie in one of the two clusters of data, but have larger variances. The second type of anomalies (83 data points) do not lie in any of the two clusters. Finally, the third group of anomalies (84 data points) form a small cluster on their own, distinct from the other two clusters. Figure 5 shows the true positive rate for this dataset and also how the true positive rate varies with the sample size (keeping the number of nearest neighbours and the proportion of anomalies fixed). The three anomaly detection methods perform similarly well when the number of nearest neighbours (here fixed to 75) is greater than the number of anomalies that form a small cluster. However when there are more anomalies in that cluster than the number of nearest neighbours used, then our method outperforms LOF and LDF as it is able to better detect these anomalies.

Results from cross-matched SDSS-UKIDSS data

We are currently applying our method to data cross-matched from the Sloan Digital Sky Survey (SDSS; York et al. (2000)) and the Large Area Survey (LAS) from the United Kingdom Infrared Telescope (UKIRT) Infrared Deep Sky Survey (UKIDSS; Lawrence et al. (2007)). SDSS observes data in the optical filters u, g, r, i and z, whereas UKIDSS LAS observes data in the near-infrared bands Y, J, H and K. We use colour ($u-g$, $g-r$, $r-i$, $i-z$, $z-Y$, $Y-J$, $J-H$, $H-K$) and morphology (concentrations in the five SDSS filters and the ClassStat variables in the four UKIDSS LAS filters) variables as input variables for our anomaly detection method. We are not using any

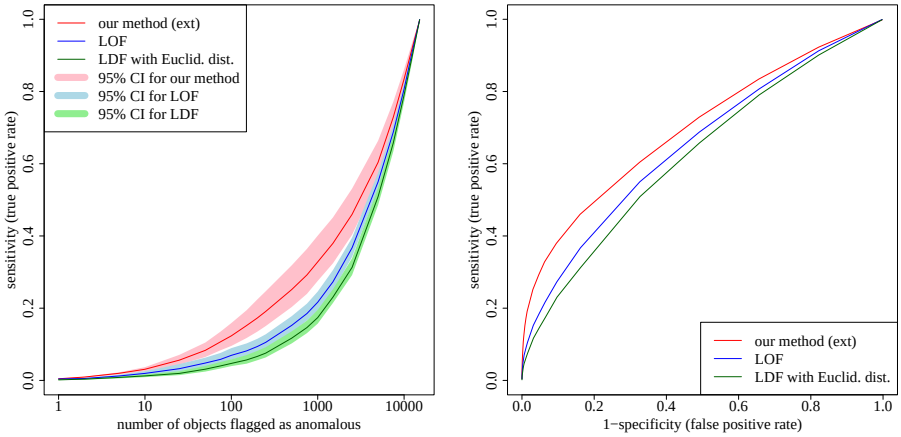

Figure 3: Sensitivity (left) and ROC curve (right) for our method, LOF and LDF for dataset 1. Anomalies are anomalous in 3 of 35 variables.

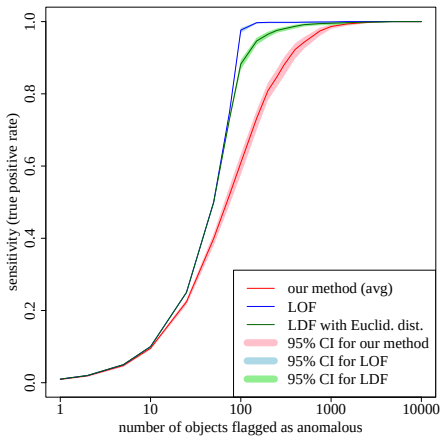

Figure 4: Sensitivity for our method, LOF and LDF for dataset 2. Anomalies do not lie in any of the two clusters of normal data points.

