# Handbook SubPCA: a galaxy subtraction algorithm

July 16, 2015

Version
Jully 2015

GDL SubPCA algorithm

**What you will find here :**

- a general overview,
- a step-by-step tutorial,

Last compilation : July 16, 2015 by Remy Joseph, Danka Paraficz, Frederic Courbin
LaTeX 2$_\varepsilon$ KOMA document class "LaTeX for Europe"
danuta.paraficz@epfl.ch,remy.joseph@epfl.ch,frederic.courbin@epfl.ch, Laboratory of Astrophysics EPFL
Swiss Federal Institute of Technology Lausanne

# Contents



# 1 Introduction and overview

This algorithms aims at subtracting central galaxies from patch images taken from a field of view or from a fits cube of stamp images using machine learning techniques. Each stamps is centered on a Galaxy. This algorithm has been designed to facilitate galaxy-galaxy strong lenses automated detection (Joseph et al. 2014) by subtracting only the central galaxy, leaving any companion or lensed arc untouched and de-blended from lens galaxy. The principle of our algorithm relies on Principal Component Analysis (PCA). The idea is to teach our machine a basis that is fit to represent sets of galaxies with similar properties while keeping it from learning possible shapes of companion images. Then, we use this basis to reconstruct stamp images of galaxies. Since our algorithm can only learn the shape of central galaxies, the reconstruction will be only partial and any feature other than the central galaxy will thus be ignored. When subtracting this partial reconstruction to the original image, only the central galaxy will be removed.

A traditional way of subtracting galaxies is to fit a two dimensional elliptical profile to the data, e.g. as done with the `galfit` software (Peng et al. 2011). While this is sufficient

to characterise the main morphological properties of galaxies, it turns out to be insufficient to detect faint arcs seen superposed on bright galaxies with a significant level of resolved structures. In particular, in the case of galaxy-galaxy lensing, residuals can often present ring or arc like structures that could easily be misclassified by an automated lens finder. One way to circumvent the problem is to build an empirical light model from the sample of galaxies itself, i.e. to use machine learning techniques like PCA ( Jollie 1986). The sparsity and the diversity in terms of shape of the lensed objects (rings, arcs, multiple images) or companions, prevents them from being well enough represented in the basis, hence allowing for an accurate separation of lenses and sources. This has already been used to find lensed sources from PCA decomposition of quasar spectra (e.g. Courbin et al. 2012; Boroson & Lauer 2010). We adopt here a similar strategy to analyse images.

## 1.1 Principal component analysis

Principal component analysis (PCA) is a procedure that converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. Fig. 1 displays examples of the full PCA procedure implemented in this code.

The PCA analysis is computed by building a matrix $X_b$ in which each row is an image of selected galaxies, reshaped as a vector. PCA is done here by singular value decomposition which is performed on the matrix of $X_b$:

$$X_b X_b^T = UWU^T \tag{1}$$

where $W$ is a diagonal and $U$ is a unitary matrix, $U^T$ is the conjugate transpose of U. The result of PCA is the estimated reshaped image. A partial reconstruction of the image is done by using only the k-first coefficient of the PCA:

$$X_b = \alpha_{set} Ei \tag{2}$$

where $Ei$ are the eigenvectors for the decomposition of $X_b$

$$Ei = X_b U^t W^{-1/2} \tag{3}$$

The decomposition of an $n \times n$ image of galaxy reshaped as a column vector, $X_{set}$ can now be decomposed via:

$$\alpha_{\rm set} = Ei^T X_{\rm set}, \tag{4}$$

where $\alpha_{\rm set}$ is a N-sized vector of PCA coefficient that represents the image $X_{\rm set}$, reshaped as a vector.

Since a field of view can present a rather large variety of galaxies in terms of morphologies, we allow the user to narrow down the sample of galaxies used both in learning process and in the subtraction. Indeed, when trying to subtract galaxies with larger fwhm using a basis built upon a population of galaxies massively small can become challenging for our algorithm and could force the user to use larger number of coefficients leading to possible representation of companion images in the stamp. To circumvent this issue, we allow the user to set cuts in size and magnitude when selecting galaxies to subtract to make sure the basis is adapted to a specific population. It is unadvised to subtract the whole sample of galaxies from a field of view since the ones with rarest morphologies will be under-represented in PCA basis leading to low quality subtraction.

