Variability search in M 31 using Principal Component Analysis and the Hubble Source Catalog

M. I. Moretti\textsuperscript{1,2,*}, D. Hatzidimitriou\textsuperscript{3,1}, A. Karampelas\textsuperscript{1,6}, K. V. Sokolovsky\textsuperscript{1,4,5}, A. Z. Bonanos\textsuperscript{1}, P. Gavras\textsuperscript{1}, M. Yang\textsuperscript{1}

\textsuperscript{1}IAASARS, National Observatory of Athens, Vas. Pavlou & I. Metaxa, 15236 Penteli, Greece
\textsuperscript{2}INAF - Osservatorio Astronomico di Capodimonte, Via Moiariello 16, 80131 Naples, Italy
\textsuperscript{3}Department of Physics, National and Kapodistrian University of Athens, Panepistimiopolis, 15784 Zografos, Greece
\textsuperscript{4}Sternberg Astronomical Institute, Moscow State University, Universitetskii pr. 13, 119992 Moscow, Russia
\textsuperscript{5}Astro Space Center of Lebedev Physical Institute, Profsoyuznaya Str. 84/32, 117997 Moscow, Russia
\textsuperscript{6}American Community Schools of Athens, 129 Aghias Paraskevis Ave. & Kazantzaki Street, Halandri, Athens GR 15234

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT
Principal Component Analysis (PCA) is being extensively used in Astronomy but not yet exhaustively exploited for variability search. The aim of this work is to investigate the effectiveness of using the PCA as a method to search for variable stars in large photometric data sets. We apply PCA to variability indices computed for light curves of 18152 stars in three fields in M 31 extracted from the Hubble Source Catalogue. The projection of the data into the principal components is used as a stellar variability detection and classification tool, capable of distinguishing between RR Lyrae stars, long period variables (LPVs) and non-variables. This projection recovered more than 90% of the known variables and revealed 38 previously unknown variable stars (about 30% more), all LPVs except for one object of uncertain variability type. We conclude that this methodology can indeed successfully identify candidate variable stars.

Key words: methods: data analysis – methods: statistical – stars: statistics – stars: variables: general – galaxies: individual: M 31

1 INTRODUCTION
An increasing number of time-domain surveys are producing large catalogs of multi-epoch and multi-band photometric data, making it ever more important to invent algorithms that efficiently detect and classify variable objects. Multi-epoch photometry from the ground is collected by surveys for optical transients (e.g. Palomar Transient Factory – PTF, Law et al. 2009; Catalina Real-Time Transient Survey – CRTS, Drake et al. 2009; All Sky Automated Survey for SuperNovae – ASAS-SN, Shappee et al. 2014), microlensing surveys (MAssive Compact Halo Object – MACHO, Griest et al. 1999; Optical Gravitational Lensing Experiment – OGLE, Udalski et al. 2003; Expérience pour la Recherche d’Objets Sombres – EROS, Kim et al. 2014; Tisserand et al. 2007), near-infrared surveys (e.g. Vista Variables in the Via Lactea – VVV, Minniti et al. 2010), the Vista near-infrared Y, J, Ks survey of the Magellanic Clouds – VMC, Cioni et al. 2011), ground-based exoplanet surveys (Trans-Atlantic Exoplanet Survey – TrES, Alonso et al. 2007; Super Wide Angle Search for Planets – SuperWASP, Butters et al. 2010), Hungarian-made Automated Telescope Network – HAT-Net, Bakos et al. 2004; the Kilodegree Extremely Little Telescope – KELT, Pepper et al. 2007) and from space by Gaia (Perryman et al. 2001; Gaia Collaboration et al. 2016; Clementini et al. 2016) and Kepler (Koch et al. 2010) with the next generation time-domain surveys running or just around the corner (among them PanSTARRS, Chambers et al. 2016; Large Synoptic Survey Telescope – LSST, Ivezic et al. 2008; the Next Generation Transit Survey – NGTS, Wheatley et al. 2017).

Identifying variable objects in a set of light curves (LCs) is non-trivial as the photometric measurements are often affected by correlated noise and outliers (Sec. 5). Traditional variability detection techniques include: identifying LCs with high scatter (e.g. Kolesnikova et al. 2005; Burdanov et al. 2016; Dutta et al. 2018) or the ones showing smooth systematic variations (Welch & Stetson 1993; Stetson 1996; Mowlavi et al. 2014), periodicity search (Fruth et al. 2014; Drake et al. 2014; Oelkers et al. 2015; Medina et al. 2018) and LC template fitting (Layden et al.
It is desirable for a variability detection algorithm to have the capability of detecting not only specific types of variable stars (e.g. Cepheids) or transient events with unique signatures (e.g. microlensing events), but also to discover all variable sources within the capabilities of a given survey. Once a variable object has been detected, it is often desired to determine its variability type from the light curve. Methods for automatically classifying LC of variable stars have been investigated by several authors (Debosscher et al. 2007, Paegert et al. 2014, Kim & Bailer-Jones 2016).

The “Hubble Catalog of Variables” (HCV; Gavras et al. 2009, 2013; Di Criscienzo et al. 2011; Fiorentino et al. 2010, 2013; Di Criscienzo et al. 2011; Bernard et al. 2013; Hoffmann et al. 2016) contains a wealth of variability information that is yet to be explored. While constructing the HCV pipeline we considered a number of statistical characteristics, referred hereafter as variability indices (Sec. 3.1), which quantify LC scatter and/or its smoothness. The ability of each index to identify variable sources depends on the variability type and observing cadence. Since the HST archive includes observations of Galactic and extragalactic fields (expected to have at least one variable source as well as active galaxies, supernovae etc.) performed with various observing cadence over various time intervals, a single variability index will not optimally find variables in all HST fields. Here we explore a technique that may solve the problem of selecting an optimal variability index by automatically finding a quasi-optimal combination of multiple variability indices with no a priori information about the types of variable objects found in a given field.

Sokolovsky et al. (2017a) have investigated the effectiveness of applying variability indices to ground-based photometry. The authors proposed two robust ways of identifying variable sources in time-series photometry of any cadence: (i) to use a combination of two indices, the correlation-based inverse von Neumann ratio and the scatter-based Interquartile Range and (ii) to use the admixture coefficient of the first principal component resulting from the principal component analysis (PCA), as an optimal linear combination of multiple variability indices. Despite ground-based data are unstable both in terms of detector performance as well as in PSF variability due to seeing, it was found that the variability-related information may be recovered from the first significant principal component, without having to identify which index is the most suitable for each data set in advance.

In the current study we extend the PCA-based methodology of variability detection and apply it to variability indices derived from HSC LC of sources in three fields in M 31. These fields constitute part of the “control sample” for the HCV, as they have been studied in terms of variability by Brown et al. (2004) and Jeffery et al. (2011). Our goal is to construct a nearly-complete list of variable sources in these fields and evaluate the applicability of the PCA-based method to LCs containing a much smaller number of epochs than those investigated by Sokolovsky et al. (2017a). This technique is not specific to the HSC, but can be applied to identify variable sources in virtually any large set of LCs.

This paper is structured as follows: Section 2 describes the data used in this study; Section 3 introduces the PCA-based variability detection method adopted for the analysis; Section 4 presents the newly discovered variables and compares them with previously reported variables in the studied fields. In Section 5 we discuss the results and in Section 6 we summarize the conclusions.

2 DATA

2.1 The Hubble Source Catalog

We extract the multi-epoch photometry from the HSC version 1 (HSCv1, Whitmore et al. 2016), which contains 80 million detections of 30 million sources. The catalog is based on images obtained with the WFPC2, ACS/WFC, WFC3/UVIS and WFC3/IR cameras on board the HST over a period of 16, 15, 8 and 8 years, respectively. As the HST data were collected with different instruments, filters and observing strategies, the photometric accuracy, data quality and cadence of observations vary greatly across the HSC, making the detection of variable objects a non-trivial task. The HSCv1 contains about 5 million sources with more than 2 measurements and about 40,000 objects with more than 10 measurements obtained with the ACS/F814W instrument/filter combination. The F814W is available also on the other cameras: WFPC2 and WFC3. The F814W magnitudes of objects listed in HSCv1 range from 15th to 26th mag for WFPC2, 15–27 mag for ACS and 16–28 mag for WFC3. The mean photometric accuracy is better than 0.10 mag, while the relative accuracy is $\simeq 0.02$ mag at best. HSCv1 is based on the Hubble Legacy Archive (HLA) Data Release 8 images.

