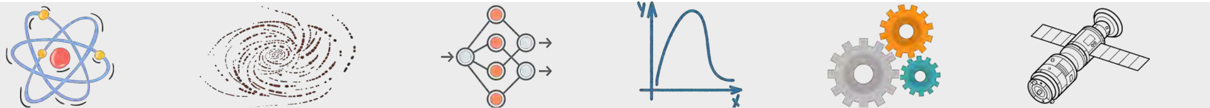
Harnessing Tailored Statistical Techniques to Discover Star Clusters

KICC Focus Meeting on Astrostatistics & Astro-ML

Rafael S. de Souza

Chair: The Cosmostatistics Initiative

University of Hertfordshire



Astrophysical Outline

- Why Star Clusters?
- The Old & The Young
- Optimizing discoveries
- How star clusters can map the galaxy and beyond

Methodological Outline

Some tailored solutions for everyday statistical problems

- Missing data imputation
- Low-rank Heteroskedastic data-denoising
- GPU Scalability
- And old-school Hierarchical Bayesian Models

Types of Star Clusters

- Young stellar objects clusters
  - Offers a glimpse into early star and planet formation processes.
  - They are independent tracers of the galactic spiral arms structure.
- Open Clusters (OCs)
  - Comprised of stars of mixed ages and higher metallicity, OCs map galactic chemical enrichment.
  - Their location helps tracing the galaxy's spiral structure and star formation history.
- Globular Clusters (GCs)
  - Old, metal-poor stars, they are relics of the early Universe, shedding light on the formation and evolution of the Milky Way.
  - Their dynamics provide constraints on dark matter.

Mapping Young Stellar Objects in the Milky Way

- YSOs live in regions of intense star formation.
- They enable to map of the galactic structure. Because they are close to the place they are born.
- Challenge is to identify them among 10^{8-9} objects observed by the Gaia space mission. With upcoming surveys, those numbers will be at least ten times larger.

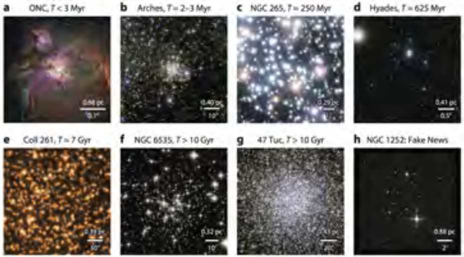


Figure 1

Images of a range of star clusters, along with NGC 1252, an object previously classified as a cluster but now known to be an asterism. The field of view in all frames is 3 pc × 3 pc, and North is up; angular sizes are indicated by scale bars. Panel *a* adapted from Robberto et al. (2013); panels *b* from NASA and ESA; panel *c* from NASA and ESA, Davide De Martin (ESA/Hubble) and Edward W. Olszewski (University of Arizona, USA); panel *d* from NASA, ESA, and STScI; panel *e* from ESO/Digitized Sky Survey; panel *f* from ESA/Hubble and NASA, Gilles Chapdelaine; panel *g* from NASA, ESA, and the Hubble Heritage (STScI/AURA)-ESA/Hubble Collaboration, J. Mack (STScI) and G. Piotto (University of Padova, Italy); panel *h* from WEBDA database, **https://www.univie.ac.at/webda/**. Abbreviation: ONC, Orion Nebula Cluster.

Mapping Young Stellar Objects in the Milky Way

- YSOs live in regions of intense star formation.
- They enable to map of the galactic structure. Because they are close to the place they are born.
- Challenge is to identify them among 10^{8-9} objects observed by the Gaia space mission. With upcoming surveys, those numbers will be at least ten times larger.

