Statistical techniques

David W. Hogg

(SCDA & NYU & MPIA)

Cosmic Dawn / 2016 June 21

Summary

- *This talk is not a review talk!*
- Use very flexible models to marginalize out calibration issues.
- Use data-driven models when you don't believe the physical models.
- Use Bayes to mix together competing models.

Church of Bayes

- A model is a **likelihood function** and **priors over nuisance parameters**.
  - $p(D \mid \theta, \alpha)$
  - $p(\alpha)$
- If you want to perform MCMC, you need priors over everything.
  - $p(\theta)$ too

Likelihood function

- The point of Bayes is to produce likelihood functions!
  - The likelihood is the thing that updates beliefs.
- This is true for both observers and theorists.
- Likelihood functions are technically **subjective**.
  - They involve decisions.
  - You use your judgement to make choices.

Pragmatism

- You can't make a measurement without a model.
  - $p(D \,|\, \theta, \alpha)$ and $p(\alpha)$
- However, often we can't afford to live the dream.
  - All other methods for making measurements can be seen as approximations to Bayes.
  - (for example: estimate and uncertainty)

Very important high-redshift science?

- *Trigger warning:* self-aggrandizement

Very important high-redshift science?

- Around 1995, with Judy Cohen I had the high-redshift record for a normal galaxy (something like $z = 0.8$), **but...**
- In 1996, Roger Blandford and I wrote a paper called "Gravitational Telescopes", **but...**
- In 1996–1998 with Smail and Cohen I had the deepest (faintest) galaxy counts in the U and R bands, and (in 2000) at 3 microns **but...**

Very important high-redshift science?

- Around 1995, with Judy Cohen I had the high-redshift record for a normal galaxy (something like $z = 0.8$), **but...**
- In 1996, Roger Blandford and I wrote a paper called "Gravitational Telescopes", **but...**
- In 1996–1998 with Smail and Cohen I had the deepest (faintest) galaxy counts in the U and R bands, and (in 2000) at 3 microns **but...**
- Lesson learned: In the high-redshift business, don't rest on your laurels!

Paradox of astrophysics

- The big secret of astronomy:

Paradox of astrophysics

- The big secret of astronomy:
- All models are wrong!
  - (strongly ruled out by the data)

Paradox of astrophysics

- The big secret of astronomy:
- All models are wrong!
  - (strongly ruled out by the data)
- All data are wrong!
  - (systematics and selections)

Combining spectroscopy and photometry

- *"I have both spectroscopy and photometry of my sources, and I want to fit models. There are so many more pixels in the spectroscopy than the photometry, if I just multiply the likelihoods, the spectroscopy dominates, the photometry is ignored, and I get wrong answers!"*

— Many astronomers

Combining spectroscopy and photometry

$$\ln p(D \mid \theta) = \underbrace{-\frac{1}{2}\sum_i \frac{[D_i - M_i(\theta)]^2}{\sigma_i^2}}_{\text{photometry}} \underbrace{-\frac{1}{2}\sum_j \frac{[D_j - M_j(\theta)]^2}{\sigma_j^2}}_{\text{spectroscopy}}$$

Combining spectroscopy and photometry

- Why do we want to upweight the photometry and downweight the spectroscopy?
  - Because we don't believe the calibration of the spectroscopy.
  - Sky subtraction, unaccounted noise sources.
- The right thing to do is to **marginalize out the calibration** and etc.
  - No reweighting is permitted!