A visit is a series of one or more exposures on a target, including overheads, that are executed in one or more consecutive orbits. The exposures are interleaved by the time required for dithering, filter change, other instrumental overheads and the time the target source is occulted by the Earth. All exposures taken within one visit are combined by the HLA into a white-light image that is used for source detection. For each filter used in a given visit, the images taken in this filter are also combined. The filter-combined images are used to measure brightness of the sources in each filter during the visit. These stacked images are used for source extraction and photometry for the purpose of removing image artifacts caused by cosmic ray hits. The SExtractor software (Bertin & Arnouts 1996) is employed to produce the HLA source lists providing magnitude measurements for each source. SExtractor is operated in dual-image mode with the white-light image used for source detection and filter-combined images used for photometry. Since the HSCv1 photometry is visit-based, the HSCv1 light curves

1 A second version of the HSC was made available during the preparation of the present manuscript.
for a specific source typically contain fewer data points compared to the respective LC presented in the literature that are usually extracted from individual images within a visit. Positions of sources detected on the visit-combined HLA images are based on the information from HST fine guidance sensors. While the fine guidance sensors have superb internal astrometric accuracy (Benedict et al. 2017), the absolute positions reported by them are limited by the position accuracy of individual guide stars. The resulting absolute position errors could be as large as $\sim 1-2$ arcsec for HST images obtained before 2005 when the Guide Star Catalog 2 became available (Lasker et al. 2008).

The HSC is using PanSTARRS, SDSS (Ahn et al. 2014) and 2MASS sources within the HST cameras’ field of view to refine the absolute astrometry of the HLA images. The lists of sources detected during multiple visits of the same sky area are cross-matched using the Bayesian technique proposed by Budavári & Lubow (2012). This results in the final absolute astrometric accuracy of better than 0.1 arcsec for most of the HSC sources (Whitmore et al. 2016).

### 2.2 M 31 fields

We selected three fields in M 31 observed by HST. These HST fields were originally investigated for variability by Brown et al. (2004) and Jeffery et al. (2011). The fields fall on the southeast minor axis (“Halo11”), on the north-east major axis (“Disk”) and in the giant stellar stream of metal rich stars (“Stream”; Ibata et al. 2001; Kirihara et al. 2017), as illustrated in Figure 1. The selection of these fields was based on (i) the large number of published variables (100 RR Lyrae stars and more than 30 Long Period Variables – LPVs) with very good astrometry and (ii) the availability of LC in two filters (F606W and F814W) in the HSCv1. Table 1 summarizes the main characteristics of these fields and the corresponding observations: coordinates, original Program ID, number of HSCv1 sources, minimum, maximum and median number of HST visits (corresponding to the number of data points in the HSCv1 LC) and time (modified Julian date, MJD) and magnitude range. In the Halo11 and the Disk fields $\sim 7000$ point sources were detected in HSCv1, while $\sim 4300$ sources were recorded in the Stream field. For the Halo11 field, the observations were obtained over a period of 40 days. The corresponding LC contain about twice as many visits as the Stream and Disk LC, which were obtained at lower cadence and over a shorter period of about 30 days (see discussion in Brown et al. 2009).

The HSCv1 sources used in our analysis satisfy the following criteria: (i) they are point sources according to the HSCv1 classification flag; (ii) they have at least 6 measurements in both the F606W and F814W LC; (iii) they are located within 240 arcsec from the center of the corresponding field; (iv) they have a SExtractor flag $\leq 7$. Sources satisfying this constraint may be affected by bright and nearby neighbors, bad pixels affecting at least 10% of the integrated area, blending with another source, or saturated pixels. We retained data points marked as saturated since in the HSCv1 the saturation level is set incorrectly. Sources within $\sim 0.5$ mag of the bright limit in either filter (see Table 1) are discarded at a later stage of the analysis to avoid false variability introduced by possible saturation (Section 4). Sources within $0.5$ mag of the faint limit in either filter (Table 1) were treated with particular care (Section 3).

With the above constraints, we ended up with a catalog of 18152 sources in the three fields (“HSCv1 based sample” hereafter). Figure 2 shows the normalized magnitude histograms for the three fields. The histograms are similar for each filter in both bands, with a slight exception around magnitudes 24-26 mag, possibly due to different stellar population properties (Brown et al. 2006, 2009; Jeffery et al. 2011; see also discussion in Section 4).

### 3 METHODOLOGY

#### 3.1 Variability indices

We characterize each LC by its mean magnitude and the values for 18 variability indices. For a variable source we expect that (i) its LC has larger scatter than the LC of a non-variable object of similar brightness (irrespective of the variability timescale) and (ii) the measurements taken close in time result in similar magnitudes (i.e. the LC is smooth) if the variability timescale is significantly longer than the LC sampling rate. The variability indices used here capture one or both of these characteristics. In addition to the measured magnitudes, some indices take into account the estimated photometric errors, as well as the order and time at which the measurements were obtained. Table 2 presents the summary of the indices used and whether errors, order

---

2 Brown et al. (2004) analyzed several fields along the south-east minor axis: one at 11, one at 21 and two at 35 kpc with respect to the M 31 center; here we refer to the first one.

3 MJD=JD-2400000.5 d

4 We discarded the only source having SExtractor flag >7 lying in the Halo11 field.
Table 1. Characteristics of the M31 fields selected for PCA. In particular, Col. 6 and 7 indicate the number of minimum-maximum, median visits in F606W and F814W filters respectively.

| Field    | RA (deg) J2000 | Dec (deg) J2000 | Program ID | HSCv1 sources | # Visits F606W | # Visits F814W | MJD range F606W | MJD range F814W |
|----------|---------------|----------------|------------|----------------|---------------|---------------|-----------------|-----------------|
| Halo11   | 11.52958      | 40.71083       | GO-9453    | 7109           | 6-27, 27      | 6-33, 33      | 52610-52650    | 52611-52650     |
| Disk     | 12.28583      | 42.75055       | GO-10265   | 6732           | 6-11, 11      | 6-16, 16      | 53359-53374    | 53350-53389     |
| Stream   | 11.07583      | 39.79222       | GO-10265   | 4311           | 6-10, 10      | 6-15, 15      | 53248-53282    | 53250-53282     |

Figure 2. Normalized distribution of the mean F606W and F814W magnitudes of HSC sources in the three fields.

Characterizing a LC with variability indices is directly analogous to the feature extraction [Nun et al. 2013; Christ, Kempa-Liehr, & Feindt 2016] performed for machine learning classification of variable stars [Debosscher et al. 2009; Richards et al. 2011; Kim & Bailer-Jones 2016]. However, here, we purposefully avoid features based on periodicity in order to have the same set of indices characterizing periodic, irregular and non-variable objects. Period search results may not be reliable given the small number of observing epochs. Typically, more than a hundred brightness measurements that randomly sample the LC are needed for a reliable period determination [Horne & Baliunas 1986; Graham et al. 2013]. A method of characterizing LCs without defining variability features is proposed by [Kügler, Gianniotis, & Polsterer 2013].

Although the individual indices are useful tools for variability search, a combination of multiple indices may be an even better variability indicator, as different indices are sensitive to different types of variability. The PCA offers a promising option to optimally combine variability indices, without having to decide a priori which indices are more suitable for a given data set. The PCA also provides a natural way to combine multi-band data for variability search, as it is possible to add features computed from the LC in all available filters.

