Statistical Analysis and Data Mining

CASOS: a Subspace Method for Anomaly Detection in High Dimensional Astronomical Databases

Journal:	*Statistical Analysis and Data Mining*
Manuscript ID:	Draft
Wiley - Manuscript type:	Research Article
Keywords:	anomaly detection, astrostatistics

CASOS: a Subspace Method for Anomaly Detection in High Dimensional Astronomical Databases

Marc Henrion[1]*, David J. Hand[1], Axel Gandy[1], Daniel J. Mortlock[2]

[1]Department of Mathematics, Imperial College London, London SW7 2AZ, U.K.
[2]Astrophysics Group, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, U.K.

Abstract

We develop a novel algorithm for detecting anomalies. Our method has been developed to suit the challenging task of detecting anomalous sources in cross-matched astronomical survey data. Our algorithm computes anomaly scores in lower-dimensional subspaces of the data and presents several advantages over existing methods: it can work directly on data with missing values, it addresses some of the problems posed by high-dimensional data spaces, it is less susceptible to a masking effect from irrelevant features, it can be easily adapted to suit specific needs and it allows an easier interpretation of why a given object has a high combined anomaly score. One drawback of our method is that it cannot detect outliers that are only apparent in high-dimensional spaces. We demonstrate the properties of our algorithm and evaluate its performance on both simulated and real datasets. We show that it is capable of outperforming state-of-the-art, full-dimensional approaches in some situations.

1 Introduction

Survey astronomy consists in observing light emitting sources and recording these measurements in data catalogues. Observations are made by telescopes which can be ground-based, air-borne or in orbit around the Earth or the Sun. Though, ideally, data would be recorded over the full light spectrum, for technological reasons 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 might 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.

Virtual Observatories (VO) aim to facilitate making use of this overlap between surveys. VOs are 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)].

*Correspondence to marc.henrion03@imperial.ac.uk.

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 data (also called outliers) from datasets as they can severely impact statistical analyses. But anomalies can be interesting in their own right. 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) is one of the main aims of astronomical surveys.

For detecting interesting unusual objects, automatic outlier detection methods are only the first step, to be followed by human examination. In the case of astronomy, anomaly detection methods can be used to select a set of candidate objects, with potentially highly unusual physical properties, for which detailed follow-up observations will be made.

Different methods are effective for detecting different kinds of anomalies in different situations, and it would be naive to expect a single method always to be best. Our approach is intended to be complementary to other methods, with properties described below.

1.1 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 can be large and high-dimensional. While the gain of information about source populations achieved by cross-matching surveys is highly desirable, the resulting, potentially massive, datasets can pose various computational and methodological problems (e.g. sparsity of high-dimensional feature spaces, feasibility problems of algorithms scaling badly with sample size and / or the number of variables, etc.).

Another property of cross-matched catalogues is that they contain many missing values. Different objects will be observed in different surveys and if a source has not been observed in a given survey, it will have no measurements for all variables from that survey. Further, 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.

In summary, we want 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 (Aggarwal et al., 2001). A first consequence of this fact is that the local density around every object is low. Since anomalies are typically defined as objects that lie in low-density regions of the data-space, the very concept of an anomaly makes less sense in higher dimensions. A second consequence is that the notion of distance becomes less meaningful in high-dimensional spaces. Beyer et al. (1999) show that the discrimination between the nearest and furthest point of a given point becomes poor in high-dimensional datasets. This particularly affects nearest neighbour based anomaly detection techniques as they rely on finding the k nearest neighbours of a given object.

- 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 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. Figure 1 illustrates a three-dimensional setting that produces an anomaly only apparent when all three dimensions are considered jointly.

- Anomalies might be anomalous in only a subset of variables (a point also illustrated, in a three-dimensional setting, by figure 1, with three anomalies being anomalous in only two of three variables, and two anomalies being outlying in one variable only). In a full-dimensional approach the anomaly score of such objects 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 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.

- For high-dimensional data, it can be difficult to see why a given source has been declared anomalous by a full-dimensional method. A lower-dimensional subspaces can make this interpretation easier.

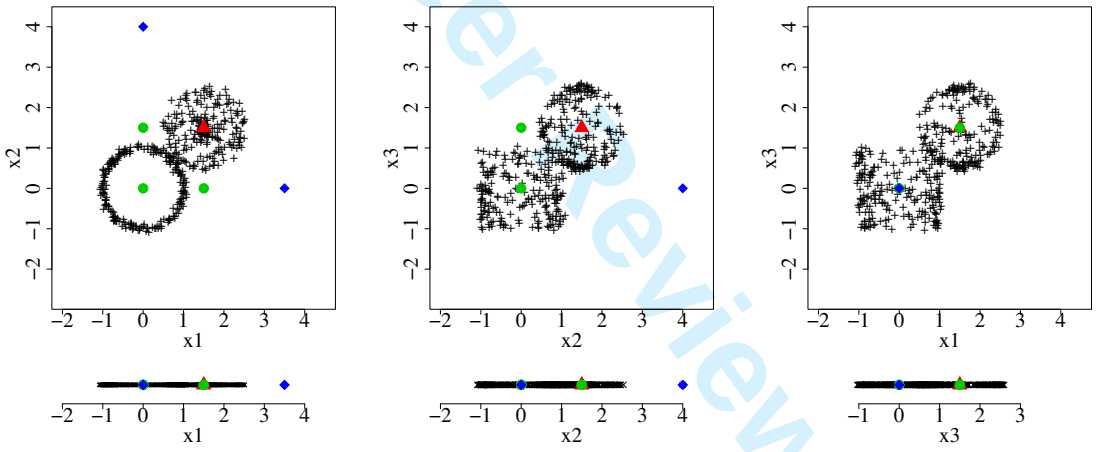

Figure 1: Two- and one-dimensional plots of the dataset that we will use as a motivating example. There are two groups of ordinary objects: a first group lying (subject to additional noise) on a sphere with radius 1 and centre $(1.5, 1.5, 1.5)$ and those lying (again subject to additional noise) on a cylinder with radius 1, height 2 and centre at the origin $(0, 0, 0)$. There are three types of anomalies: those that can be discerned in either one- (blue), two- (green) or three-dimensional (red) subsets of the original data variables.

1.2 Existing and similar work

Our method reduces the problem of detecting anomalies over a multi-dimensional feature space to one of detecting anomalies in many lower-dimensional subspaces. Further, our method will use a local density, nearest neighbour based anomaly score calculation algorithm (but this could be replaced by any algorithm computing a numerical anomaly score). In particular we will use 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 anomaly score. 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 computing the neighbourhood set of a given object in an incremental fashion. While this improves LOF, it is also much more computationally intensive and has only been shown to outperform LOF on contrived datasets (such as straight lines). Local Density Factor (LDF; Latecki et al. (2007)) is very similar to LOF, but uses kernel density functions to compute density estimates, rather than 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 specified. In practice, it can prove difficult to set this parameter. LOCI (Papadimitriou et al., 2003) improves on LOF by computing radius-dependent anomaly scores [referred to as multi-granularity deviation factors (MDEF) by the authors] for all objects for various radii. If, for any object, its MDEF at any radius exceeds three times the standard deviation of the MDEF at that radius, the object is flagged as an anomaly. LOCI thus produces binary anomaly / non-anomaly labels. Theoretically, for any object, the MDEF scores at the various radii could be combined (e.g. by averaging, selecting the largest, ...) to produce anomaly scores. How best to combine the various MDEF scores for each object is, however, not clear. Angle-based outlier detection (ABOD; Kriegel et al. 2008), computes, for a given object, the variance of the weighted angles between the difference vectors of the point to any other two points in the dataset. The underlying assumption is that for objects within a cluster, these angles differ widely, resulting in a large variance, whereas for outliers the angles do not vary much. The complexity of this approach is cubic in sample size and even a faster, nearest neighbours based version, fastABOD, described by the authors has complexity worse than the brute-force LOF algorithm. Furthermore, this method is not applicable to datasets with missing values.