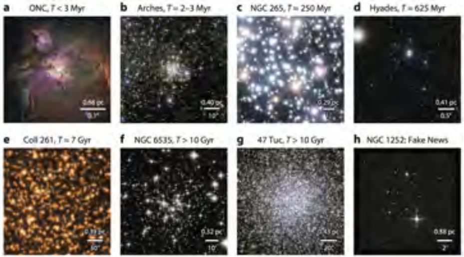


Figure 1

Images of a range of star clusters, along with NGC 1252, an object previously classified as a cluster but now known to be an asterism. The field of view in all frames is 3 pc × 3 pc, and North is up; angular sizes are indicated by scale bars. Panel *a* adapted from Robberto et al. (2013); panels *b* from NASA and ESA; panel *c* from NASA and ESA, Davide De Martin (ESA/Hubble) and Edward W. Olszewski (University of Arizona, USA); panel *d* from NASA, ESA, and STScI; panel *e* from ESO/Digitized Sky Survey; panel *f* from ESA/Hubble and NASA, Gilles Chapdelaine; panel *g* from NASA, ESA, and the Hubble Heritage (STScI/AURA)-ESA/Hubble Collaboration, J. Mack (STScI) and G. Piotto (University of Padova, Italy); panel *h* from WEBDA database, **https://www.univie.ac.at/webda/**. Abbreviation: ONC, Orion Nebula Cluster.

Mapping Young Stellar Objects in the Milky Way

- YSOs live in regions of intense star formation.
- They enable to map of the galactic structure. Because they are close to the place they are born.
- Challenge is to identify them among 10^{8-9} objects observed by the Gaia space mission. With upcoming surveys, those numbers will be at least ten times larger.

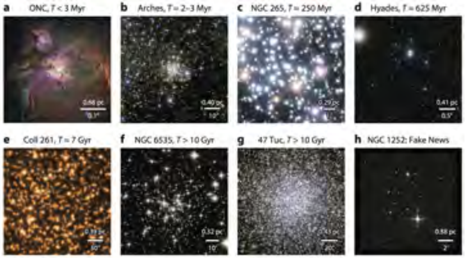
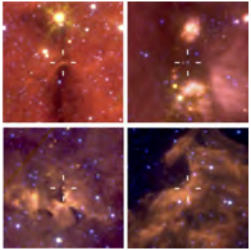
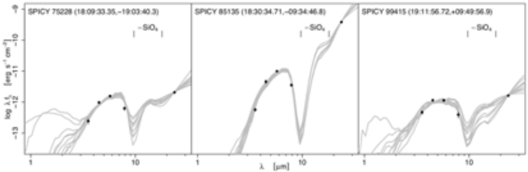


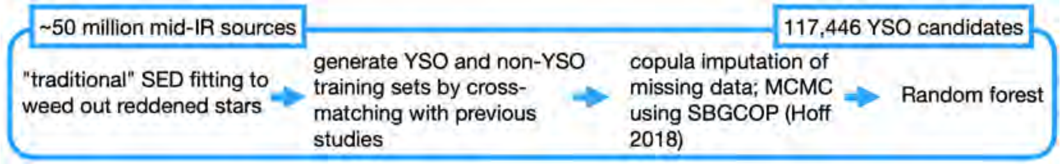
Figure 1

Images of a range of star clusters, along with NGC 1252, an object previously classified as a cluster but now known to be an asterism. The field of view in all frames is 3 pc × 3 pc, and North is up; angular sizes are indicated by scale bars. Panel *a* adapted from Robberto et al. (2013); panels *b* from NASA and ESA; panel *c* from NASA and ESA, Davide De Martin (ESA/Hubble) and Edward W. Olszewski (University of Arizona, USA); panel *d* from NASA, ESA, and STScI; panel *e* from ESO/Digitized Sky Survey; panel *f* from ESA/Hubble and NASA, Gilles Chapdelaine; panel *g* from NASA, ESA, and the Hubble Heritage (STScI/AURA)-ESA/Hubble Collaboration, J. Mack (STScI) and G. Piotto (University of Padova, Italy); panel *h* from WEBDA database, **https://www.univie.ac.at/webda/**. Abbreviation: ONC, Orion Nebula Cluster.