3.2 Principal Component Analysis

The PCA [Pearson 1901] is extensively used in Astronomy (e.g. Bailer-Jones et al. 1998; Re Fiorentin et al. 2007; Yip et al. 2003; Karampelas et al. 2012; Steiner et al. 2009). It has also been employed for variable star detection using multi-band LCs [Suárezes et al. 2013; Sokolovsky et al. 2017a], as first suggested by Eyer (2006). PCA linearly and orthogonally transforms a data set of \( m \) quantities (where each data point is represented by a vector, \( x_j \), in the \( m \)-dimensional space) onto a new set of \( m \) uncorrelated axes (the eigenvectors of the variance-covariance matrix of the data), where the data variance is being emphasized. These eigenvectors are called the principal components (PCs).
Table 2. Variability indices used as an input for the PCA.

| Index | Ref. | Errors | Order | Time |
|-------|------|--------|-------|------|
| Scatter-based indices |
| reduced $\chi^2$ statistic ($\chi^2_{\text{red}}$) | (a) | ✓ | ✓ | ✓ |
| weighted standard deviation ($\sigma$) | (b) | ✓ | ✓ | ✓ |
| median abs. deviation (MAD) | (c) | ✓ | ✓ | ✓ |
| interquartile range (IQR) | (d) | ✓ | ✓ | ✓ |
| robust median statistic (RobMS) | (e) | ✓ | ✓ | ✓ |
| norm. excess variance ($\sigma^2_{\text{XX}}$) | (f) | ✓ | ✓ | ✓ |
| norm. peak-to-peak amp. ($\nu$) | (g) | ✓ | ✓ | ✓ |
| Stetson’s $K$ index | (h) | ✓ | ✓ | ✓ |
| Correlation-based indices |
| Stetson’s $J$ index | (h) | ✓ | ✓ | ✓ |
| weighted Stetson’s $J$ (time) index | (i) | ✓ | ✓ | ✓ |
| clipped Stetson’s $J$ (clip) index | (d) | ✓ | ✓ | ✓ |
| Stetson’s $L$ index | (h) | ✓ | ✓ | ✓ |
| weighted Stetson’s $L$ (time) index | (i) | ✓ | ✓ | ✓ |
| clipped Stetson’s $L$ (clip) index | (d) | ✓ | ✓ | ✓ |
| excursions ($E_x$) | (j) | ✓ | ✓ | ✓ |
| autocorrelation ($l_1$) | (k) | ✓ | ✓ | ✓ |
| inv. von Neumann ratio ($1/\eta$) | (l) | ✓ | ✓ | ✓ |
| $S_B$ statistic | (m) | ✓ | ✓ | ✓ |

References: (a) [de Diego 2010], (b) [Dutta et al. 2013], (c) [Zhang et al. 2016], (d) [Sokolovsky et al. 2017a], (e) [Rose & Hintz 2007], (f) [Nandra et al. 1993], (g) [Brown et al. 1981], (h) [Stetson 1990], (i) [Fruth et al. 2012], (j) [Parks et al. 2014], (k) [Kim et al. 2011], (l) [Shin et al. 2008], (m) [Figuera Jaimes et al. 2013].

Each observation $x_j$ of the original data is expressed as

$$x_j = \sum_{i=1}^{m} a_{j,i} \cdot PC_i$$  \tag{1}$$

where $a_i$ is the admixture coefficient of the principal component $PC_i$. The coefficients $a_i$ are the coordinates of the data point $x_j$ in the new axes.

PCs are ordered so that $PC_1$ accounts for the highest data variance, $PC_2$ (uncorrelated to $PC_1$) incorporates most of the remaining variance and so on. Thus, the sequence from $PC_1$ to $PC_m$, assumed to be represented by the data variance, extends from widespread, to rare, to noise. The amount of the variance corresponding to each PC is the respective eigenvalue $\lambda_i$ divided by the sum of all the eigenvalues. This relation originates from the diagonalization of the data variance-covariance matrix $C$, achieved through the PCA implementation. Each element $c_{ij}$ of $C$ is the covariance between the $i$-th and the $j$-th variables. If $i = j$, then $c_{ii}$ is the variance of the $i$-th variable. The variance-covariance matrix $C'$ of the transformed data is diagonal, having $c_{ij} = 0$ and $c_{ii} = \lambda_i$. The larger the $\lambda_i$, the higher the variance represented by $PC_i$.

Since the lower-order PCs do not represent any useful information, each observation can be approximately expressed by replacing $m$ with $r$ in Eq. (1) where $r < m$. There is no standard procedure for deciding how many principal components should be kept. If $r = 2$ or $r = 3$, the data can be visualized in two or three dimensions, using the corresponding admixture coefficients $a_1$, $a_2$ and $a_3$.

The resulting dimensionality reduction/data compression is among the advantages of the PCA. The PCA is an unsupervised (no need for training data) and non-parametric (no need for tuning) procedure that provides a linear decomposition. On the other hand, PCA may not perform very well when the processes dominating the original data are not linear, although alternatives do exist such as kernel-PCA (e.g. [Ishida & de Souza 2013]) and robust-PCA (e.g. [de Souza et al. 2014]). PCA is also data-dependent, thus requiring a standardization of the original data, if the data variables express different characteristics (Sec. 5.3).

3.3 PCA implemented in the M 31 fields

The input characteristics for the PCA were the 18 variability indices listed in Table 2 and the mean magnitude computed for the LCs in each filter (F606W and F814W), resulting in a total of 38 parameters, for all sources detected in the three selected M 31 fields (Halo11, Disk and Stream). The variability indices were first standardized to zero average and unit variance. Standardization is necessary since the variance-enhancing PCA is data dependent and the input characteristics express different quantities. High values of an index may result in numerically large variance, although this index may not contain useful variability information. The standardized index values express the deviation from the average in multiples of the standard deviation, making the various indices comparable. By construction, large values of the variability indices (subsequently, positive values well above the zero average of the standardized indices) are expected to correspond to variable stars.

The scree plot presented in Figure 3 illustrates the percentage of the total variance that corresponds to the top 10 most significant PCs.

High variance is caused by high values of the variability indices of variable sources, thus linking the most significant principal components to variable sources. About 60% of the data variance is captured by the first two PCs (average value for the three fields).

3.4 Selection of candidate variables

The detectability of variable sources hinges on the following assumptions: (a) The sample contains variables and therefore the variability indices by construction contain variability information from the LC. (b) Variability is mainly encoded in $PC_1$ and $PC_2$ and their respective values (admixture coefficients $a_1$ and $a_2$.) The high values of the variability indices of the variable stars and their respective high variance “feed” the most significant principal components, since PCA preferentially highlights variance. Therefore, the variability information is expected to be encoded in the most significant principal components. (c) Variable stars are found in sparse regions on the $a_1$–$a_2$ plane, while constant stars have $a_1$ and $a_2$ values close to zero. Additionally, the observational fact that the majority of stars appear to be constant (at the level of photometric accuracy expected from HSC), ensures that they will form the distinct dense area in the $a_1$–$a_2$ plane, making the variables more easily distinguishable. Thus, sparseness in the $a_1$–$a_2$ plane emerges from variable stars being rare and having wider $a_1$ and $a_2$ ranges compared to the constant stars.
Scree plot representing the variances of the 10 most significant Principal Components.

By construction, $D_{\text{inside}}$ characterizes constant stars, as the vast majority of sources inside the box are expected to be non-variables. On the other hand, most of the stars contained in the outer box are expected to be variables. Thus, for the $k$-th inner box star to be a variable candidate, its distance $d_k$ is expected to have a value closer to $D_{\text{inside}}$ rather than to $D_{\text{outside}}$, that is $|d_k - D_{\text{inside}}| < |d_k - D_{\text{outside}}|$. The resulting candidate variables are marked with black dots on the $a_1$–$a_2$ selection plane for M 31 Halo11 (upper left panel of Figure 3). These candidate variables are validated and compared against published variables in Section 4. The candidate variables are also marked on higher order admixture coefficient planes (remaining panels of Figure 3) which are further discussed in Section 5.1. Similar plots are available for the Disk and Stream fields in the Appendix (Figures A3, A4).

4 RESULTS

The two step selection procedure described in the previous section, resulted in a total of 156 candidate variables in the Halo11 field (2.2% of the total number of sources), 192 in the Disk field (2.9%) and 88 in the Stream field (2.0%). Figures 3, 6, 7 show the color magnitude diagrams (CMD) for the three fields, respectively. The sources lie on the red giant (RGB), subgiant (SGB), asymptotic giant (AGB) and horizontal branches (HB). In most cases, main sequence stars lie below the detection limit, with the exception of the Disk field where they are detected at $\sim$23rd magnitude, indicating the presence of a younger population (see also Jeffery et al. 2011, Brown et al. 2004, 2009). The location of the candidates on the CMD is also shown (black points in Figures 3, 6, 7). Most of the candidate variables are associated with the HB where one expects to find RR Lyrae stars, with the AGB where one would expect to find LPVs and with the supergiant region. Several candidates lie on the RGB and SGB and might be eclipsing binaries. We consider as reliable variable candidates stars fainter than 21.0 mag in F606W (20.5 in F814W), corresponding to the saturation limit derived from the raw HST images.