DB, LOF, COF, LDF, LOCI and fastABOD compute distances between objects in order to compute anomaly scores. 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. ABOD and fastABOD also compute scalar products between difference vectors, which, again, requires data with no missing values.

Methods that reduce the dimensionality before checking for anomalies, usually use principal component analysis (PCA; Jolliffe 2002). PCA projects the data onto directions of decreasing variance so that the first principal component corresponds to the direction of maximum variance. However, PCA cannot be performed if the data contain missing values. As we wish to analyse datasets containing missing values, PCA based techniques are not applicable.

Seidl et al. (2009) use subspace clustering to look for anomalies. However, this approach suffers from the usual drawbacks of clustering based techniques (not optimised for outlier detection, sensitive to the definition and the number of clusters,...) and, at present, is unable to accommodate data with missing values. Lazarevic and Kumar (2005) select a random number of subspaces of random dimensionality and compute LOF scores for all objects in these subspaces. An overall anomaly score is obtained by combining the ranks of the objects' LOF scores in the different subspaces, or alternatively, but summing all anomaly scores for all objects. The random selection of subspaces means that some anomalies can go undetected. Also the two proposed anomaly score combination procedures are not suited for data with missing values. Kriegel et al. (2009) aim to detect anomalies in axis-parallel

subspaces of the data. For each object, they define a reference set of points that will be used to assess its outlierness. This outlierness is then assessed in the subspace consisting of the data variables in which the reference set objects exhibit low variance. An object's anomaly score is then the normalised distance to the subspace hyperplane of its reference set. This technique is highly reliant on the choice of reference objects, especially if an object is anomalous in more than one given subspace. Also, at present this technique is not able to handle data with missing values. Aggarwal and Yu (2005) discretise features into ϕ bins with each bin covering an equal number of objects. The authors the fix a dimensionality $1 \leq D \leq N_x$ and aim to find the most sparse D-dimensional cubes among the $\binom{N_x}{D}\phi^D$ cubes given by the grid defined on the data variables. This method computes binary anomaly labels rather than anomaly scores, but can be applied to datasets with missing values as for each object we only need to consider the subspaces in which it has been observed. In practice we have found this approach to be very sensitive to the choice of the parameters ϕ and D.

2 CASOS: 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 in the dataset
N_x - the number of variables of the dataset
D - the maximum dimensionality of the subspaces ($1 \leq D \leq N_x$)
AS - anomaly score (we assume the more anomalous an object is, the higher its AS)
MV - missing value

Algorithm 1 Proposed approach

```
1. for i in 1 : D
2.       for j in 1 : (N_x choose i)
3.            compute ASs 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
```

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 problems posed by high-dimensional feature spaces (commonly referred to as the "curse of dimensionality"). In practice, any AS computation algorithm can be used with our approach. For this work we have used LOF to compute anomaly scores.

2.1 Combination functions and required properties

The key step in our approach is step 6 in algorithm 1 above. It is by – sensibly – combining, for each object, the ASs 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 ASs from the observed subspaces then objects with many observed variables are more likely to have high ASs 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:

$N_D = \binom{N_x}{D}$, the number of distinct subspaces of dimension D in a N_x-dimensional dataset

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

$\mathcal{G} = \{X \,|\, X \text{ an } n \times N_D \text{ AS matrix}\}$, the set of all $n \times N_D$ 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 3-5 below, it is termed *well-behaved*.

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

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

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

This property guarantees that objects with many missing values have combined ASs 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 ASs. For example, if we were to combine ASs by summing all the non-missing ASs for each object, then an object with, say, 5 non-missing ASs 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. At the end of section 2.1.1 we will illustrate how this property, and indeed properties 2-5, are satisfied for a few examples of combination functions.

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

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

Together with property 1, this property means an object has a missing combined AS if and only if all of its subspace-specific ASs 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 (ASs inequality for comparable objects)

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

$$\begin{cases} X_{i_1,j} \leq X_{i_2,j} & \forall j \in \{k \in \{1,\dots,N_D\} | X_{i_1,k} \neq \text{NA}, X_{i_2,k} \neq \text{NA}\} \\ X_{i_1,j} = X_{i_2,j} = \text{NA} & \forall j \in \{k \in \{1,\dots,N_D\} | X_{i_1,k} = \text{NA} \text{ or } X_{i_2,k} = \text{NA}\} \end{cases}$$

we have that

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

This property simply means that if an object's ASs 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 ASs of other objects)

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

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

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 ASs, in particular by increasing one of its ASs (i.e. by making that object more anomalous in one subspace), then the combined ASs 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 (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 5 will be called *continuous*.

2.1.1 Examples of combination functions

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

- Selecting the highest AS:

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

- Averaging the ASs:

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

- Averaging the top N ASs:

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

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

- Sum of the excess above a certain quantile:
For each $j \in 1, \ldots, N_D$ let $q_j^{(1-\alpha)}$ be the $(1-\alpha)$ quantile of the ASs recorded for subspace j. For

all j, we subtract $q_j^{(1-\alpha)}$ from the ASs for that subspace. Finally, for each object, we sum the non-negative values.

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

where $(X)^+ = \max(X, 0)$.

- 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 ASs 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 ASs for that subspace. Finally, for each object, we sum the non-negative values.

$$\rho^{(midquant)}{}_i(X) = \sum_{X \in S_i} \left[\left(X - q_j^{(1-\alpha_1)}\right)^+ - 2\left(X - q_j^{(1-\alpha_2)}\right)^+\right]^+$$

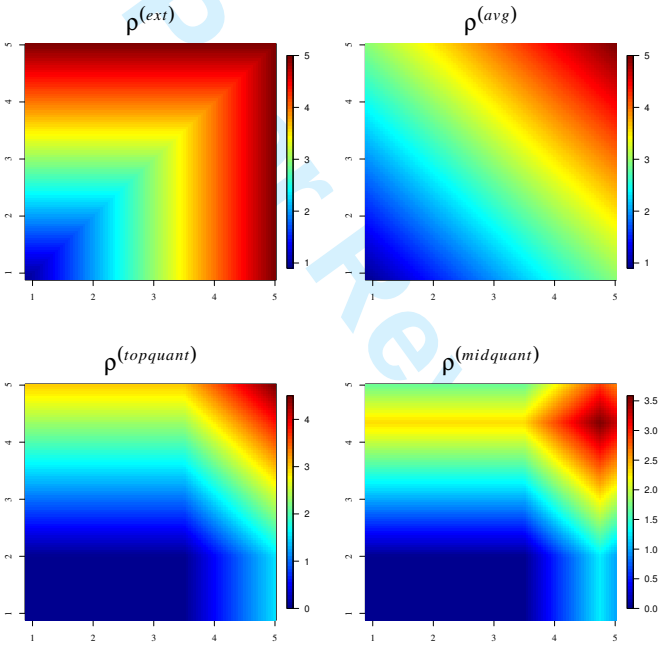

Figure 2: Examples of combination functions; the axes represent two different ASs and the colour scale indicates the magnitude of the combined AS.

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.