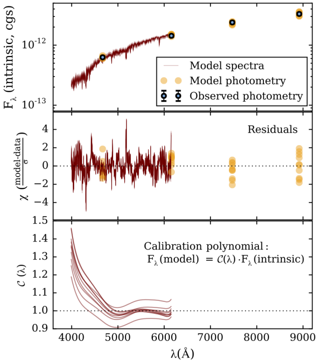
Combining spectroscopy and photometry

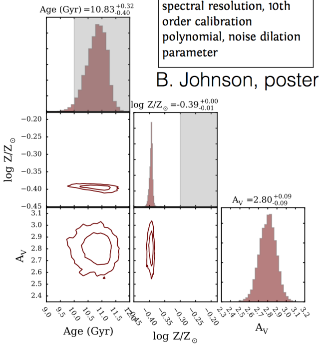
$$\ln p(D\,|\,\theta,\alpha) = -\frac{1}{2}\underbrace{\sum_i \frac{[D_i - M_i(\theta)]^2}{\sigma_i^2}}_{\text{photometry}} - \frac{1}{2}\underbrace{\sum_j \frac{[D_j - M_j(\theta,\alpha)]^2}{\sigma_j^2}}_{\text{spectroscopy}}$$

$$p(D\,|\,\theta) = \int p(D\,|\,\theta,\alpha)\,p(\alpha)\,\mathrm{d}\alpha$$

http://www.github.com/bd-j/prospector

Milky Way Globular Cluster NGC6553, fitting combined photometry & spectroscopy





simultaneously fit but not shown: mass, redshift, spectral resolution, 10th order calibration polynomial, noise dilation parameter

B. Johnson, poster #18

Dust constraint comes from photometry. Age and metallicity from spectrum.

Basic Idea:

Combine spectroscopic and photometric likelihood

Flexible model for spectroscopic calibration (“correlated noise”)

$$\ln p(d_s, d_p \,|\, \theta, \alpha, b) = \ln p_{spec}(d_s \,|\, \theta, \alpha) + \ln p_{phot}(d_p \,|\, \theta, b)$$

d_s: spec data

d_p: phot data

Physical model e.g.

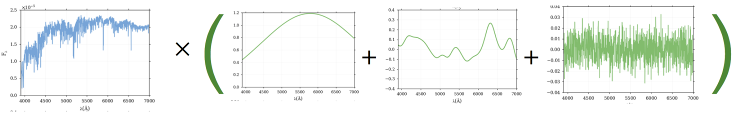
- age
- metallicity
- redshift

Spec calibration and noise params

- low order polynomial
- moderate scale (~100Å) wiggles as covariant noise
- small scale wiggles (~2Å) for model imperfections
- additive jitter
- multiplicative jitter

Photometry noise params

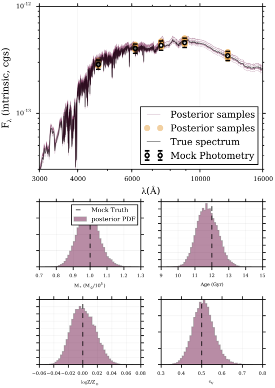
- multiplicative jitter



Prospector: Mock validation

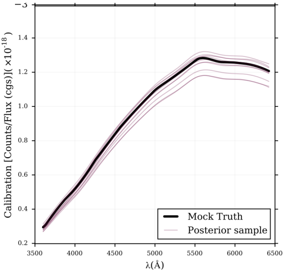
Simulated globular clusters: "Simple" Stellar Populations

(using FSPS and DFM's python-fsps)



With mock *uncalibrated* data we

1) recover a non-trivial calibration vector,
2) marginalize over calibration uncertainty,
3) and obtain dust constraints



Combining spectroscopy and photometry

(Johnson *et al.*, in prep)

- The flexible calibration vector marginalization effectively down-weights the “shape” of the spectrum.
- The procedure **obviates spectrophotometric calibration!**

Summary

- *This talk is not a review talk!*
- Use very flexible models to marginalize out calibration issues.
- Use data-driven models when you don't believe the physical models.
- Use Bayes to mix together competing models.

Models are wrong—Approaches

- Use no model: Data-driven
- Use every model ever made!
- Hybrid approaches

Data-driven models (my personal usage)

- Make use of things you **strongly believe**:
  - noise model & instrument resolution
  - causal structure (shared parameters)
- Capitalize on huge amounts of data.
- Use an exceedingly flexible model.
- (Concepts of train, validate, and test.)
- (Every situation will be **bespoke**.)

