UPMASK
Unsupervised Photometric Membership Assignment in Stellar Klusters

Alberto Krone-Martins
André Moitinho de Almeida
Introduction

• Why and what? Or, the problem...

• How good is our solution? Or, results from simulations...

• How we do it? Or, the methods...

• The real life...
Why and what? Or, the problem...
In a given field, segregate the field and cluster stars.

Why and what? Or, the problem...
Why and what? Or, the problem...

• Stellar Cluster membership

   • Usual **data**:

   • Positions;

   • **Basic, but huge contamination!**
Why and what? Or, the problem...

• Stellar Cluster membership

• Usual data:
  • Positions;
  • Positions and parallaxes;

• Basic, but huge contamination!

• Perfect! But parallaxes...
Why and what? Or, the problem...

- Stellar Cluster membership

- Usual data:
  - Positions;
  - Positions and parallaxes;
  - Proper-motions;
  - Basic, but huge contamination!
  - Perfect! But parallaxes...
  - Good, but contamination, hard to measure.
Why and what? Or, the problem...

• Stellar Cluster membership

• Usual data:
  
  • Positions;
  
  • Positions and parallaxes;
  
  • Proper-motions;
  
  • Radial-velocities;
  
  • Basic, but huge contamination!
  
  • Perfect! But parallaxes...
  
  • Good, but contamination, hard to measure.
  
  • Good, but contamination and expensive.
### Why and what? Or, the problem...

- **Stellar Cluster membership**

  - **Usual data:**
    - Positions;
    - Positions and parallaxes;
    - Proper-motions;
    - Radial-velocities;
    - Photometry (with isochrones).

  - **Basic, but huge contamination!**
    - Perfect! But parallaxes...
  - **Good, but contamination, hard to measure.**
  - **Good, but contamination and expensive.**
  - **Cheap! But contamination and strong assumptions.**
Why and what? Or, the problem...

- Stellar Cluster membership with easily available, large scale data
  - Positions;
  - Positions and parallaxes;
  - Proper-motions;
  - Radial-velocities;
  - Photometry.
  - Basic.
  - Perfect!
  - Good.
  - Good.
  - Cheap!
Why and what? Or, the problem...

- Stellar Cluster membership with easily available, large scale data
  - Positions;  **Basic.**
  - Positions and parallaxes;  **Perfect!**
  - Proper motions;  **Good.**
  - Radial velocities;  **Good.**
  - Photometry.  **Cheap!**
Why and what? Or, the problem...

• What are the weakest assumptions one can make?
  
  • The member stars are located in the same region of the space;
  
  • The member stars share common characteristics.

• This means...
  
  • Cluster stars should be clustered in arbitrary parameter spaces;
  
  • Field stars may not.

• UPMASK heuristic method
Why and what? Or, the problem...

- Using stellar positions and stellar photometry from a given field, segregate the field and cluster stars.

- Avoid to make strong assumptions.
How good is our solution? Or, the results from simulations...
How good is our solution?

- Simulations (position + UBVRI photometry):
  - Realistic clusters: MASSCLEAN (Popescu & Hanson, 2009)
  - Realistic galaxy (field stars): TRILEGAL (Girardi et al., 2005)

- Grid:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age [$10^\text{n yr}$]</td>
<td>7.1, 8.0, 8.3, 8.7, 8.8, 9.0, 9.2, 9.5</td>
</tr>
<tr>
<td>Initial Mass [$M_{\text{sun}}$]</td>
<td>500, 1000, 2500, 5000, 10000</td>
</tr>
<tr>
<td>Distance [kpc]</td>
<td>0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0</td>
</tr>
</tbody>
</table>
How good is our solution?
How good is our solution?
How good is our solution?

- True Positive Rate @ P>90%
- Contamination
How good is our solution?

