Object classification and the
determination of stellar parameters

Coryn A.L. Bailer-Jones
Max-Planck-Institut für Astronomie, Heidelberg

Philip Willemsen, Torsten Kaempf
Sternwarte der Universität Bonn
Classification objectives

To determine the intrinsic properties of individual objects observed by Gaia

• discrete source classification (DSC)
  – star, galaxy, quasar, asteroid, supernova etc.

• (stellar) astrophysical parameter estimation (APE)
  – $T_{\text{eff}}$, log g, [Fe/H], extinction ($A_V$), [$\alpha$/Fe], CNO, $V_{\text{rot}}$, activity

• identification of (unresolved) binaries

• identification of new types of objects

• Classification working group (ICAP)

$\Rightarrow$ catalogue of astrometric and astrophysical information
## Gaia data

<table>
<thead>
<tr>
<th>Data</th>
<th>Dimensions</th>
<th>No. sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBP</td>
<td>5-6 filters</td>
<td>all</td>
</tr>
<tr>
<td>MBP</td>
<td>ca. 10 filters</td>
<td>all (minus crowding)</td>
</tr>
<tr>
<td>RVS</td>
<td>~350 bins</td>
<td>G(&lt;15) (bright for classification quality)</td>
</tr>
<tr>
<td>parallaxes</td>
<td></td>
<td>some (sufficient accuracy)</td>
</tr>
<tr>
<td>time domain</td>
<td></td>
<td>all</td>
</tr>
</tbody>
</table>
## From Gaia data to APs

<table>
<thead>
<tr>
<th>Data</th>
<th>Derived quantity</th>
<th>Assumption/Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>normalized SED</td>
<td>Teff, log g, [Fe/H], A( )</td>
<td>atmospheric model: APs $\in$ SED inversion (data) model</td>
</tr>
<tr>
<td>G or abs. SED, A( )</td>
<td>L (Luminosity)</td>
<td>$2.5 \log L = f(G, \text{abs. SED}) = A - 5 \log$</td>
</tr>
<tr>
<td>L, Teff, log g, R</td>
<td>R (Radius)</td>
<td>$\log L \propto 2 \log R + 4 \log \text{Teff}$</td>
</tr>
<tr>
<td>M, R, [Fe/H]</td>
<td>M (Mass)</td>
<td>$\log g \propto \log M - 2 \log R$</td>
</tr>
<tr>
<td>t (Age)</td>
<td></td>
<td>evolutionary model</td>
</tr>
</tbody>
</table>

**SED = spectral energy distribution, from BBP, MBP, (RVS)**  
$= \text{parallax, } A = \text{extinction}$
Colour-colour diagrams: SDSS

- Black triangle = solar [Fe/H] dwarf
- Green = low [Fe/H] dwarf
- Red = solar [Fe/H] giant
- Blue = low [Fe/H] giant
- Black square = other

Simulated SDSS photometry of BaSeL 2.2 SEDs

Note: uncalibrated magnitudes (zero point offset)
Principle of AP estimation

Supervised approach:

- match data against pre-classified templates (i.e. with known APs)
- synthetic or real template data
- inverse mapping $\Rightarrow$ potential non-uniqueness (AP degeneracy)
Multidimensional regression

- local vs. global methods
- “curse of dimensionality” $\Leftrightarrow$ structured regression
- global, non-linear methods
  - train model on representative data through error minimization
  - e.g. ANN, SVM, MARS
- Examples: ANNs trained/tested on simulated BBP/MBP stellar data to estimate $T_{\text{eff}}$, log g, [Fe/H], extinction ($A_V$)
Model performance: $T_{\text{eff}}$
Model performance: $[\text{Fe/H}]$

Norm. counts vs. $[\text{Fe/H}]$ (black=true, red=stalnet)

Note the different vertical scales

Photon counts

$[\text{Fe/H}]$
AP estimation errors

red = mean error (systematic error)   blue = std. dev. about mean (random error)
APE performance from MBP

Accuracy varies strongly as a function of the 4 APs and magnitude
⇒ results averaged over wide AP ranges misleading

<table>
<thead>
<tr>
<th></th>
<th>( \sigma(A_V) )</th>
<th>( \sigma([\text{Fe/H}]) )</th>
<th>( \sigma(\text{logg}) )</th>
<th>( \sigma(T_{\text{eff}}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mag</td>
<td>dex</td>
<td>dex</td>
<td>%</td>
</tr>
<tr>
<td>KV, G=15, ( A_V = 0 ), [Fe/H] = +0.1..-2</td>
<td>~0.1</td>
<td>0.05–0.25</td>
<td>0.2</td>
<td>1–2</td>
</tr>
<tr>
<td>KV, G=20, ( A_V = 0 )</td>
<td>0.1–0.7</td>
<td>0.1–0.35</td>
<td>0.3–0.5</td>
<td>2–5</td>
</tr>
<tr>
<td>KV, G=20, ( A_V = 6 )</td>
<td></td>
<td>0.7</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>KIII, G=15, ( A_V = 0 )</td>
<td></td>
<td></td>
<td>0.2–0.4</td>
<td>2.5–4</td>
</tr>
<tr>
<td>KIII, G=15, ( A_V = 6 )</td>
<td></td>
<td></td>
<td>0.7–0.8</td>
<td>8</td>
</tr>
<tr>
<td>AIII, G=15, ( A_V = 0 ), [Fe/H] ~ 0</td>
<td>0.08</td>
<td>0.4</td>
<td>0.03</td>
<td>4</td>
</tr>
<tr>
<td>BV, G=15, ( A_V = 6 ), [Fe/H] ~ 0</td>
<td>0.3</td>
<td>0.35</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Willemsen, Kaempf, Bailer-Jones (2003)
AP error estimates

Bootstrap uncertainty estimates using neural networks (G=15)

Detection of unresolved binaries

- single star/binary classification
- IMF: Kroupa (2001) + more RGs
- composite MBP photometry
- $P(\text{BINARY}) = 50\%$ (prior)
- Support Vector Machine Classifier

Performance is a strong function of component luminosity ratio ($L_r$) and magnitude ($G$)

see poster 4.17
Issues

- model dependent physical (supervised) scheme
  - additional empirical (unsupervised) scheme?
- calibration of models
  - e.g. synthetic spectra + Gaia observations of known stars
- model training
  - regularization
- problem of “strong” vs. “weak” APs
  - systematic/correlated errors
- dealing with AP degeneracy
  - data space partitioning?
- heterogeneous data
  - optimal combination of BBP, MBP, RVS etc.
Optimization of Gaia filter systems

- optimize the input data ⇒ optimize photometric system (PS) for doing APE
- parametrize PS
- define figure-of-merit (FoM) of PS performance
- optimize FoM with respect to PS parameters (evolutionary algorithm)
- see Bailer-Jones (2004, A&A) and poster 4.3
Summary

- Objective: to determine the intrinsic properties of individual objects observed by Gaia
- wide range of objects/APs  \( \bigcup \) challenging problem
- multidimensional data analysis methods required
- several algorithmic issues to be addressed
  - degeneracies, training data (calibration)
- numerous open issues
  - discrete source classification, galaxy classification, treatment of peculiar/rare stars, empirical (unsupervised) classification, optimal combination of data (MBP, BBP, RVS, parallaxes etc.)
- several relevant posters (2.6, 2.10, 4.8, 4.9, 4.13, 4.17, 5.16, 5.19)