YSO data: Spectral Energy Distribution

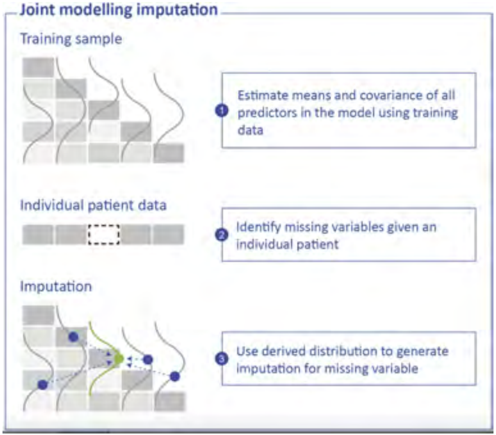






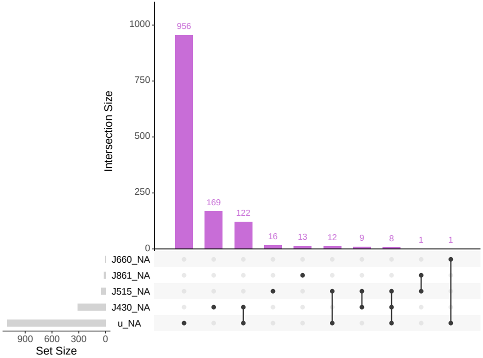
First issue: Missing data

- Most off-the-shelf approaches assume missingness at random:
- An alternative is to learn the joint distribution from the complete data, which often requires assumptions about the joint density



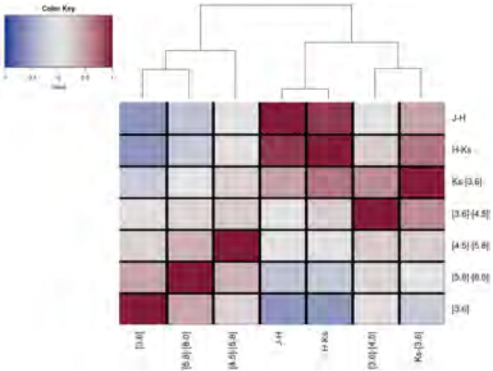
First issue: Missing data

- Astronomical data shows non-trivial missing patterns



First issue: Missing data

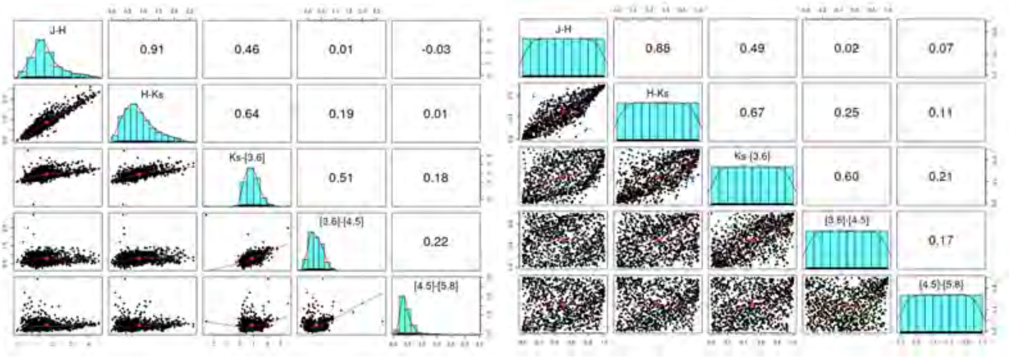
- How can we take advantage of the data's correlated structure for arbitrary marginal distributions?



Multiple Imputation via Copulas

Sklar's Theorem: Let F be a p-dimensional joint distribution function with marginals $F_1, \ldots, F_p$. Then there exists a copula C with uniform marginals such that

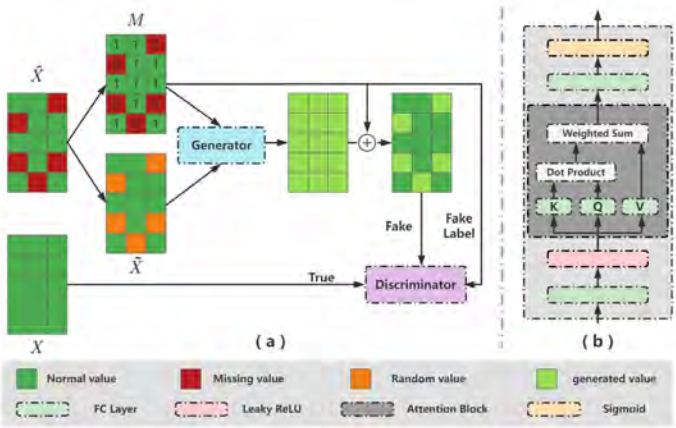
$$F(x_1, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p))$$



Multiple Imputation via Generative Adversarial Networks

MIGAN employs a self-attention mechanism, which learns a sparse representation of the relevant features for a given task (de Souza et al, in prep). Initially used for images, can be adapted to Astronomical catalogues.