4.1 Recovery of known variables

To compare the list of variable stars published by Brown et al. (2004) and Jeffery et al. (2011), to our list of candidate variables (Section 3.4) and evaluate the effectiveness of our variability selection technique, we first need to ascertain whether a published variable is present in the HSCv1 based sample, as defined in [23]. Some of the published variables may be excluded due to differences in image processing and source extraction strategies used to create the HSC and the ones employed by Brown et al. (2004) and Jeffery et al. (2011). For the same reason some of the sources that could not be used in the original studies may become useful for variability search through the HSC.

Most of the published variables have an HSCv1 counterpart within about 0.6 arcsec, indicating that the published and the HSCv1 astrometry are in good agreement. For RR Lyrae variables the validity of HSCv1 counterparts to the published variables was confirmed by visual inspection of the LC folded with the published period. In the case
Figure 4. The location of candidate variables (black points) on the admixture coefficient planes for M31 Halo11. The known LPVs (red open squares), RR Lyrae stars (small blue open circles), RR Lyrae star candidates (black open diamonds), eclipsing binaries (black open triangles), the dwarf Cepheid (black open star), the anomalous Cepheids (black open pentagons), the post-AGB star (black asterisk) are labeled, along with all the sources (grey dots). The newly discovered variables are marked with big black open circles. The solid box corresponds to the initial step of the selection process (step (i) in Sec. 3.4), while the dashed box indicates the region where the second step (ii) of the process was applied. Insets present the most crowded areas.

Table 3. Published variables in the three M31 HST fields.

| Field  | Published variables | Reference | Listed in HSCv1 | Recovered variables |
|--------|---------------------|-----------|-----------------|---------------------|
| Halo11 | 115                 | Brown et al. (2004) | 88 | 78 (89%) |
| Disk   | 21                  | Jeffery et al. (2011) | 21 | 21 (100%) |
| Stream | 24                  | Jeffery et al. (2011) | 22 | 21 (95%) |
| All    | 160                 | Brown et al. (2004) | 131 | 120 (92%) |

Halo11 field

In the Halo11 field, 53 out of 55 published RR Lyrae variables have a confirmed counterpart in the HSCv1. One RR Lyrae star (V167) has not been included as it had fewer than 6 data points in the HSCv1 F814W filter LC. RR Lyrae variable V124 did not pass the point source selection, due to blending with a nearby source.

HSCv1 counterparts were confirmed for 27 out of 32 published LPVs (all 32 with amplitudes higher than 0.1 mag in both filters). Five LPVs (V14, V18, V19, V21, V25) have no HSCv1 counterpart within 1 arcsec, possibly due to their location near the globular cluster SKHB 312 (Holland et al. 1997), where severe blending can be expected. The Brown et al. (2004) list of variables also contains two anomalous Cepheid (aCep) candidates (V60, V118), one post-AGB (V84) and two RR Lyrae star candidates (V29, V6), all successfully matched within the HSCv1.

of LPVs published by Brown et al. (2004), no time series data were provided by the authors, thus the matching was performed based only on the positional coincidence and visual comparison of HSCv1 LCs with the LC plots from the paper. Table 3 summarizes the comparison results listing the number of published variables, the number of published variables with a counterpart within the HSCv1 and the number of published variables recovered in this work. Figures 4, 5, 6, 7 also show the known variables highlighted with different symbols as indicated in the legend of the upper right panel of Figure 4.

Figure 8 shows the LC for two RR Lyrae variables, one in the Halo11 and one in the Stream field. Literature LCs in Halo11 are provided without errors; according to Brown et al. (2004), they have typical photometric uncertainty of 0.03 mag in F606W and 0.04 mag in F814W. The literature data have been converted from STMAG to ABMAG to match the HSCv1 photometric system, by applying an appropriate aperture correction using the following equations: $\text{ABMAG}_{F606W} = \text{STMAG}_{F606W} - 0.169 + \text{APC}_{F606W}$, where $\text{APC}_{F606W} = 0.248$, and $\text{ABMAG}_{F814W} = \text{STMAG}_{F814W} - 0.840 + \text{APC}_{F814W}$, where $\text{APC}_{F814W} = 0.292$ (Brown et al. 2004). A small photometric discrepancy of few hundreds of a mag is still present, likely due to the charge transfer efficiency loss correction applied by Brown et al. (2004), but not in HSCv1, as comprehensively described in Whitmore et al. (2016).

5 https://archive.stsci.edu/hst/hsc/help/hsc_use_case_1.html

6 BLBOO in the GCVS nomenclature (Samus et al. 2017).
Figure 5. Color-magnitude distribution of sources in the Halo11 field. All sources are represented with grey dots, the PCA variable candidates are shown with black points. The new and known variables are labeled as in Figure 4. The inset presents the HB region.

Figure 6. Same as Fig. 5 but for the Disk field. The literature variables are from Jeffery et al. (2011); the classical Cepheid is marked by a reversed open triangle.

Figure 7. Same as Fig. 5 but for the Stream field. The literature variables are from Jeffery et al. (2011).

Due to the shallower magnitude limit of HSCv1, we could not identify counterparts for the fainter variables published by Brown et al. (2004). Out of 8 eclipsing binary candidates only two have been matched with HSCv1 (V64 and V173). Out of 15 dwarf Cepheids only one could be reliably identified (V48) on the basis of its coordinates and location on the CMD, although the HSCv1 LC is too noisy to compare reliably with the published one.

A total of 88 known variables in the Halo11 field (53 confirmed RR Lyrae stars and 2 candidates, 27 LPVs, 2 possible anomalous Cepheids, 1 post AGB, 2 eclipsing binary candidates and 1 possible dwarf Cepheid) have HSCv1 LCs that pass our selection criteria and are expected to be recovered by our variability detection technique. Our procedure has yielded a total of 156 candidate variables when using both filters, and 128 (167) using only F606W (F814W) data. Out of the 88 known variables in the input catalog, 78 (89%) have been recovered by the variability selection method when using both filters, 90% when using only the F606W filter and 86% when using only the F814W filter. The recovery rate is higher for the F606W data compared to the F814W data, mainly due to the 1.5 times higher amplitude of RR Lyrae variables in F606W than in F814W. The variables not recovered are faint (two eclipsing binary candidates, V64 and V173, and one dwarf Cepheid, V48) or low amplitude variables (a post-AGB star V84, an LPV V158, four low amplitude RR Lyrae variables of RRc type, V40, V76, V80 and V137, and one possible RR Lyrae variable, V29). Some of these sources are also affected by problems

Dwarf Cepheids are short-period pulsating variables below the Horizontal Branch on the main sequence (δ Scuti stars, DSCT) and SGB (SX Phoenicis stars, SXPHE).
such as possible blending, vicinity to the edge of an image, or, to a diffraction spike of a bright neighboring source.

**Disk field**

Jeffery et al. (2011) searched for short period variables discovering 21 RR Lyrae stars and two Cepheids in the Disk field. No reliable counterparts were found in the HSCv1 based sample for two Jeffery et al. (2011) RRc variables with IDs 15256 (due to a nebulosity) and 21631 (blending with a brighter source). We have HSCv1 LCs for 19 known RR Lyrae variables and 2 Cepheids in the Disk field.

The selection process yielded 192 candidate variables using both filters, which include all of the 21 known variables, thus yielding a 100% recovery rate. For comparison, 266 (202) candidates were found using only the F606W (F814W) filter, with a recovery rate of 100% (90%) of known variables.

**Stream field**

Jeffery et al. (2011) discovered 24 RR Lyrae stars in the Stream field. One of the RR Lyrae stars (ID 409) is missing from the HSCv1 based sample as it only had 5 visits in F606W filter, probably due to its location near the edge of the frame. No reliable counterpart was found for the RR Lyrae variable ID 8544, probably due to the adverse effect of a nearby nebulosity on the source extraction procedure employed for creating the HSCv1. We hence have 22 RR Lyrae stars in the Stream field successfully matched with HSCv1.

