Hierarchical modeling

David W. Hogg
Center for Cosmology and Particle Physics, New York University
and
Max-Planck-Institut für Astronomie, Heidelberg

2013 July 26
**XDQSO** target selection (Bovy *et al.*, 1011.6392)

```
20.0 \leq i < 20.2  \ 33958 \ objects
```

```
20.0 \leq i < 20.2  \ 34471 \ objects
```
**XDQSO target selection** (Bovy et al., 1011.6392)
**XDQSO target selection** (Bovy et al., 1011.6392)

The diagram illustrates the process of XDQSO target selection, which involves a model with features $q_n$, $q_m$, $X_n$, $X_m$, $x_n$, and $x_m$. The model appears to be split into 'train' and 'test' sections, with the central variable $\theta$ serving as a connecting node. The input $\alpha$ suggests a selection process that may involve some form of prioritization or targeting.
collaborators

- Jo Bovy (IAS)
- Brendon Brewer (Auckland)
- Rob Fergus (NYU)
- Dan Foreman-Mackey (NYU)
- Jonathan Goodman (NYU)
- Dustin Lang (CMU)
eccentricities
eccentricity inference, usual story

\[ \omega_n \equiv (\kappa_n, T_n, \phi_n, e_n, \varpi_n) \]

\[ \nu_{nj} = V_n + g_\omega(t_{nj}) + E_{nj} \]

\[ -2 \ln p(D_n|\omega_n) = Q + \sum_{j=1}^{M_n} \ln(\sigma^2_{nj} + S_n^2) + \sum_{j=1}^{M_n} \frac{[V_n + g_\omega(t_{nj}) - \nu_{nj}]^2}{\sigma^2_{nj} + S_n^2} \]

\[ p(\omega_n|D_n) = \frac{1}{Z_n} p(D_n|\omega_n) p_0(\omega_n) , \]

where \( p_0(\omega_n) \) is some “uninformative” prior like flat in some parameters, \( 1/x \) in others.
Of course you don’t know the priors on exoplanet properties! What if you think there might be some family of priors $p(\omega_n|\alpha)$, parameterized by some $\alpha$? Could you find the best $\alpha$?

$$p(\{D_n\}_{n=1}^N | \alpha) = \prod_{n=1}^N \int d\omega_n p(D_n | \omega_n) p(\omega_n | \alpha) .$$

This is still a likelihood, but we have marginalized out the properties of every exoplanet—these are “nuisance” parameters in this formulation.
Say all you get, for each exoplanet, are $K$ samples drawn from an uninformative prior. What then? Importance sampling.

$$p(\omega_n|\alpha) \equiv \frac{f_\alpha(e_n) \, p_0(\omega_n)}{p_0(e_n)}$$

$$\int d\omega_n \, p_0(\omega_n|D_n) \, F(\omega_n) \approx \frac{1}{K} \sum_{k=1}^{K} F(\omega_{nk})$$

$$p(\{D_n\}_{n=1}^{N} | \alpha) \approx \prod_{n=1}^{N} \frac{1}{K} \sum_{k=1}^{K} \frac{f_\alpha(e_{nk})}{p_0(e_{nk})}$$

(Concept of an “interim prior”.)
distribution inference demo: ML estimates—bad

300 stars / truth

300 stars / ML estimates
distribution inference demo: Hierarchical inference good!
The marginalized likelihood is large when there is high prior probability in locations where there is high likelihood.

When likelihoods are broad, the best prior is the most concentrated prior that is “consistent with” all individual-object likelihood functions.

The operation is a **heteroskedastic deconvolution**.

(in modern parlance, a “deconvolution” is always the result of fitting a generative or forward model)
We can infer the true distribution even with extremely noisy measurements.

This is an extreme form of deconvolution.

Depends crucially on having full—and accurate—likelihood or posterior information.