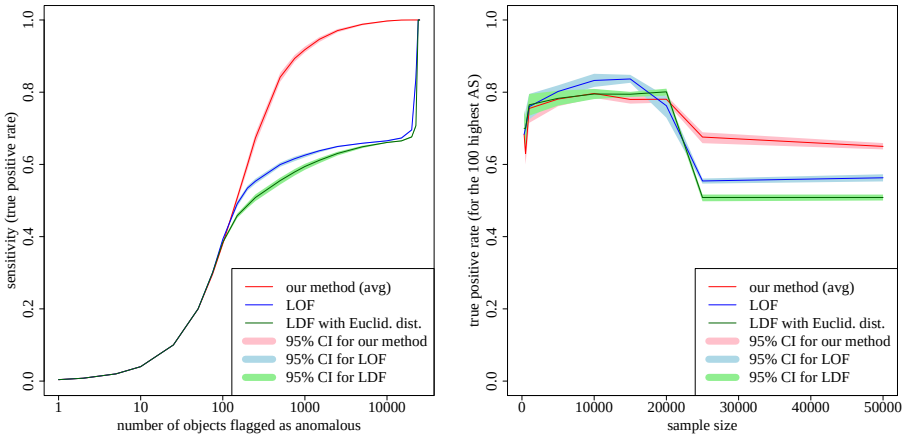

Figure 5: The left-hand-side figure shows the sensitivity for our method, LOF and LDF for dataset 3. The right-hand-side figure shows the results for different sample sizes. There are three different types of anomalies. The number of nearest neighbours used with all three methods is fixed to 75.

magnitude variables directly as in that case most bright sources would have high AS simply because there are few bright sources.

Results from these data will be presented at the conference.

REFERENCES (RÉFERENCES)

Aggarwal, C., Hinneburg, A., and Keim, D. (2001). On the surprising behavior of distance metrics in high dimensional space. In Van den Bussche, J. and Vianu, V., editors, *Database Theory ICDT 2001*, volume 1973 of *Lecture Notes in Computer Science*, pages 420–434. Springer Berlin / Heidelberg.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*. John Wiley & Sons, 3rd edition.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: Identifying density-based local outliers. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 29(2):93–104.

Hawkins, D. (1980). *Identification of Outliers*. Chapman and Hall, London.

Knorr, E. M., Ng, R. T., and Tucakov, V. (2000). Distance-based outliers: algorithms and applications. *The VLDB Journal*, 8:237–253.

Latecki, L., Lazarevic, A., and Pokrajac, D. (2007). Outlier detection with kernel density functions. In Perner, P., editor, *Machine Learning and Data Mining in Pattern Recognition*, volume 4571 of *Lecture Notes in Computer Science*, pages 61–75. Springer Berlin / Heidelberg.

Lawrence, A. et al. (2007). The ukirt infrared deep sky survey (ukidss). *Monthly Notices of the Royal Astronomical Society*, 379:1599–1617(19).

Tang, J., Chen, Z., chee Fu, A. W., and Cheung, D. (2001). A robust outlier detection scheme for large data sets. In *In 6th Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, pages 6–8.

York, D. G. et al. (2000). The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120:1579–1587.

ABSTRACT (RÉSUMÉ)

Modern astronomical surveys, in particular cross-matched databases from virtual observatories, are very large datasets (hundred of thousands to millions and even billions of objects), which are high-dimensional (from a dozen variables up to a few hundred) and which often contain large numbers of missing values (due to sources emitting light at different wavelengths and faint sources not being detected in all filter passbands). The objects most interesting for astronomers are typically very rare, very faint and have one or several features that set them apart from the other sources in the survey. Indeed common stars and galaxies are fairly well-understood and it are objects right at the detection limits of the different surveys or objects that have peculiar astrophysical properties which drive much of the astrophysical research. Therefore anomaly detection tools are vital for finding such potential interesting sources. However the size of the datasets involved, the high dimensionality and above all the large numbers of missing values present severe challenges to existing anomaly detection methods. We propose a novel approach which works by computing, for each object, anomaly scores in lower dimensional subspaces and then combining these scores to a unique score for each source. Working in subspaces allows us to work around the curse of dimensionality and deal very intuitively with missing values. As a result our method allows direct comparisons of sources, even if they have been observed in quite different sets of variables. We will discuss several ways of combining anomaly scores and look at various properties of our approach. The proposed approach is very flexible and can be used with most anomaly score computation methods.