A critical step in the PCA reconstruction is the choice of the number of PCA coefficients. If all of the coefficients are used, the reconstruction will include elements of the basis that

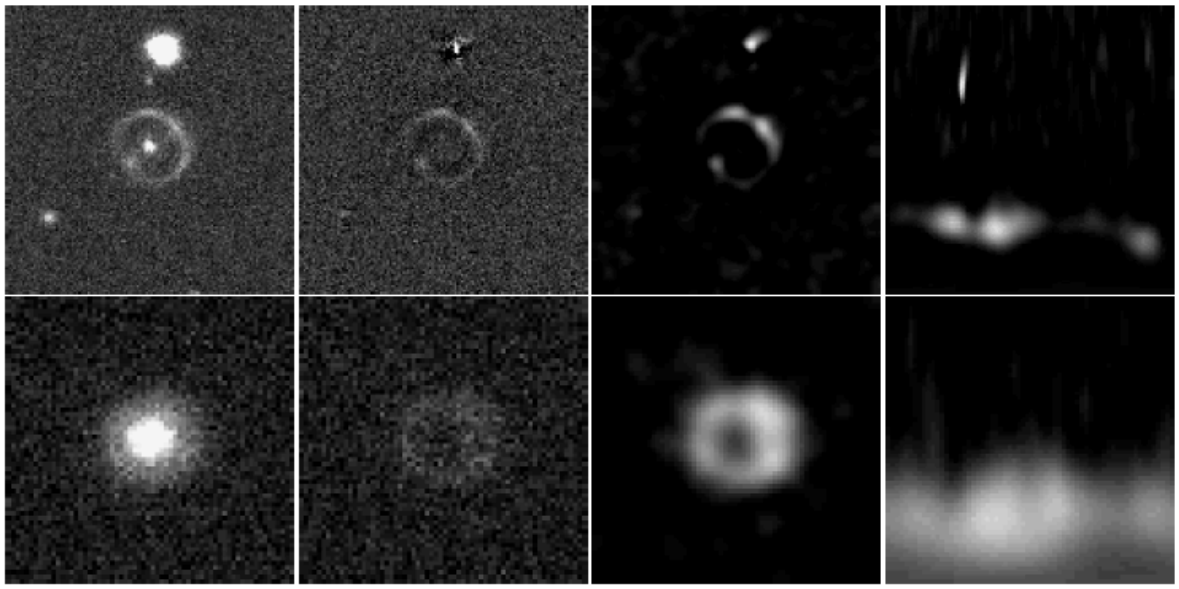
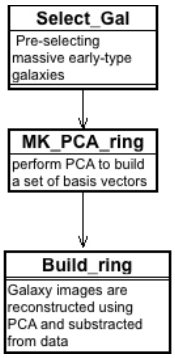


**Figure 1 :** Illustration of the ring finding process for two Einstein rings. For each row, from left to right are shown 1- an Einstein ring with its lens galaxy, 2- the lensed ring after PCA subtraction of the foreground galaxy, 3- the result of curvelet denoising, 4- the polar transform of the ring revealing a well visible horizontal line which position along the y-axis gives a measurement of the radius of the Einstein ring.

represent the noise, hence resulting in an over-fitting of the data and to an apparent smoothing of the residual image obtained after subtraction of the galaxy. This can be damaging when trying to detect faints rings and arcs. Conversely, if the number of coefficients is insufficient the central galaxy will be only partially removed leaving significant and undesired structures in the residual image.

## 1.2 A walkthrough

PCA-lens finder is a working environment, defining a not-too-tight structure with robust scripts. SubPCA consists of an ordered list of independent scripts, each of whom is kept as simple as possible, typically applying a specific task ("run SExtractor", "build PCA lens model", ...) to every concerned image.




The steps conducted by SubPCA sre as such:

1. Pre-select the galaxies with a predefined range of shape parameters (size, ellipticities, magnitudes, etc.) using Sextractor.
2. Subtract all the selected galaxies from FoV (SExtractor).
3. Rotate all the images to 0 position angle. This is necessary to ensure similarity between galaxies and hence make the basis more representative of the range of galaxies.
4. Build a "typical" lens model for each magnitude range by performing a PCA to build a set of basis vectors.
5. Reconstruct of the central galaxies using only the first elements of the PCA basis
6. Subtract it from the original images.