To show that $\rho^{(ext)}$, $\rho^{(avg)}$ and $\rho^{(topN)}$ satisfy properties 3-5 it is enough to show that $\rho^{(topN)}$ satisfies them, as in the extreme cases for N (i.e. averaging all non-missing ASs or simply selecting the highest AS), $\rho^{(topN)}$ is equivalent to $\rho^{(avg)}$ and $\rho^{(ext)}$, respectively. $\rho^{(topN)}$ satisfies property 3 trivially (since $\forall N$, if $x_1 \leq y_1, \ldots, x_N \leq y_N$ then $(x_1 + \ldots + x_N)/N \leq (y_1 + \ldots + y_N)/N$). Property 4 is also trivially met as $\rho^{(topN)}$ only involves the ASs of the given object and none of the ASs of the other objects (so that equality of the combined ASs holds $\forall i \neq i_0$ in property 4). To show that $\rho^{(topN)}$

is continuous we need to show that $\rho^{(topN)}$ is continuous over $\mathbb{R}^{N_D}, \forall N_D$ and $\forall N = 1, \ldots, N_D$. The ordering operation does not affect which numerical values the top N ASs can take, and so we only need to show that averaging N values is continuous. This again is trivial and so $\rho^{(topN)}$ is continuous, as defined by property 5.

$\rho^{(topquant)}$ satisfies property 3, since $\forall N_D$, $x_1 \leq y_1, \ldots, x_{N_D} \leq y_{N_D} \Rightarrow (x_1 + \ldots + x_{N_D}) \leq (y_1 + \ldots + y_{N_D})$ and since subtracting the same value $q_j^{(1-\alpha)}$ from x_j and y_j (respectively setting $x_j = 0$ or $x_j = y_j = 0$), for each $j = 1, \ldots, N_D$, does not change those inequalities. Property 4 holds for $\rho^{(topquant)}$, since $X_{i_0,j_0} \leq \tilde{X}_{i_0,j_0}$ implies that the quantile $q_{j_0}^{(1-\alpha)}$ is higher for $\tilde{X}$ and so, even though all the other elements of X and $\tilde{X}$ have the same numerical values, we subtract higher values from the $\tilde{X}_{i,j_0}$ (and set more of them to zero). This means that $\rho_i(X) \geq \rho_i(\tilde{X})\ \forall i \neq i_0$. For i_0, since $X_{i_0,j_0} \leq \tilde{X}_{i_0,j_0}$, and all other values are equal, and since the difference $\tilde{X}_{i_0,j_0} - X_{i_0,j_0}$ is higher than the corresponding difference in quantiles, we have that $\rho_{i_0}(X) \leq \rho_{i_0}(\tilde{X})$ and so property 4 holds. That $\rho^{(topquant)}$ is continuous in the two-dimensional case is clear from figure 2 and this easily extends to higher-dimensional cases.

$\rho^{(midquant)}$ is not well-behaved as it does not satisfy properties 3 and 4. It is, however, continuous (see figure 2). 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 ASs above the second quantile equal to zero for $\rho^{(midquant)}$.

As we have already explained, property 1 needs to be satisfied to guarantee comparability of the combined ASs and property 2 guarantees non-missing combined ASs 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 3-4, combination functions such as $\rho^{(midquant)}$ above allow one to effectively discard ASs which are too extreme and focus on sources with consistently high but not extreme ASs. 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 5 would not be needed.

3 Properties of CASOS

Having defined CASOS in section 2, we can now look at further properties of our approach.

Flexibility

Through the choice of combination function, our method can very easily be adapted to specific needs. For instance $\rho^{(avg)}$ will find objects which either have very large ASs in some subspaces, or which have consistently high ASs. However, $\rho^{(avg)}$ can be affected by the masking effect from irrelevant features since it averages ASs over all available subspaces. $\rho^{(ext)}$ will be less affected by irrelevant features, as it is enough for an object to have a high AS in a single subspace to have a high combined AS. However, if a dataset contains both noise objects (e.g. cosmic rays in astronomical datasets) and objects which are physically anomalous, both $\rho^{(ext)}$ and $\rho^{(avg)}$ will result in a high combined AS for noise sources, as their measured values are often highly outlying. This can, in turn, affect the detection of true, non-noise anomalies. $\rho^{(midquant)}$ can adjust for this, as it essentially ignores too extreme ASs during the combination step. Thus one can design a combination function which will suit whatever beliefs

one might hold about a particular datasets. All combination functions should be checked, however, to see that they satisfy properties 1 and 2.

Transition between one-variable-at-a-time and full-dimensional approach

The case when $D = 1$ corresponds to a one-variable-at-a-time approach, whereas $D = N_x$ is equivalent to the full-dimensional approach. Thus our approach can be regarded as an intermediary approach between these two, including them as special cases.

In the astronomy setting, if we only use magnitude variables (i.e. measures of brightness), then the one-dimensional AS vectors will be of little use as they will merely flag up very bright and / or very faint objects. Indeed, the individual magnitude variables have low-density regions only at the upper and lower ends of their range. CASOS can be used with $D = 1$ to check the dataset for objects with physically impossible values, but such a quality-control step should, ideally, have been performed prior to the actual data analysis.

Approximate approach (if required)

As we will see in section 3.2, for given N_x and D, CASOS has to compute ASs in $\binom{N_x}{D}$ subspaces. For large N_x and $D \simeq N_x/2$, this can become computationally prohibitive. One solution to this problem consists in sacrificing exactness for speed and, similar to Lazarevic and Kumar (2005), compute ASs in a random subset of subspaces only.

However, for this work, we wish to have exact results and have therefore not investigated such an approximate approach further.

Interpretability

As we noted previously, for high-dimensional data, it can be difficult to see why a given source has been declared anomalous by a full-dimensional method. With CASOS it is possible to check the individual anomaly scores that led to a high combined anomaly score. Thus the subspaces in which a given source is particularly anomalous can be singled out. If the dimensionality of these subspaces is low (e.g. $D = 2$ or $D = 3$), then it is possible to pinpoint the exact reason why the combined anomaly score computed by CASOS is large.

3.1 Analysis of the motivating example

The dataset from the motivating example introduced in section 1.1 features six anomalies, labelled $o1, o2, o3, o4, o5$ and $o6$, and it also contains missing values. $o1$ appears anomalous only when all three data variables are considered jointly. $o2, o3$ and $o4$ are two-dimensional anomalies, but $o3$ has a missing value in one of the data variables. Finally $o5$ and $o6$ can be seen to be anomalous by considering only one variable, but for $o5$ only two attributes have been recorded.

We apply CASOS with $D = 1, 2, 3$, the latter being equivalent to the full-dimensional LOF approach (Breunig et al., 2000), LDF (Latecki et al., 2007), LOCI (Papadimitriou et al., 2003), the method described in Aggarwal and Yu (2005) and fastABOD (Kriegel et al., 2008) to this dataset and list the results in table 1. For CASOS we have used $\rho^{(avg)}$ as combination function. ABOD was too slow even for this small, low-dimensional dataset, hence why we have used fastABOD instead. For all methods, except Aggarwal and Yu (2005) which is not a nearest neighbour approach, we have used $k = 10$ nearest neighbours (considering each point to be the nearest neighbour of itself). For LDF we have used bandwidth $h = 1$ and for the method from Aggarwal and Yu (2005), we have discretised each feature into three bins and we have calculated sparsity scores for two-dimensional, rectangular regions. We have used the threshold recommended by Aggarwal and Yu (2005) to flag anomalies.

Table 1: Various anomaly detection methods applied to the dataset described in section 1.1. Checkmarks indicate successful detection.

a: objects with the 6 highest ASs have been flagged as anomalies

b: objects with the 4 highest ASs (resp. smallest ASs, for fastABOD) have been flagged as anomalies

c: for binary anomaly / non-anomaly labels, we report those objects flagged as anomalies; note that LOCI gave 1 false-positive result whereas the method from Aggarwal and Yu (2005) gave 41 false-positives

anomaly	$o1$	$o2$	$o3$	$o4$	$o5$	$o6$
type	3D	2D	2D + MV	2D	1D + MV	1D
[a]CASOS, $D = 1$					✓	✓
[a]CASOS, $D = 2$		✓	✓	✓	✓	✓
[b]CASOS, $D = 3$ / LOF	✓	✓				✓
[b]LDF						✓
[c]LOCI	✓	✓				✓
[c]Aggarwal and Yu (2005)			✓	✓	✓	✓
[b]fastABOD	✓	✓		✓		✓

Note that CASOS and the method from Aggarwal and Yu (2005) have been applied to the full dataset, whereas LDF, LOCI and fastABOD have only been applied to the reduced dataset of 373 objects with no missing values. Also note that CASOS with $D = 3$, i.e. LOF, has been applied to the full dataset, but only provides anomaly scores for the data with no missing values. Likewise, CASOS with $D = 2$ and the method from Aggarwal and Yu (2005) only provide meaningful results for the subset of objects with at least two observed variables. For all AS-based methods we have flagged as many objects as there are anomalies in the dataset (six for the full dataset; four for the reduced dataset).

