Three data analysis problems

Andreas Zezas

University of Crete
CfA
Two types of problems:

- Fitting
- Source Classification
Fitting: complex datasets

Galaxy Spectral Energy Distribution (SED)

- Best Model
- Photometric Flux
Fitting: complex datasets

Maragoudakis et al. in prep.
Fitting: complex datasets

Galaxy Spectral Energy Distribution (SED)

- **Wavelength (Å)**
- **$F_{\nu}$ (mJy)**

![Graph showing the Galaxy Spectral Energy Distribution (SED)]
Fitting: complex datasets

Iterative fitting may work, but it is inefficient and confidence intervals on parameters not reliable.

How do we fit jointly the two datasets?

VERY common problem!
Problem 2

Model selection in 2D fits of images
A primer on galaxy morphology

Three components:

spheroidal

\[ I(R) = I_e \exp \left[ -7.67 \left( \frac{R}{R_e} \right)^{1/4} - 1 \right] \]

exponential disk

\[ I(R) = I_0 \exp \left( \frac{r}{r_h} \right) \]

and nuclear point source (PSF)
Fitting: The method

Use a generalized model

\[ I(R) = I_e \exp \left[ -k \left( \frac{R}{R_e} \right)^{1/n} - 1 \right] \]

\( n = 4 : \) spheroidal
\( n = 1 : \) disk

Add other (or alternative) models as needed

Add blurring by PSF

Do \( \chi^2 \) fit (e.g. Peng et al., 2002)

\[ \chi^2_{\nu} = \frac{1}{N_{\text{dof}}} \sum_{x=1}^{n_x} \sum_{y=1}^{n_y} \frac{(\text{flux}_{x,y} - \text{model}_{x,y})^2}{\sigma^2_{x,y}} \]

\[ \text{model}_{x,y} = \sum_{\nu=1}^{n_f} f_{\nu,x,y}(\alpha_1 \ldots \alpha_n) \]
Fitting: The method

Typical model tree

\[ I(R) = I_e \exp \left[ -k \left( \frac{R}{R_e} \right)^{1/n} - 1 \right] \]

\[ \chi^2 \]
Fitting: Discriminating between models

Generally $\chi^2$ works

BUT:

Combinations of different models may give similar $\chi^2$

How to select the best model?

Models not nested: cannot use standard methods

Look at the residuals
Fitting: Discriminating between models
Fitting: Discriminating between models

Excess variance

\[ \sigma_{XS}^2 = \sigma_{obj}^2 - \sigma_{sky}^2 \]

Best fitting model among least \( \chi^2 \) models
the one that has the lowest exc. variance
Fitting: Examples

Figure 1: Results of the fits for the object MCG10-17-019, performed using the models presented in this section. For this specific target, the best-fit model (selected as described in Section ??) turned out to be the Sérsic + psfAgn + exDisk.

3.2 2D fit of SFRS galaxies

We used the results of the single Sérsic fits as first-guess parameters to implement more complex models, intended to separately account for different galaxy components. In particular, we wanted to model bulges and disks with Sérsic and exponential (exDisk) profiles respectively, and AGNs with a PSF component. The evaluation of the fit results, and the disk/bulge separation procedure based on them will be represented in Section 3.4. These additional, more complex models were:

▶ Sérsic + PSF
▶ Sérsic + exDisk
▶ Sérsic + exDisk + PSF

An accurate estimate of the first-guess input magnitude of each model component turned out to be critical to guarantee the convergence of these fits. We calculated the brightness of each component by re-distributing the integrated flux of the Sérsic model: in the case of Sérsic + PSF, the PSF component was initially attributed 1/10 of the total flux; in the other cases the flux was re-distributed equally among the components.

No constraints were applied in this stage, while contaminating objects were fit as above.