## 1.3 Practical organization

The PCA-lens finder scripts come in a single tarball. The scripts to be executed by the user are:

```
select_gal
mk_PCA_ring
build_ring
```

```
export GDL_PATH=/home/epfl/your_path/IDL_code:\$GDL_PATH
```

Soon a more general version should be available that would allow for a one command execution of the code, but the current architecture allows more freedom in the way users can handle the different steps of the PCA subtraction. Indeed, each step will write its main results in a fits file that is opened by the next routine. This allows shortcuts such as creating a PCA basis with a given set of images and then using it on an other field of view without having to reprocess the whole analysis.

You need to have also:

- The science image (and its weight image) where you want to find lenses.
- Sextrator configuration and parameter files: `correct_temp.sex, correct_temp.param, field.sex, field.param, default.conv, default.nnw`.

`FIELD.PARAM` should have:
`NUMBER, MAG_AUTO,X_IMAGE, Y_IMAGE, A_IMAGE, B_IMAGE, THETA_IMAGE`
`CORRECT_TEMP.PARAM` should have:
`NUMBER, MAG_AUTO,X_IMAGE, Y_IMAGE`

Computational time is an important parameter to consider. Building the PCA basis involves finding the eigenvectors and the eigenvalues of a $N^2 \times N_{\rm gal}$ matrix, where `N` is the number of pixels per stamp and where $N_{\rm gal}$ is the number of stamps in the training set.

# 2 Requirements and dependencies

SubPCA focused on the ESO scisoft distribution (http://www.eso.org/sci/data-processing/software/scisoft/). Installing this is the easiest solution to meet most of the requirements.

## 2.1 IDL/GDL

First off all, we need IDL or free GNU version of it, GDL. In case of IDL ideally version 8.1 (as comes with a recent scisoft). But older version should also work fine. GDL newest version is advised. All IDL/GDL codes that are called in by SubPCA should in principle be included in the GDL code tarball that you can download. If not, please let one of the authors know.

## 2.2 SExtractor

Source Extractor, by Emmanuel Bertin (Bertin et al. 1996), is an essential tool for SubPCA to work properly. It has been successfully tested on the latest V2.19.5 version that you can download at http://www.astromatic.net/software/sextractor

# 3 Select Galaxies

Selects galaxies from full scientific images or already clipped galaxy patches in order to build a set of galaxies suitable for ring extraction by PCA method. Part of this selection is manual. In order to make sure the sample of galaxies used in the training set is representative enough of the galaxies to subtract, we advise the user to take a look at the distribution of ellipticities and magnitude and set parameters accordingly. 3 parameters: `low, high and magmax` that are detailed below, allow to control the selection function. This script therefore needs to be ran few times if one wants to create couple of sets of galaxies with different sizes and brightness (eg. two sets of small/faint, and large/bright galaxies). We advise to have at least a hundred or more galaxies selected. Users should keep in mind that a subselection can be applied later on to select the most adequate stamps to build the basis. Therefore it is reasonable to have a few hundreds of galaxies selected in order to account for this.

The script creates image stamps of the selected galaxies, and rotates all the galaxies so that their major axes are all aligned and centered in their image stamp. The rotation is performed using a polynomial transformation and a bilinear interpolation. It writes the result in a `fits` file. (25 minutes maximum for a large sample of galaxies)
SYNTAX:

```
select_gal,path, fitsfileout, low, high, magmax, img=img, n1, n2,
/patches, /havesex, ngal
```

EXAMPLE:

```
select_gal,'./','Galaxies_4_5_25.fits',4,5,25,img='FoV.fits',
100.,100,/havesex
```

1. Executes `sextractor` over whole the FoV.
2. From `sextractor` output catalog `field_objects.txt` it selects galaxies that match the input criteria (`low, high and magmax`) in order to form an homogeneous set of galaxies in terms of size and magnitude for the PCA.

3. Runs sextractor on small patches to find the exact location of the centroid and find possible bright companions around central galaxies.

4. Rotates the galaxies so that every patch will have galaxies with the same inclination.

5. Writes the properties of the selected galaxies in a file : position, number, original alignment in select_gals.txt and selected galaxies in a fits file.

## 3.1 Inputs

### 3.1.1 Mandatory

- PATH the path where to look for the FoV or the patches to be processed. .
- LOW, HIGH : respectively lower and higher thresholds (in pixels) of semimajor axis selection of galaxies as measured by sextractor.
- MAGMAX : maximum magnitude allowed for galaxies to be selected, as measured by sextractor.
- N1, N2 sizes of the stamps to extract from the field of view. These stamps must be at least sqrt(2)*NPCASTAMP, NPCASTAMP being the size of the stamps used in the final PCA. Keep in mind that stamp size needs to be big enough to fit the rotated galaxy, therefore, bigger HIGH means bigger N1, N2 (obligatory).

