Machine-Learning Atmospheric Retrieval

Kevin Heng (University of Bern, Switzerland)

Marquez-Neila, Fisher, Sznitman & Heng

MY GROUP: (Auclair-Desrotour), Deitrick, Fisher, Grimm, (Guzman Mesa), (Hakim), Hoeijmakers, Kitzmann, (Li), Malik, Oreshenko, Tsai, (?)
The challenge of atmospheric retrieval

Sophistication / Physical realism

Computational feasibility
The challenge of atmospheric retrieval

Spectrum → Forward model → Posterior distributions of parameters
The challenge of atmospheric retrieval

- Can be checked, made public, highly reproducible
- Can be made arbitrarily sophisticated
- Use model grids from other research groups
Random forest method of machine learning

decision tree + boot-strapping
**Decision trees: basics**

**Principle:** maximise information entropy gain
What the heck does any of this have to do with a spectrum?
What the heck does any of this have to do with a spectrum?

**Random forest method:**
quantify non-linear relationships between data points in a measured spectrum against underlying parameters of a model grid
My mind struggles with multi-dimensional, non-linear correlations. But the random forest does just fine!
Every regression tree is a “voter” who predicts the outcome with some error. Random forest of trees mitigates error with combined prediction. Ensemble of “voters” produces posterior distributions.

Train on 80,000 spectra
Test on 20,000 spectra
Compare predicted versus real values (red dashed line)

Information content analysis

THOR
Non-hydrostatic GCM
(Mendonca, Grimm et al.)

HELIOS
Radiative transfer
(Malik)

HELIOS-K
Opacity calculator
(Grimm)

VULCAN
Disequilibrium chemistry
(Tsai)

HELA
Machine-learning retrieval
(Marquez-Neila & Fisher)

LX-MIE
Mie scattering
(Kitzmann)

FastChem
Equilibrium chemistry
(Kitzmann & Stock)

HELIOS-R2
Nested-sampling retrieval
(Kitzmann)

MY GROUP: (Auclair-Desrotour), Deitrick, Fisher, Grimm, (Guzman Mesa), (Hakim), Hoeijmakers, Kitzmann, (Li), Malik, Oreshenko, Tsai, (?)
CSH Fellowships
Exoplanets, Solar System, geophysics, astrochemistry, etc
2-4 positions available this fall (AAS job register)
3 years of full independence
Including 5000 CHF (~5000 USD) individual funds per year
Backup Slides
This rule always applies:

GARBAGE IN,
GARBAGE OUT
Decision Tree

- Discrete data
- Binary questions
- Maximise entropy gain

Regression Tree

- Continuous data
- Threshold questions
- Minimise total variance
P = 10 mbar
Resolution = 5 cm$^{-1}$ (wavenumber)

![Graph showing opacity vs. wavelength for P = 10 mbar and resolution = 5 cm$^{-1}$ with curves for H$_2$O, HCN, and NH$_3$]
<table>
<thead>
<tr>
<th><strong>Random-forest retrieval</strong></th>
<th><strong>Regular retrieval</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-computed model grid</td>
<td>On-the-fly model grid</td>
</tr>
<tr>
<td>Model grid is cumulative</td>
<td>Model grid usually discarded</td>
</tr>
<tr>
<td>Arbitrary sophistication</td>
<td>Typically simplified</td>
</tr>
<tr>
<td>Utilize grids from multiple groups</td>
<td>...</td>
</tr>
<tr>
<td>Fast</td>
<td>...</td>
</tr>
<tr>
<td>[train: ~minutes</td>
<td>interpret: ~milli-seconds]</td>
</tr>
<tr>
<td>Demonstrated for $\sim10^4$ data points</td>
<td>...</td>
</tr>
<tr>
<td>Information content analysis “for free”</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>Model selection via Bayesian evidence</td>
</tr>
</tbody>
</table>