As an example, we report the fit results for the object NGC 3306 in Figure 3.4, where we show the data images (left panels), the models (central panels), and the model-subtracted images (“residuals”; right panels), for the models described in this section. The fit statistics for each model are reported in Table 3.3. This specific target contains an AGN (see Table 3.5), and the best-fit model (selected as described in Section 3.4) turned out to be the Sérsic + psfAgn + exDisk. The best-model fit results for the entire SFRS sample and the corresponding statistics are shown in Figure B.1.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2_\nu$</th>
<th>$\sigma^2_{\chi S}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sérsic</td>
<td>1.107</td>
<td>1.722(0.120)</td>
</tr>
<tr>
<td>Sérsic + psfAgn</td>
<td>1.107</td>
<td>1.657(0.118)</td>
</tr>
<tr>
<td>Sérsic + exDisk</td>
<td>1.107</td>
<td>1.770(0.121)</td>
</tr>
<tr>
<td>Sérsic + psfAgn + exDisk</td>
<td>1.106</td>
<td>1.472(0.113)</td>
</tr>
</tbody>
</table>

Table 3.3: Statistical results of the GALFIT fits for the object NGC 3306, performed using the models presented in this section.

(1) Fit model.
(2) Reduced $\chi^2$.
(3) Excess variance (described in Section 3.4).

Bonfini et al. in prep.
Fitting: Problems

However, method not ideal:

It is not calibrated

Cannot give significance

Fitting process computationally intensive

Require an alternative, robust, fast, method
Problem 3

Source Classification

(a) Stars
Classifying stars

Relative strength of lines discriminates between different types of stars

Currently done “by eye” or by cross-correlation analysis
Classifying stars

Would like to define a quantitative scheme based on strength of different lines.
Classifying stars

Maravelias et al. in prep.
Classifying stars

Not simple….

- Multi-parameter space
- Degeneracies in parts of the parameter space
- Sparse sampling
- Continuous distribution of parameters in training sample (cannot use clustering)
- Uncertainties and intrinsic variance in training sample
Problem 3

Source Classification

(b) Galaxies
Classifying galaxies

Ho et al. 1999
Classifying galaxies

Kewley et al. 2001

Kewley et al. 2006
Classifying galaxies

Basically an empirical scheme

- Multi-dimensional parameter space
- Sparse sampling - but now large training sample available
- Uncertainties and intrinsic variance in training sample

→ Use observations to define locus of different classes
Classifying galaxies

- Uncertainties in classification due to
  - measurement errors
  - uncertainties in diagnostic scheme
- Not always consistent results from different diagnostics

➔ Use ALL diagnostics together
➔ Obtain classification with a confidence interval

Maragoudakis et al in prep.
Classification

• Problem similar to inverting hardness ratios to spectral parameters

• But more difficult
  • We do not have well defined grid
  • Grid is not continuous

<table>
<thead>
<tr>
<th>N_H</th>
<th>0.250-0.500</th>
<th>0.125-0.250</th>
<th>0.075-0.125</th>
<th>0.050-0.075</th>
<th>0.025-0.050</th>
<th>0.010-0.025</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.75-2.00</td>
<td>11.36%</td>
<td>13.93%</td>
<td>3.35%</td>
<td>1.00%</td>
<td>0.53%</td>
<td>0.24%</td>
</tr>
<tr>
<td>1.50-1.75</td>
<td>5.56%</td>
<td>13.70%</td>
<td>5.99%</td>
<td>2.34%</td>
<td>1.70%</td>
<td>0.67%</td>
</tr>
<tr>
<td>1.25-1.50</td>
<td>1.80%</td>
<td>7.76%</td>
<td>5.61%</td>
<td>3.11%</td>
<td>2.82%</td>
<td>1.56%</td>
</tr>
<tr>
<td>1.00-1.25</td>
<td>0.38%</td>
<td>2.71%</td>
<td>2.87%</td>
<td>2.26%</td>
<td>2.33%</td>
<td>1.58%</td>
</tr>
<tr>
<td>0.75-1.00</td>
<td>0.07%</td>
<td>0.54%</td>
<td>0.82%</td>
<td>0.75%</td>
<td>1.00%</td>
<td>0.81%</td>
</tr>
<tr>
<td>0.50-0.75</td>
<td>0.01%</td>
<td>0.09%</td>
<td>0.15%</td>
<td>0.18%</td>
<td>0.23%</td>
<td>0.17%</td>
</tr>
</tbody>
</table>

Taeyoung Park’s thesis
Summary

• Model selection in multi-component 2D image fits
• Joint fits of datasets of different sizes
• Classification in multi-parameter space
  • Definition of the locus of different source types based on sparse data with uncertainties
  • Characterization of objects given uncertainties in classification scheme and measurement errors

All are challenging problems related to very common data analysis tasks.

Any volunteers?