We note that only methods capable of handling data with missing values are able to detect anomalies $o3$ and $o5$, which both have a missing value. This is a clear advantage for such methods. Further, we note that only CASOS (with $D = 2$), the method from Aggarwal and Yu (2005) and fastABOD are capable of detecting anomaly $o4$. CASOS with $D = 3$ would have detected this anomaly if we had chosen to flag the top five anomaly candidates, rather than the top four. However, since in real datasets we do not know a priori which are the anomalies, CASOS has failed to detect this anomaly for the given threshold that we applied. Since CASOS with $D = 2$ and the method from Aggarwal and Yu (2005), which we have also used in a two-dimensional setting, can detect this anomaly, this suggests that the detection of $o4$ is affected by the additional feature (which is irrelevant for declaring $o4$ as anomalous) when we work in the full three-dimensional space. We also note that an inherently three-dimensional anomaly such as $o1$ can only be detected in a full-dimensional approach. It is worth pointing out the problem that can be posed by binary anomaly / non-anomaly labels: while LOCI returns only one false-positive, the method from Aggarwal and Yu (2005) yields a total of 45 potential anomalies, of which only four turn out to be true anomalies. Finally we note that of the six methods compared in table 1, CASOS with $D = 2$ works best, which shows that lower-dimensional approaches can outperform their full-dimensional counterparts.

We conclude that this example shows both the advantages (ability to handle data with missing values, ability – for some datasets at least – to outperform full-dimensional approaches, ability to avoid the masking effect from irrelevant features) and limitations (unability to detect full-dimensional anomalies) of our approach.

3.2 Complexity / speed

The number of subspaces required to work through by our approach is $\sum_{i=1}^{D} \binom{N_x}{i}$. If we compute ASs only for D-dimensional subsets of the original data variables, we still have to work through $\binom{N_x}{D}$ subspaces. Depending on D and N_x, this can be very large! Figure 3 shows the number of subspaces as a function of D (the maximum dimensionality of subspaces) and N_x (the number of features).

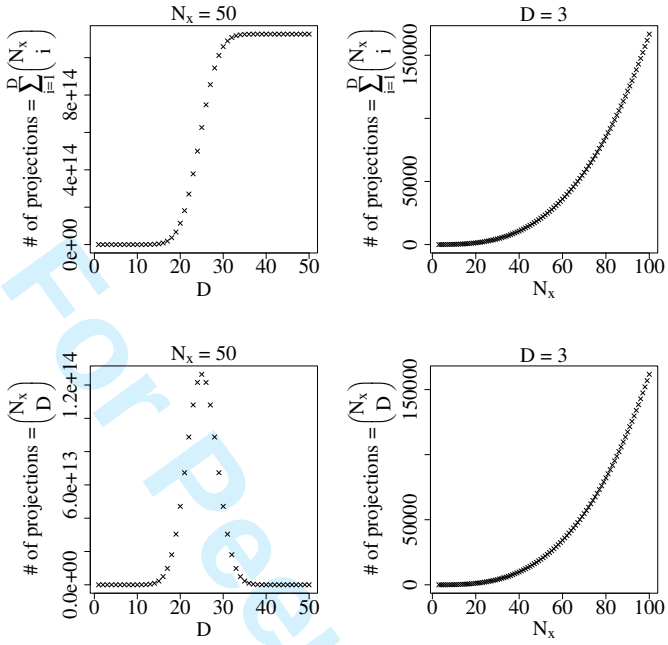

Figure 3: number of subspaces as a function of D and N_x

The choice of D is obviously dependent on N_x and the computational resources available. If $N_x > 100$, choosing $D > 3$ would probably be unwise for most systems.

For $D \leq N_x/2$, the complexity of the proposed approach is $O\left(D \cdot k \cdot n \cdot \log(n) \cdot {N_x}^D\right)$, which is clearly dominated by the $O({N_x}^D)$ factor for large values of D and / or N_x.
[For $D > N_x/2$, the complexity is $O\left(D \cdot k \cdot n \cdot \log(n) \cdot {N_x}^{N_x - D}\right)$.]

Figure 4 shows the average time required by different anomaly detection methods for a set of multivariate normal data with zero mean vector and identity covariance matrix. For each set of parameters (sample size, number of variables, percentage of missing values and number of nearest neighbours) we have generated 8 datasets from the distribution with these parameters and then averaged the wall clock times taken for each method.

The only two methods that can handle data with missing values are CASOS and the method from Aggarwal and Yu (2005). Looking at figure 4 and comparing just these two methods, CASOS is seen to be faster than the method from Aggarwal and Yu (2005), at least for the given choices of combination function ρ and number of bins per variable. In particular the method from Aggarwal and Yu (2005) scales worse with sample size. However, the method from Aggarwal and Yu (2005) is not a nearest neighbour technique and hence its speed does not vary when k is varied. Both methods, while able to address the problems posed to anomaly detection by high-dimensional techniques are exponential in the number of variables. This, at present, is a severe practical limitation. Looking at the effect of the percentage of missing values, it is clear that, apart from a steep initial decrease in speed, both methods perform faster as the percentage of missing values increases.

Comparing to full-dimensional techniques, LOF is seen to be the fastest method overall, with LDF second fastest. The speed of LDF, however, deteriorates rapidly with increasing number of nearest

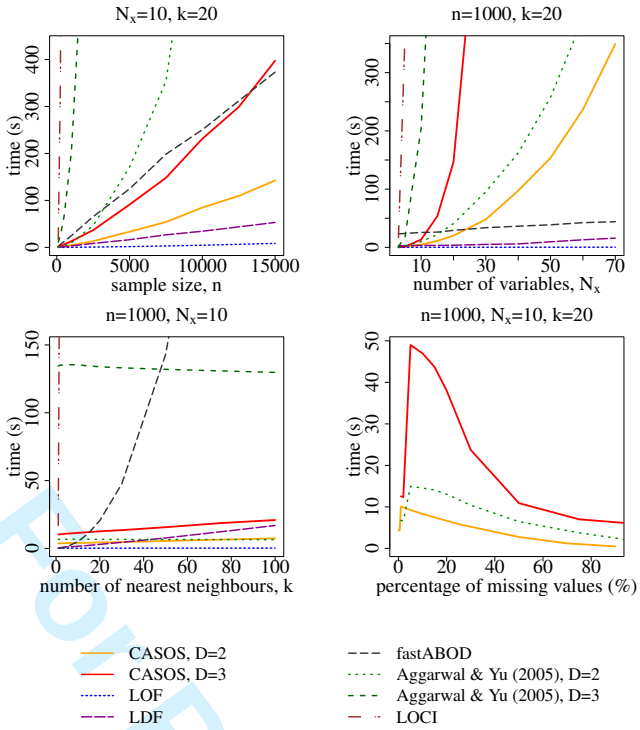

Figure 4: The average time required by different anomaly detection techniques, including CASOS, to analyse a data of multivariate normal data with zero mean vector and identity covariance matrix. CASOS has been used with combination function $\rho^{(avg)}$ and, for the method from Aggarwal and Yu (2005), we have used four bins per variables.

neighbours. fastABOD scales well with the number of dimensions, but is exponential in the number of nearest neighbours. fastABOD and CASOS are both of order $O(n \cdot \log n)$, for sample size n. We found the exact LOCI algorithm to be extremely slow.