### 3.1.2 Optional

- IMG: name of the Field of View (in a fits file) where to select galaxies from (mandatory if option /HAVESEX is set).

## 3.2 Keywords

- /PATCHES if set, the program will look for a listfits.txt file with the names of every stamps to use. Both the listfits.txt file and the stamps should be located in the PATH folder (mandatory if patches are to be used instead of FoVs).
- /HAVESEX if set, the program will perform selection on a field of view given by IMG.

## 3.3 Outputs

select_gal writes selected galaxies in a fits cube (FITSFILEOUT) in the current folder.

- FITSFILEOUT : name of the output file (mandatory).

  There are two files produced by sextractor

  - teste_final_apertures_cris.fits - Output of sextractor, checkimage. It is an image presenting selected galaxies with two radii visible. Good to know: as soon as you specify in your sextractor file either AUTO (Kron) parameters or PETRO parameters for the output catalog, these apertures will be drawn in the checkimage (APERTURES). The Petrosian is usually the outer radius.
  - field_objects.txt - the output catalog.

## 3.4 External Calls

Calls `sextractor`. The following input files are required to run sextractor and provided in the tarball:

- `field.sex`, `field.param` - `sextractor` configuration file for whole FoV
  `NUMBER, X_IMAGE, Y_IMAGE, A_IMAGE, B_IMAGE,THETA_IMAGE`
- `correct_temp.sex`, `correct_temp.param` - `sextractor` configuration file for galaxies stamps
  `NUMBER, X_IMAGE, Y_IMAGE`

# 4 Make PCA

Performs PCA on a subset of clean, single galaxies. It selects only galaxies with no bright companions or with companions far away from the center of light, as any companions are possible sources of artefacts. If we do not remove such galaxies PCA would create fake lensed objects or conversely remove part of the lensed object at the end of the process. Selecting only single galaxies results of course in reducing the size of the PCA basis.
SYNTAX:

```
mk_PCA_ring,path, fitsfilein, /training_set, T_set=T_set,
/gauss, fwhm=fwhm, /select, /iset, input_set=S_base, prefix=prefix
```

EXAMPLE:

```
mk_PCA_ring,'./','Galaxies_4_5_25.fits',/select
```

1. Builds the images used to form the basis.
2. Multiplies by a gaussian profile, if required.
3. Selects or rejects candidates for the basis.
   `r lt fwhm && gal(x[j], y[j]) gt gal[n1/2,n2/2]/2`
4. Builds covariance matrix and its eigenvalues.

## 4.1 Inputs

### 4.1.1 Mandatory

- `PATH` : Working folder (usually same as in `select_gal.pro`).
- `FITSFILEIN` : name of file containing patches of galaxies (created by `select_gal.pro`).

### 4.1.2 Optional

- `FWHM` : fwhm of the gaussian that is to be convolved if `/GAUSS` set.
- `T_set` : set that can be used to build the PCA basis thus not relying anymore on the original data only.
- `PREFIX` : adds a prefix to Eigen files when writing.

- INPUT_SET=S_BASE : if /ISET is set, S_base is used to perform PCA analysis. This means the algorithm does not have to calculate the PCA basis but uses one already built.

## 4.2 Keywords

- /GAUSS : if set, convolves the images in the set by a gaussian with the given FWHM.
- /TRAINING_SET if set, the program expects T_SET to be a set of images as an array of line vectors (each line vector being a reshaped galaxy). This set will be used to form the basis. If not set, mk_pca_ring will look for suitable candidates to form the basis in the input file.
- /SELECT : if set, the images used to build the basis will be selected using only galaxies with no or week companions. S_base and S_set will be different in this case. It usually improves the quality of subtractions but increases computational time (optional, but advised).
- /ISET : similar as /TRAINING_SET. In this case S_BASE is used as a basis for the PCA and a set to decompose. This is a way to skip the select_gal.pro process if sets are already selected by the user.

## 4.3 Outputs

- Eigen_vec_rmring.fits : vectors of the generated PCA basis.
- Eigen_val_rmring.fits : coefficients of the decomposition of the vectors in S_SET over the generated PCA basis.