The variability selection procedure yielded 88 candidate variables using both filters, out of which 21 are in common with the Jeffery et al. (2011) list of 22 known variables in the input catalog (95% recovery rate). For comparison, 124 (96) candidates were found using only F606W (F814W) filter, with a recovery rate of 86% (91%). The only known variable not recovered is an RRc type RR Lyrae variable (ID 9975) of low amplitude (∼0.2 mag in F606W and ∼0.15 in F814W).

### 4.2 New variables

Candidates that were not matched with known variables were validated using the following procedure:

(i) visual inspection of the images in both filters, in order to identify probable sources of error such as: blending; proximity to a diffraction spike of another source; cosmic rays; bad column or hot pixel; proximity to the frame edge;

(ii) inspection of the LC for the presence of outliers or inconsistencies in the appearance of variability in the two filters.

(iii) lack of obvious similarities with LC of other sources in the field, which may indicate systematic errors in photometry.

The source position on the CMD was used to confirm the variability type inferred from the LC, but it was not considered as a criterion when deciding whether a given source is variable or not. With the exception of a single source, 22221197, all new candidate variables that passed our visual inspection, were found to lie in either the RR-Lyrae or the LPV region of the CMD.

Table 4 summarizes the variability search results listing the field name, the total number of HSCv1 sources passing the initial quality cuts, the number of candidates obtained with the variable stars selection procedure, the number of confirmed variables together with the percentage of recovered known variables, the number of candidates that were rejected. The process of visual inspection is subjective by its nature.

---

Photometry of sources near frame edge in the HSCv1 may be affected by incorrect sky background estimation.

---

Figure 8. Light curves of two RR Lyrae stars from HSCv1 (black) and literature (grey) from Halo11 (upper panel) and the Stream (lower panel). Both are folded with the period provided in the literature.
nature. When presented with apparent low-amplitude variability we had to decide if a neighboring source is bright and close enough to corrupt the LC or if the LC is sufficiently different form some other LCs in the field. For about a half of the rejected candidates listed in Table 4, we had to make these judgment calls while the other half did not pass the saturation-magnitude cuts or had severe imaging problems.

The list of newly identified variables is presented in Table 5. For each variable the table lists the HSCv1 ID, Field, equatorial coordinates, F814W average magnitudes, number of visits in F814W, F606W average magnitudes, number of visits in F606W, classification and comments. The LCs of the new variables are presented in Fig. 5, Fig. 6 and 7, which show the candidates and the confirmed new variables on the CMD together with all sources.

### Halo11 field

A total of 78 new candidate variables have been identified. Out of these, 13 are rejected because they are brighter than 21.0 mag in F606W and/or 20.5 mag in F814W. We visually inspected the images, LC and CMD location of 65 candidates: 30 of these are false detections: 13 (5 of which are close to globular cluster SKHB 312) are affected by blending; 15 are close to the frame edge; 1 is close to diffraction spikes of a nearby bright star; 1 is surrounded by a halo in some images. We were able to confirm 7 new variables, all LPVs. The results of visual inspection for the remaining candidates were inconclusive, so they were not included in the list of confirmed variables. The same procedure was followed for the candidates yielded by the selection procedure from the single-filter data resulting in two more confirmed LPVs.

### Disk field

A total of 171 candidate new variables were identified using the two filters. Out of these, 15 are rejected as being possibly saturated. We visually inspected the remaining 156 candidates: 62 are affected by image problems of several types: 15 are blended with nearby sources, 31 are close to the frame edge, 10 are close to a nebulosity or surrounded by a halo in some images, 3 are very close diffraction spikes of nearby sources, 3 are close to hot pixels. We were able to confirm 20 new variables: 19 LPVs (one of which at the faint limit, see Table 5) and one with uncertain variability type. Visual inspection did not result in a firm classification for 74 sources.

### Stream field

A total of 67 new candidate variables have been identified using both filters. Out of these, 11 have been rejected as possibly saturated. We hence visually inspected 56 candidates: 11 are affected by image problems such as 2 blended, 6 lying on the edge/very close to the edge, 2 very close diffraction spikes of nearby sources, 1 close to hot pixels. We were able to confirm 9 new variables, all LPVs (see Table 5). No other variables could be confirmed after inspecting candidates selected from the single-filter analysis.

The Stream field observations present a 20 day gap that complicates the analysis and classification of the LCs. In this case, we relied mostly on the shape of the LC in the first part of the time series. In some dubious cases (e.g. 27664093, see Figure A7), we visually inspected the LCs for the nearby sources to rule out photometric problems.

## 5 DISCUSSION

The application of PCA to variability indices that characterize LCs allowed us to statistically select candidate variable sources using the admixture coefficients of the first two PCs, $a_1$ and $a_2$. The recovery rate of known variables in the fields studied has been more than $90\%$ in all cases, while 38 new variables were identified. We also confirmed that various issues that can affect the photometry, such as saturated pixels, blending, proximity to diffraction spikes, cosmic rays etc., significantly increase the number of false positives, thus making visual inspection of the automatically selected candidates imperative.

The need for visual inspection of candidate variables is common for all types of variability search (e.g. Bernard et al. 2010; Cusano et al. 2013; Abbas et al. 2014; Ramsay et al. 2014; Klagyivik et al. 2016; Pawlak et al. 2016). The underlying problem is that corrupted measurements cannot be reliably identified in all cases. Individual outlier measurements in a LC can be removed by sigma-clipping (e.g. Kim & Baier-Jones 2016) or tolerated thanks to the use of robust variability features (Pérez-Ortiz et al. 2017). However, these measures will not solve the problem if a given object has a large fraction of its measurements corrupted. Despite the progress in automatic rejection of imaging artifacts (Fruchter & Hook 2002; Desai et al. 2016), visual inspection of images (Melchior et al. 2016) and LCs remains a valuable quality control tool. The goal of an automated variability detection algorithm is to minimize the percentage of false candidates passed to the visual inspection stage (Fruth et al. 2014). The PCA-based variability search is doing well in this regard when compared to other variability detection techniques applied to the same LC data (Sec. 5.3).

### 5.1 Physical Interpretation of the Principal Components

Each PC is a linear combination of all the input variability indices listed in Table 2:

$$PC_i = c_{i,1}X_{eq,1} + c_{i,2}\sigma_{L2} + c_{i,3}\text{MAD}_{l3} + c_{i,4}\text{IQRI4} + \ldots$$

where $c_{i,1}, c_{i,2}, c_{i,3}, c_{i,4} \ldots$ are the coefficients determining the contribution of each index to the $i$th principal component, $PC_i; L1, L2, L3, \ldots$ are the unit vectors setting the
Table 5. Characteristics of new variable stars ordered by F814W average magnitudes in each field.