4 Empirical Evaluation

We have evaluated the performance of CASOS on both simulated and real datasets and we have used the following metrics to assess the performance of the different methods:

- the completeness, also known as the detection or true positive rate

$$\text{completeness} = \frac{\#\text{ of true anomalies flagged as anomalies}}{\#\text{ of true anomalies}}$$

- the efficiency

$$\text{efficiency} = \frac{\#\text{ of true anomalies flagged as anomalies}}{\#\text{ of flagged objects}}$$

The completeness gives the fraction of true anomalies that have been detected, whereas the efficiency is the fraction of true anomalies within a set of top anomaly candidates. Different applications prioritise these two measurements differently. Sometimes it is better to accept low efficiency in order to detect all interesting objects (i.e. achieve high completeness), sometimes it is vital not to waste resources on false positives, in which case achieving high efficiency is paramount.

The simulated datasets that we present illustrate two situations in which a lower-dimensional approach such as CASOS can outperform full-dimensional approaches, and we also present one dataset simulated so as to be similar to cross-matched astronomical survey data.

4.1 Performance on simulated datasets

4.1.1 Dataset 1

A first simulated dataset consists of 5100 sources, of which 100 anomalies, with 30 data attributes. For the non-anomalous sources, all variables have been sampled independently from a standard normal $\mathcal{N}(0,1)$ distribution. The 100 anomalies are anomalous in 3 variables only, and these variables have been sampled independently from a normal distributions with larger variance [$\mathcal{N}(0,4)$].

To assess the average performance of the different methods for datasets generated in the way described above, we have sampled 20 such datasets and computed the average performance for the different methods.

We have used $k = 25$ for all methods. CASOS has been used with $\rho^{(avg)}, \rho^{(ext)}$ and $\rho^{(topquant)}$, and we have set $D = 2$. Figure 5 summarises the results.

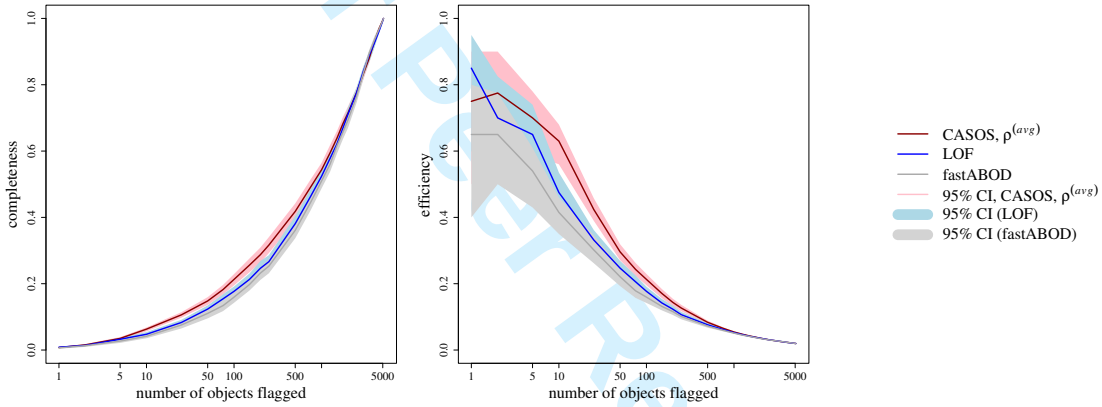

Figure 5: CASOS, LOF and fastABOD applied to 5100 simulated data, of which 100 have been simulated to be anomalous; $k = 25$ for all methods and $D = 2$ for CASOS.

CASOS is seen to marginally outperform both LOF and fastABOD for this data. Note that, for clarity, we have only plotted the results for CASOS with $\rho^{(avg)}$. Using $\rho^{(ext)}$ or $\rho^{(topquant)}$ produces slightly worse performance for CASOS, but still better than LOF, on average. However only for $\rho^{(avg)}$ the difference in performance with LOF is significant as indicated by the non-overlap (for some numbers of objects flagged as anomalous) of the 95% confidence region.

We conclude that in the case of irrelevant features (only 3 variables relevant for anomaly detection in this case), CASOS is able to outperform full-dimensional methods.

4.1.2 Dataset 2

A second simulated dataset consists of $25,250$, 60–dimensional data, of which 250 are anomalies. The non-anomalous objects are consisting of two groups: 16667 are sampled from a 60–dimensional multivariate normal distribution with parameters means $\mu_1 = (1, 1, \ldots, 1)^T$ and covariance matrix Σ_1, and the remaining 8333 objects are sampled from another multivariate normal distribution, this time with means $\mu_2 = (-1, -1, \ldots, -1)^T$ and covariance matrix Σ_2. Both covariance matrices have non-zero off-diagonal elements.

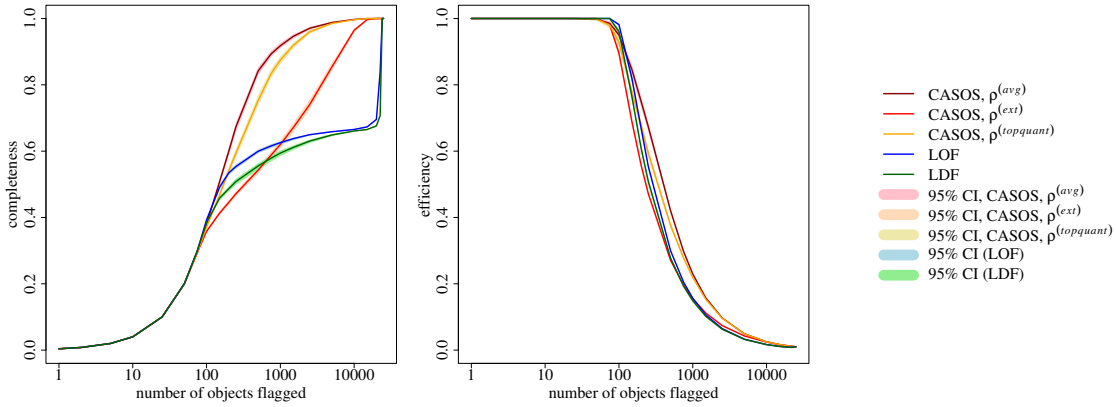

Figure 6: CASOS, LOF and LDF applied to 25, 250 simulated data, of which 250 have been simulated to be anomalous (with three types of anomalies); $k = 75$ for all methods and $D = 2$ for CASOS.

There are three types of anomalies, all of which are anomalous in only half of the data variables. The first type of anomalies (83 objects) have been sampled identically as the non-anomalous objects, but have then had independent $\mathcal{N}(0, 1)$ noise added in each of the first 30 variables. The second group of anomalies (83 objects), have been sampled identically to the non-anomalous objects but then have had their first 30 variables replaced by values sampled independently from $\mathcal{N}(0, 4)$, so that these objects effectively do not lie within the two groups of non-anomalous objects. Finally, the third group of anomalies (84 objects) have, once again, been sampled identically to the non-anomalous objects, but then have had their first 30 variables replaced by variables sampled from a 30–dimensional multivariate normal distribution with means $\mu_3 = (4, 4, \ldots, 4)^T$ and covariance matrix $\Sigma_3 = \frac{1}{16} \cdot I_{30}$, where I_{30} is the 30–dimensional identity matrix, so that these objects form a tight cluster well outside the feature space populated by most other objects.

We have used $k = 75$ for all methods and $D = 2$ for CASOS.