Both of these files are necessary for the subtraction step to come.

# 5 Build Ring

Builds images of galaxies according to their decomposition and compares them to the original images. The quality of the PCA reconstruction depends on 3 main factors: 1- the range in galaxy sizes (set in `select_gal`), 2- the presence of companions near the galaxies used to build the PCA basis (hence `base_cleaning`), 3- the number of PCA coefficients (`nvecini`).

To estimate the number of PCA components, it is advised to carry out several reconstructions with an increasing number of PCA coefficients. Stop adding coefficients when reaching an acceptable quality, i.e. when there is no residual above the noise level. At the end of the execution the build_ring routine displays a a value q that indicates the quality of the removal. One should aim for a value of q as close as possible to 1-1.5.

The routine will look for Eigen_vec_rmring.fits and Eigen_val_rmring.fits files in the current directory. They should have been created when executing the mk_pca_ring routine.

SYNTAX:

```
build_ring, path, fitsfilein, fitsfileout, /training_set, T_set = T_set,
nvecini, /base_cleaning
```

EXAMPLE:

```
build_ring,'./','Galaxies_4_5_25.fits','Residuals_4_5_25_75.fits',75,/base_cleaning
```

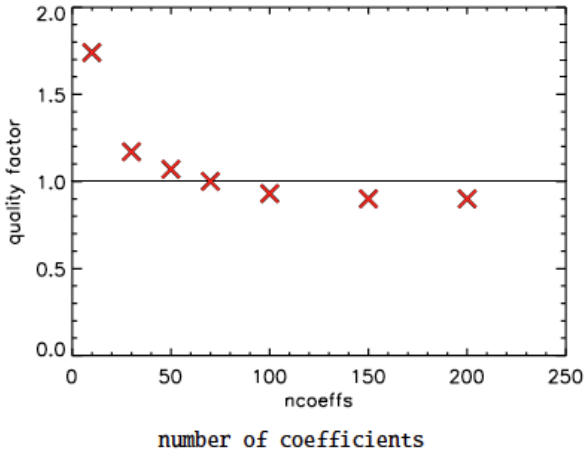



**Figure 2 :** Quality factor as a function of the number of coefficients used in the reconstruction. Only 50-70 coefficients are needed to reach $q \sim 1$ in the case of CFHT images from Stripe82.

1. Original images are cut in patches of the same size as the reconstructed images.
2. Images with a limited number of coefficients are rebuild.
3. If asked, the PCA is performed a second time on a set in which polluting companions have been removed (takes more computational time, but results in better quality subtraction).
4. Creates a clean version of the set of images (no polluting companion) by subtracting the residuals once smoothed to the original image.
   `isse = transpose(psf_gaussian(ndimen = 2, npixel = 10, fwhm = 0.5))`
5. Smooths the residuals.
6. Finally, produces clean images without central galaxies.

## 5.1 Inputs

### 5.1.1 Mandatory

- `PATH` : path to file containing patches of galaxies (the usual thing).
- `FITSFILEIN` : name of file containing patches of galaxies (created by `select_gal.pro`).
- `FITSFILEOUT` : the name of the output file where the residuals will be written.
- `NVECINI` : limiting number of coefficients that rebuilds images. You can begin with 50-100 coefficients. At the end, the script will return a value for q that indicates the quality of reconstruction. $1 < q < 1.5$ is usually good. If $q > 1.5$, try again with more coefficients.

### 5.1.2 optional

- T_SET : if /TRAINING_SET is set T_SET is expected to provide a set of images with central galaxies to be subtracted on an already existing PCA basis. (optional)

## 5.2 Keywords

- /TRAINING_SET : if set, T_SET needs to be provided.
- /BASE_CLEANING : if set, a second iteration of PCA will be performed on cleaner images of galaxies. Indeed, once central galaxies have been reconstructed and subtracted, the resulting residuals are smoothed (by convolving the images with a gaussian profile) and subtracted from the original images leaving only the central galaxies intact. These newly formed images are used to form a new basis for PCA analysis. Then the reconstruction/-subtraction process is repeated producing images free from their central galaxy with a limited number of coefficients.

## 5.3 Outputs

- Residuals : clean images without central galaxies.
- QFACTOR : factor of reconstruction is reported in the screen. This factor is the median value of chi square between original and reconstructed images computed in the central region of images. Values between 1 and 1.5 are considered good reconstructions.