The Cannon: label transfer for stars

- A few of your stars have good labels (from somewhere).
- Can you use this to label the other stars?
- Why would you want to do this?

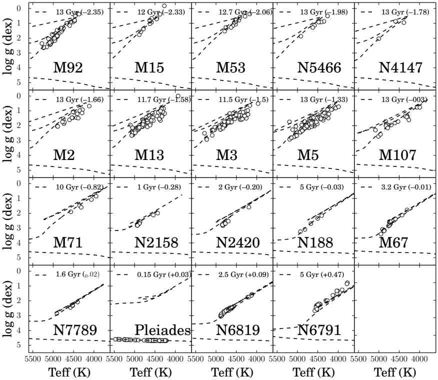
The Cannon: label transfer for stars

- A few of your stars have good labels (from somewhere).
- Can you use this to label the other stars?
- Why would you want to do this?
  - you don't have good models at your wavelengths?
  - you want two surveys to be on the same "system"?
  - you have some stars at high SNR, some at low SNR?
  - you spent human time on some stars but can't on all?

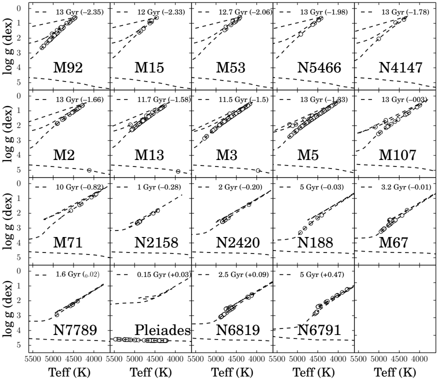
The Cannon: model

- $\ln p(\boldsymbol{f}_n \,|\, \boldsymbol{\ell}_n, \boldsymbol{\theta})$
- **training step**: optimize w.r.t. parameters $\boldsymbol{\theta}$ at fixed labels $\boldsymbol{\ell}$ using training-set data
  - linear least squares
  - every wavelength λ treated independently
- **test step**: optimize w.r.t. labels $\boldsymbol{\ell}$ at fixed parameters $\boldsymbol{\theta}$ using test-set (survey) data
  - non-linear optimization
  - every star treated independently

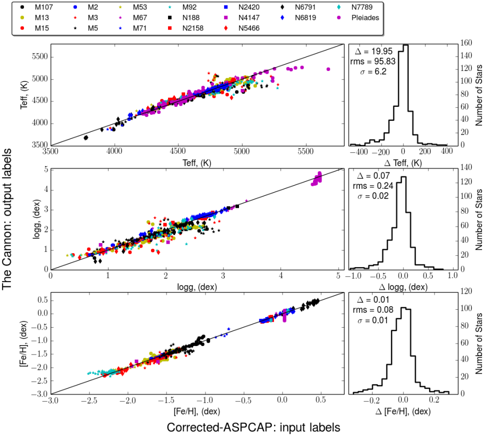
The Cannon (Ness *et al.*, 1501.07604)



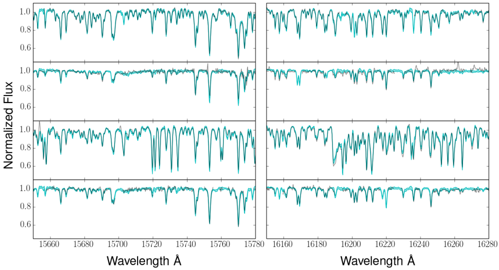
The Cannon (Ness *et al.*, 1501.07604)