Figure 6 shows the completeness and efficiency for this dataset. As can be seen all methods perform similarly well, until more than 166 anomaly candidates are considered. For higher values of flagged objects, the performance of LOF and LDF decreases sharply relative to CASOS' performance. Since we have used $k = 75 < 84$, LOF and LDF are not able to detect the 84 anomalies that form a tight cluster of anomalies. For each such object, LOF and LDF consider only objects that lie also in this cluster of anomalies and thus consider these objects to be non-anomalous. CASOS however, by working in lower-dimensional subspaces, is able to 'capture', at least in some subspaces, objects outside this cluster of anomalous objects. As a result the objects within the cluster get much larger anomaly scores and are flagged as anomalous.

Figure 7 illustrates this point further: on the left-hand-side plot we keep all parameters (including the proportion of anomalies) fixed, and only vary the sample size. Once the sample size gets so large that the number of anomalies exceeds the number of nearest neighbours, the performance of all methods decreases sharply. However for CASOS, at least with $\rho^{(avg)}$ and $\rho^{(topquant)}$, this decrease is much less than for LOF and LDF. On the right-hand-side plot we keep all parameters fixed, but vary the number of nearest neighbours used with all methods. Again, CASOS outperforms LOF and LDF for small k.

We conclude that CASOS is less dependent on a the choice of k as full-dimensional approaches such as LOF and LDF.

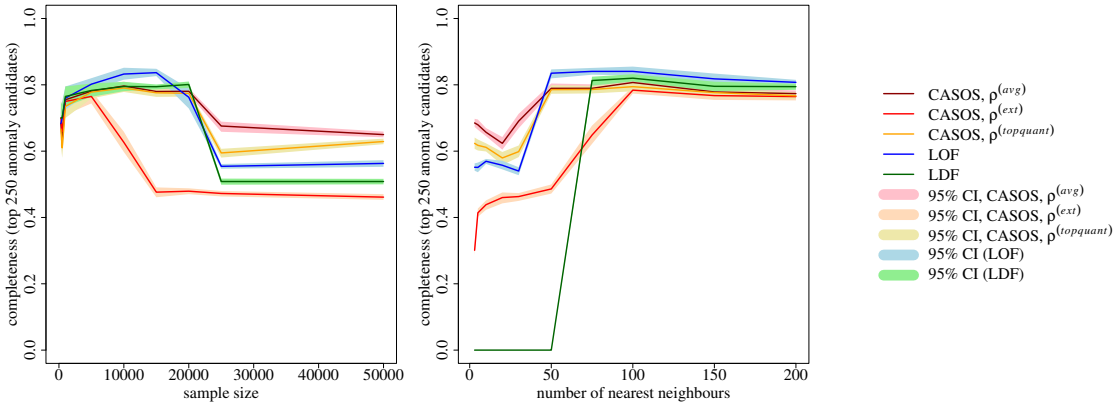

Figure 7: CASOS, LOF and LDF applied to 25, 250 simulated data, of which 250 have been simulated to be anomalous (with three types of anomalies); $D = 2$ for CASOS. Performance is shown for varying sample size and k.

4.1.3 Dataset 3: simulated, cross-matched astronomical survey data

In section 4.2.1 we will analyse a dataset obtained by cross-matching survey data from the Sloan Digital Sky Survey (SDSS; York et al. 2000) and the UKIRT Infrared Deep Sky Survey (UKIDSS; Lawrence et al. 2007). To evaluate how CASOS performs on such data, we have first simulated cross-matched SDSS-UKIDSS data.

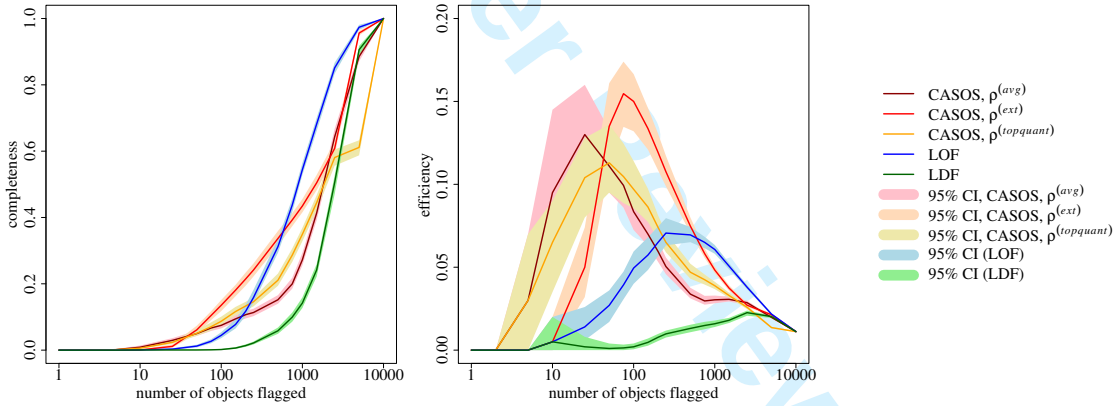

Figure 8: CASOS, LOF and LDF applied to 10, 100 simulated, cross-matched SDSS-UKIDSS data, of which 100 have been simulated to be anomalous; $k = 100$ for all methods and $D = 7$ for CASOS.

The cross-matched dataset that we will analyse in section 4.2.1 has 17 measured variables: 8 colour variables and 9 morphology statistics. In astronomy, magnitudes measure the apparent brightness of sources. Large magnitude values correspond to faint sources, whereas small values correspond to bright sources. Colours are obtained by taking differences of magnitudes from different bands (technically they are ratios of fluxes). It is a convention in astronomy to report magnitudes as the filter names (i.e. u rather than m_u, etc.).

SDSS observes astronomical sources in five optical filter passbands: u, g, r, i, z and UKIDSS uses four near-infrared bands: Y, J, H and K. The 8 colour variables we have used are $u - g$, $g - r$, $r - i$, $i - z$, $z - Y$, $Y - J$, $J - H$ and $H - K$. UKIDSS measures an extendedness statistics called `ClassStat`, which is derived from the curves-of-growth of sources. `ClassStat` characterises the shape of a source as it appears on an image of the sky and is a useful statistic to differentiate between point-like sources (i.e. stars) and "fuzzier" objects (i.e. galaxies). SDSS does not record this statistic,

but a similar statistic can be obtained by computing the difference between the point-spread function (PSF) magnitude and the best fit galaxy profile model magnitude for each source. This statistic is usually referred to as concentration (e.g. Scranton et al. 2002).

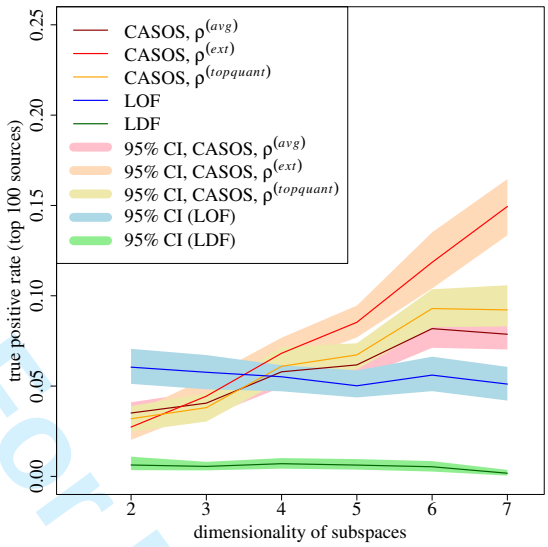

Figure 9: Same data simulation mechanism as in figure 8, but with D varied for CASOS.

For simplicity, to generate the SDSS data, we have simulated `ClassStat` variables for the five SDSS bands, rather than concentrations. The exact details of the data simulation algorithm are described in Henrion et al. (2011).