Performed by “forward modeling”.
distribution inference demo: Small samplings

300 stars / truth

300 stars / 50 samples per star / inferred distribution
distribution inference demo: Small sample

- 30 stars / truth
- 30 stars / ML estimates

Frequency $p(e)$ vs eccentricity $e$ for 30 stars.
distribution inference demo: Still good!
distribution inference demo: Truly hierarchical
\[
f_{\theta}(t_n), \quad r, \quad x(t), \quad t_0, a, e, \varpi, \delta i, M_p, \quad R_*, f_*, \{I_k\}, \quad \sigma_n^2, \quad f_{\text{obs}; n}, \quad t_{\exp n}
\]
Generated using Bart v0.0.2-2f9b755

Time [Hours Since Transit]
**Bart** (Foreman-Mackey *et al*., forthcoming)

- built on very successful *emcee* package (Foreman-Mackey *et al*., 1202.3665)
- designed for exoplanet measurement and discovery of false positives
- Gaussian Process models for stellar variability

```python
import bart

# Initialize a planet.
planet = bart.Planet(r=0.01, a=21.3, t0=3.85)
planet.parameters += [bart.parameters.Parameter(r"r", "r"),
    bart.parameters.LogParameter(r"a", "a")]

# Initialize the star.
ldp = bart.kepler.fiducial_ldp(teff=6438, logg=4.28, feh=0.0)
star = bart.Star(mass=planet.get_mstar(12.4138), ldp=ldp)

# Set up the system.
system = bart.PlanetarySystem(star)
system.parameters.append(bart.parameters.CosParameter(r"i", "iobs"))
system.add_planet(planet)

# Add data and fit.
system.add_dataset(bart.KeplerDataset("path/to/kepler/data/lc.fits"))
system.fit(2000)
```
stack 003

Radial velocity (km/s)

time (d)
ersatz stack

radial velocity (km/s)

time (d)
ersatz stacked and binned

radial velocity \( (\text{km} s^{-1}) \)

\[ \begin{array}{c}
\text{time (d)} \\
0 & 200 & 400 & 600 & 800 & 1000
\end{array} \]

The graph shows a nearly constant radial velocity over time, with fluctuations around zero.
Once again, imagine you think there might be some family of priors $p(\omega_n|\alpha)$, parameterized by some $\alpha$, that describes the full population.

$$
p(\{D_n\}_{n=1}^{N} | \alpha) = \prod_{n=1}^{N} \int d\omega_n p(D_n|\omega_n) p(\omega_n|\alpha)
$$

The fact that each internal $p(D_n|\omega_n)$ contains no clear peak (no clear object detection at all) doesn’t change anything!
ersatz
The marginalized likelihood is large when there is high prior probability in locations where there is high likelihood.

When likelihoods are broad, the best prior is the most concentrated prior that is “consistent with” all individual-object likelihood functions.

The operation is a **heteroskedastic deconvolution**.
  - (in modern parlance, a “deconvolution” is always the result of fitting a generative or forward model)
hierarchical inference: What does it require?

- accurate likelihood functions
  - accurate noise models, or \textit{parameterized} noise models
- fast inference
  - self-tuning MCMC (like \textit{emcee}; Foreman-Mackey \textit{et al.}, 1202.3665)
  - robustness to multimodal likelihood functions
- concept of self-calibration
  - calibration and noise parameters are not different from astrophysical parameters
- racks and racks of metal
  - (it can’t be done in “map–reduce” framework)
astronomical source detection (Brewer et al., 1211.5805)

▶ the usual story:
  ▶ take images
  ▶ use noise and calibration parameters to make a catalog of “detected sources”
  ▶ perform scientific analyses on the catalog

▶ the hierarchical approach:
  ▶ a scientific hypothesis implies a source distribution; this is a parameterized prior over sources
  ▶ images are generated by sources, plus noise and calibration functions
  ▶ infer high-level parameters by hierarchical inference
  ▶ (involves sampling in “catalog space”)
Fig. 1.— The two simulated images used to test our methodology. **Left:** An image containing $\sim100$ stars. **Right:** An image containing $\sim1000$ stars.
Fig. 6.— The cumulative luminosity functions (number of stars above a given flux, as a function of flux) produced by the Bayesian method (several posterior samples shown) and SExtractor (for various values of the threshold parameters), compared with the actual cumulative LF. Both methods correctly identify the fluxes at the bright end, with some uncertainty due to overlapping sources. However, at the lower end SExtractor is unable to detect all of the stars whereas the true CLF is typical of the posterior distribution.
The Tractor (Lang et al., forthcoming)

- replacing the SDSS Catalog with
  - a maximum-likelihood model
  - a posterior sampling in catalog space
  - a callable likelihood function
    (see also Hogg & Lang, “Telescopes don’t make catalogs!”)
  - a hierarchical model
conclusions

- hierarchical inference permits
  - predictions of good data built entirely from bad data
  - learning of noise-deconvolved distribution functions
  - measurement of populations “too faint to detect”
  - improvements to inferences using full-population information

- hierarchical inference requires
  - enormous amounts of computation (in general)
  - extremely good understanding of the noise