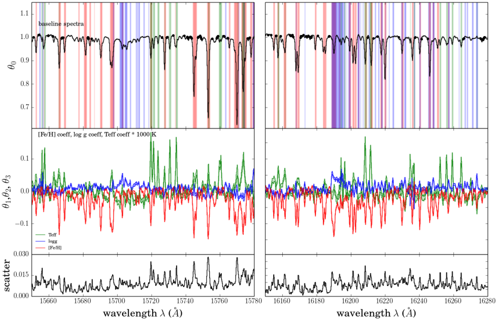
The Cannon (Ness *et al.*, 1501.07604)



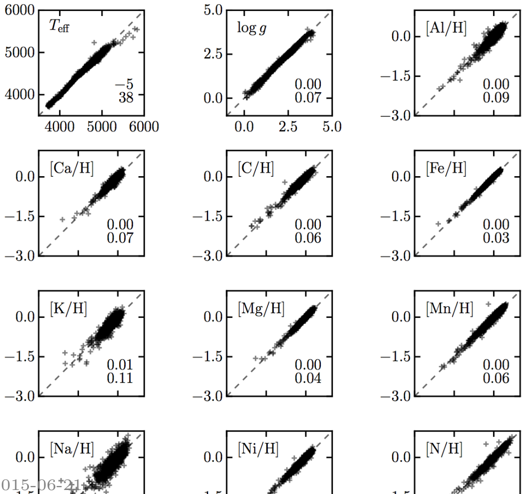
The Cannon (Ness *et al.*, 1501.07604)



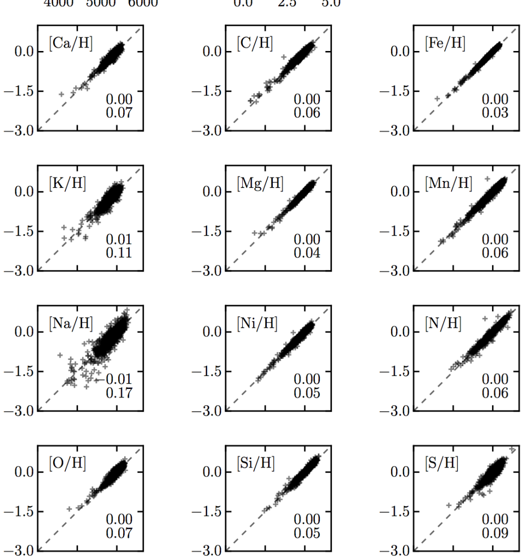
The Cannon (Ness *et al.*, 1501.07604)



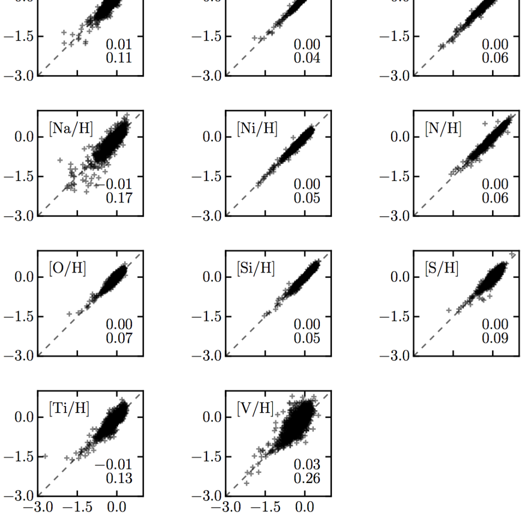
The Cannon 2: (Casey *et al.*, 1603.03040)