| MatchID | Field | RA (deg) J2000 | Dec (deg) J2000 | F814W (mag) | # Visits F814W | F606W (mag) | # Visits F606W | Type | Comments |
|---------|-------|----------------|----------------|-------------|---------------|-------------|---------------|------|----------|
| 27284400 | Halo11 | 11.550205 | 40.692210 | 20.586 | 32 | 22.026 | 27 | LPV | — |
| 27296061 | Halo11 | 11.517188 | 40.681656 | 21.465 | 33 | 23.454 | 27 | LPV | — |
| 27276420 | Halo11 | 11.518473 | 40.714756 | 21.528 | 33 | 23.064 | 27 | LPV | — |
| 27285008 | Halo11 | 11.553176 | 40.694355 | 21.587 | 31 | 23.353 | 27 | LPV | — |
| 27273342 | Halo11 | 11.540703 | 40.740814 | 21.635 | 31 | 23.506 | 27 | LPV | — |
| 27272426 | Halo11 | 11.552447 | 40.733310 | 21.671 | 33 | 23.316 | 27 | LPV | — |
| 27307954 | Halo11 | 11.502041 | 40.674786 | 21.834 | 17 | 23.806 | 13 | LPV | F606W only |
| 27283624 | Halo11 | 11.523931 | 40.608988 | 21.932 | 32 | 23.184 | 27 | LPV | F606W only |
| 27289942 | Halo11 | 11.504722 | 40.696518 | 22.117 | 33 | 24.310 | 27 | LPV | — |
| 22196858 | Disk | 12.289636 | 42.773460 | 20.882 | 16 | 23.270 | 11 | LPV | — |
| 22197833 | Disk | 12.296558 | 42.763620 | 21.333 | 16 | 24.384 | 11 | LPV | — |
| 22211719 | Disk | 12.291348 | 42.761390 | 21.363 | 16 | 22.572 | 11 | LPV | — |
| 22206385 | Disk | 12.283233 | 42.737830 | 21.379 | 16 | 22.310 | 11 | LPV | — |
| 22208821 | Disk | 12.296735 | 42.758934 | 21.468 | 16 | 23.315 | 11 | LPV | — |
| 22209236 | Disk | 12.301403 | 42.709558 | 21.487 | 16 | 23.414 | 11 | LPV | — |
| 22211391 | Disk | 12.318888 | 42.740982 | 21.618 | 16 | 23.468 | 11 | LPV | — |
| 22210219 | Disk | 12.296622 | 42.762190 | 21.672 | 16 | 22.573 | 11 | LPV | — |
| 22209784 | Disk | 12.253314 | 42.765840 | 21.718 | 16 | 24.904 | 11 | LPV | — |
| 22206501 | Disk | 12.294667 | 42.731520 | 21.760 | 12 | 25.256 | 11 | LPV | — |
| 22217682 | Disk | 12.296862 | 42.739853 | 21.800 | 16 | 23.691 | 11 | LPV | — |
| 22203253 | Disk | 12.269516 | 42.742680 | 21.816 | 16 | 23.324 | 11 | LPV | — |
| 22212431 | Disk | 12.262894 | 42.727220 | 21.826 | 16 | 23.603 | 11 | LPV | — |
| 22219207 | Disk | 12.290964 | 42.729600 | 21.832 | 16 | 23.294 | 11 | LPV | — |
| 22210620 | Disk | 12.250249 | 42.762253 | 21.845 | 16 | 23.980 | 11 | LPV | — |
| 22204580 | Disk | 12.321586 | 42.736580 | 21.890 | 16 | 24.706 | 11 | LPV | — |
| 22208241 | Disk | 12.298845 | 42.753922 | 21.922 | 16 | 24.031 | 11 | LPV | — |
| 22208319 | Disk | 12.278159 | 42.736460 | 22.163 | 12 | 25.869 | 8 | LPV | — |
| 22259268 | Disk | 12.277175 | 42.731316 | 22.977 | 16 | 26.348 | 10 | LPV | faint limit in F606W |
| 22221197 | Disk | 12.268279 | 42.742805 | 25.637 | 16 | 26.134 | 11 | other | — |
| 27663814 | Stream | 11.041388 | 39.788070 | 21.490 | 15 | 23.836 | 10 | LPV | — |
| 27661867 | Stream | 11.066272 | 39.778630 | 21.498 | 15 | 24.351 | 10 | LPV | — |
| 27654057 | Stream | 11.091002 | 39.781937 | 21.624 | 15 | 23.641 | 10 | LPV | — |
| 27659783 | Stream | 11.071672 | 39.786972 | 21.704 | 15 | 24.404 | 9 | LPV | — |
| 27664545 | Stream | 11.096615 | 39.765050 | 21.979 | 15 | 23.580 | 10 | LPV | — |
| 27665958 | Stream | 11.089661 | 39.768350 | 21.980 | 15 | 24.179 | 10 | LPV | — |
| 27664093 | Stream | 11.099528 | 39.772636 | 22.296 | 15 | 24.089 | 10 | LPV | — |
| 27630664 | Stream | 11.091218 | 39.797176 | 22.421 | 15 | 23.014 | 10 | LPV | — |
| 27619606 | Stream | 11.088532 | 39.802055 | 22.441 | 15 | 24.915 | 10 | LPV | — |

Figure 9. Sample LCs of the new variables in the Halo11 field. All LC plots are available in Appendix A.
directions of the $\chi^2_{cdot}$, $\sigma$, MAD, IQR . . . axes in the variability index space. As there are different ranges for the numerical values of the variability indices, they are standardized (as described in Sec. 3.3) before entering the equation (2). Figure 10 presents the contributions, $c_{i,j}$, of the variability indices to the first six principal components for the M31 Halo11 PCA results. The dashed line indicates zero contribution of an index to the PC (the values of indices near that line have no effect on the PC value). The larger the distance of an index from the zero contribution line, the more it contributes to this PC. A qualitatively similar relative contribution of the variability indexes to the first few PCs is found in the Disk and Stream fields (Figures 11 and 12).

Most variability indices contribute significantly to PC1. This is expected, since by construction the indices highlight variability. PC1 is dominated by three scatter-based indices that take into account the estimated photometric errors ($\chi^2$, RoMS, $\sigma_{nx}$; see Table 2) and several correlation-based indices ($L$, $J$, $S_B$, $E_X$, $1/\eta$). The remaining scatter-based indices ($IQR$, MAD, $\sigma$; the first two being robust measures of scatter not relying on photometric errors) contribute to PC1 as well, but they are the main contributors to PC2 together with the magnitude. The admixture coefficient of PC1, $a_1$, has a large positive value for a bright source showing smooth high-amplitude lightcurve. The value of $a_2$ is large and negative for a faint source with a lightcurve that is not smooth and is showing a large scatter that is not driven by a few outlier points. PC3–PC6 show no easily interpretable patterns, while together they account for 15\% of the total data variance. These components are expected to represent rare information encoded into the variability indices. $K$ and $l_1$ that do not contribute to PC1 and PC2 show more significant contribution to the higher order PCs and in particular PC5. As discussed in Section 3.4 (Table 3, the $K$ index (being a robust measure of kurtosis of the distribution of magnitudes in a lightcurve) cannot identify known variables when used on its own ($F_1 = 0$; c.f. Friedrich, Koenig, & Wicenec 1997). It does not contribute to PC1 and contributes little to PC2 (Fig. 11).

It is interesting to note that although in PC1 and PC2 the variability indices computed in the two filters show similar behavior, this is not the case for the higher order PCs. Therefore, the higher order PCs highlight differences between LCs in two filters (color changes).

As the PC2 is dominated by the indices that are robust measures of scatter while the correlation-based indices contribute less to the PC2, it becomes possible to separate short and long period variables on the $a_1$–$a_2$ admixture coefficient plots. RR Lyrae variables (marked with blue circles in Fig. 11) have short periods and uncorrelated LCs, for the cadence of the available observations, thus yielding low absolute values for $a_1$ and higher values for $a_2$. On the contrary, LPVs (marked with red squares in Fig. 11) show smooth LCs with a high degree of correlation between measurements taken close in time and hence higher values of $a_1$. As it can be clearly seen in the left panel of Figure 11 the two types of variables are clearly separated in the $a_1$–$a_2$ plane. The two first PCs are effective not only in separating variable from constant sources, but also in distinguishing RR Lyrae variables from LPVs, without the need of constructing a CMD. In general, “fast” and “slow” (compared to the observing cadence) variables occupy separate locations in the $a_1$–$a_2$ plane. There are only few variables of types other than RR Lyrae and LPV in the studied fields to check if further type separation is possible using admixture coefficients of higher-order PCs.

Many candidate variable sources that were not confirmed by our visual inspection of LCs and images occupy a distinct region on the $a_1$–$a_2$ plane (for the Halo 11 field $a_1 \lesssim 100$ and $a_2 \gtrsim 12$. C.f. Figure 14). The Disk and Stream fields show a similar location of RR Lyrae stars, LPVs and (some) false candidates in the admixture coefficient space (see Appendix).

### 5.2 Selection efficiency on the $a_1$–$a_2$ plane

As described in Section 3.4, the selection process was performed in two steps. Examination of the resulting candidates from the first selection step alone, indicated a recovery level of known variables of $\approx 80\%$, which is significantly lower than the recovery level achieved when using the fine-tuning process (for the Halo 11 field as an example). On the other hand, the initial selection yields about half the artifacts selected with fine-tuning. The artifacts that are not included in the initial selection are mostly caused by image problems, while the majority of the artifacts common in both selections are bright sources (probably affected by saturated pixels). All new variables were selected as such in both steps. Therefore, selection of variables using only the first step described in Section 5.1 provides lower completeness (as indicated by the lower recovery rate of known variables) but higher purity (i.e. lower incidence of artifacts).

