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 $10^8 - 10^9$ objects observed by the Gaia space mission. With upcoming surveys, those numbers will be at least ten times larger.
• 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.
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.
YSO data: Spectral Energy Distribution

- ~50 million mid-IR sources
- "traditional" SED fitting to weed out reddened stars
- generate YSO and non-YSO training sets by cross-matching with previous studies
- copula imputation of missing data; MCMC using SBGCO (Hoff 2018)
- 117,446 YSO candidates
- Random forest
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?
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

$$F(x_1, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p))$$
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.
MIGAN employs a self-attention mechanism, which learns a non-local sparse representation of the data.
The MICE Algorithm

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.53</td>
<td>1.50</td>
<td>1.25</td>
<td></td>
</tr>
<tr>
<td>0.24</td>
<td>0.75</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>0.95</td>
<td>1.24</td>
<td>1.46</td>
<td></td>
</tr>
<tr>
<td>0.23</td>
<td>0.57</td>
<td>1.28</td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>0.42</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>0.89</td>
<td>1.23</td>
<td>1.45</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.53</td>
<td>1.50</td>
<td>1.25</td>
<td></td>
</tr>
<tr>
<td>0.24</td>
<td>0.75</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>0.95</td>
<td>1.24</td>
<td>1.46</td>
<td></td>
</tr>
<tr>
<td>0.23</td>
<td>0.57</td>
<td>1.28</td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>0.42</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>0.89</td>
<td>1.23</td>
<td>1.45</td>
<td></td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.53</td>
<td>1.50</td>
<td>1.25</td>
<td></td>
</tr>
<tr>
<td>0.24</td>
<td>0.75</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>0.95</td>
<td>1.24</td>
<td>1.46</td>
<td></td>
</tr>
<tr>
<td>0.23</td>
<td>0.57</td>
<td>1.28</td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>0.42</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>0.89</td>
<td>1.23</td>
<td>1.45</td>
<td></td>
</tr>
</tbody>
</table>

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

\[ R^2 = 0.9545 \]

\[ R^2 = 0.9906 \]

\[ R^2 = 0.9911 \]

\[ R^2 = 0.8771 \]
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

The largest catalogue of YSOs (~ 200,000) in the Milky Way midplane
For each YSO association

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

$$p_{\text{clust}}(\varpi_i, \mu_{\ell}^*, i, \mu_b, i | \varpi_0, \mu_{\ell}^*, 0, \mu_b, 0) =$$

$$\phi(\varpi_i | \varpi_0, \sigma_{\varpi}^2) \cdot f(\mu_{\ell}^*, i | \mu_{\ell}^*, 0, \sigma_{\mu_{\ell}^*}^2, \nu_\mu) \cdot f(\mu_b, i | \mu_b, 0, \sigma_{\mu_b}^2, \nu_\mu),$$

where $\theta = (\varpi_0, \mu_{\ell}^*, 0, \mu_b, 0)$ are the mean astrometric values for the cluster,

$x_i = (\varpi_i, \mu_{\ell}^*, i, \mu_b, i)$ are the measured values for the $i$th star, $\sigma_i$ are corresponding uncertainties, $\phi$ denotes a Gaussian distribution, and $f$ denotes a $t$-distribution.
Mapping the Spiral Arms with YSOs

- YSOs are independent tracers of Spiral Arm Structure
We have identified a new structure near the Sagittarius arm
We then compared it with other independent tracers such as dust maps and masers to confirm the structure was not an artifact.
Our analysis provided the first evidence of a high-pitch angle structure in the galactic spiral arms.
SPICY byproducts

- 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

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

- Uncertainty aware PCA
- Data-denoising
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

- It utilizes Torch and Pytorch for GPU acceleration.
- QRPCA behaves similarly to standard implementations in R and python.
- It is 10-20× faster than sklearn and prcomp.
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.
• 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 comparison to the known GC.
• We are systematically performing spectroscopic follow-up, which has borne fruit so far
• An analysis of their spatial distribution suggests possible evidence for a bridge between M81 and M82, which is currently under investigation.