The Ca

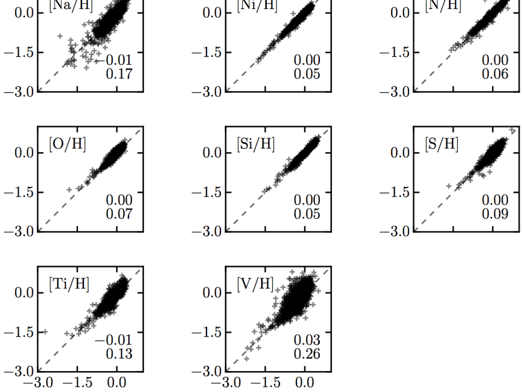

4000 5000 6000
0.0 2.5 5.0
[Ca/H]
0.00
0.07
[C/H]
0.00
0.06
[Fe/H]
0.00
0.03
[K/H]
0.01
0.11
[Mg/H]
0.00
0.04
[Mn/H]
0.00
0.06
[Na/H]
−0.01
0.17
[Ni/H]
0.00
0.05
[N/H]
0.00
0.06
[O/H]
0.00
0.07
[Si/H]
0.00
0.05
[S/H]
0.00
0.09
0.0
−1.5
−3.0

The Ca



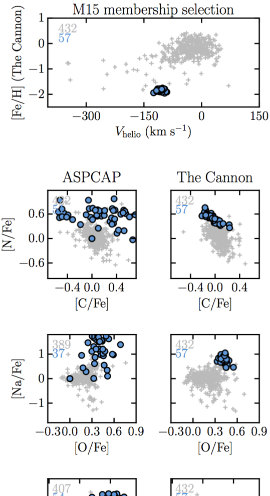
0.01
0.11
0.00
0.04
0.00
0.06
[Na/H]
−0.01
0.17
[Ni/H]
0.00
0.05
[N/H]
0.00
0.06
[O/H]
0.00
0.07
[Si/H]
0.00
0.05
[S/H]
0.00
0.09
[Ti/H]
−0.01
0.13
[V/H]
0.03
0.26
0.0
−1.5
−3.0
−3.0 −1.5 0.0

The Ca

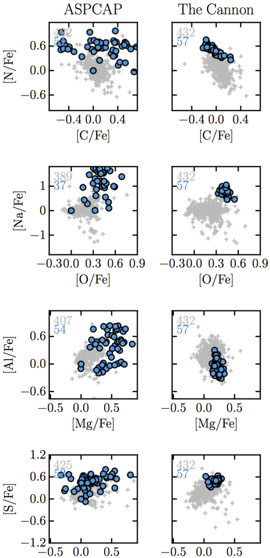


[Na/H]
−0.01
0.17
[Ni/H]
0.00
0.05
[N/H]
0.00
0.06
[O/H]
0.00
0.07
[Si/H]
0.00
0.05
[S/H]
0.00
0.09
[Ti/H]
−0.01
0.13
[V/H]
0.03
0.26
0.0
−1.5
−3.0

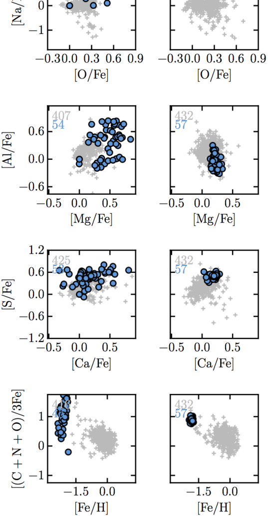
The Cannon 2: (Casey *et al.*, 1603.03040)



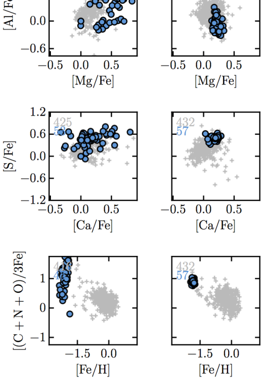
The Cannon 2: (Casey *et al.*,



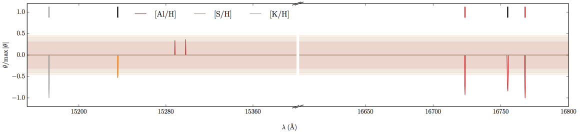
The Cannon 2: (Casey *et al.*,



The Cannon 2: (Casey *et al.*,



The Cannon 2: (Casey *et al.*, 1603.03040)



Data-driven models

- Huge successes (*e.g.*, accuracy, precision, adoption of *The Cannon*).
- Rarely will you meet the **training-data requirements**.
- Their output can be **hard to interpret**

Summary

- *This talk is not a review talk!*
- Use very flexible models to marginalize out calibration issues.
- Use data-driven models when you don't believe the physical models.
- Use Bayes to mix together competing models.