Application of the fine-tuning step, provides higher completeness, but lower purity (thus increasing the number of sources that need to be expert validated). If the number of input lightcurves is very large, one may prefer to use only the first candidate step to reduce the number of candidates that has to be expected (increase the purity of the candidate list at the cost of its completeness).

### 5.3 Reduced number of source and visits

The PCA-based variability detection method proposed here is a statistical method, that requires a substantial sample of sources (variable and constant) in order to identify candidate variables. Here we investigate the applicability of the method, when the size of the sample is reduced significantly. We divided the Halo11 sample into subsamples of different sizes, ranging from 6500 to 500 sources, in steps of 500. We then applied the PCA variability detection method for each subsample and calculated the recovery rate of known variables from Brown et al. (2004) that happen to be included in the specific subsample. The tests were repeated 100 times for each subsample size, each time randomly selecting the specified number of sources from the full sample. The average recovery rate was not affected significantly. This test indicates that the method can be applied even to relatively small samples with as few as 500 sources.

An additional sensitivity test was performed to evaluate the efficiency of our method when the number of data points in the LC is reduced. The experiment was conducted using the Halo11 LCs. Epochs were randomly omitted from the LCs and the PCA and selection analysis repeated each time, following the same steps as described in Section 3. The
Figure 10. Coefficients $c_{i,j}$ determining relative contributions of variability indices to the first six principal components for the M31 Halo11 data set; see equation (2). The indices computed using the F606W LCs are shown with green triangles and the ones computed using the F814W LCs are shown as red squares. The dashed line indicates zero contribution of an index to the PC.

The experiment was repeated for 30, 20, 10 and 5 data points in the LC. We do not consider LCs with smaller number of points, as the variability indices (Table 2), while being mathematically defined, are expected to lose their predictive power.

The decrease of the number of points in the LC expectedly increases the size of the constant star locus on the $a_1-a_2$ plane, thus degrading the detectability of variables that lie close to the constant star locus in the original selection plane of Figure $3$ (upper left panel). When only 5 points are retained in the LC, additional 10 known variables are not recovered, reducing the recovery rate to $\approx 77\%$, without appreciably increasing the false variable detection. This is still an acceptable result, indicating that the method can be successful even with a small number of epochs in the LC.

5.4 Comparison with a single-index search

We compare our new variability detection technique relying on identification of isolated points in the $(a_1, a_2)$ plane (presented in Section 4) to the conventional methods discussed by Sokolovsky et al. (2017a). These methods identify candidate variables as objects having a value of a single variability index (Table 2) above some magnitude-dependent threshold (Fig. 14). The indices are computed independently for F606W and F814W LCs. The authors also suggested to use $a_1$ as a composite variability index. We test this approach (Fig. 14 top panel) combining in $a_1$ for each object all the variability indexes computed using its LCs in F606W and F814W bands.

To evaluate the quality of a variability search in a given set of LCs we need to know the ground truth: which objects are variable and which are not. As a working approximation of this we use a list of variables found by Brown et al. (2004) and Jeffery et al. (2011) cross-matched with the HSC (Table 3) together with the results of our visual inspection of candidate variables identified with the $(a_1, a_2)$ technique (Sec. 4.2). Following Sokolovsky et al. (2017a); Pashchenko, Sokolovsky, & Gavras (2017) we adopt the $F_1$-score (van Rijsbergen 1974) as the success metric of variability detection. The $F_1$-score

$$F_1 = \frac{2(C \times P)}{C + P}$$

is defined through the completeness $C$ (also known as “recall”) and purity $P$ (“precision”) of the list of candidate variables (Graham et al. 2014):

$$C = \frac{\text{Number of variables in the list}}{\text{Total number of variables in the field}}$$

$$P = \frac{\text{Number of variables in the list}}{\text{Total number of candidates in the list}}$$

$F_1 = 1$ for the perfect selection of variables when one retrieves all the true variables and no false candidates while $F_1 = 0$ if no true variables got into the list of candidates.

Table 5 compares the $F_1$-scores reached by the PCA-based and the conventional single-index techniques in the three investigated fields. For each individual index we use a $3\sigma$ threshold to select candidate variables. Despite a smaller median number of visits (Table 1), for most individual variability indices the F606W LCs result in higher $F_1$-scores than the F814W LCs. This may be attributed to the fact that pulsating stars tend to have higher variability amplitudes at shorter wavelengths and are therefore easier to distinguish from non-variable stars.
The PCA-based search results in consistently high $F_1$ values compared to candidate selection based on most individual indices, however the PCA $F_1$ is the highest one only in the Disk field. In the Halo11 field, the PCA search is outperformed by MAD, $J$ (clip) and $L$ (clip) computed on F606W LCs. In the Stream field MAD computed on the F814W LCs results in the $F_1$ value higher than the one we obtain with the PCA, while IQR (F814W) results in only a slightly lower $F_1$. The MAD computed for F606W LCs results in much lower $F_1$ values compared to MAD–F814W in the Stream field, while in the Halo11 field it is the opposite: MAD–F606W has the $F_1$-score much higher than MAD–F814W. Our PCA variability search based on two admixture coefficients ($a_1$, $a_2$) for all three fields results in higher $F_1$ values compared to the PCA-based approach relying on a magnitude dependent cut in $a_1$ considered by Sokolovsky et al. (2017a).

It is hard to predict which index will be the most efficient variability indicator in a given data set. The efficiency of an index in identifying variables depends on variability type, observing cadence, percentage of outlier measurements and the level of correlated (systematic) noise in the data. The authors suggested IQR and $1/\eta$ as the indices that tend to perform well on diverse test data, while not necessarily

\begin{table}
\centering
\caption{Variability detection efficiency ($F_1$-score).}
\label{tab:variability_detection}
\begin{tabular}{lccc}
\hline
Index & Filter & Halo11 & Disk & Stream \\
\hline
PCA-based search & & & & \\
(a_1, a_2) & 0.628 & 0.434 & 0.595 \\
$a_1$ & 0.564 & 0.373 & 0.530 \\
\hline
Scatter-based indices & & & & \\
$\chi^2_{\text{red}}$ & F606W & 0.419 & 0.154 & 0.260 \\
$\sigma_{\text{red}}$ & F814W & 0.344 & 0.151 & 0.316 \\
$\sigma$ & F606W & 0.532 & 0.230 & 0.412 \\
$\sigma$ & F814W & 0.402 & 0.180 & 0.440 \\
MAD & F606W & 0.657 & 0.308 & 0.442 \\
MAD & F814W & 0.467 & 0.281 & 0.633 \\
IQR & F606W & 0.617 & 0.347 & 0.488 \\
IQR & F814W & 0.516 & 0.380 & 0.593 \\
RoMS & F606W & 0.602 & 0.304 & 0.454 \\
RoMS & F814W & 0.528 & 0.322 & 0.547 \\
$\sigma_{\text{NXS}}^2$ & F606W & 0.381 & 0.137 & 0.258 \\
$\sigma_{\text{NXS}}^2$ & F814W & 0.314 & 0.126 & 0.290 \\
$v$ & F606W & 0.447 & 0.195 & 0.371 \\
v & F814W & 0.350 & 0.146 & 0.323 \\
$K$ & F606W & 0.000 & 0.000 & 0.000 \\
$K$ & F814W & 0.000 & 0.000 & 0.000 \\
$\chi^2_{\text{NXS}}$ & F606W & 0.314 & 0.126 & 0.290 \\
$\chi^2_{\text{NXS}}$ & F814W & 0.314 & 0.126 & 0.290 \\
\hline
Correlation-based indices & & & & \\
$J$ & F606W & 0.551 & 0.429 & 0.447 \\
$J$ & F814W & 0.396 & 0.354 & 0.444 \\
$J$(time) & F606W & 0.494 & 0.361 & 0.467 \\
$J$(time) & F814W & 0.373 & 0.295 & 0.410 \\
$J$(clip) & F606W & 0.671 & 0.381 & 0.452 \\
$J$(clip) & F814W & 0.566 & 0.335 & 0.552 \\
$L$ & F606W & 0.553 & 0.329 & 0.379 \\
$L$ & F814W & 0.404 & 0.336 & 0.354 \\
$L$(time) & F606W & 0.481 & 0.372 & 0.467 \\
$L$(time) & F814W & 0.391 & 0.298 & 0.378 \\
$L$(clip) & F606W & 0.660 & 0.376 & 0.504 \\
$L$(clip) & F814W & 0.557 & 0.379 & 0.567 \\
$E_x$ & F606W & 0.505 & 0.158 & 0.246 \\
$E_x$ & F814W & 0.379 & 0.193 & 0.232 \\
l_1 & F606W & 0.271 & 0.107 & 0.000 \\
l_1 & F814W & 0.068 & 0.037 & 0.000 \\
l_1/\eta & F606W & 0.193 & 0.099 & 0.055 \\
l_1/\eta & F814W & 0.176 & 0.091 & 0.056 \\
$S_B$ & F606W & 0.257 & 0.099 & 0.146 \\
$S_B$ & F814W & 0.232 & 0.104 & 0.121 \\
\hline
\end{tabular}
\end{table}

The $F_1$-score is defined by equation (3); see Table 2 for additional information on the variability indices.