Multiple Imputation via Generative Adversarial Networks

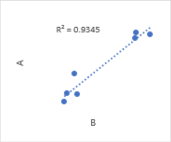


MIGAN employs a self-attention mechanism, which learns a non-local sparse representation of the data.

The MICE Algorithm

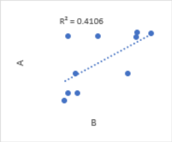
Missing data is in red. There is a strong correlation between A and B, so let's try to impute A using B and C.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
	0.80	
0.95	1.24	1.46
0.23	0.57	
0.90		1.28
0.15	0.42	
0.47	0.54	0.63
	1.14	
0.89	1.23	1.45



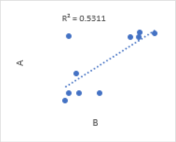
Missing data is filled in randomly. This dillutes the correlations, but allows us to impute using all available data.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.90	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	0.46	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.47	1.14	1.28
0.89	1.23	1.45



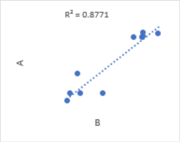
A random forest is used to predict A with B and C. Notice the correlation between A and B improved.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.24	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	0.46	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.89	1.14	1.28
0.89	1.23	1.45

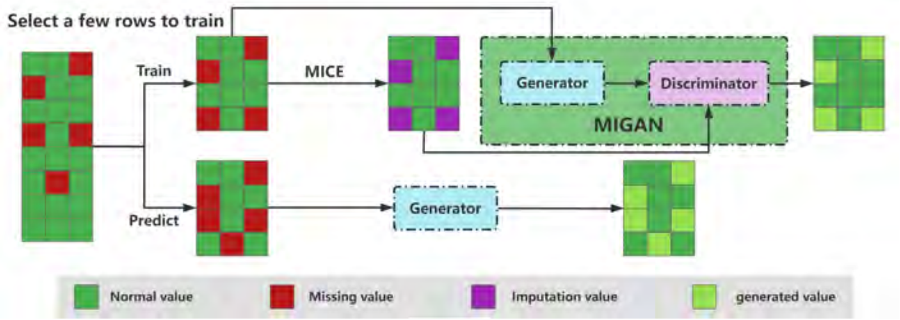


After Imputing B using A and C, we have achieved a correlation between A and B much closer to the original data.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.24	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	1.24	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.89	1.14	1.28
0.89	1.23	1.45

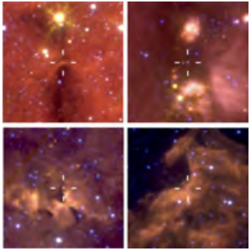


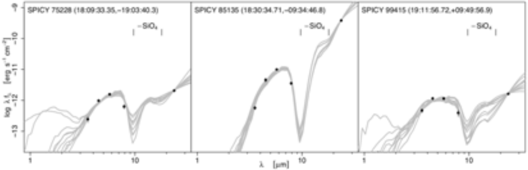
MIGAN as Emulator

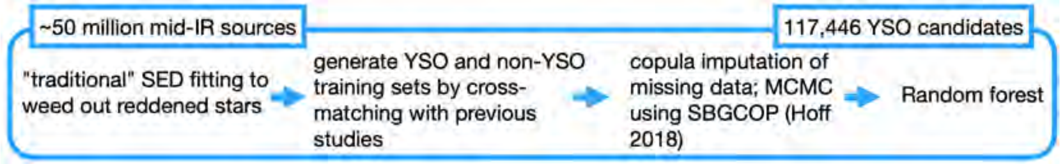


MIGAN also enables to user to mimic a particular model of choice as e.g. Multiple Imputation via Chained Equations.