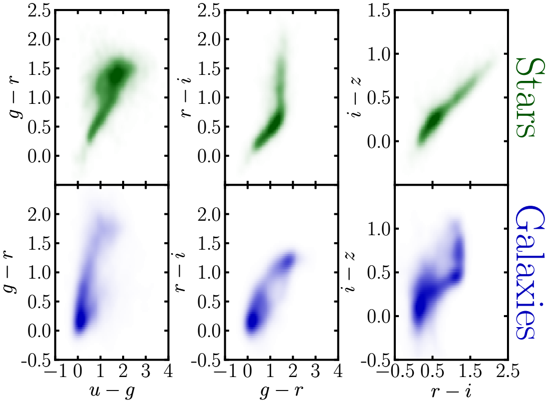
How to decide among different models?

- Different spectral template sets are different and give different answers!
- **Don't decide.**
- Throw all models in, let Hierarchical Bayes sort them out.

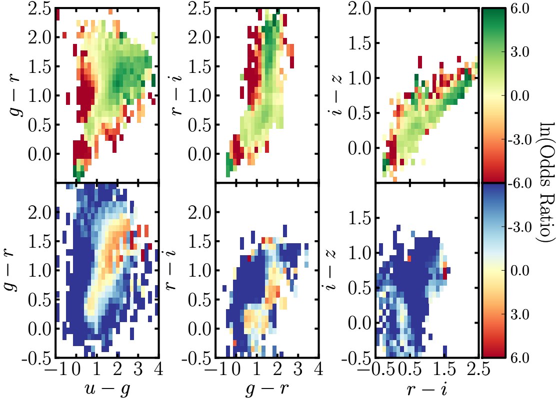
How to decide among different models?

- Every model template gets an adjustable prior probability!
- Learn the prior for every template **and** the posterior template for every star.
- Bayes goes for **parsimony**.
  - the **data decide** which models to keep
  - we did this in the star–galaxy separation context
  - (Fadely *et al.*, 1206.4306)

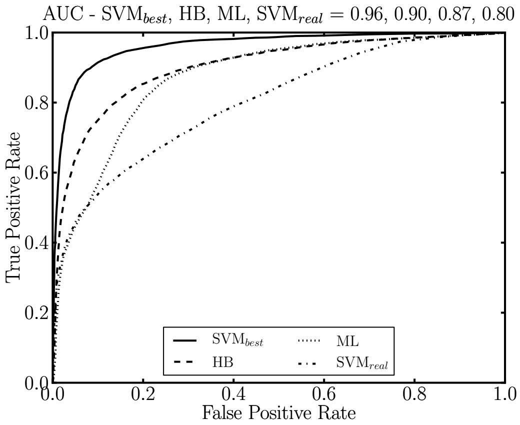
Star–galaxy separation (Fadely *et al.*, 1206.4306)



Star–galaxy separation (Fadely *et al.*, 1206.4306)



Star–galaxy separation (Fadely *et al.*, 1206.4306)



All models are wrong!

- What is the future?
- It is the **combination of these ideas**:
  - Hierarchical Bayes will **trim the list of models** to the models that work well.
  - Data-driven models will capture the **information in the residuals**.
  - Something very flexible will be used to **marginalize out calibration issues**.

BUT

- None of this obviates **building better physical models**!
  - Physical models are **interpretable** and **the basis of everything we know**.
  - All statistical projects are **ultimately in the service of improving the physical models**.

Summary

- *This talk is not a review talk!*
- Use very flexible models to marginalize out calibration issues.
- Use data-driven models when you don't believe the physical models.
- Use Bayes to mix together competing models.