We have not included any measurement of the apparent brightness of sources, as otherwise anomaly detection methods will flag a large proportion of very bright sources. Such sources are rare, hence it is good that they are detected by the anomaly detection techniques, however, they are not interesting in the context of finding anomalous astronomical objects, as such sources have been observed by many previous surveys and are well-studied.

For each of 20 runs, we have simulated a set of $10,000$ sources, with an additional 100 anomalous sources. Anomalies have been simulated such as to represent compact galaxies: sources with colours of galaxies, but morphologies of stars. We have used $k = 100$ for all methods and $D = 7$ for CASOS. Figure 8 plots the completeness and efficiency for this dataset.

While no method performs really well (best achieved efficiency of ~ 0.15), CASOS performs best if low numbers of top anomalies are considered (i.e. at most the top ~ 500 sources). This corresponds to the realistic case, where astronomers use anomaly detection as a tool to guide the investigation and will have to select a set of anomaly candidates for follow-up observations to confirm their exact nature. As such observations are expensive, only a small subset of sources will be selected for follow-up. For the top 100 anomaly candidates CASOS, especially when used with $\rho^{(ext)}$, can be seen to detect about three times as many true anomalies as the full-dimensional approach.

Figure 9 shows how the performance of CASOS varies with D. For computational reasons, we have not evaluated the performance of CASOS for $D > 7$. While the performance is seen to be increasing with D from $D = 2$ to $D = 7$, it will eventually decrease with increasing D to equal the performance of full-dimensional LOF when $D = N_x = 17$. Thus for this dataset the optimal dimensionality to detect anomalies lies between 7 and 17.

4.2 Performance on real datasets

We have tested CASOS on two real, astronomical datasets.

4.2.1 Cross-matched SDSS-UKIDSS data

In section 4.1.3 we described the variables recorded by a cross-matched SDSS-UKIDSS dataset. We use data from a region called SDSS Stripe 82, for which there is a good overlap between SDSS and UKIDSS, and for which SDSS made multiple observations, thus providing much more information on the detected sources. After applying various data quality filters, our final dataset consists of 109,368 sources. In UKIDSS, if a source is too faint in a given band to be detected, it will have missing values for the variables from that band. In SDSS however, if a source is detected in any of the band, the SDSS data processing pipeline will re-extract the measurements for any band in which the source has not been detected. Thus the SDSS data contains no missing values. However as those measurements contain little useful information, we have re-introduced missing values in the SDSS data, by setting those detections to missing for which the signal–to–noise ratio is less than 5.

Due to the missing values contained in this dataset it has not been possible to apply full-dimensional approaches such as LOF, LDF or fastABOD to this dataset. All results are for CASOS (with $D = 2$ and $\rho^{(avg)}$) only.

Figure 10 shows images in each of the four UKIDSS bands (Y, J, H, K) for the top three anomaly candidates. While, from these images, it is difficult to say why the first two objects have high combined anomaly scores, the photometry for the third source looks as if it could have been affected by noise artifacts. While such sources are physically not interesting, we would expect any anomaly detection technique to detect such sources.

Figure 11 shows low-resolution spectra for six of the top '100 anomaly candidates. The spectra have been obtained by using the measured SDSS and UKIDSS magnitudes. Many of the anomaly candidates that CASOS reported exhibit the characteristics of the top row spectra in figure 11: some of the SDSS magnitudes are extremely faint compared to the other measured magnitudes. These are in fact the re-extracted measurements whenever a source was too faint to be detected in a given band. While we have tried to re-instate the original missing values, we have not been able to do so for all sources. Again, the physical properties of the majority of such sources will not exhibit any true anomalousness. However any anomaly detection method should flag sources with such measured data, as they are anomalous in the feature space that has been analysed.

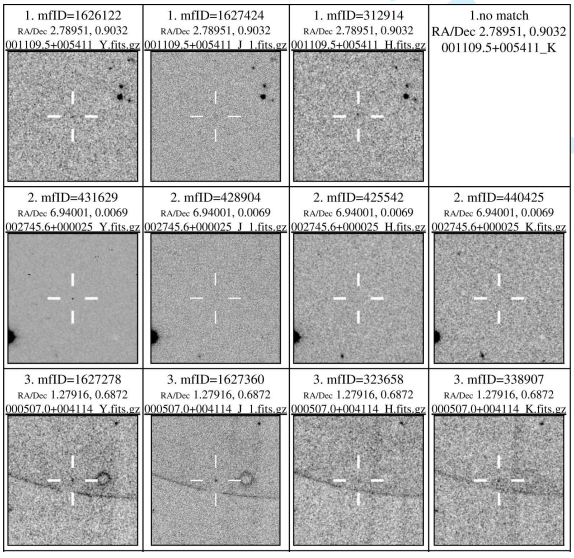

Figure 10: Top three anomaly candidates as observed in the UKIDSS Y, J, H and K bands. Images obtained from the WFCAM Science Archive (Hambly et al. 2008; `http://surveys.roe.ac.uk/wsa/`).

Some of the other spectra are more interesting. For instance, the spectrum in the middle of the

bottom row shows a source which is much brighter in the optical SDSS bands than in the near-infrared UKIDSS bands. For other sources, such as the ones whose spectra are the shown on the left and right of the bottom row in figure 11, it is not possible to explain why they have been flagged by looking at their spectra alone.

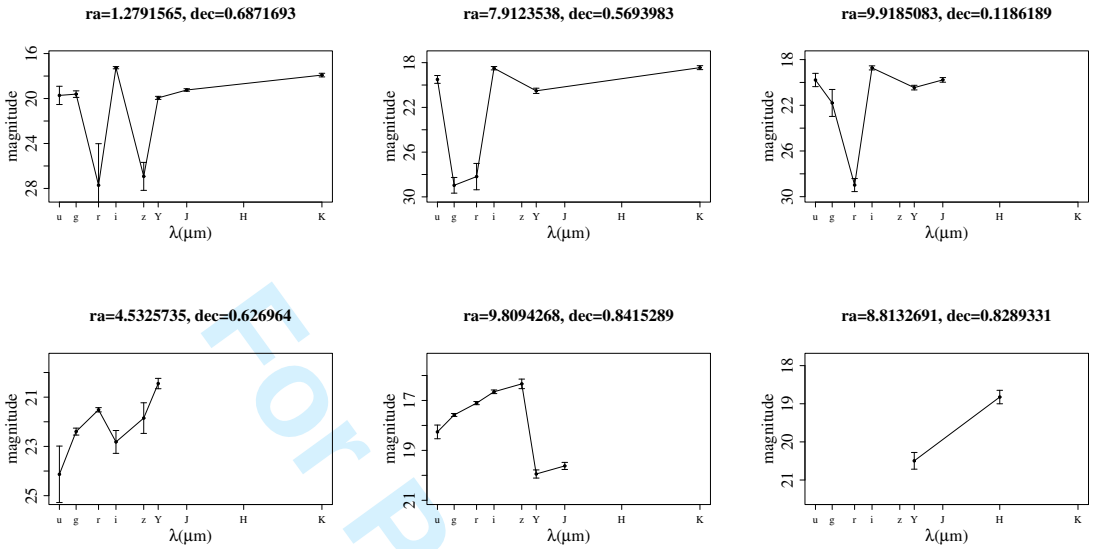

Figure 11: Low-resolution spectra (made up from the original SDSS and UKIDSS survey data) for six of the top 100 anomaly candidates from the cross-matched SDSS-UKIDSS data. The error bars represent the errors reported by the SDSS and UKIDSS data processing pipelines.

We conclude that CASOS does show some potential on cross-matched survey data, but, as we stated in section 1, any sources flagged by CASOS would need to be observed to greater detail with follow-up observations. However it will require much more focused input from astronomers to get meaningful anomaly candidate lists, as otherwise CASOS will flag up noisy sources, or sources with weird, but uninteresting measurements (such as the artificially low SDSS magnitudes).