YSO search pipeline



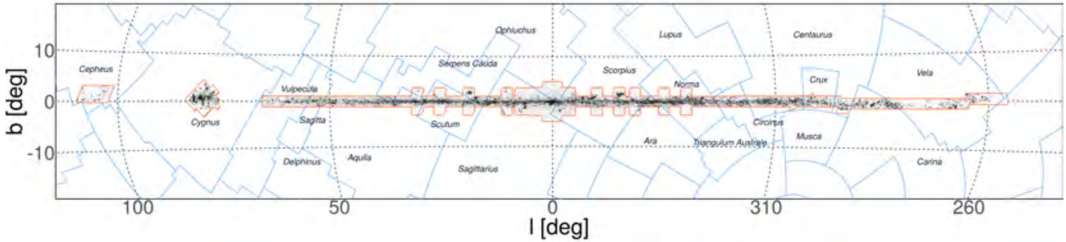




The SPitzer/IRAC Candidate YSO Catalog

COIN SPICY

The largest catalogue of YSOs ($\sim$ 200,000) in the Milky Way midplane

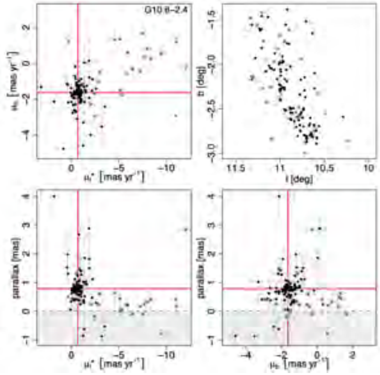


For each YSO association

For star i of a cluster, the probability distribution is,

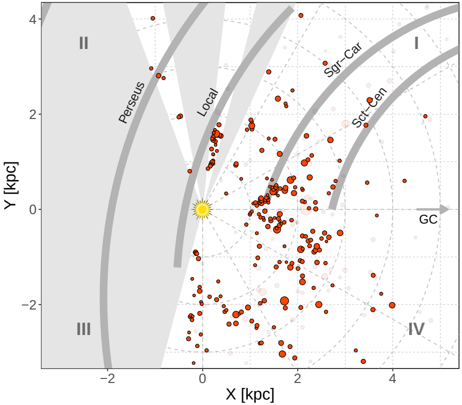
$$p_{\text{clust}}(\varpi_i, \mu_{\ell\star,i}, \mu_{b,i} | \varpi_0, \mu_{\ell\star,0}, \mu_{b,0}) =$$
$$\phi(\varpi_i | \varpi_0, \sigma^2_{\varpi_i}) \cdot f(\mu_{\ell\star,i} | \mu_{\ell\star,0}, \sigma^2_{\mu_{\ell\star,0}}, \nu_\mu) \cdot$$
$$f(\mu_{b,i} | \mu_{b,0}, \sigma^2_{\mu_{b,0}}, \nu_\mu),$$

where $\theta = (\varpi_0, \mu_{\ell\star,0}, \mu_{b,0})$ are the mean astrometric values for the cluster, $x_i = (\varpi_i, \mu_{\ell\star,i}, \mu_{b,i})$ are the measured values for the ith star, σ_i are corresponding uncertainties, ϕ denotes a Gaussian distribution, and f denotes a t-distribution.



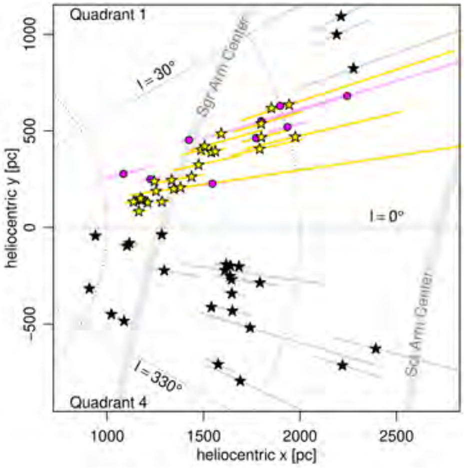
Mapping the Spiral Arms with YSOs

- YSOs are independent tracers of Spiral Arm Structure



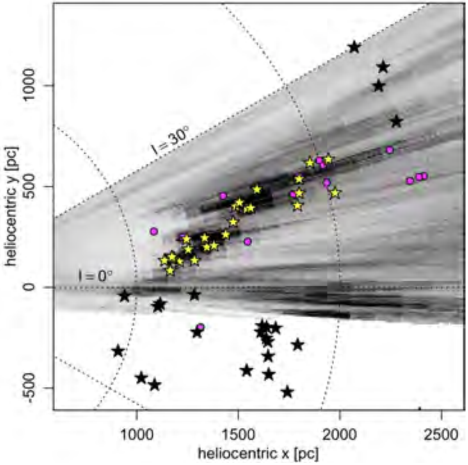
Mapping the Spiral Arms with YSOs

- We have identified a new structure near the Sagittarius arm



Mapping the Spiral Arms with YSOs

- We then compared it with other independent tracers such as dust maps and masers to confirm the structure was not an artifact