- The fraction of simulated members that were recovered @ P>90%

- Age, distance, initial mass
How we do it? Or, the methods...
UPMASK concept

• Weak assumptions only
  • *Cluster stars should be clustered in arbitrary parameter spaces;*
  • *Field stars may not.*

• Data-driven approach
  • *Be simple, avoid pre-conceived models: let the data tell us what is in there.*

• Take the errors into account

• Allow for missing data
The UPMASK Kernel

- Photometric data
- Spatial data

Cluster/Field segregated data
The UPMASK Kernel

Photometric data → PCA → 4 PCs
Spatial data → Cluster/Field segregated data
The UPMASK Kernel

Photometric data → PCA → 4 PCs → Clustering algorithm (k-means) → Cluster/Field segregated data

Spatial data
The UPMASK Kernel

Photometric data → PCA

Spatial data → 4 PCs

Clustering algorithm (k-means)

Select photometric clusters which are not compatible with random distributions using spatial data (KDE2D)

Cluster/Field segregated data
The UPMASK Kernel

Photometric data → PCA → 4 PCs → Clustering algorithm (k-means) → Select photometric clusters which are not compatible with random distributions using spatial data (KDE2D) → is the list empty, or is it the same as last iteration?

Cluster/Field segregated data
The UPMASK Kernel

- Photometric data
- Spatial data

1. **Photometric data** → **PCA**
2. **PCA** → **4 PCs**
3. **4 PCs** → **Clustering algorithm (k-means)**
4. **Clustering algorithm (k-means)** → **Select photometric clusters which are not compatible with random distributions using spatial data (KDE2D)**
5. **Select photometric clusters which are not compatible with random distributions using spatial data (KDE2D)** → **is the list empty, or is it the same as last iteration?**
6. **is the list empty, or is it the same as last iteration?** → **No** → **Cluster/Field segregated data**
7. **is the list empty, or is it the same as last iteration?** → **Yes** → **Assign the remaining stars as members of the stellar cluster, while all the others to the field**
The UPMASK outer iterations

Original data → Draw random values from the photometric data and error models

Photometric & spatial data

Cluster/Field data with “probabilities”
The UPMASK outer iterations

Original data → Draw random values from the photometric data and error models → Photometric & spatial data

and optionally missing data

Cluster/Field data with “probabilities”
The UPMASK outer iterations

Original data → Draw random values from the photometric data and error models → Photometric & spatial data → Run the UPMASK Kernel → and optionally missing data

Cluster/Field data with “probabilities”
The UPMASK outer iterations

Original data → Draw random values from the photometric data and error models

Photometric & spatial data

Run the UPMASK Kernel

Max iterations reached? → No

Cluster/Field data with “probabilities”
The UPMASK outer iterations

1. Original data
2. Draw random values from the photometric data and error models
3. Photometric & spatial data
4. Run the UPMASK Kernel
5. Max iterations reached?
   - Yes: Compute membership frequencies
   - No: and optionally missing data
How good is our solution?

Contamination in original data (%)

True Positive Rate at P>90%

500 M_{sun}

1000 M_{sun}

2500 M_{sun}

5000 M_{sun}

log10(Age) [yr]

Distance [kpc]

P>90%
But what about the real life?
Haffner 16

- Typical stellar cluster...
Haffner 16
Haffner 16

UPMASK (Cluster Prob)
Haffner 16

Double stars

UPMASK (Cluster Prob)
Haffner 16

Double stars

PMS Signature?
Haffner 16
Haffner 10 + Czernik 29

- Two stellar clusters, projected near each other
Haffner 10 + Czernik 29

- Two stellar clusters, projected near each other
Haffner 10 + Czernik 29
Haffner 10 + Czernik 29
Haffner 10 + Czernik 29
Issues

• Variable reddening can “fake a cluster”
  
  • *UPMASK could be also used for mapping variable redning fields!*

• Highly compute intensive:
  
  • @12-cores, ~20 minutes for ~5000 stars (UBVRI + XY)

• Holes in the CMD diagrams
Conclusions

• The UPMASK heuristic method...
  
  • can make use of arbitrary data (positions, parallaxes, proper-motions, radial velocities, photometry in any filters);

  • has good purity and completeness;

  • can assist the automatic data analysis of stellar clusters... but also on the traditional, visual data analysis.

  • can be used to automatically map variable reddening on observational fields.

• The UPMASK code will be available through CRAN.
Thank you!