4.2.2 Quasar candidate dataset

We have also applied CASOS to a set of 12,074 pre-selected quasar candidates. This dataset is described in greater detail in Mortlock et al. (2011). The dataset contains 5 colour variables: $i-z, z-Y, Y-J, J-H$ and $H-K$. Objects have been pre-selected to lie in a certain region in $i-Y$ vs. $Y-J$ space (shown on figure 12).

The dataset contains 7 confirmed high-redshift quasars. The aim was to see if CASOS would be able to detect these. We have used CASOS with $D=2$ and $\rho^{(avg)}$. Figure 12 shows the top 120 anomaly candidates from CASOS.

Flagging the top 1% of sources as anomalous (121 sources), CASOS detects one of the seven high-redshift quasars. The probability of this occurring by chance is 0.0679. However when the top 10% of sources (1,207 sources) are flagged as anomalous, then CASOS is able to detect five of the seven quasars. The probability of this occurring by chance is now $1.765 \cdot 10^{-4}$.

So this shows that there is some potential in CASOS. However, again, more guided input from astronomers will be needed, as the efficiencies (1/121, respectively 5/1207) are very low. Still, without using any additional information, CASOS managed to reduce the problem of finding 7 high-redshift quasars in 12,074 sources to finding 5 quasars in 1,207 sources. Comparing this to the result from

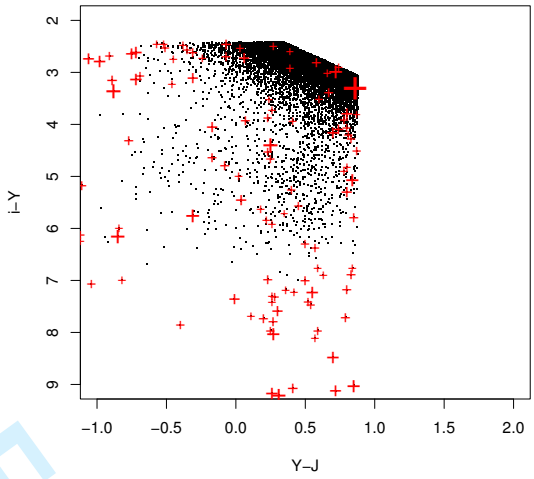

Figure 12: 12,074 pre-selected quasar candidates (compare to figure 1 from Mortlock et al. 2011). Red crosses indicate the top 120 anomaly candidates, with the size of the red crosses proportional to the combined anomaly score.

Mortlock et al. (2011), which was obtained by using Bayesian model comparison, CASOS is impressively competitive.

5 Conclusion

We have introduced a novel algorithm, CASOS, which performs anomaly detection in lower-dimensional subspaces of the data. the advantages of this algorithm are multiple:

- ability to directly use data with missing values
- addresses some of the problems of high-dimensional data spaces (such as the breakdown of the notion of anomaly and distance)
- less susceptible to the masking effect from irrelevant features
- the choice of combination function adds flexibility to adapt the method to the requirements of a particular dataset
- better interpretability

We should, however, also stress again that CASOS has the disadvantage that it will not be able to detect outliers which are only apparent in multivariate spaces with significant numbers of variables. But we believe such situations are rare, and normally outliers will be apparent in low dimensional spaces.

We have shown that CASOS can outperform state-of-the-art, full-dimensional methods, such as LOF, LDF and fastABOD an simulated datasets.

We have applied CASOS to three real datasets, in particular a set of cross-matched SDSS-UKIDSS data. While the results for the astronomy datasets look promising, CASOS needs to be supervised more closely by astronomers in order to get meaningful results, which would justify the costs of follow-up observations.

Acknowledgments

The results presented here would not have been possible without the efforts of the many people involved in the SDSS and UKIDSS projects.
We have implemented CASOS in the statistical programming language R (`http://www.r-project.org/`). Full source code for CASOS and for generating the various datasets described in this article are available from the authors upon request.
Marc Henrion was supported by an EPSRC research studentship, and David Hand was partially supported by a Royal Society Wolfson Research Merit Award.

References

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.

Aggarwal, C. C. and Yu, P. S. (2005). An effective and efficient algorithm for high-dimensional outlier detection. *The VLDB Journal*, 14(2):211–221.

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

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbor" meaningful? In Beeri, C. and Buneman, P., editors, *Database Theory ICDT99*, volume 1540 of *Lecture Notes in Computer Science*, pages 217–235. Springer Berlin / Heidelberg.

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.

Hambly, N. C. et al. (2008). The wfcam science archive. *Monthly Notices of the Royal Astronomical Society*, 384:637–662.

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

Henrion, M., Mortlock, D. J., Hand, D. J., and Gandy, A. (2011). A bayesian approach to star-galaxy classification. *Monthly Notices of the Royal Astronomical Society*, 412:2286–2302.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer, 2nd edition.

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

Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. (2009). Outlier detection in axis-parallel subspaces of high dimensional data. In Theeramunkong, T., Kijsirikul, B., Cercone, N., and Ho, T.-B., editors, *Advances in Knowledge Discovery and Data Mining*, volume 5476 of *Lecture Notes in Computer Science*, pages 831–838. Springer Berlin / Heidelberg.

Kriegel, H.-P., Schubert, M., and Zimek, A. (2008). Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*.

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., Warren, S. J., Almaini, O., Edge, A. C., Hambly, N. C., Jameson, R. F., Lucas, P., Casali, M., Adamson, A., Dye, S., Emerson, J. P., Foucaud, S., Hewett, P., Hirst, P., Hodgkin, S. T., Irwin, M. J., Lodieu, N., McMahon, R. G., Simpson, C., Smail, I., Mortlock, D., and Folger, M. (2007). The ukirt infrared deep sky survey (ukidss). *Monthly Notices of the Royal Astronomical Society*, 379:1599–1617(19).

Lazarevic, A. and Kumar, V. (2005). Feature bagging for outlier detection. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 157–166, New York, NY, USA. ACM.

Mortlock, D. J., Patel, M., Warren, S. J., Hewett, P. C., Venemans, B. P., and McMahon, R. G. (2011). Probabilistic selection of high-redshift quasars. *Monthly Notices of the Royal Astronomical Society*, in press.

Papadimitriou, S., Kitagawa, H., Gibbons, P. B., and Faloutsos, C. (2003). Loci: Fast outlier detection using the local correlation integral. In *Proceedings of the IEEE 19th International Conference on Data Engineering (ICDE'03)*. IEEE Computer Society.

Scranton, R., Johnston, D., Dodelson, S., Frieman, J. A., Connolly, A., Eisenstein, D. J., Gunn, J., Hui, L., Jain, B., Kent, S., Loveday, J., Narayanan, V., Nichol, R. C., O'Connell, L., Scoccimarro, R., Sheth, R. K., Stebbins, A., Strauss, M. A., Szalay, A., Szapudi, I., Tegmark, M., Vogeley, M., Zehavi, I., Annis, J., Bahcall, N. A., Brinkman, J., Csabai, I., Hindsley, R., Ivezic, Z., Kim, R. S. J., Knapp, G. R., Lamb, D. Q., Lee, B., Lupton, R. H., McKay, T., Munn, J., Peoples, J., Pier, J., Richards, G. T., Rockosi, C., Schlegel, D., Schneider, D. P., Stoughton, C., Tucker, D. L., Yanny, B., and York, D. G. (2002). Analysis of systematic effects and statistical uncertainties in angular clustering of galaxies from early sloan digital sky survey data. *The Astrophysical Journal*, 579(1):48–75.

Seidl, T., Müller, E., Assent, I., and Steinhausen, U. (2009). Outlier detection and ranking based on subspace clustering. In Koch, C., König-Ries, B., Markl, V., and van Keulen, M., editors, *Uncertainty Management in Information Systems*, number 08421 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.

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.