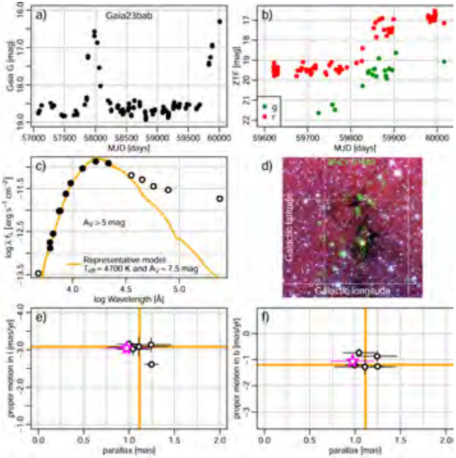
Mapping the Spiral Arms with YSOs

- Our analysis provided the first evidence of a high-pitch angle structure in the galactic spiral arms

SPICY byproducts

COIN SPICY

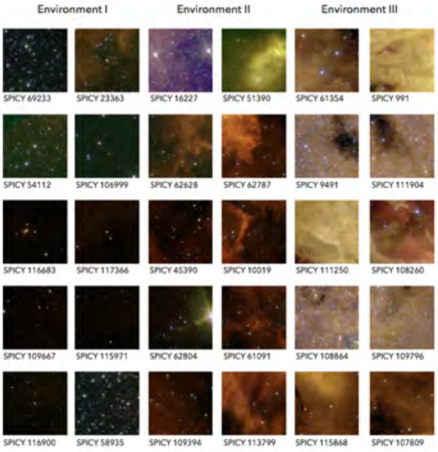
- Hundreds of thousands Light-curves (Time-Series)
- The light curve of Gaia23bab (=SPICY 97589) suggests the presence of an accretion outburst.
- These still scarce class of objects play a significant role in our understanding of star and planetary system formation.



SPICY byproducts

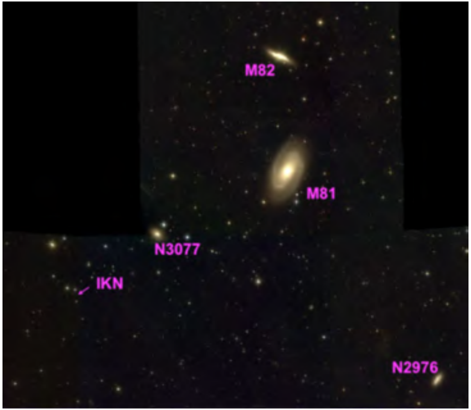
COIN SPICY

- 117,224 stamps of star forming regions
  - → Computer vision
  - → Fourier and Wavelets Analysis
  - → Marked Point Process



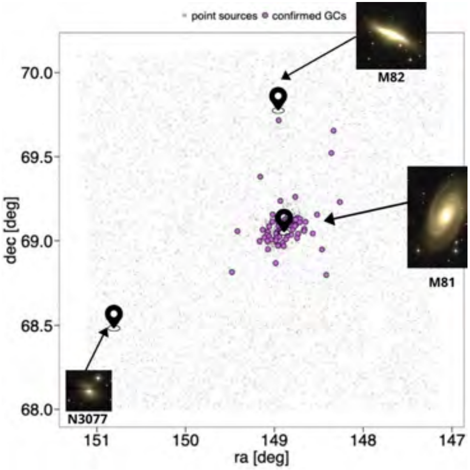
Searching for Extragalactic Globular Clusters

- Approximate figures
- Dwarf galaxies: 0 - 10 GCs
- Disk Galaxies 10s - 100s GCs
- Elliptical Galaxies 100s - 10k GCs
- Unsurprisingly GCs are usually targeted around E/S0 galaxies, because of large numbers and easier detection



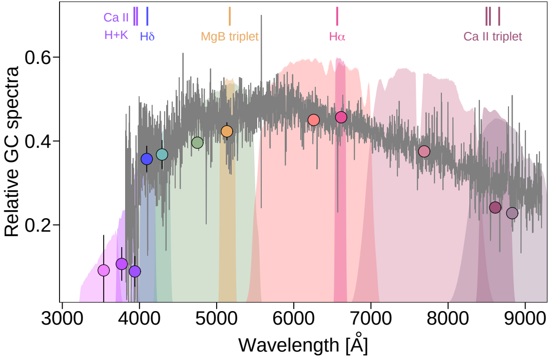
Searching for Extragalactic Globular Clusters