![Figure 11](image.png)

Figure 11. Candidate variable selection in the Halo11 field above magnitude-dependents cuts (solid line) in $a_1$, IQR and $1/\eta$ (Table 2).
being the best indices for any given data set. In the Halo11, Disk and Stream data sets considered here, IQR is among the best variability indices, while $1/\eta$ is among the worst ones. This is easily understood as $1/\eta$ being the measure of correlation (smoothness) of a lightcurve is unable to detect the numerous RR Lyrae variables in these data sets. The observing cadence is long compared to RR Lyrae periods, making their LCs appear smooth only when folded with the correct period (Fig. 3). Only the LPVs which have smooth LCs plotted as a function of time can be identified with $1/\eta$ (and other correlation-based indices).

The PCA-based search results in higher $F_1$ values compared to both IQR and $1/\eta$ for the three studied fields. MAD, another outlier resistant index, may show both higher and lower $F_1$-scores compared to both IQR and the PCA-based search, depending of field and filter. This suggests that, unless it is somehow known a priori which variability index is the best one for the studied data set, it is more efficient (in terms of archiving a higher $F_1$-score) to compute multiple variability indices and combine them via the PCA rather than use the “safe” indices, MAD, IQR and $1/\eta$.

While the PCA constructs a linear combination of indices, machine-learning may help to find useful nonlinear combinations of indices resulting in higher $F_1$-scores (Pashchenko, Sokolovsky, & Gavras [2017]). However, a variability search based on supervised machine learning requires a representative training set of LCs pre-classified as variable or non-variable by some other means and thus cannot be used for a blind variability search in a small field. Training a machine learning classifier on LCs obtained with one survey and then applying this classifier to LCs from another survey (“knowledge transfer”), while being in principle possible, remains an unsolved problem in practice. Variability search based on unsupervised machine learning may be a promising approach for conducting blind variability surveys (e.g. Shin et al. 2009, 2013; Mackenzie, Pichara, & Protopapas 2014).

6 CONCLUSIONS

We investigate a new method of variable object detection in a large set of light curves, which is based on principal component analysis. Each light curve is characterized by its mean magnitude and a set of variability indices (Table 2) that are used as the input for the PCA. Candidate variable objects are identified as outliers in the plane of admixture coefficients $a_1-a_2$ corresponding to the two most significant principal components. The proposed method is suitable for (and most efficient in) large sets of LCs. It requires no a priori information about the type of the searched variable objects. Instead it relies on assumptions that variable objects are rare and the indices listed in Table 2 are able to capture variability information – the assumptions that hold true for most ground-based and many space-based observations in optical and near-infrared bands. This methodology can indeed successfully identify variable stars not only in the HSC database, but also in ground based data with different sample sizes and epochs as discussed in Section 5.3. There is no need to preselect the “best” variability indices for different samples, since PCA outputs their optimal linear combinations. Human intervention is still necessary to validate the resulting candidate variables (see discussion in 5.4).

The present algorithm is performed in the framework of the HCV, which will be available in 2018.

The method is verified using 18152 LCs of stars in 3 fields in M31 extracted from the Hubble Source Catalogue. We recovered almost 90% of the known variables reported by Brown et al. (2004) and Jeffery et al. (2011). We found 38 new variable stars, among which 37 LPVs and one object of an uncertain variability type (Table 5). This demonstrates that the Hubble Source Catalogue, despite its shallower depth and reduced time resolution compared to what can be obtained with dedicated manual analysis of the archival HST images, includes many previously unknown variable objects in fields previously studied for variability.

ACKNOWLEDGEMENTS

We acknowledge financial support by the European Space Agency (ESA) under the “Hubble Catalog of Variables” program, contract No. 4000112940. This work uses the HSC, based on observations made with the NASA/ESA Hubble Space Telescope and obtained from the Hubble Legacy Archive, which is a collaboration between the Space Telescope Science Institute (STScI/NASA), the Space Telescope European Coordinating Facility (ST-ECF/ESAC/ESA) and the Canadian Astronomy Data Centre (CADC/NRC/CSA). This research has made use of NASA’s Astrophysics Data System. We acknowledge support of the HCV by Danny Lennon and Brad Whitmore (B.W.), useful discussions by B.W., Ioannis Bellas-Velidis, Ektoras Pouliasis, Zoi Spetsieri and critical comments on the manuscript by Vassilis Charmandaris.

REFERENCES

Abbas M. A., Grebel E. K., Martin N. F., Burgett W. S., Flewelling H., Wainscoat R. J., 2014, MNRAS, 441, 1230
Ahn C. P., et al., 2014, ApJS, 211, 17
Alonso, R., Brown, T. M., Charbonneau, D., et al. 2007, Transiting Extrasolar Planets Workshop, 366, 13
Angeloni R., et al., 2014, A&A, 567, A100
Auvergne, M., Bodin, P., Boisnard, L., et al. 2009, A&A, 506, 411
Bailer-Jones, C. A. L., Irwin, M., & von Hippel, T. 1998, MNRAS, 298, 361
Bakos, G.; Noyes, R. W., Kovács, G., et al. 2004, PASP, 116, 266
Benedict G. F., McArthur B. E., Nelan E. P., Harrison T. E., 2017, PASP, 129, 012001
Bernard E. J., et al., 2010, ApJ, 712, 1259
Bernard E. J., et al., 2013, MNRAS, 432, 3047
Bertin, E., & Arnouts, S. 1996, A&A Suppl., 117, 393
Bonanos, A. Z., & Stanek, K. Z. 2003, ApJL, 591, L111
Borucki, W. J., Koch, D., Basri, G., et al. 2010, Science, 327, 977
Brown, L. M. J., Robson, E. I., Gear, W. K., & Smith, M. G. 1989, ApJ, 340, 150
Brown, T. M., Ferguson, H. C., Smith, E., et al. 2003, ApJL, 592, L17
APPENDIX A: DISK AND STREAM FIELDS

In this appendix we show similar figures as those included in the main part of the paper but for the Disk and Stream fields. The presentation and discussion provided for field Halo11 are also applicable to the Disk and Stream fields unless specifically indicated. We also present the LCs of all new variables identified in the Halo11 (Fig. A5), Disk (Fig. A6) and Stream (Fig. A7) fields.
Figure A1. The first six principal components for the M31 Disk data set. The dashed line indicates a zero contribution of an index to the PC. Green triangles correspond to F606W, red squares to F814W.

Figure A2. Same as Fig. A1 but for the Stream field.
Figure A3. Same as Figure 4 but for the M 31 Disk. The candidate variables obtained with the method described in Section 3.4 are highlighted with black points, while newly discovered variables are marked with big black open circles. Small blue open circles, the black open pentagon and reversed black open triangle highlight RR Lyrae stars, anomalous and classical Cepheid variables, respectively, from Jeffery et al. (2011).

Figure A4. Same as Figure 4 but for the M 31 Stream. The candidate variables obtained with the method described in Section 3.4 are highlighted with black points, while newly discovered variables are marked with big black open circles. Small blue open circles highlight RR Lyrae variables from Jeffery et al. (2011).
Figure A5. Light curves for the new variables in the Halo11 field, ordered by F814W magnitude.
Figure A6. Light curves for the new variables in the Disk field, ordered by F814W magnitude.
Figure A6 – continued
Figure A7. Light curves for the new variables in the Stream field, ordered by F814W magnitude.