- To help mitigate this bias, we start a campaign to search for GCs around Spirals
- Only 105 confirmed GCs around the region (spectroscopic + HST data)



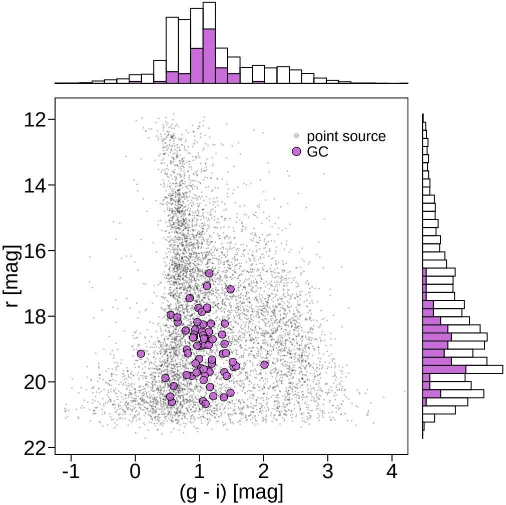
Searching for Extragalactic Globular Clusters

- Data from S-PLUS - a ongoing survey mapping about 9300 square degrees of the southern sky with an optical 12-bands.
- The figure shows a typical GC SED and Spectra



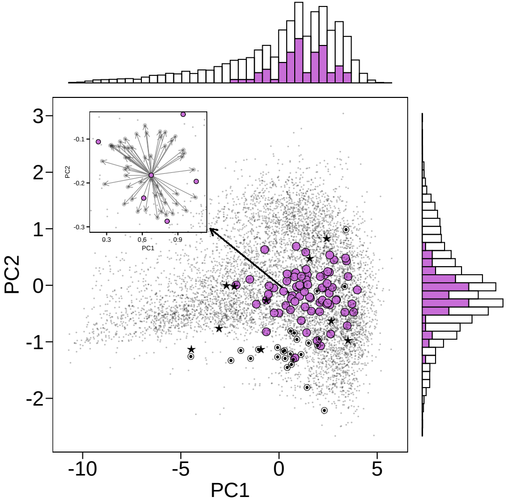
Photometric Selection - 7.2K point sources

- A traditional GC selection would apply color-magnitude cuts around regions of known GCs



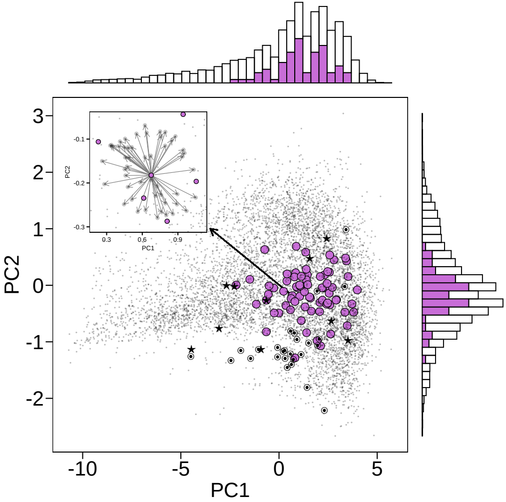
Photometric Selection - 7.2K point sources

- Going a bit further, we can just apply a Principal Components Analysis



Photometric Selection - 7.2K point sources

- But what about handling heteroskedastic errors with known variance?
- Off-the-shelf packages often don't account for errors in measurements



Yonder: Low-rank data denoising

RNAAS RESEARCH NOTES OF THE AAS

OPEN ACCESS

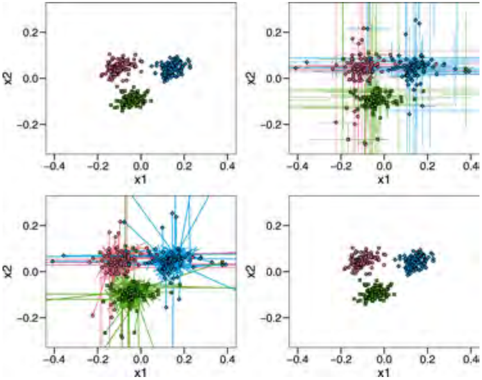
Yonder: A Python Package for Data Denoising and Reconstruction

Peng Chen (彭琛)[1] and Rafael S. de Souza[2]

Published March 2022 • © 2022. The Author(s). Published by the American Astronomical Society.
Research Notes of the AAS, Volume 6, Number 3
Citation Peng Chen and Rafael S. de Souza 2022 *Res. Notes AAS* **6** 51

Figures • References •

- Uncertainty aware PCA
- Data-denoising



PCA Scalability


Contents lists available at ScienceDirect

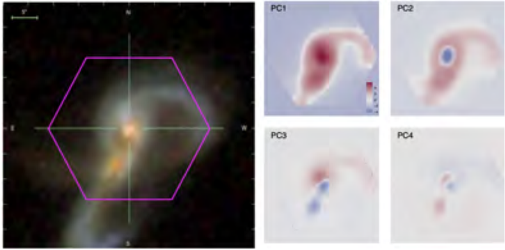
Astronomy and Computing

Full length article

qrpca: A package for fast principal component analysis with GPU acceleration

R. S. de Souza, X. Quanfeng, S. Shen, C. Peng, Z. Mu

- I was somewhat dissatisfied with the standard Python and R implementations of PCA, particularly when applied to IFUs (data cubes).
- I developed a QR-based PCA package.



PCA Scalability

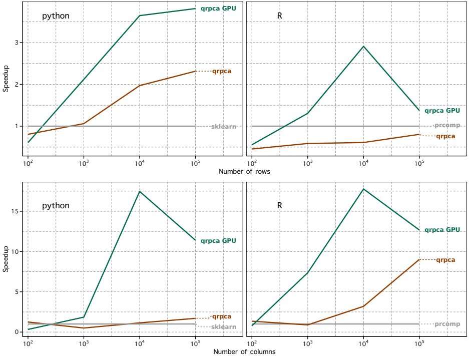

Astronomy and Computing

Full length article

qrpca : A package for fast principal component analysis with GPU acceleration

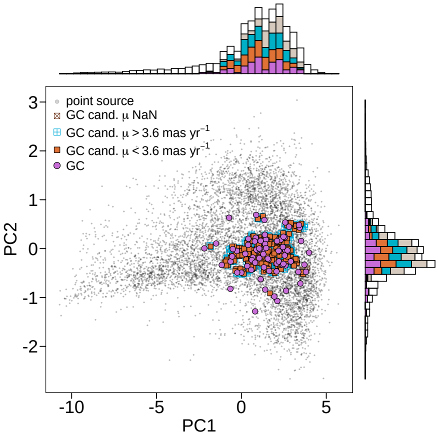
R. S. de Souza, X. Quanfeng, S. Shen, C. Peng, Z. Mu

- It utilizes Torch and Pytorch for GPU acceleration.
- QRPCA behaves similarly to standard implementations in `R` and `python`
- It is 10-20× faster then `sklearn` and `prcomp`



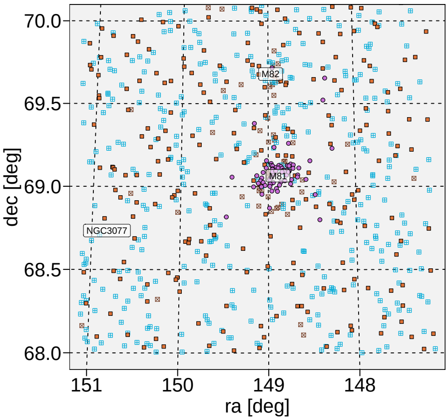
Back to Extragalactic GCs

- After employing our customized pre-processing, including imputation, denoising, proper motion cuts, and a propensity score matching, we compiled an initial list of 640 GC candidates out of 7k sources.



Back to Extragalactic GCs

- The first compilation of extragalactic GCs around the triplet.
- In the figure orange stands for GCs with lower proper motions, while cyans are higher in comparision to the known GC.
- We are systematically performing spectroscopic follow-up, which has borne fruit so far



Back to Extragalactic GCs

- An analysis of their spatial distribution suggests possible evidence for a bridge between M81 and M82, which is currently under investigation